SIAM J. CONTROL OPTIM.
Vol. 59, No. 6, pp. 4081–4109

POLICY OPTIMIZATION FOR $\mathcal{H}_2$ LINEAR CONTROL WITH $\mathcal{H}_\infty$ ROBUSTNESS GUARANTEE: IMPLICIT REGULARIZATION AND GLOBAL CONVERGENCE*

KAIQING ZHANG†, BIN HU†, AND TAMER BAŞAR†

Abstract. Policy optimization (PO) is a key ingredient for modern reinforcement learning. For control design, certain *constraints* are usually enforced on the policies to optimize, accounting for stability, robustness, or safety concerns on the system. Hence, PO is by nature a *constrained (nonconvex) optimization* in most cases, whose global convergence is challenging to analyze in general. More importantly, some constraints that are safety-critical, e.g., the closed-loop stability, or the $\mathcal{H}_\infty$-norm constraint that guarantees the system robustness, can be difficult to enforce on the controller being learned as the PO methods proceed. In this paper, we study the convergence theory of PO for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ *robustness guarantee*. This general framework includes *risk-sensitive* linear control as a special case. One significant new feature of this problem, in contrast to the standard $\mathcal{H}_2$ linear control, namely, linear quadratic regulator problems, is the *lack of coercivity* of the cost function. This makes it challenging to guarantee the *feasibility*, namely, the $\mathcal{H}_\infty$ robustness, of the iterates. Interestingly, we propose two PO algorithms that enjoy the *implicit regularization* property, i.e., the iterates preserve the $\mathcal{H}_\infty$ robustness automatically, as if they are regularized. Furthermore, despite the nonconvexity of the problem, we show that these algorithms converge to a certain *globally optimal* policy with *globally sublinear* rates, without getting stuck at any other possibly suboptimal stationary points, and with *locally (super)linear* rates under additional conditions. To the best of our knowledge, our work offers the first results on the implicit regularization property and global convergence of PO methods for robust/risk-sensitive control.

Key words. policy optimization, robust control, global convergence, implicit regularization, learning for control

AMS subject classifications. 93-XX, 90-XX

DOI. 10.1137/20M1347942

1. Introduction.

Recent years have seen tremendous success of reinforcement learning (RL) in various sequential decision-making applications [40] and continuous control tasks [26]. Interestingly, most successes hinge on the algorithmic framework of *policy optimization* (PO), umbrellaing policy gradient (PG) methods [41], actor-critic methods [24], trust-region methods [39], etc. This inspires an increasing interest in studying the convergence theory, especially global convergence to optimal policies, of PO methods; see recent progress in both classical RL contexts [1, 6], and continuous control benchmarks [10, 15, 29, 47].

In general, PO methods solve RL problems under the framework of constrained optimization $\min_{K\in\mathcal{K}} \mathcal{J}(K)$, where K is the parameter of the policy/controller, $\mathcal{J}(K)$ is the cost function the agent needs to minimize, and $\mathcal{K}$ denotes the feasible set of

*Received by the editors June 24, 2020; accepted for publication (in revised form) May 7, 2021; published electronically November 1, 2021.

https://doi.org/10.1137/20M1347942

Funding: K. Zhang and T. Başar were supported in part by the U.S. Army Research Laboratory (ARL) Cooperative Agreement W911NF-17-2-0196, and in part by the Office of Naval Research (ONR) MURI Grant N00014-16-1-2710. B. Hu was supported by the National Science Foundation (NSF) award CAREER-2048168 and the 2020 Amazon Research Award.

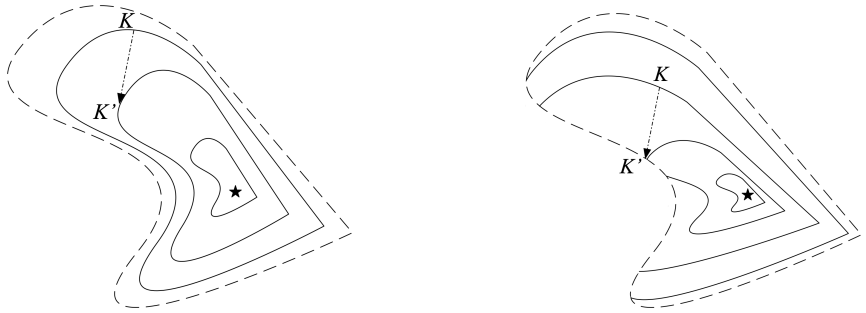
†Departmemt of Electrical and Computer Engineering and Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (kzhang66@illinois.edu, binhu7@illinois.edu, basar1@illinois.edu).

K.[1] For instance, in the standard continuous control task, linear quadratic regulator (LQR), the controller is parameterized as $u_t = -Kx_t$, the cost is $\mathcal{J}(K) := \sum_{t=0}^{\infty} \mathbb{E}[x_t^\top Q x_t + u_t^\top R u_t]$, and $\mathcal{K}$ is the set of K such that the system is stabilizing under K. Such a constrained optimization problem is generally nonconvex, even for the simple LQR problems [10, 15]. To ensure the feasibility of K on the fly as PO methods proceed, projection of the iterates onto $\mathcal{K}$ seems to be natural. However, such a projection may not be computationally efficient or even tractable. For example, projection onto the stability constraint in LQR problems can hardly be computed, as the set $\mathcal{K}$ therein is well known to be nonconvex [15]. Fortunately, such a projection is not needed when PG-based methods are used to solve LQR, as $\mathcal{J}(K)$ therein has a *coercive* property, i.e., the cost grows up to infinity as K approaches the boundary of $\mathcal{K}$ [10]. Hence, the intuition behind this avoidance of projection is that, as long as the cost is decreased along the iteration, the iterates stay in $\mathcal{K}$ and remain stabilizing. Such a result is *algorithm-agnostic*, in the sense that it is independent of the algorithms adopted, as long as they follow *any descent* directions of the cost.

Besides the stability constraint, another commonly used one in the control literature is the $\mathcal{H}_\infty$*-norm constraint*, which plays a fundamental role in robust control [2, 14, 48] and risk-sensitive control [16, 44]. Such a constraint can be used to guarantee *robust stability/performance* of the closed-loop systems when model uncertainty is present. Compared with LQR under the stability constraint, control synthesis under the $\mathcal{H}_\infty$ constraint leads to a fundamentally different optimization landscape, which has not been fully investigated yet. In this paper, we take an initial step towards understanding the theoretical aspects of policy-based RL methods on robust/risk-sensitive control problems with such a constraint. Specifically, we establish a convergence theory for PO methods on $\mathcal{H}_2$ linear control problems with $\mathcal{H}_\infty$ constraints, referred to as *mixed* $\mathcal{H}_2/\mathcal{H}_\infty$ *state-feedback control design* in the robust control literature [16, 22, 23]. As the name suggests, the goal of mixed design is to find a robust stabilizing controller that minimizes an upper bound for the $\mathcal{H}_2$-norm, subject to that the $\mathcal{H}_\infty$-norm on a certain input-output channel is less than a prespecified value. This general framework also includes risk-sensitive linear control, modeled as linear exponential quadratic Gaussian (LEQG) [21, 44] problems as a special case. In addition, this framework is also closely related to maximum-entropy $\mathcal{H}_\infty$ control [16] and zero-sum linear quadratic (LQ) dynamic games [4].

Two challenges exist in the analysis of PO methods for mixed design problems. First, by definition, the $\mathcal{H}_\infty$-norm constraint is defined in the frequency domain, and is hard to impose by directly projecting onto $\mathcal{K}$, especially when the system model is unknown in RL. Nevertheless, *preserving* the $\mathcal{H}_\infty$-norm constraint as the controller updates is critical in practice, as the violation of it can be catastrophic for real systems. Second, more importantly, compared to LQR, the cost of mixed design is no longer coercive, as illustrated in Figure 1(b) (and formally established later). Therefore, the decrease in cost cannot guarantee the feasibility of the iterate, as the cost remains *finite* around the boundary of $\mathcal{K}$. There may not even exist a constant step size that induces global convergence to the optimal policy. In this paper, we show that two PO methods can indeed preserve the robustness constraint along the iterations, and enjoy global convergence guarantees.

[1] Hereafter, we will mostly adhere to the terminologies and notational convention in the control literature, which are equivalent to, and can be easily translated to, those in the RL literature, e.g., cost versus reward, control versus action, etc.



(a) LQR (b) Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Control

FIG. 1. *Comparison of the landscapes of LQR and mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design that illustrates the difficulty of analyzing the latter. The dashed lines denote the boundaries of the constraint sets $\mathcal{K}$. For* (a) *LQR, $\mathcal{K}$ is the set of all linear stabilizing state-feedback controllers; for* (b) *mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control, $\mathcal{K}$ is set of all linear stabilizing state-feedback controllers satisfying an extra $\mathcal{H}_\infty$ constraint on some input-output channel. The solid lines represent the contour lines of the cost $\mathcal{J}(K)$. K and K' denote the control gain of two consecutive iterates; $\star$ denotes the global optimizer.*

Contribution. Our key contributions are threefold: First, we study the landscape of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problems, and propose three PG-based methods, inspired by those for LQR [15]. Second, we prove that two of them (the Gauss–Newton and the natural PG) enjoy the *implicit regularization* property, i.e., the iterates are automatically biased to satisfy the required $\mathcal{H}_\infty$ constraint. Third, we establish the global convergence of those two PO methods to the *globally optimal* policy with *globally sublinear* and *locally (super)linear* rates under certain conditions, despite the nonconvexity of the problem. In particular, the two policy search directions always lead convergence to a certain global optimum, without getting stuck at any other possibly suboptimal stationary points/local optima. To the best of our knowledge, our work appears to be the first studying the global convergence theory of PO methods for *learning for robust/risk-sensitive* control. Due to space limitations, we focus on the discrete-time settings in this paper. We have also established a set of results for continuous-time settings, and detailed numerical simulations, which can be found in the companion report [46]. We have included some of the simulation results in the appendix, to illustrate the computation efficiency of our PO methods. We highlight the most related literature as follows, and refer to [46] for a more detailed one.

Related work. The history of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design dates back to the seminal works [5, 23] for continuous-time and [22, 31] for discrete-time settings, respectively. A nonsmooth constrained optimization perspective for solving general output-feedback mixed design problems was adopted in [2], with a proximity control algorithm designed to handle the constraints explicitly, and convergence guarantees to stationary-point controllers. Numerically, there also exist other packages for multiobjective $\mathcal{H}_2/\mathcal{H}_\infty$ control [3] that are based on nonsmooth nonconvex optimization. However, in spite of achieving impressive numerical performance, these methods have no theoretical guarantees for either the global convergence or the $\mathcal{H}_\infty$-norm constraint violation. It is also not yet clear how these methods can be made *model-free*. Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control can also be unified with risk-sensitive linear control [16, 44], maximum-entropy $\mathcal{H}_\infty$ control [16], and zero-sum dynamic games [4, 21]. Recently, first-order optimization methods have also been applied to finite-horizon risk sensitive nonlinear control, but the control inputs (instead of the policy) are treated as decision variables [37]. Convergence to stationary points was shown therein. Besides

these direct controller/policy search methods, general mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control can also be tackled via Youla-parameterization-based approaches [8, 12, 36, 38], which lead to convex programming problems that can be solved numerically. However, actual implementation of these approaches either requires a finite-horizon *truncation* of system impulse responses, which loses optimality guarantees [8], or require solution of (a large enough sequence of [12, 36]) semidefinite programs or linear matrix inequalities with lifted dimensions [12, 36, 38], which may not be computationally efficient for large-scale dynamical systems. More importantly, it is not yet clear how to implement these approaches in the data-driven regime, without identifying the model. In contrast, the direct search methods can easily be made model-free; see, e.g., the methods in [15, 29] for LQR. Another related line of work is on PO for LQR, which stemmed from the adaptive policy iteration algorithm in [9]. Lately, studying the global convergence of PG-based methods for LQR [10, 15, 17, 29, 30, 43] has drawn increasing attention. Starting from the seminal work [15], where the optimization landscape was studied, [29, 30] have improved the sample complexity of [15]. However, no robustness was concerned in these LQR ($\mathcal{H}_2$ control) formulations. More recently, [17, 43] have extended the work to the case with multiplicative noises, as one way to improve the controller's robustness.

2. Preliminaries. We first formulate mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control with a single input-output channel as a constrained PO problem. We note that this problem covers risk-sensitive linear control [21] as a special case [16]. We provide some new results on this connection in [46, sect. 2]. Consider the linear system

$$x_{t+1} = Ax_t + Bu_t + Dw_t, \quad z_t = Cx_t + Eu_t, \tag{2.1}$$

where $x_t \in \mathbb{R}^m, u_t \in \mathbb{R}^d$ denote the states and controls, respectively, $w_t \in \mathbb{R}^n$ is the disturbance, $z_t \in \mathbb{R}^l$ is the controlled output, and A, B, C, D, E are matrices of proper dimensions. It has been shown in [22] that a linear time-invariant (LTI) state-feedback controller (without memory) suffices to achieve the optimal performance of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design under this *state-feedback* information structure.[2] Hence, it suffices to consider only a stationary state-feedback controller parameterized as $u_t = -Kx_t$. With this parameterization, the transfer function from the disturbance w_t to the output z_t can be represented as

$$\left[\begin{array}{c|c} A - BK & D \\ \hline C - EK & 0 \end{array}\right]. \tag{2.2}$$

In common with [4, 16, 23], we make the following assumption on (A, B, C, D, E).

Assumption 2.1. There exists some $R > 0$ such that $E^\top [C \;\; E] = [0 \;\; R]$.

Assumption 2.1 is fairly standard, which clarifies the exposition substantially by normalizing the control weighting and eliminating cross-weightings between control signal and state [4]. Hence, the transfer function in (2.2) has the equivalent form[3] of

$$\mathcal{T}(K) := \left[\begin{array}{c|c} A - BK & D \\ \hline (C^\top C + K^\top RK)^{1/2} & 0 \end{array}\right]. \tag{2.3}$$

[2]For discrete-time settings, if both the (exogenous) disturbance w_t and the state x_t are available, i.e., under the *full-information* feedback case, LTI controllers may not be optimal [22]. Interestingly, for continuous-time settings, LTI controllers are indeed optimal [23].

[3]Strictly speaking, the transfer functions for (2.2) and (2.3) are equivalent in the sense that the values of $\mathcal{T}^\sim(K)\mathcal{T}(K)$ are the same for all the points on the unit circle.

Hence, robustness of the designed controller can be guaranteed by the constraint on the $\mathcal{H}_\infty$-norm, i.e., $\|\mathcal{T}(K)\|_\infty < \gamma$ for some $\gamma > 0$. The intuition behind the constraint, which follows from the small gain theorem [45], is that the constraint on $\|\mathcal{T}(K)\|_\infty$ implies that the closed-loop system is *robustly stable* in that any stable transfer function Δ satisfying $\|\Delta\|_{\ell_2\to\ell_2} < 1/\gamma$ may be connected from z_t back to w_t without destablizing the system. For more background on $\mathcal{H}_\infty$ control, see [4, 48]. For notational convenience, we define the feasible set of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design as

$$\mathcal{K} := \big\{K \,\big|\, \rho(A - BK) < 1, \text{ and } \|\mathcal{T}(K)\|_\infty < \gamma\big\}. \tag{2.4}$$

We note that the set $\mathcal{K}$ may be unbounded.

In addition to the constraint, the objective of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design is usually an upper bound of the $\mathcal{H}_2$ norm of the closed-loop system. Let $\mathcal{J}(K)$ be the cost function of mixed design. Then the common forms of $\mathcal{J}(K)$ include [31]

$$\mathcal{J}(K) = \mathrm{Tr}(P_K DD^\top), \tag{2.5}$$

$$\mathcal{J}(K) = -\gamma^2 \log\det(I - \gamma^{-2} P_K DD^\top), \tag{2.6}$$

$$\mathcal{J}(K) = \mathrm{Tr}\left[D^\top P_K (I - \gamma^{-2} DD^\top P_K)^{-1} D\right], \tag{2.7}$$

where P_K is the solution to the following Riccati equation

$$(A - BK)^\top \widetilde{P}_K (A - BK) + C^\top C + K^\top R K - P_K = 0 \tag{2.8}$$

with $\widetilde{P}_K$ defined as

$$\widetilde{P}_K := P_K + P_K D(\gamma^2 I - D^\top P_K D)^{-1} D^\top P_K. \tag{2.9}$$

All three objectives in (2.5)–(2.7) are upper bounds of the $\mathcal{H}_2$-norm [31]. In particular, cost (2.5) has been adopted in [5, 18], which resembles the standard $\mathcal{H}_2$/LQG (linear quadratic Gaussian) control objective, but with P_K satisfying a Riccati equation instead of a Lyapunov equation. Cost (2.6) is related to maximum entropy $\mathcal{H}_\infty$-control; see the detailed relationship between the two in [32]. In addition, cost (2.7) can also be connected to the cost of LQG using a different Riccati equation [31, Remark 2.7]. As $\gamma \to \infty$, the costs in all (2.5)–(2.7) reduce to the cost for LQG, i.e., $\mathcal{H}_2$ control design problems.

In sum, the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design can be formulated as

$$\min_K \quad \mathcal{J}(K) \qquad \text{s.t.} \quad K \in \mathcal{K} \tag{2.10}$$

with $\mathcal{J}(K)$ and $\mathcal{K}$ defined in (2.5)–(2.7) and (2.4), respectively.

A key element of the above formulation is the feasibility constraint (2.4), which is characterized in the frequency domain and hence hard to directly enforce over K in PO. Interestingly, by using a significant result in robust control theory, i.e., *the bounded real lemma* [4, Chapter 1], [35, 48], constraint (2.4) can be related to Riccati equation/inequality as follows.

LEMMA 2.2 (bounded real lemma). *Consider a discrete-time transfer function $\mathcal{T}(K)$ defined in* (2.3)*, then the following three conditions are equivalent:*

- *The controller K lies in $\mathcal{K}$ defined in* (2.4)*, i.e., $\rho(A-BK) < 1$ and $\|\mathcal{T}(K)\|_\infty < \gamma$.*

- *The Riccati equation* (2.8) *admits a unique stabilizing solution* $P_K \geq 0$ *such that* (i) $I - \gamma^{-2} D^\top P_K D > 0$*;* (ii) $(I - \gamma^{-2} P_K D D^\top)^{-\top}(A - BK)$ *is stable.*
- *There exists some* $P > 0$*, such that*

$$I - \gamma^{-2} D^\top P D > 0, \tag{2.11}$$
$$(A - BK)^\top \widetilde{P}(A - BK) - P + C^\top C + K^\top R K < 0,$$

where $\widetilde{P} := P + PD(\gamma^2 I - D^\top P D)^{-1} D^\top P$.

The three conditions in Lemma 2.2 will be frequently used in the ensuing analysis. Note that the unique stabilizing solution to (2.8) for any $K \in \mathcal{K}$, is also *minimal*, if the pair $(A - BK, D)$ is stabilizable; see [34, Theorem 3.1]. This holds here since any $K \in \mathcal{K}$ is stabilizing. More comments on Lemma 2.2 are provided in section A.1.

3. Landscape and algorithms. In this section, we investigate the optimization landscape of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design, and develop PO algorithms for solving (2.10) provably. In particular, we focus on the representative example cost $\mathcal{J}(K)$ in (2.6), as it coincides with the objective of LEQG, a fundamental setting in risk-sensitive control (cf. [46, sect. 2]). Although we focus here on the LEQG cost function, the techniques and most of the results below apply to other types of costs as well; see the brief discussions in later sections, and more detailed ones in [46].

3.1. Optimization landscape. We start by showing that the mixed-design problem in (2.10) lacks convexity and coercivity.

LEMMA 3.1 (nonconvexity and no coercivity). *The discrete-time mixed* $\mathcal{H}_2/\mathcal{H}_\infty$ *design problem* (2.10) *is nonconvex. Moreover, the cost functions* (2.6) *is not coercive. Particularly, as* $K \to \partial\mathcal{K}$*, where* $\partial\mathcal{K}$ *is the boundary of the constraint set* $\mathcal{K}$*, the cost* $\mathcal{J}(K)$ *does not necessarily approach infinity.*

The proof of Lemma 3.1 is deferred to section A.2. Note that the results above also apply to the other two objectives, (2.5) and (2.7), the proofs of which are not included here due to page limitation, and can be found in [46, Lemmas 3.1, 3.2]. In particular, we show nonconvexity by an easily constructed example that the convex combination of two control gains K and K' in $\mathcal{K}$ may no longer lie in $\mathcal{K}$. Similar nonconvexity of the constraint set also exists in LQR problems [10, 15], and has been recognized as one of the main challenges for solving PO via gradient-based methods.

More importantly, we have also constructed a simple example to show the lack of coercivity for (2.10). Note that the landscape of LQR has the desired property of being coercive [10, Lemma 3.7], which played a significant role in the analysis of PO methods for LQR. The intuition for this result is that for given $K \in \mathcal{K}$, the *policy evaluation* equation for mixed design problems is a Riccati equation (see (2.8) (a quadratic equation of P_K in a 1-dimensional case), while for LQR, the policy evaluation equation is a Lyapunov equation, which is essentially linear. Hence, some additional condition on K is required for the existence of the solution, which can be more *restrictive* than the conditions on K and P_K that yield a finite $\mathcal{J}(K)$. In this case, the existence condition of the solution characterizes the boundary of $\mathcal{K}$, which leads to a well-defined P_K, and thus a finite value of $\mathcal{J}(K)$, even when K approaches the boundary $\partial\mathcal{K}$.

The lack of coercivity turns out to be the main challenge when analyzing the stability/feasibility of PO methods here, in contrast to LQR problems. Detailed discussion on this is provided in section 4.1. The illustration in Figure 1 in section 1 of the landscapes of two problems was actually based on Lemma 3.1. We then show

the differentiability of $\mathcal{J}(K)$ at each K within $\mathcal{K}$, and provide the closed form of the PG.

LEMMA 3.2 (PG expression). *The cost $\mathcal{J}(K)$ defined in* (2.6) *is differentiable in K for any $K \in \mathcal{K}$, and the PG has the following form:*

$$\nabla \mathcal{J}(K) = 2\big[(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A\big]\Delta_K,$$

where $\Delta_K \in \mathbb{R}^{m\times m}$ is a matrix given by

$$\Delta_K := \sum_{t=0}^{\infty} \big[(I - \gamma^{-2} P_K D D^\top)^{-\top}(A - BK)\big]^t D (I - \gamma^{-2} D^\top P_K D)^{-1} D^\top \cdot \big[(A - BK)^\top (I - \gamma^{-2} P_K D D^\top)^{-1}\big]^t, \tag{3.1}$$

and $\widetilde{P}_K$ is defined in (2.9).

The proof of Lemma 3.2 is provided in section A.3. Similar expressions for PG can also be derived for the other two objectives (2.5) and (2.7), with only the form of Δ_K being different, e.g., see our LQ zero-sum game paper [47], where the cost used was essentially the one in (2.5) in the trace form. Note that Lemma 3.2 also implies some property on the landscape of $\mathcal{J}(K)$. Specifically, if $\Delta_K > 0$ is full rank, then $\nabla \mathcal{J}(K) = 0$ admits a unique solution $K = (R + B^\top \widetilde{P}_K B)^{-1} B^\top \widetilde{P}_K A$, which corresponds to the unique global optimum. Otherwise, if $\Delta_K \geq 0$ is not full rank, there can be multiple stationary points. Yet, one global optimum is still of the same form. We formally establish this in the following proposition proved in section A.4.

PROPOSITION 3.3. *Suppose that the discrete-time mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design admits a global optimal solution $K^* \in \mathcal{K}$; then, one such solution has the form of $K^* = (R + B^\top \widetilde{P}_{K^*} B)^{-1} B^\top \widetilde{P}_{K^*} A$. Additionally, if the pair $\big((I - \gamma^{-2} P_K D D^\top)^{-\top}(A - BK), D\big)$ is controllable at some stationary point of $\mathcal{J}(K)$ such that $\nabla \mathcal{J}(K) = 0$, then this is the unique stationary point, and corresponds to the unique global optimizer K^*.*

We note that the landscape result above can also be shown for the other two objectives (2.5) and (2.7). In fact, the key in proving Proposition 3.3 is to show that P_{K^*} is *matrixwise* minimal in the positive semidefinite sense for all P_K with $K \in \mathcal{K}$. Note that since the objectives (2.5) and (2.7) are both monotonically nondecreasing in the eigenvalues of P_K, one can verify that K^* is also the global optimizer. Besides (2.5) and (2.7), such an argument also applies to other objectives nondecreasing in the eigenvalues of P_K. Note that K^* may not be the *unique* global minimizer without the controllability assumption. Although the controllability assumption is standard [31], and is also satisfied automatically by LEQG problems (see our [46, sect. 3]), we will show next that our PO methods can find the global optimum K^* even without this assumption.

3.2. PO algorithms. Consider three PG-based methods as follows. For simplicity, we define

$$E_K := (R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A. \tag{3.2}$$

We also suppress the iteration index, and use K and K' to represent the control gain

before and after one-step of the update.

$$K' = K - \eta \nabla \mathcal{J}(K) = K - 2\eta E_K \Delta_K, \tag{3.3}$$

PG:

$$K' = K - \eta \nabla \mathcal{J}(K) \Delta_K^{-1} = K - 2\eta E_K, \tag{3.4}$$

Natural PG:

$$K' = K - \eta (R + B^\top \widetilde{P}_K B)^{-1} \nabla \mathcal{J}(K) \Delta_K^{-1} = K - 2\eta (R + B^\top \widetilde{P}_K B)^{-1} E_K, \tag{3.5}$$

Gauss–Newton:

where $\eta > 0$ is the step size. The updates are motivated by the PO updates for solving LQR problems [10, 15], but with P_K therein replaced by $\widetilde{P}_K$. The natural PG update is related to the gradient over a Riemannian manifold, while the Gauss–Newton update is one type of quasi-Newton update; see [10] for further justifications on the updates. In particular, with $\eta = 1/2$, the Gauss–Newton update (3.5) can be viewed as the *policy iteration* update for infinite-horizon mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design. To enable a model-free RL update rule, the (natural) PG directions can be estimated by sampling the trajectories of the system, as well as the cost functions, without estimating system parameters; see examples in [15, 29, 47]. More discussions on the model-free versions of these updates can be found in [46, sects. 3 and 6].

4. Theoretical results. In this section, we study the convergence of the PO methods proposed in section 3.

4.1. Implicit regularization. The first key challenge in the convergence analysis for PO methods, is to ensure that the iterates remain *feasible* as the algorithms proceed, hopefully without the use of *projection*. This is especially significant in mixed design problems, as the feasibility here means *robust stability*, the violation of which can be catastrophic in practical *online* control design. We formally define the concept of *implicit regularization* to describe this feature.

DEFINITION 4.1 (implicit regularization). *For mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design problem* (2.10)*, suppose an iterative algorithm generates a sequence of control gains $\{K_n\}$. If $K_n \in \mathcal{K}$ for all $n \geq 0$, this algorithm is called* regularized*; if it is regularized without projection onto $\mathcal{K}$ for any $n \geq 0$, this algorithm is called* implicitly regularized.

One possible way for the iterates to remain feasible is to keep shrinking the step size, whenever the next iterate goes outside $\mathcal{K}$, following for example the Armijo rule. However, as the cost $\mathcal{J}(K)$ is not necessarily smooth (see Lemma 5.1 and its discussion later), it may not converge within a finite number of iterations [20, Theorem 3.2]. Another option is to project the iterate onto $\mathcal{K}$. Nonetheless, it is challenging to perform projection onto the $\mathcal{H}_\infty$-norm constraint set in a data driven manner.

For LQR problems, due to the coercivity of the cost that as K approaches the boundary of the stability/feasibility region $\{K \in \mathbb{R}^{d\times m} \,|\, \rho(A - BK) < 1\}$, i.e., as $\rho(A - BK) \to 1$, the cost blows up to infinity, and due to the fact that the cost is continuous with respect to K, the lower-level set of the cost is compact [28] and is contained within the stability region. Hence, there is a strict separation between any lower-level set of the cost and the set $\{K \in \mathbb{R}^{d\times m} \,|\, \rho(A - BK) \geq 1\}$. There thus exists a *constant step size* such that as long as the initialization control is stabilizing, the iterates along the path remain stabilizing and keep decreasing the cost.

In contrast, for mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problems, lack of coercivity invalidates the argument above, as the control approaching the robustness constraint boundary $\partial\mathcal{K}$ may incur a finite cost, and the descent direction may still drive the iterates out of the feasibility region. In addition, there may not exist a strict separation between all the

lower-level sets of the cost and the complementary set $\mathcal{K}^c$. This difficulty has been illustrated in Figure 1 in the introduction. Interestingly, we show in the following theorem that the natural PG and Gauss–Newton methods in (3.4)–(3.5) enjoy the implicit regularization feature, with certain *constant* step sizes.

THEOREM 4.2 (implicit regularization property). *For any control gain $K \in \mathcal{K}$ with $\|K\| < \infty$, suppose that the step size η satisfies the following:*

- *Natural PG* (3.4)*: $\eta \le 1/(2\|R + B^\top \widetilde{P}_K B\|)$;*
- *Gauss–Newton* (3.5)*: $\eta \le 1/2$.*

Then the K' obtained from (3.4)–(3.5) *also lies in $\mathcal{K}$. Equivalently, K' is stabilizing, i.e., $\rho(A-BK') < 1$, and satisfies that* (i) *there exists a solution $P_{K'} \ge 0$ to the Riccati equation* (2.8)*;* (ii) $I-\gamma^{-2}D^\top P_{K'}D > 0$*;* (iii) $\rho\big((I-\gamma^{-2}P_{K'}DD^\top)^{-\top}(A-BK')\big) < 1$.

Proof sketch. The general idea, contrary to the coercivity-based idea that works for any descent direction, is that we focus on the feasibility of K' after an update along *certain directions*: either (3.4) or (3.5). By the bounded real lemma, i.e., Lemma 2.2, the feasibility condition for K', if K' is stabilizing, is equivalent to the existence of $P > 0$ such that the linear matrix inequalities (LMIs) in (2.11) hold for K'. Moreover, it is straightforward to see that such a $P > 0$, if it exists, satisfies $(A - BK')^\top P(A - BK') - P < 0$, which can be used to show that K' is stabilizing [7]. Thus, it now suffices to find such a P.

To show this, we first study the case with step sizes being the upper bound in the theorem, i.e., $\eta = 1/2$ for Gauss–Newton and $\eta = 1/(2\|R + B^\top \widetilde{P}_K B\|)$ for natural PG. As the solution to the Riccati equation (2.8) under K, $P_K \ge 0$ satisfies $I - \gamma^{-2}D^\top P_K D > 0$, the first LMI in (2.11). Hence, it may be possible to perturb P_K to obtain a $P > 0$ such that the equality in (2.8) becomes a strict inequality of the second LMI in (2.11), while preserving the first LMI. Moreover, if K' is not too far away from K, such a perturbed P_K should also work for K'. Such an observation motivates the use of P_K as the candidate of P for the LMIs in (2.11) under K'.

Indeed, it can be shown that substituting $P = P_K$ makes the second LMI in (2.11) under K' *nonstrict*, namely, the left-hand side (LHS) ≤ 0; see (5.5) in the detailed proof. To make it strict, consider the perturbed $P = P_K + \alpha\bar{P}$ for some $\alpha > 0$, where $\bar{P} > 0$ is the solution to some Lyapunov equation

$$(A - BK)^\top (I - \gamma^{-2}DD^\top P_K)^{-\top} \bar{P} (I - \gamma^{-2}DD^\top P_K)^{-1}(A - BK) - \bar{P} = -I. \tag{4.1}$$

Such a Lyapunov equation (4.1) always admits a solution $\bar{P} > 0$, since $K \in \mathcal{K}$ implies that $(I - \gamma^{-2}DD^\top P_K)^{-1}(A - BK)$ is stable. The intuition of choosing (4.1) is as follows. First, the LHS of the second LMI in (2.11) under K' can be separated as

(4.2)

$$\begin{aligned}
&(A - BK')^\top \widetilde{P}(A - BK') - P + C^\top C + K'^\top R K' \\
&\quad = \underbrace{[(A - BK')^\top \widetilde{P}(A - BK') - (A - BK)^\top \widetilde{P}(A - BK)] + K'^\top R K' - K^\top R K}_{①} \\
&\qquad + \underbrace{(A - BK)^\top \widetilde{P}(A - BK) - P + C^\top C + K^\top R K}_{②}.
\end{aligned}$$

By some algebra, the first term ① is of order $o(\alpha)$. Since for small α,

$$\widetilde{P} = \widetilde{P}_K + (I - \gamma^{-2} P_K D D^\top)^{-1}(\alpha \bar{P})(I - \gamma^{-2} D D^\top P_K)^{-1} + o(\alpha),$$

this, combined with the Riccati equation (2.8) and (4.1), makes the second term ② $= -\alpha I + o(\alpha)$. Hence, there exists small enough $\alpha > 0$ such that ① + ② < 0, ensuring that the updated K' is feasible. Last, by the linearity of LMIs, any interpolation of K' with a smaller step size is also feasible/robustly stable, completing the proof. ∎

The detailed proof of Theorem 4.2 is deferred to section 5.1. Recall from Lemma 3.2 that for the other two objectives (2.5) and (2.7), the only difference of the PG form is in the form of Δ_K. Notice that our update rules of natural PG and Gauss–Newton are independent of this Δ_K. Hence, the implicit regularization results in Theorem 4.2, in terms of ensuring $K \in \mathcal{K}$, also apply to the other objectives. Note that theoretically, it is not clear yet if vanilla PG enjoys implicit regularization. It is possible that vanilla PG does not preserve the robustness constraint. Hence, hereafter, we only focus on the global convergence of natural PG method (3.4) and Gauss–Newton method (3.5).

4.2. Global convergence. The term *global convergence* here refers to two notions: (i) the convergence performance of the algorithms starting from *any feasible initialization* point $K_0 \in \mathcal{K}$; (ii) convergence to the global optimal policy under certain conditions. We formally establish the results for the natural PG (3.4) and Gauss–Newton (3.5) updates in the following theorem.

THEOREM 4.3 (global sublinear convergence). *Suppose that $K_0 \in \mathcal{K}$ and $\|K_0\| < \infty$. Then, under the step size choices*[4] *as in Theorem* 4.2*, both updates* (3.4) *and* (3.5) *converge to the global optimum $K^* = (R + B^\top \widetilde{P}_{K^*} B)^{-1} B^\top \widetilde{P}_{K^*} A$ in that the average of $\{\|E_{K_n}\|_F^2\}$ over iterations converges to zero with $O(1/N)$ rate.*

The proof of Theorem 4.3 is detailed in section 5.2. The key idea is to first quantify the difference in P_K (both its upper and lower bounds) for any two control gain matrices K and K' (cf. Lemma 5.1). Essentially it delivers some "almost smoothness" of P_K: the leading terms in both upper and lower bounds depend on $\|K' - K\|$ with the remaining terms being on the order of $o(\|K' - K\|)$ if K' is close to K. Then, by substituting two consecutive iterates K', K' and the upper bound in Lemma 5.1, the sublinear global convergence follows similarly to the analysis of smooth nonconvex optimization using the descent lemma [11, 33]. Note that the luxury of conducting this analysis is attributed to the implicit regularization result in Theorem 4.2, which already ensures that the next iterate is always robustly stable. This is the key difference from any smooth nonconvex optimization and LQR analysis. Similarly to the discussion after Theorem 4.2, the global convergence results also apply to the other two objectives (2.5) and (2.7), as the update rules natural PG and Gauss–Newton, as well as the performance metric (the average of $\{\|E_{K_n}\|_F^2\}$), are all independent of Δ_K.

We note that the controllability assumption made in Proposition 3.3 is not required for the global convergence here. Remarkably, there might be multiple stationary points such that $\nabla \mathcal{J}(K) = 0$, while the two specific search directions (3.4) and (3.5) provably avoid the other local minima, and always converge to the global optimum K^* that minimizes P_K in the positive semidefinite sense. This can be viewed as another implication of implicit regularization, in that (3.4) and (3.5) always bias the iterates towards a certain global optimal solution, without getting stuck at other stationary points that are possibly suboptimal. The key reason is that, without using

[4] In fact, for natural PG (3.4), it suffices to require $\eta \le 1/(2\|R + B^\top \widetilde{P}_{K_0} B\|)$ for the initial K_0.

the curvature information in Δ_K, these two PO methods can converge to the specific and optimal stationary point such that $E_K = 0$, instead of an arbitrary one. In contrast to the results for LQR [15], only globally *sublinear* $O(1/N)$, instead of *linear*, convergence rates can be obtained so far. This $O(1/N)$ rate of the (iteration average) gradient norm square matches the *global* convergence rate of gradient descent and second order algorithms to stationary points for nonconvex problems [11].

We can obtain faster local (super)linear rates, and this is formalized as follows.

THEOREM 4.4 (local (super)linear convergence). *Suppose that the conditions in Theorem* 4.3 *hold, and additionally* $DD^\top > 0$ *holds. Then, under the step size choices as in Theorem* 4.3*, both updates* (3.4) *and* (3.5) *converge to the optimal control gain* K^* *with* locally linear *rate, in the sense that the objective* $\{\mathcal{J}(K_n)\}$ *defined in* (2.6) *converges to* $\mathcal{J}(K^*)$ *with a linear rate. In addition, if* $\eta = 1/2$*, the Gauss–Newton update* (3.5) *converges to* K^* *with a locally* Q-quadratic *rate.*

Proof of the above theorem is deferred to section 5.3. Key to the locally faster rates is that the property of *gradient dominance* [33] holds locally around the global optimum, which can be proved by substituting the lower bound of two consecutive P_K's in Lemma 5.1. Note that this lower bound only holds locally. The gradient dominance property has been shown to hold globally for LQR problems [15], and also hold locally for zero-sum LQ games [47]. The Q-quadratic rate mirrors the rate of Gauss–Newton with $\eta = 1/2$ (essentially policy iteration) for LQR problems [10, 19].

5. Proofs of the main results. Now we provide proofs for the main results.

5.1. Proof of Theorem 4.2. To show that K' lies in $\mathcal{K}$, we first argue that it suffices to find some $P > 0$ such that

$$I - \gamma^{-2} D^\top P D > 0, \qquad \text{and} \tag{5.1}$$

$$(A - BK')^\top \widetilde{P}(A - BK') - P + C^\top C + (K')^\top R K' < 0, \tag{5.2}$$

where $\widetilde{P} := P + PD(\gamma^2 I - D^\top P D)^{-1} D^\top P$. By the Schur complement, showing (5.1)–(5.2) is also equivalent to showing

$$\begin{bmatrix} (A - BK')^\top P (A - BK') - P + C^\top C + K'^\top R K' & (A - BK')^\top P D \\ D^\top P (A - BK') & -(\gamma^2 I - D^\top P D) \end{bmatrix} < 0. \tag{5.3}$$

We denote the LHS of (5.2) by $-M$. Thus, (5.1) and (5.2) imply

$$\begin{aligned}(A - BK')^\top P (A - BK') - P = -M - C^\top C - (K')^\top R K' - (A - BK')^\top P D \\ \cdot (\gamma^2 I - D^\top P D)^{-1} D^\top P (A - BK') \le -M < 0,\end{aligned}$$

which shows that K' is stabilizing, i.e., $\rho(A - BK') < 1$ [7]. Thus, Lemma 2.2 can be applied to K'. Then, (5.1) and (5.2) are identical to (2.11), which further shows that $\|\mathcal{T}(K')\|_\infty < \gamma$ and hence $K' \in \mathcal{K}$. Hereafter we will focus on finding such a $P > 0$.

We first find such a P for the Gauss–Newton update (3.5) with step size $\eta = 1/2$. Specifically, we have

$$K' = K - (R + B^\top \widetilde{P}_K B)^{-1} E_K = (R + B^\top \widetilde{P}_K B)^{-1} B^\top \widetilde{P}_K A. \tag{5.4}$$

Since $P_K \ge 0$ satisfies the conditions (i)–(iii), and K and K' are close to each other, we can choose P_K as a candidate for P. Hence, by the Riccati equation (2.8), the

LHS of (5.2) can be written as

(5.5)

$$\begin{aligned}
&[A - B(R + B^\top \widetilde{P}_K B)^{-1} B^\top \widetilde{P}_K A]^\top \widetilde{P}_K [A - B(R + B^\top \widetilde{P}_K B)^{-1} B^\top \widetilde{P}_K A] - P_K + C^\top C \\
&\quad + [(R + B^\top \widetilde{P}_K B)^{-1} B^\top \widetilde{P}_K A]^\top R [(R + B^\top \widetilde{P}_K B)^{-1} B^\top \widetilde{P}_K A] \\
&= -\big[(R + B^\top \widetilde{P}_K B)^{-1} B^\top \widetilde{P}_K A - K\big]^\top \\
&\quad \cdot (R + B^\top \widetilde{P}_K B)\big[(R + B^\top \widetilde{P}_K B)^{-1} B^\top \widetilde{P}_K A - K\big] \le 0,
\end{aligned}$$

where we substitute K' from (5.4), and the last equation is by completing the squares.

Now we need to perturb P_K to obtain a P such that (5.5) holds with a *strict* inequality. To this end, we define $\bar{P} > 0$ as the solution to the Lyapunov equation

$$(A - BK)^\top (I - \gamma^{-2} DD^\top P_K)^{-\top} \bar{P} (I - \gamma^{-2} DD^\top P_K)^{-1} (A - BK) - \bar{P} = -I, \tag{5.6}$$

and let $P = P_K + \alpha \bar{P} > 0$ for some $\alpha > 0$. By Lemma 2.2, $(I - \gamma^{-2} DD^\top P_K)^{-1}(A - BK)$ is stable, and thus the solution $\bar{P} > 0$ exists. For (5.1) to hold, we need a small $\alpha > 0$ to satisfy

$$\alpha D^\top \bar{P} D < \gamma^2 I - D^\top P_K D. \tag{5.7}$$

Also, the LHS of (5.2) can be written as in (4.2), with ①, ② being defined therein. We aim to find some $\alpha > 0$ such that ① + ② < 0. Note that $\widetilde{P}$ can be written as

(5.8)

$$\begin{aligned}
\widetilde{P} &= [I - \gamma^{-2}(P_K + \alpha \bar{P}) DD^\top]^{-1} (P_K + \alpha \bar{P}) \\
&= \big[(I - \gamma^{-2} P_K DD^\top)^{-1} + (I - \gamma^{-2} P_K DD^\top)^{-1} (\alpha \gamma^{-2} \bar{P} DD^\top)(I - \gamma^{-2} P_K DD^\top)^{-1} \\
&\quad + o(\alpha)\big](P_K + \alpha \bar{P}) = \widetilde{P}_K + (I - \gamma^{-2} P_K DD^\top)^{-1} (\alpha \bar{P})(I - \gamma^{-2} DD^\top P_K)^{-1} + o(\alpha),
\end{aligned}$$

where the first equation follows by definition, and the second one uses the fact that $(X + Y)^{-1} = X^{-1} - X^{-1} Y X^{-1} + o(\|Y\|)$, for a small perturbation Y around the matrix X. Thus, ① can be written as

(5.9)

$$\begin{aligned}
① &= -K'^\top B^\top \widetilde{P} A - A^\top \widetilde{P} B K' + K'^\top (R + B^\top \widetilde{P} B) K' + K^\top B^\top \widetilde{P} A + A^\top \widetilde{P} B K \\
&\quad - K^\top (R + B^\top \widetilde{P} B) K \\
&\le -K'^\top B^\top \widetilde{P} A - A^\top \widetilde{P} B K' + K'^\top (R + B^\top \widetilde{P} B) K' + A^\top \widetilde{P} B (R + B^\top \widetilde{P} B)^{-1} B^\top \widetilde{P} A \\
&= \big[(R + B^\top \widetilde{P} B)^{-1} B^\top \widetilde{P} A - K'\big]^\top (R + B^\top \widetilde{P} B)\big[(R + B^\top \widetilde{P} B)^{-1} B^\top \widetilde{P} A - K'\big],
\end{aligned}$$

where the inequality follows by completing squares. By substituting in K' from (5.4), we further have

$$\begin{aligned}
① \le &\big[(R + B^\top \widetilde{P} B)^{-1} B^\top \widetilde{P} A - (R + B^\top \widetilde{P}_K B)^{-1} B^\top \widetilde{P}_K A\big]^\top (R + B^\top \widetilde{P} B) \\
&\cdot \big[\underbrace{(R + B^\top \widetilde{P} B)^{-1} B^\top \widetilde{P} A}_{③} - (R + B^\top \widetilde{P}_K B)^{-1} B^\top \widetilde{P}_K A\big].
\end{aligned} \tag{5.10}$$

Note that by (5.8), we have

$$\begin{aligned}③ &= \big[(R+B^\top \widetilde{P}_K B)^{-1} - (R+B^\top \widetilde{P}_K B)^{-1}B^\top (I-\gamma^{-2}P_K DD^\top)^{-1}(\alpha\bar{P})(I-\gamma^{-2}DD^\top P_K)^{-1}\\ &\quad \cdot B(R+B^\top \widetilde{P}_K B)^{-1} + o(\alpha)\big]B^\top\big[\widetilde{P}_K + (I-\gamma^{-2}P_K DD^\top)^{-1}(\alpha\bar{P})(I-\gamma^{-2}DD^\top P_K)^{-1}\\ &\quad + o(\alpha)\big]A = (R+B^\top \widetilde{P}_K B)^{-1}B^\top \widetilde{P}_K A + O(\alpha) + o(\alpha),\end{aligned}$$

which can be combined with (5.10) to show ① $= o(\alpha)$.

Moreover, by (5.8), ② can be written as

$$\begin{aligned}e② &= (A-BK)^\top\big[\widetilde{P}_K + (I-\gamma^{-2}P_K DD^\top)^{-1}(\alpha\bar{P})(I-\gamma^{-2}DD^\top P_K)^{-1}\\ &\quad + o(\alpha)\big](A-BK) - P + C^\top C + K^\top RK\\ &= (A-BK)^\top (I-\gamma^{-2}P_K DD^\top)^{-1}(\alpha\bar{P})(I-\gamma^{-2}DD^\top P_K)^{-1}\\ &\quad \cdot (A-BK) - \alpha\bar{P} + o(\alpha)\\ &= -\alpha I + o(\alpha),\end{aligned} \tag{5.11}$$

where the first equation uses (5.8), the second one uses the Riccati equation (2.8), and the last one uses (5.6). Therefore, for small enough $\alpha > 0$ such that ① + ② < 0, and it also satisfies (5.7), there exists some $P > 0$ such that both (5.1) and (5.2) hold for K' obtained with step size $\eta = 1/2$. On the other hand, such a P also makes the LMI (5.3) hold for K, i.e.,

$$\begin{bmatrix} (A-BK)^\top P(A-BK) - P + C^\top C + K^\top RK & (A-BK)^\top PD \\ D^\top P(A-BK) & -(\gamma^2 I - D^\top PD)\end{bmatrix} < 0, \tag{5.12}$$

as now ① in (4.2) is null, and the same α above makes ② < 0. For $\eta \in [0, 1/2]$, define $K_\eta = K + 2\eta(K'-K)$. By linearity, convexity of the quadratic form, and combining (5.3) and (5.12), one can show that (5.12) also holds for K_η for any $\eta \in [0, 1/2]$. Thus, K_η satisfies the conditions (i)–(iii) in the theorem.

Now we prove a similar result for the natural PG update (3.4). Recall that

$$K' = K - 2\eta[(R+B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A]. \tag{5.13}$$

As before, we first choose $P = P_K$. Then, the LHS of (5.2) under K' is written as

$$\begin{aligned}&(A-BK')^\top \widetilde{P}_K(A-BK') - P_K + C^\top C + K'^\top RK'\\ &\quad = (K'-K)^\top (R+B^\top \widetilde{P}_K B)[K' - (R+B^\top \widetilde{P}_K B)^{-1}B^\top \widetilde{P}_K A]\\ &\quad\quad + [K - (R+B^\top \widetilde{P}_K B)^{-1}B^\top \widetilde{P}_K A]^\top (R+B^\top \widetilde{P}_K B)(K'-K),\end{aligned} \tag{5.14}$$

where the equation holds by adding and subtracting $[K-(R+B^\top \widetilde{P}_K B)^{-1}B^\top \widetilde{P}_K A]^\top (R+B^\top \widetilde{P}_K B)[K' - (R+B^\top \widetilde{P}_K B)^{-1}B^\top \widetilde{P}_K A]$. Substituting (5.13) into (5.14) yields

$$\begin{aligned}&(A-BK')^\top \widetilde{P}_K(A-BK') - P_K + C^\top C + K'^\top RK'\\ &\quad = -2\eta[(R+B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A]^\top [(R+B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A]\\ &\quad\quad + 4\eta^2[(R+B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A]^\top (R+B^\top \widetilde{P}_K B)[(R+B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A]\\ &\quad\quad - 2\eta[(R+B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A]^\top [(R+B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A].\end{aligned} \tag{5.15}$$

By requiring the step size η to satisfy

$$\eta \le \frac{1}{2\|R + B^\top \widetilde{P}_K B\|}, \tag{5.16}$$

we can bound (5.15) as

$$\begin{aligned}&(A - BK')^\top \widetilde{P}_K (A - BK') - P_K + C^\top C + K'^\top R K' \\ &\le -2\eta \cdot [(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A]^\top [(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A] \le 0,\end{aligned} \tag{5.17}$$

namely, letting $P = P_K$ leads to the desired LMI that is not strict.

Now suppose that $P = P_K + \alpha \bar{P}$ for some $\alpha > 0$, where $\bar{P} > 0$ is the solution to (5.6). Note that α still needs to satisfy (5.7). Also, the LHS of (5.2) can be separated into ① and ② as in (4.2). From the LHS of the inequality in (5.9), we have

$$\begin{aligned}① &= -2\eta[(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A]^\top [(R + B^\top \widetilde{P} B)K' - B^\top \widetilde{P} A] \\ &\quad - 2\eta[(R + B^\top \widetilde{P} B)K - B^\top \widetilde{P} A]^\top [(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A] \\ &\le -2\eta[(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A]^\top [(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A] \\ &\quad - 2\alpha\eta[(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A]^\top [(B^\top \bar{P} B)K' - B^\top \bar{P} A] \\ &\quad - 2\alpha\eta[(B^\top \bar{P} B)K - B^\top \bar{P} A]^\top [(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A],\end{aligned}$$

where the first step follows by adding and subtracting $[(R + B^\top \widetilde{P} B)^{-1} B^\top \widetilde{P} A - K]^\top (R + B^\top \widetilde{P} B)[(R + B^\top \widetilde{P} B)^{-1} B^\top \widetilde{P} A - K']$, and the definition of K' in (5.13). The second step follows by separating $\widetilde{P}$ as $\widetilde{P}_K + \alpha \bar{P}$ in a similar manner, and by using (5.16), (5.15), and (5.17). Moreover, notice that

$$\begin{aligned}&- 2\alpha\eta[(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A]^\top [(B^\top \bar{P} B)K' - B^\top \bar{P} A] - 2\alpha\eta[(B^\top \bar{P} B)K - B^\top \bar{P} A]^\top \\ &\quad \cdot [(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A] \\ &\le 2\eta[(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A]^\top [(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A] + 2\alpha^2\eta[(B^\top \bar{P} B)K' \\ &\quad - B^\top \bar{P} A]^\top [(B^\top \bar{P} B)K' - B^\top \bar{P} A] - 4\alpha\eta^2[(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A]^\top (B^\top \bar{P} B) \\ &\quad \cdot [(R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A].\end{aligned}$$

Therefore, we can finally show

$$① \le 2\alpha^2\eta[(B^\top \bar{P} B)K' - B^\top \bar{P} A]^\top [(B^\top \bar{P} B)K' - B^\top \bar{P} A]. \tag{5.18}$$

By assumption $\|K\| < \infty$ and $P_K \ge 0$ exists; we know that $\widetilde{P}_K$ is bounded, and so is K' obtained from (5.13) using a finite step size η. Also, $\bar{P}$ has a bounded norm. Thus, ① in (5.18) is $o(\alpha)$ and there exists a small enough $\alpha > 0$ such that ①+② < 0, since from (5.11), ② $= -\alpha I + o(\alpha)$. In other words, there exists some $P > 0$ such that (5.3) holds for K' obtained from (5.13) with step size satisfying (5.16). This completes the proof of the first argument. Last, by Lemma 2.2, we equivalently have that the conditions (i)–(iii) in the theorem hold for K', which completes the proof. ∎

5.2. Proof of Theorem 4.3. We first introduce the following lemma that can be viewed as the counterpart of the *cost difference lemma* in [15]. Unlike the equality relation given in the lemma in [15], we establish both lower and upper bounds for the difference of two matrices $P_{K'}$ and P_K. The proof of the lemma is provided in section A.5.

LEMMA 5.1 (cost difference lemma). *Suppose that both $K, K' \in \mathcal{K}$. Then, we have the following upper bound:*

$$P_{K'} - P_K \le \sum_{t\ge 0}[(A-BK')^\top (I-\gamma^{-2}P_{K'}DD^\top)^{-1}]^t\big[-(K-K')^\top E_K - E_K^\top (K-K') + (K-K')^\top (R+B^\top \widetilde{P}_K B)(K-K')\big][(I-\gamma^{-2}P_{K'}DD^\top)^{-\top}(A-BK')]^t, \tag{5.19}$$

where E_K is defined in (3.2). *If additionally $\rho\big((A-BK')^\top (I-\gamma^{-2}P_K DD^\top)^{-1}\big) < 1$, then we also have the lower bound:*

$$P_{K'} - P_K \ge \sum_{t\ge 0}[(A-BK')^\top (I-\gamma^{-2}P_K DD^\top)^{-1}]^t\big[-(K-K')^\top E_K - E_K^\top (K-K') + (K-K')^\top (R+B^\top \widetilde{P}_K B)(K-K')\big][(I-\gamma^{-2}P_K DD^\top)^{-\top}(A-BK')]^t. \tag{5.20}$$

Now we show the global convergence of two PO updates.

Gauss–Newton: Recall that for the Gauss–Newton update, $K' = K - 2\eta(R + B^\top \widetilde{P}_K B)^{-1}E_K$. By Theorem 4.2, K' also lies in $\mathcal{K}$ if $\eta \le 1/2$. Then, by the upper bound in (5.19), we know that if $\eta \in [0, 1/2]$,

$$P_{K'} - P_K \le (-4\eta + 4\eta^2)\sum_{t\ge 0}[(A-BK')^\top (I-\gamma^{-2}P_{K'}DD^\top)^{-1}]^t \cdot \Big[E_K^\top (R+B^\top \widetilde{P}_K B)^{-1}E_K\Big][(I-\gamma^{-2}P_{K'}DD^\top)^{-\top}(A-BK')]^t \le 0, \tag{5.21}$$

which implies the monotonic decrease of P_K (matrix wise) along the update. Since P_K is lower bounded, such a monotonic sequence of $\{P_{K_n}\}$ along the iterations must converge to some $P_{K_\infty} \in \mathcal{K}$. Now we show that this P_{K_∞} is indeed P_{K^*}. By multiplying both sides of (5.21) with any matrix $M > 0$, and then taking the trace, we have that if $\eta \in [0, 1/2]$,

$$\begin{aligned}
\operatorname{Tr}(P_{K'}M) - \operatorname{Tr}(P_K M) &\le (-4\eta + 4\eta^2)\operatorname{Tr}\Big\{\sum_{t\ge 0}[(A-BK')^\top (I-\gamma^{-2}P_{K'}M)^{-1}]^t \\
&\quad \cdot \Big[E_K^\top (R+B^\top \widetilde{P}_K B)^{-1}E_K\Big][(I-\gamma^{-2}P_{K'}M)^{-\top}(A-BK')]^t M\Big\} \\
&\le -2\eta \operatorname{Tr}\big[E_K^\top (R+B^\top \widetilde{P}_K B)^{-1}E_K M\big] \\
&\le \frac{-2\eta\sigma_{\min}(M)}{\sigma_{\max}(R+B^\top \widetilde{P}_K B)}\operatorname{Tr}(E_K^\top E_K) \\
&\le \frac{-2\eta\sigma_{\min}(M)}{\sigma_{\max}(R+B^\top \widetilde{P}_{K_0} B)}\operatorname{Tr}(E_K^\top E_K),
\end{aligned} \tag{5.22}$$

where the second inequality follows by keeping only the first term in the infinite summation of positive definite matrices, the third one uses that $\operatorname{Tr}(PA) \ge \sigma_{\min}(A)\operatorname{Tr}(P)$, and the last one is due to the monotonic decrease of P_K, and the monotonicity of $\widetilde{P}_K$ with respect to P_K with $K_0 \in \mathcal{K}$ being the initialization of K at iteration 0. From iterations $n = 0$ to $N-1$, replacing M by I, summing over both sides of (5.22), and dividing by N, we further have

$$\frac{1}{N}\sum_{n=0}^{N-1}\operatorname{Tr}(E_{K_n}^\top E_{K_n}) \le \frac{\sigma_{\max}(R+B^\top \widetilde{P}_{K_0} B)\cdot\big[\operatorname{Tr}(P_{K_0}) - \operatorname{Tr}(P_{K_\infty})\big]}{2\eta\cdot N},$$

namely, the sequence $\{K_n\}$ converges to the stationary point K such that $E_K = 0$ with $O(1/N)$ rate. By Proposition 3.3, this stationary point corresponds to one global optimum K^*.

Natural PG: Recall that the natural PG update follows $K' = K - 2\eta E_K$. By Theorem 4.2, K' also lies in $\mathcal{K}$ if $\eta \leq 1/(2\|R + B^\top \widetilde{P}_K B\|)$. By the upper bound (5.19), this step size yields that

$$P_{K'} - P_K \leq \sum_{t\geq 0}[(A - BK')^\top (I - \gamma^{-2} P_{K'} DD^\top)^{-1}]^t \left[-4\eta E_K^\top E_K + 2\eta E_K^\top E_K\right]$$
$$\cdot [(I - \gamma^{-2} P_{K'} DD^\top)^{-\top} (A - BK')]^t \leq 0, \tag{5.23}$$

which also implies the matrixwise monotonic decrease of P_K along the update. Suppose the convergent matrix is P_{K_∞}. As before, multiplying both sides of (5.23) by $M > 0$, and taking the trace, yields

$$\mathrm{Tr}(P_{K'} M) - \mathrm{Tr}(P_K M) \leq -2\eta \,\mathrm{Tr}\left(E_K^\top E_K M\right) \tag{5.24}$$

for any $M > 0$, where the inequality follows by only keeping the first term in the infinite summation. Letting $M = I$, summing up (5.24) from $n = 0$ to $n = N - 1$, and dividing by N, we conclude that

$$\frac{1}{N}\sum_{n=0}^{N-1} \mathrm{Tr}(E_{K_n}^\top E_{K_n}) \leq \frac{\mathrm{Tr}(P_{K_0}) - \mathrm{Tr}(P_{K_\infty})}{2\eta \cdot N},$$

namely, $\{K_n\}$ converges to the stationary point K such that $E_K = 0$ with an $O(1/N)$ rate, which is also the global optimum. In addition, since $\{P_{K_n}\}$ is monotonically decreasing, it suffices to require the step size $\eta \in [0, 1/(2\|R + B^\top \widetilde{P}_{K_0} B\|)]$. ∎

5.3. Proof of Theorem 4.4. To ease the analysis, we show the convergence rate of a surrogate value $\mathrm{Tr}(P_K DD^\top)$. This is built upon the following relationship between the objective value $\mathcal{J}(K)$ and $\mathrm{Tr}(P_K DD^\top)$.

LEMMA 5.2. *Suppose that both $K, K' \in \mathcal{K}$, and $P_K \geq P_{K'}$. Then, it follows that*

$$\mathcal{J}(K) - \mathcal{J}(K') \leq \|(I - \gamma^{-2} D^\top P_K D)^{-1}\| \cdot [\mathrm{Tr}(P_K DD^\top) - \mathrm{Tr}(P_{K'} DD^\top)].$$

Proof. First, by Sylvester's determinant theorem, $\mathcal{J}(K)$ can be rewritten as

$$\mathcal{J}(K) = -\gamma^2 \log\det(I - \gamma^{-2} P_K DD^\top) = -\gamma^2 \log\det(I - \gamma^{-2} D^\top P_K D).$$

By the mean value theorem, for any (A, B) with $\det(A), \det(B) > 0$, we have $\log\det(A) = \log\det(B) + \mathrm{Tr}[(B + \tau(A - B))^{-1}(A - B)]$ for some $0 \leq \tau \leq 1$. This leads to

$$\begin{aligned}\mathcal{J}(K) - \mathcal{J}(K') &= -\gamma^2 \log\det(I - \gamma^{-2} D^\top P_K D) + \gamma^2 \log\det(I - \gamma^{-2} D^\top P_{K'} D)\\ &= \mathrm{Tr}[X D^\top (P_K - P_{K'}) D]\\ &\leq \|X\| \cdot [\mathrm{Tr}(D^\top P_K D) - \mathrm{Tr}(D^\top P_{K'} D)] = \|X\| \cdot [\mathrm{Tr}(P_K DD^\top) - \mathrm{Tr}(P_{K'} DD^\top)],\end{aligned}$$

where $X = (I - \gamma^{-2}\tau D^\top P_{K'} D - \gamma^{-2}(1 - \tau) D^\top P_K D)^{-1}$, and the inequality uses the facts $P_K \geq P_{K'}$ and $\mathrm{Tr}(PA) \leq \|A\| \cdot \mathrm{Tr}(P)$ for any real symmetric $P \geq 0$. Note that by $P_K \geq P_{K'}$,

$$X \leq (I - \gamma^{-2} D^\top P_K D)^{-1} \Longrightarrow \|X\| \leq \|(I - \gamma^{-2} D^\top P_K D)^{-1}\|.$$

This completes the proof. ∎

Lemma 5.2 implies that in order to show the convergence of $\mathcal{J}(K)$, it suffices to study the convergence of $\mathrm{Tr}(P_K DD^\top)$, as long as $\|(I-\gamma^{-2}D^\top P_K D)^{-1}\|$ is bounded along the iterations. This is indeed the case since by (5.21) and (5.23), P_K is monotone along both updates (3.4) and (3.5). By induction, if $K_0 \in \mathcal{K}$, i.e., $I-\gamma^{-2}D^\top P_{K_0} D > 0$, then $I-\gamma^{-2}D^\top P_{K_n} D \geq I-\gamma^{-2}D^\top P_{K_0} D > 0$ holds for all iterations $n \geq 1$. This further yields that for all $n \geq 1$, $\|(I-\gamma^{-2}D^\top P_{K_n} D)^{-1}\| \leq \|(I-\gamma^{-2}D^\top P_{K_0} D)^{-1}\|$, namely, $\|(I-\gamma^{-2}D^\top P_K D)^{-1}\|$ is uniformly bounded.

Now we show the local linear convergence rate of $\mathrm{Tr}(P_K DD^\top)$. By (5.20), for any K' such that $(I-\gamma^{-2}P_K DD^\top)^{-\top}(A-BK')$ is stabilizing, we have

$$P_{K'} - P_K \geq \sum_{t\geq 0}[(A-BK')^\top (I-\gamma^{-2}P_K DD^\top)^{-1}]^t \big[-E_K^\top (R+B^\top \widetilde{P}_K B)^{-1} E_K \big]$$

$$\cdot [(I-\gamma^{-2}P_K DD^\top)^{-\top}(A-BK')]^t, \tag{5.25}$$

where the inequality follows from completion of squares. By taking traces on both sides of (5.25), and letting $K' = K^*$, we have

$$\begin{aligned}\mathrm{Tr}(P_K DD^\top) - \mathrm{Tr}(P_{K^*} DD^\top) &\leq \mathrm{Tr}\left[E_K^\top (R+B^\top \widetilde{P}_K B)^{-1} E_K\right] \cdot \|\mathcal{W}_{K,K^*}\| \\ &\leq \frac{\mathrm{Tr}\left(E_K^\top E_K\right)}{\sigma_{\min}(R)} \cdot \|\mathcal{W}_{K,K^*}\|, \end{aligned} \tag{5.26}$$

where $\mathcal{W}_{K,K^*}$ is defined as

$$\mathcal{W}_{K,K^*} := \sum_{t\geq 0}[(I-\gamma^{-2}P_K DD^\top)^{-\top}(A-BK^*)]^t DD^\top [(A-BK^*)^\top (I-\gamma^{-2}P_K DD^\top)^{-1}]^t.$$

Note that $K^* \in \mathcal{K}$ and thus $(I-\gamma^{-2}P_{K^*} DD^\top)^{-\top}(A-BK^*)$ is stabilizing. Let $\epsilon := 1-\rho\big((I-\gamma^{-2}P_{K^*} DD^\top)^{-\top}(A-BK^*)\big)$, and note that $\epsilon > 0$. By the continuity of P_K, and that of $\rho(\cdot)$ [42], there exists a ball $\mathcal{B}(K^*, r) \subseteq \mathcal{K}$, centered at K^* with radius $r > 0$, such that for any $K \in \mathcal{B}(K^*, r)$,

$$\rho\big((I-\gamma^{-2}P_K DD^\top)^{-\top}(A-BK^*)\big) \leq 1-\epsilon/2 < 1. \tag{5.27}$$

Gauss–Newton: By Theorem 4.3, $\{K_n\}$ approaches K^*. Thus, there exists some $K_n \in \mathcal{B}(K^*, r)$. Let $K = K_n$ and thus $K' = K_{n+1}$. Replacing M in (5.22) by $DD^\top > 0$ and combining (5.26), we have

$$\begin{aligned}&\mathrm{Tr}(P_{K'} DD^\top) - \mathrm{Tr}(P_K DD^\top) \\ &\quad \leq \frac{-2\eta\sigma_{\min}(DD^\top)\sigma_{\min}(R)}{\sigma_{\max}(R+B^\top \widetilde{P}_{K_0} B)\|\mathcal{W}_{K,K^*}\|}[\mathrm{Tr}(P_K DD^\top) - \mathrm{Tr}(P_{K^*} DD^\top)],\end{aligned}$$

which further implies that

$$\begin{aligned}&\mathrm{Tr}(P_{K'} DD^\top) - \mathrm{Tr}(P_{K^*} DD^\top) \\ &\leq \left(1 - \frac{2\eta\sigma_{\min}(DD^\top)\sigma_{\min}(R)}{\sigma_{\max}(R+B^\top \widetilde{P}_{K_0} B)\|\mathcal{W}_{K,K^*}\|}\right) \cdot [\mathrm{Tr}(P_K DD^\top) - \mathrm{Tr}(P_{K^*} DD^\top)].\end{aligned} \tag{5.28}$$

(5.28) shows that the sequence $\{\mathrm{Tr}(P_{K_{n+p}} DD^\top)\}$ decreases to $\mathrm{Tr}(P_{K^*} DD^\top)$ starting from some $K_n \in \mathcal{B}(K^*, r)$. By continuity, there must exist a close enough K_{n+p} such that the lower-level set $\{K \mid \mathrm{Tr}(P_K DD^\top) \leq \mathrm{Tr}(K_{n+p} DD^\top)\} \subseteq \mathcal{B}(K^*, r)$. Hence,

starting from K_{n+p}, the iterates will never leave $\mathcal{B}(K^*, r)$. By (5.27), $\mathcal{W}_{K,K^*}$, as the unique solution to the Lyapunov equation

$$[(I-\gamma^{-2}P_K DD^\top)^{-\top}(A-BK^*)]\mathcal{W}_{K,K^*}[(I-\gamma^{-2}P_K DD^\top)^{-\top}(A-BK^*)]+DD^\top = \mathcal{W}_{K,K^*},$$

must have its norm bounded by some constant $\overline{\mathcal{W}}_r > \|DD^\top\|$ for all $K \in \mathcal{B}(K^*, r)$. Replacing $\|\mathcal{W}_{K,K^*}\|$ in (5.28) by $\overline{\mathcal{W}}_r$ gives the uniform local linear contraction of $\{\mathrm{Tr}(P_{K_n}DD^\top)\}$, which gives the local linear rate of $\{\mathcal{J}(K_n)\}$ by Lemma 5.2.

In addition, by the upper bound (5.19) and $E_{K^*}=0$, we have

(5.29)
$$\begin{aligned}&\mathrm{Tr}(P_{K'}DD^\top)-\mathrm{Tr}(P_{K^*}DD^\top)\\ &\le \mathrm{Tr}\Big\{\sum_{t\ge 0}[(A-BK')^\top(I-\gamma^{-2}P_{K'}DD^\top)^{-1}]^t\big[(K'-K^*)^\top(R+B^\top\widetilde{P}_{K^*}B)\\ &\qquad\cdot(K'-K^*)\big][(I-\gamma^{-2}P_{K'}DD^\top)^{-\top}(A-BK')]^t DD^\top\Big\}.\end{aligned}$$

For $\eta = 1/2$, suppose that some $K = K_n \in \mathcal{B}(K^*, r)$. Then, $K' = K_{n+1} = (R+B^\top\widetilde{P}_K B)^{-1}B^\top\widetilde{P}_K A$ yields that

$$\begin{aligned}K'-K^* = (R+B^\top\widetilde{P}_K B)^{-1}B^\top(\widetilde{P}_K-\widetilde{P}_{K^*})B(R+B^\top\widetilde{P}_{K^*}B)^{-1}B^\top\widetilde{P}_K A\\ +[(R+B^\top\widetilde{P}_{K^*}B)^{-1}B^\top(\widetilde{P}_K-\widetilde{P}_{K^*})A].\end{aligned}$$

Moreover, notice that

$$\begin{aligned}\widetilde{P}_K-\widetilde{P}_{K^*} &= (I-\gamma^{-2}P_K DD^\top)^{-1}P_K-(I-\gamma^{-2}P_{K^*}DD^\top)^{-1}P_{K^*}\\ &= (I-\gamma^{-2}P_{K^*}DD^\top)^{-1}\gamma^{-2}(P_K-P_{K^*})DD^\top(I-\gamma^{-2}P_K DD^\top)^{-1}\\ &\quad+(I-\gamma^{-2}P_{K^*}DD^\top)^{-1}(P_K-P_{K^*}),\end{aligned}\tag{5.30}$$

which, combined with (5.3), gives

$$\|K'-K^*\|_F \le c\cdot\|P_K-P_{K^*}\|_F\tag{5.31}$$

for some constant $c>0$. Combining (5.29) and (5.31) yields

$$\mathrm{Tr}(P_{K'}DD^\top)-\mathrm{Tr}(P_{K^*}DD^\top)\le c'\cdot[\mathrm{Tr}(P_K DD^\top)-\mathrm{Tr}(P_{K^*}DD^\top)]^2$$

for some constant c'. Note that from some $p\ge 0$ such that K_{n+p} onwards never leaves $\mathcal{B}(K^*, r)$, the constant c' is uniformly bounded, which proves the Q-quadratic convergence rate of $\{\mathrm{Tr}(P_{K_n}DD^\top)\}$, and thus the rate of $\{\mathcal{J}(K_n)\}$, around K^*.

Natural PG: Replacing M in (5.24) by $DD^\top>0$ and combining (5.24) and (5.26) yield

$$\mathrm{Tr}(P_{K'}DD^\top)-\mathrm{Tr}(P_{K^*}DD^\top)\le\left(1-\frac{2\eta\sigma_{\min}(R)}{\|\mathcal{W}_{K,K^*}\|}\right)\cdot[\mathrm{Tr}(P_K DD^\top)-\mathrm{Tr}(P_{K^*}DD^\top)].$$

Using a similar argument as above, one can establish the local linear rate of $\{\mathcal{J}(K_n)\}$ with a different contracting factor. This concludes the proof. ∎

6. Conclusions. In this paper, we have investigated the convergence theory of policy optimization methods for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$-norm robustness guarantees. Viewed as a constrained nonconvex optimization, this problem was addressed by policy optimization methods with provable convergence to the global optimal policy. More importantly, we showed that the proposed policy optimization methods enjoy the implicit regularization property, despite the lack of coercivity of the cost function. We expect the present work to serve as an initial step toward further understanding of RL algorithms on robust/risk-sensitive control tasks. Interesting future research directions include developing the sample complexity of the proposed methods in the model-free setting, investigating the implicit regularization property of other policy optimization methods, extending the results to the settings beyond linear time-invariant systems and state-feedback controllers, and studying the policy optimization landscape for $\mathcal{H}_\infty$ control synthesis.

Appendix A. Supplementary proofs.

A.1. On the proof of Lemma 2.2. This lemma just states a variant of the well-known bounded real lemma, and we omit the details here. The only subtlety worth mentioning is that the second condition automatically ensures that K is a stabilizing controller due to the well-known fact that one can perturb the solution P_K to obtain a solution for the strict matrix inequality in the third condition.

A.2. Proof of Lemma 3.1. Due to space limitation, we refer to [46, sect. 3.1] for a proof of nonconvexity of the problem, with an easily constructed example.

Note that $\|K\|$ may be unbounded for $K \in \mathcal{K}$. We show via counterexamples that for K with $\|K\| < \infty$, the cost does not necessarily go to infinity as K approaches the boundary of $\mathcal{K}$. Suppose $DD^\top > 0$ is full rank. For cost $\mathcal{J}(K)$ of form (2.5), it remains finite as long as P_K is finite. By Lemma 2.2, $I - \gamma^{-2} D^\top P_K D > 0$ always holds for $K \in \mathcal{K}$. Thus, $\lambda_{\max}(P_K)$ also has to be finite.

For cost $\mathcal{J}(K)$ of the forms (2.6) and (2.7), with $DD^\top > 0$, it is finite if both P_K is finite and $I - \gamma^{-2} D^\top P_K D > 0$ is nonsingular. The first condition is not violated as already shown above. We now show via a 1-dimensional example that the second condition is not violated either as $K \to \partial\mathcal{K}$. In fact, the Riccati equation (2.8) that defines P_K becomes a quadratic equation for the 1-dimensional case:

$$D^2 P_K^2 - [\gamma^2 - (A - BK)^2\gamma^2 + (C^2 + RK^2)D^2]P_K + (C^2 + RK^2)\gamma^2 = 0. \tag{A.1}$$

Thus, it is possible that the condition for the *existence* of solutions to the quadratic equations is *more restrictive* than the conditions on P_K in the bounded real lemma. Specifically, the solutions have the following form:

$$\begin{aligned} P_K = {} & \frac{\gamma^2 - (A - BK)^2\gamma^2 + (C^2 + RK^2)D^2}{2D^2} \\ & \pm \frac{\sqrt{[\gamma^2 - (A - BK)^2\gamma^2 + (C^2 + RK^2)D^2]^2 - 4D^2(C^2 + RK^2)\gamma^2}}{2D^2}. \end{aligned} \tag{A.2}$$

Denote the discriminant of (A.1) by $\blacklozenge$, and let $\blacklozenge = 0$ admit solutions. Note

$$\begin{aligned} 1 - \gamma^{-2} D^2 P_K &= 1 - \frac{1 - (A - BK)^2 + \gamma^{-2}(C^2 + RK^2)D^2}{2} \pm \frac{\gamma^{-2}\sqrt{\blacklozenge}}{2} \\ &= \frac{1 + (A - BK)^2}{2} - \frac{\gamma^{-2}(C^2 + RK^2)D^2}{2} \pm \frac{\gamma^{-2}\sqrt{\blacklozenge}}{2}, \end{aligned}$$

which, as $\blacklozenge \to 0$, can be greater than 0 with small enough D and large enough γ. Additionally, if the choices of A, B, C, D, R, γ ensure that $(A-BK)(1-\gamma^{-2}P_K D^2)^{-1} < 1$, then such a $K \in \mathcal{K}$. This way, as K approaches the boundary of $\{K \,|\, \blacklozenge \geq 0\}$, it is also approaching $\partial\mathcal{K}$, while the value of P_K approaches $[\gamma^2 - (A-BK)^2\gamma^2 + (C^2+RK^2)D^2]\cdot(2D^2)^{-1}$, a finite value. The above argument can be verified numerically by choosing $A = 2.75$, $B = 2$, $C^2 = 1$, $R = 1$, $D^2 = 0.01$, $\gamma = 0.2101$. In this case, $1-\gamma^{-2}D^2P_K \to 0.2354 > 0$ and $(A-BK)(1-\gamma^{-2}P_K D^2)^{-1} \to 0.9998 < 1$ if $K \to 1.2573$, which is the value that makes $\blacklozenge \to 0$. However, the corresponding $P_K \to [\gamma^2-(A-BK)^2\gamma^2+(C^2+RK^2)D^2]\cdot(2D^2)^{-1} = 3.3752 > 0$, a finite value that also satisfies $1-\gamma^{-2}D^2P_K > 0$. Hence, both the costs in (2.6) and (2.7) approach a finite value, which completes the proof of Lemma 3.1. ∎

A.3. Proof of Lemmas 3.2. Note that $\mathcal{J}(K)$ defined in (2.6) is differentiable with respect to P_K, provided that $\det(I-\gamma^{-2}P_K DD^\top) > 0$. This holds for any $K \in \mathcal{K}$ since by Lemma 2.2

$$I-\gamma^{-2}D^\top P_K D > 0 \Rightarrow \det(I-\gamma^{-2}D^\top P_K D) = \det(I-\gamma^{-2}P_K DD^\top) > 0,$$

where we have used Sylvester's determinant theorem that $\det(I+AB) = \det(I+BA)$. Thus, it suffices to show that P_K is differentiable with respect to K.

Recall that

$$\widetilde{P}_K = P_K + P_K D(\gamma^2 I - D^\top P_K D)^{-1}D^\top P_K = (I-\gamma^{-2}P_K DD^\top)^{-1}P_K, \tag{A.3}$$

where the second equation uses the matrix inversion lemma, and define the operator $\Psi : \mathbb{R}^{m\times m}\times\mathbb{R}^{d\times m} \to \mathbb{R}^{m\times m}$ as

$$\Psi(P_K, K) := C^\top C + K^\top RK + (A-BK)^\top \widetilde{P}_K(A-BK).$$

Note that Ψ is continuous with respect to both P_K and K, provided that $\gamma^2 I - D^\top P_K D > 0$. Also note that the Riccati equation (2.8) can be written as

$$\Psi(P_K, K) = P_K. \tag{A.4}$$

Notice the fact that for any matrices A, B, and X with proper dimensions,

$$\mathrm{vec}(AXB) = \big(B^\top \otimes A\big)\mathrm{vec}(X). \tag{A.5}$$

Thus, by vectorizing both sides of (A.4), we have

$$\begin{aligned}\mathrm{vec}\big(\Psi(P_K,K)\big) &= \mathrm{vec}(C^\top C + K^\top RK) + \mathrm{vec}\big((A-BK)^\top \widetilde{P}_K(A-BK)\big)\\ &= \mathrm{vec}(C^\top C + K^\top RK) + \big[(A-BK)^\top \otimes (A-BK)^\top\big]\\ &\quad\cdot \mathrm{vec}\big((I-\gamma^{-2}P_K DD^\top)^{-1}P_K\big) = \mathrm{vec}(P_K).\end{aligned} \tag{A.6}$$

By defining $\widetilde{\Psi} : \mathbb{R}^{m^2}\times\mathbb{R}^{dm} \to \mathbb{R}^{m^2}$ as a new mapping such that $\widetilde{\Psi}\big(\mathrm{vec}(P_K), \mathrm{vec}(K)\big) := \mathrm{vec}\big(\Psi(P_K, K)\big)$, the fixed-point equation (A.6) can be rewritten as

$$\widetilde{\Psi}\big(\mathrm{vec}(P_K), \mathrm{vec}(K)\big) = \mathrm{vec}(P_K). \tag{A.7}$$

Since vec is a linear mapping, it now suffices to show that $\mathrm{vec}(P_K)$ is differentiable with respect to $\mathrm{vec}(K)$. To this end, we apply the implicit function theorem [25] on the fixed-point equation (A.7). To ensure the applicability, we first note that the set

$\mathcal{K}$ defined in (2.4) is an open set. In fact, by Lemma 2.2, for any $K \in \mathcal{K}$, there exists some $P > 0$ such that the two LMIs in (2.11) hold. Since the inequality is strict, there must exists a small enough ball around K such for any K' in the ball, the LMIs still hold. Hence, the set $\mathcal{K}$ is open by definition.

Moreover, by the chain rule of matrix differentials [27, Theorem 9], we know that

$$\begin{aligned}&\frac{\partial \mathrm{vec}\big((I-\gamma^{-2}P_KDD^\top)^{-1}P_K\big)}{\partial \mathrm{vec}^\top(P_K)}\\ &\quad=(P_K\otimes I)\cdot\frac{\partial \mathrm{vec}[(I-\gamma^{-2}P_KDD^\top)^{-1}]}{\partial \mathrm{vec}^\top(P_K)}+I\otimes(I-\gamma^{-2}P_KDD^\top)^{-1},\end{aligned}\tag{A.8}$$

where I denotes the identity matrix of proper dimension.

Now we claim that

$$\begin{aligned}&\frac{\partial \mathrm{vec}[(I-\gamma^{-2}P_KDD^\top)^{-1}]}{\partial \mathrm{vec}^\top(P_K)}\\ &\quad=[(\gamma^{-2}DD^\top)\cdot(I-\gamma^{-2}P_KDD^\top)^{-1}]\otimes(I-\gamma^{-2}P_KDD^\top)^{-1}.\end{aligned}\tag{A.9}$$

To show this, we compare the element at the $[(j-1)m+i]$th row and the $[(l-1)m+k]$th column of both sides of (A.9) with $i,j,k,l\in[m]$, where both sides are matrices of dimensions $m^2\times m^2$. On the LHS, notice that

$$\begin{aligned}&\frac{\partial(I-\gamma^{-2}P_KDD^\top)^{-1}}{\partial[P_K]_{k,l}}\\ &\quad=(I-\gamma^{-2}P_KDD^\top)^{-1}\cdot\frac{\partial(\gamma^{-2}P_KDD^\top)}{\partial[P_K]_{k,l}}\cdot(I-\gamma^{-2}P_KDD^\top)^{-1},\end{aligned}$$

which follows from $(F^{-1})'=-F^{-1}F'F^{-1}$ for some matrix function F. Also,

$$\frac{\partial(\gamma^{-2}P_KDD^\top)}{\partial[P_K]_{k,l}}=\gamma^{-2}\begin{bmatrix}\multicolumn{3}{c}{\overline{\qquad\qquad}}\\ [DD^\top]_{l,1} & \cdots & [DD^\top]_{l,m}\\ \multicolumn{3}{c}{\overline{\qquad\qquad}}\end{bmatrix}\leftarrow k\text{th row},$$

where only the kth row is nonzero and is filled with the lth row of $DD^\top$. Due to these two facts, we have

$$\begin{aligned}&\left[\frac{\partial \mathrm{vec}[(I-\gamma^{-2}P_KDD^\top)^{-1}]}{\partial \mathrm{vec}^\top(P_K)}\right]_{(j-1)m+i,(l-1)m+k}=\frac{\partial[(I-\gamma^{-2}P_KDD^\top)^{-1}]_{i,j}}{\partial[P_K]_{k,l}}\\ &\quad=\gamma[(I-\gamma^{-2}P_KDD^\top)^{-1}]_{i,k}\cdot\sum_{q=1}^{m}[DD^\top]_{l,q}\cdot[(I-\gamma^{-2}P_KDD^\top)^{-1}]_{q,j}.\end{aligned}\tag{A.10}$$

On the right-hand side of (A.9), we have

$$\begin{aligned}&\big[[(\gamma^{-2}DD^\top)\cdot(I-\gamma^{-2}P_KDD^\top)^{-1}]\otimes(I-\gamma^{-2}P_KDD^\top)^{-1}\big]_{(j-1)m+i,(l-1)m+k}\\ &\quad=[(\gamma^{-2}DD^\top)\cdot(I-\gamma^{-2}P_KDD^\top)^{-1}]_{l,j}\cdot[(I-\gamma^{-2}P_KDD^\top)^{-1}]_{i,k},\end{aligned}\tag{A.11}$$

due to the definition of the Kronecker product and the fact that the matrix

$$(\gamma^{-2}DD^\top)\cdot(I-\gamma^{-2}P_KDD^\top)^{-1}=D(\gamma^2I-D^\top P_KD)^{-1}D^\top\tag{A.12}$$

is symmetric. Thus, (A.10) and (A.11) are identical for any (i,j,k,l), proving (A.9).

By substituting (A.9) into (A.8), we have

$$\begin{aligned}&\frac{\partial \mathrm{vec}\big((I-\gamma^{-2}P_KDD^\top)^{-1}P_K\big)}{\partial \mathrm{vec}^\top(P_K)}\\&\quad=[I+(\gamma^{-2}P_KDD^\top)(I-\gamma^{-2}P_KDD^\top)^{-1}]\otimes(I-\gamma^{-2}P_KDD^\top)^{-1}\\&\quad=(I-\gamma^{-2}P_KDD^\top)^{-1}\otimes(I-\gamma^{-2}P_KDD^\top)^{-1},\end{aligned}$$

where the first equation uses the facts that $(A\otimes B)(C\otimes D)=(AC)\otimes(BD)$ and $(A\otimes B)+(C\otimes B)=(A+C)\otimes B$, and the last one uses the matrix inversion lemma. By (A.8), we can thus write the partial derivative of $\widetilde{\Psi}\big(\mathrm{vec}(P_K),\mathrm{vec}(K)\big)$ as

$$\begin{aligned}&\frac{\partial\widetilde{\Psi}\big(\mathrm{vec}(P_K),\mathrm{vec}(K)\big)}{\partial\mathrm{vec}^\top(P_K)}=\big[(A-BK)^\top\otimes(A-BK)^\top\big]\cdot\frac{\partial\mathrm{vec}\big((I-\gamma P_KDD^\top)^{-1}P_K\big)}{\partial\mathrm{vec}^\top(P_K)}\\&\quad=\big[(A-BK)^\top(I-\gamma P_KDD^\top)^{-1}\big]\otimes\big[(A-BK)^\top(I-\gamma P_KDD^\top)^{-1}\big].\end{aligned}$$

Therefore, the partial derivative

$$\begin{aligned}&\frac{\partial\big[\widetilde{\Psi}\big(\mathrm{vec}(P_K),\mathrm{vec}(K)\big)-\mathrm{vec}(P_K)\big]}{\partial\mathrm{vec}^\top(P_K)}\\&\quad=\big[(A-BK)^\top(I-\gamma^{-2}P_KDD^\top)^{-1}\big]\otimes\big[(A-BK)^\top(I-\gamma^{-2}P_KDD^\top)^{-1}\big]-I,\end{aligned}$$

which is invertible, since the eigenvalues of $[(A-BK)^\top(I-\gamma^{-2}P_KDD^\top)^{-1}]\otimes[(A-BK)^\top(I-\gamma^{-2}P_KDD^\top)^{-1}]$ are the products of the eigenvalues of $(A-BK)^\top(I-\gamma^{-2}P_KDD^\top)^{-1}$, and the matrix $(A-BK)^\top(I-\gamma^{-2}P_KDD^\top)^{-1}$ has spectral radius less than 1 for all $K\in\mathcal{K}$. Also, since $\widetilde{\Psi}\big(\mathrm{vec}(P_K),\mathrm{vec}(K)\big)-\mathrm{vec}(P_K)$ is continuous with respect to both $\mathrm{vec}(P_K)$ and $\mathrm{vec}(K)$, by the implicit function theorem [25], we know that there exists an open neighborhood around $\mathrm{vec}(P_K)$ and $\mathrm{vec}(K)$ (thus including $\mathrm{vec}(P_K)$ and $\mathrm{vec}(K)$), so that $\mathrm{vec}(P_K)$ is a continuously differentiable function with respect to $\mathrm{vec}(K)$, and so is P_K with respect to K, in the neighborhood. Note that this holds for any $K\in\mathcal{K}$. This proves the differentiability of $\mathcal{J}(K)$ at all $K\in\mathcal{K}$.

Now we establish the form of the PG. By Lemma 2.2, we know that for any $K\in\mathcal{K}$, $(A-BK)^\top(I-\gamma^{-2}P_KDD^\top)^{-1}$ is stable and $I-\gamma^{-2}D^\top P_KD>0$. Therefore, the expression Δ_K in (3.1) exists, and so does the expression for $\nabla\mathcal{J}(K)$. We then verify the expressions by showing the form of the directional derivative $\nabla_{K_{ij}}\mathcal{J}(K)$, i.e., the derivative with respect to each element K_{ij} in the matrix K. By definition of $\mathcal{J}(K)$ in (2.6), we have

$$\begin{aligned}\nabla_{K_{ij}}\mathcal{J}(K)&=-\gamma^2\,\mathrm{Tr}\left\{(I-\gamma^{-2}P_KDD^\top)^{-\top}[\nabla_{K_{ij}}(I-\gamma^{-2}P_KDD^\top)]^\top\right\}\\&=\mathrm{Tr}\left[(I-\gamma^{-2}P_KDD^\top)^{-1}\nabla_{K_{ij}}(P_KDD^\top)\right],\end{aligned}\tag{A.13}$$

where the first equality follows from the chain rule and the fact that $\nabla_X\log\det X=X^{-\top}$, and the second one follows from the fact that $\mathrm{Tr}(A^\top B^\top)=\mathrm{Tr}(BA)^\top=\mathrm{Tr}(BA)=\mathrm{Tr}(AB)$. Furthermore, since $DD^\top$ is independent of K, and $\mathrm{Tr}(ABC)=\mathrm{Tr}(BCA)$, we obtain from (A.13) and (A.12) that

$$\begin{aligned}\nabla_{K_{ij}}\mathcal{J}(K)&=\mathrm{Tr}\left[(I-\gamma^{-2}P_KDD^\top)^{-1}\nabla_{K_{ij}}P_K\cdot DD^\top\right]\\&=\mathrm{Tr}\left[\nabla_{K_{ij}}P_K\cdot D(I-\gamma^{-2}D^\top P_KD)^{-1}D^\top\right].\end{aligned}\tag{A.14}$$

Now we establish the recursion of $\nabla_{K_{ij}}\mathcal{J}(K)$ using the Riccati equation (2.8). Letting $M := D(I-\gamma^{-2}D^\top P_K D)^{-1}D^\top$, we have from (2.8), (A.3), and (A.14) that

$$\begin{aligned}\nabla_{K_{ij}}\mathcal{J}(K) &= \operatorname{Tr}\left(\nabla_{K_{ij}}P_K\cdot M\right)\\ &= (2RKM)_{ij} - \left[2B^\top\widetilde{P}_K(A-BK)M\right]_{ij} + \operatorname{Tr}\left[(A-BK)^\top(\nabla_{K_{ij}}\widetilde{P}_K)(A-BK)M\right],\end{aligned}\tag{A.15}$$

where on the right-hand side (RHS) of (A.15), the first term is due to that $\nabla_K \operatorname{Tr}(K^\top RKM) = 2RKM$ for any positive definite (and thus symmetric) matrix M, the second term is the gradient with $\widetilde{P}_K$ fixed, and the third term is the gradient with $A-BK$ fixed.

In addition, by taking the derivative on both sides of (A.3), we have

$$\begin{aligned}\nabla_{K_{ij}}\widetilde{P}_K &= (I-\gamma^{-2}P_KDD^\top)^{-1}\nabla_{K_{ij}}P_K\cdot\gamma^{-2}DD^\top(I-\gamma^{-2}P_KDD^\top)^{-1}P_K\\ &\quad + (I-\gamma^{-2}P_KDD^\top)^{-1}\nabla_{K_{ij}}P_K\\ &= (I-\gamma^{-2}P_KDD^\top)^{-1}\nabla_{K_{ij}}P_K\cdot[D(\gamma^2I-D^\top P_KD)^{-1}D^\top P_K+I]\\ &= (I-\gamma^{-2}P_KDD^\top)^{-1}\cdot\nabla_{K_{ij}}P_K\cdot(I-\gamma^{-2}DD^\top P_K)^{-1},\end{aligned}\tag{A.16}$$

where the first equation uses the fact that $\nabla_X(P^{-1}) = -P^{-1}\cdot\nabla_X P\cdot P^{-1}$, the second one uses (A.12), and the last one uses the matrix inversion lemma. Notice that $(I-\gamma^{-2}DD^\top P_K)^{-1} = (I-\gamma^{-2}P_KDD^\top)^{-\top}$. Thus, (A.16) can be written as

$$\nabla_{K_{ij}}\widetilde{P}_K = (I-\gamma^{-2}P_KDD^\top)^{-1}\cdot\nabla_{K_{ij}}P_K\cdot(I-\gamma^{-2}P_KDD^\top)^{-\top}.\tag{A.17}$$

Substituting (A.17) into (A.15) yields the following recursion:

$$\begin{aligned}\nabla_{K_{ij}}\mathcal{J}(K) &= (2RKM)_{ij} - \left[2B^\top\widetilde{P}_K(A-BK)M\right]_{ij}\\ &\quad + \operatorname{Tr}\Big[\nabla_{K_{ij}}P_K\cdot\underbrace{(I-\gamma^{-2}P_KDD^\top)^{-\top}(A-BK)M(A-BK)^\top(I-\gamma^{-2}P_KDD^\top)^{-1}}_{M_1}\Big].\end{aligned}$$

By performing a recursion on $\operatorname{Tr}\left(\nabla_{K_{ij}}P_K\cdot M_1\right)$, and combining all the i,j terms into a matrix, we obtain the form of the gradient given in Lemma 3.2. ∎

A.4. Proof of Proposition 3.3. The proof is based on a game-theoretic perspective on the problem. First, for any $K\in\mathcal{K}$, by [4, Theorem 3.7], with A, B, D, Q, R, γ therein being replaced by $A-BK$, 0, D, $Q+K^\top RK$, R, γ here, we obtain that the Riccati equation in (2.5) corresponds to the generalized algebraic Riccati equation (3.52b) in [4], for this auxiliary game. By Lemma 2.2, the solution $P_K\geq 0$ satisfies (3.53) in [4]. Recall that P_K is the unique stabilizing solution to (2.5), and is thus also minimal if $(A-BK,D)$ is stabilizable [34, Theorem 3.1], which is indeed the case since $K\in\mathcal{K}$ is stabilizing. Hence, by [4, Theorem 3.7(ii), (iv)], the controller and the disturbance that attain the upper value of the game have, respectively, the forms of $u_t=0$ and $w_t = (\gamma^2I-D^\top P_KD)^{-1}D^\top P_K(A-BK)x_t$ for all t. This shows that in the original game with A, B, D, Q, R, γ (as defined in [4, Chapter 3.7]), and with a fixed $K\in\mathcal{K}$, the maximizing disturbance has the form w_t as above, and the value under the pair $(K,-(\gamma^2I-D^\top P_KD)^{-1}D^\top P_K(A-BK))$ is indeed $x_0^\top P_Kx_0$. By again applying [4, Theorem 3.7] to the original game, we know that the value is $x_0^\top P_{K^*}x_0$, and is achieved by the optimal controller $u_t^* = -K^*x_t$ and the maximizing disturbance $w_t^* = [(\gamma^2I-D^\top P_{K^*}D)^{-1}D^\top P_{K^*}(A-BK^*)]x_t$ with K^* being defined in the proposition. By the definition of the value of the game, we know that $x_0^\top P_Kx_0\geq x_0^\top P_{K^*}x_0$

for any $K \in \mathcal{K}$. As the above arguments hold for any x_0, we know that $P_K \geq P_{K^*}$. Finally, notice that for $K, K^* \in \mathcal{K}$, $0 < I - \gamma^{-2} D^\top P_K D \leq I - \gamma^{-2} D^\top P_{K^*} D$ (cf. Lemma 2.2). By $\det(I - \gamma^{-2} P_K D D^\top) = \det(I - \gamma^{-2} D^\top P_K D)$, we know that $\mathcal{J}(K) \geq \mathcal{J}(K^*)$ for any $K \in \mathcal{K}$. This completes the proof for the first half of the proposition.

For the second half of the proposition, note that $\Delta_K \geq 0$ since $I - \gamma^{-2} D^\top P_K D > 0$ for any $K \in \mathcal{K}$ by Lemma 2.2. Also, since $(I - \gamma^{-2} D^\top P_K D)^{-1} \geq I$, we know that

(A.18)
$$\Delta_K \geq \sum_{t=0}^{\infty} \left[(I - \gamma^{-2} P_K D D^\top)^{-\top} (A - BK)\right]^t D D^\top \left[(A - BK)^\top (I - \gamma^{-2} P_K D D^\top)^{-1}\right]^t.$$

By [48, Lemma 21.2], the RHS of (A.18) is always positive definite, since $((I - \gamma^{-2} P_K D D^\top)^{-\top} (A - BK), D)$ is controllable, i.e.,

$$\left((A - BK)^\top (I - \gamma^{-2} P_K D D^\top)^{-1}, D^\top\right)$$

is observable. Thus, $\Delta_K > 0$ is full rank. By the optimality condition $\nabla \mathcal{J}(K) = 0$, it follows that $K^* = (R + B^\top \widetilde{P}_{K^*} B)^{-1} B^\top \widetilde{P}_{K^*} A$ is the unique stationary point, which is thus the unique global optimizer. This completes the proof of Proposition 3.3. ∎

A.5. Proof of Lemma 5.1. We start with the following helper lemma.

LEMMA A.1. *Suppose that $K, K' \in \mathcal{K}$. Then we have that $I - \gamma^{-2} P_{K'} D D^\top$ is invertible, and*

$$(I - \gamma^{-2} P_{K'} D D^\top)^{-1} (P_K - \gamma^{-2} P_{K'} D D^\top P_{K'}) (I - \gamma^{-2} D D^\top P_{K'})^{-1} \leq \widetilde{P}_K. \tag{A.19}$$

Proof. First, since $K, K' \in \mathcal{K}$, by Lemma 2.2, $I - \gamma^{-2} D^\top P_{K'} D > 0$ is invertible. Thus, $\det(I - \gamma^{-2} P_{K'} D D^\top) = \det(I - \gamma^{-2} D^\top P_{K'} D) \neq 0$, namely, $I - \gamma^{-2} P_{K'} D D^\top$ is invertible. Then the desired fact is equivalent to

$$\begin{aligned} P_K - \gamma^{-2} P_{K'} D D^\top P_{K'} &\leq (I - \gamma^{-2} P_{K'} D D^\top) \widetilde{P}_K (I - \gamma^{-2} D D^\top P_{K'}) \\ &= \widetilde{P}_K - \gamma^{-2} P_{K'} D D^\top \widetilde{P}_K - \gamma^{-2} \widetilde{P}_K D D^\top P_{K'} + \gamma^{-4} P_{K'} D D^\top \widetilde{P}_K D D^\top P_{K'} \end{aligned}$$

which can be further simplified as

$$\begin{aligned} &(\widetilde{P}_K - P_K) - \gamma^{-2} P_{K'} D D^\top \widetilde{P}_K - \gamma^{-2} \widetilde{P}_K D D^\top P_{K'} + \gamma^{-2} P_{K'} D D^\top P_{K'} \\ &\quad + \gamma^{-4} P_{K'} D D^\top \widetilde{P}_K D D^\top P_{K'} \geq 0. \end{aligned} \tag{A.20}$$

By $\widetilde{P}_K = (I - \gamma^{-2} P_K D D^\top)^{-1} P_K$ and (A.12), we have

$$\gamma^{-2} P_{K'} D D^\top \widetilde{P}_K = \gamma^{-2} P_{K'} D D^\top (I - \gamma^{-2} P_K D D^\top)^{-1} P_K = P_{K'} D (\gamma^2 I - D^\top P_K D)^{-1} D^\top P_K.$$

Thus, it follows that

$$\begin{aligned} &(\widetilde{P}_K - P_K) - \gamma^{-2} P_{K'} D D^\top \widetilde{P}_K - \gamma^{-2} \widetilde{P}_K D D^\top P_{K'} \\ &\quad = (P_K - P_{K'}) D (\gamma^2 I - D^\top P_K D)^{-1} D^\top (P_K - P_{K'}) - P_{K'} D (\gamma^2 I - D^\top P_K D)^{-1} D^\top P_{K'}. \end{aligned}$$

Therefore, (A.20) is equivalent to

$$\begin{aligned} &(P_K - P_{K'}) D (\gamma^2 I - D^\top P_K D)^{-1} D^\top (P_K - P_{K'}) - P_{K'} D (\gamma^2 I - D^\top P_K D)^{-1} D^\top P_{K'} \\ &\quad + \gamma^{-2} P_{K'} D D^\top P_{K'} + \gamma^{-4} P_{K'} D D^\top \widetilde{P}_K D D^\top P_{K'} \geq 0. \end{aligned}$$

Given the fact $\gamma^2 I > D^\top P_K D$ and another fact that

(A.21)
$$-P_{K'}D(\gamma^2 I - D^\top P_K D)^{-1}D^\top P_{K'} + \gamma^{-2}P_{K'}DD^\top P_{K'} + \gamma^{-4}P_{K'}DD^\top \widetilde{P}_K DD^\top P_{K'} = 0,$$

we know that the above inequality holds and hence our lemma is true. To show that (A.21) holds, it suffices to apply the matrix inversion lemma, i.e.,

$$(\gamma^2 I - D^\top P_K D)^{-1} = \gamma^{-2}I + \gamma^{-4}D^\top(-\gamma^{-2}P_K DD^\top + I)^{-1}P_K D = \gamma^{-2}I + \gamma^{-4}D^\top \widetilde{P}_K D,$$

where the first equation uses the matrix inversion lemma. The proof is complete. ∎

By the definition of $\widetilde{P}_K$ in (2.9) and the Riccati equation (2.8), we have

(A.22)
$$\begin{aligned}
P_{K'} &= (A - BK')^\top \widetilde{P}_{K'}(A - BK') + C^\top C + (K')^\top RK' \\
&= (A - BK')^\top (I - \gamma^{-2}P_{K'}DD^\top)^{-1}(P_K - \gamma^{-2}P_{K'}DD^\top P_{K'})(I - \gamma^{-2}P_{K'}DD^\top)^{-\top} \\
&\quad \cdot (A - BK') + (A - BK')^\top (I - \gamma^{-2}P_{K'}DD^\top)^{-1}(P_{K'} - P_K)(I - \gamma^{-2}P_{K'}DD^\top)^{-\top} \\
&\quad \cdot (A - BK') + C^\top C + (K')^\top RK'.
\end{aligned}$$

By (A.19) in Lemma A.1, we further have

$$\begin{aligned}
P_{K'} - P_K &\le C^\top C + (K')^\top RK' + (A - BK')^\top \widetilde{P}_K(A - BK') - P_K \\
&\quad + (A - BK')^\top (I - \gamma^{-2}P_{K'}DD^\top)^{-1}(P_{K'} - P_K)(I - \gamma^{-2}P_{K'}DD^\top)^{-\top}(A - BK').
\end{aligned}$$

By induction, we can apply the above inequality iteratively to show that

(A.23)
$$\begin{aligned}
P_{K'} - P_K \le \sum_{t\ge 0}&[(A - BK')^\top (I - \gamma^{-2}P_{K'}DD^\top)^{-1}]^t \big[C^\top C + (K')^\top RK' \\
&+ (A - BK')^\top \widetilde{P}_K(A - BK') - P_K\big][(I - \gamma^{-2}P_{K'}DD^\top)^{-\top}(A - BK')]^t.
\end{aligned}$$

On the other hand, we have

(A.24)
$$\begin{aligned}
&C^\top C + (K')^\top RK' + (A - BK')^\top \widetilde{P}_K(A - BK') - P_K \\
&= C^\top C + (K' - K + K)^\top R(K' - K + K) + (A - BK - B(K' - K))^\top \widetilde{P}_K(A - BK \\
&\quad - B(K' - K)) - P_K \\
&= (K' - K)^\top \Big((R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A\Big) \\
&\quad + \Big((R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A\Big)^\top (K' - K) \\
&\quad + (K' - K)(R + B^\top \widetilde{P}_K B)(K' - K),
\end{aligned}$$

which can be substituted into (A.23) to obtain the upper bound in (5.19).

For the lower bound (5.20), note that the conditions in Lemma A.1 also hold here when the roles of K and K' are interchanged. Thus, we have

$$(I - \gamma^{-2}P_K DD^\top)^{-1}(P_{K'} - \gamma^{-2}P_K DD^\top P_K)(I - \gamma^{-2}DD^\top P_K)^{-1} \le \widetilde{P}_{K'},$$

which gives a lower bound on the RHS of (A.22) directly as

(A.25)

$$
\begin{aligned}
P_{K'} - P_K &= (A - BK')^\top \widetilde{P}_{K'}(A - BK') - P_K + C^\top C + (K')^\top R K' \\
&\geq (A - BK')^\top \big[(I - \gamma^{-2} P_K D D^\top)^{-1}(P_{K'} - \gamma^{-2} P_K D D^\top P_K)(I - \gamma^{-2} D D^\top P_K)^{-1}\big] \\
&\quad \cdot (A - BK') - P_K + C^\top C + (K')^\top R K' \\
&= (A - BK')^\top \underbrace{\big[(I - \gamma^{-2} P_K D D^\top)^{-1}(P_K - \gamma^{-2} P_K D D^\top P_K)(I - \gamma^{-2} D D^\top P_K)^{-1}\big]}_{\widetilde{P}_K} \\
&\quad \cdot (A - BK') - P_K + (A - BK')^\top \big[(I - \gamma^{-2} P_K D D^\top)^{-1}(P_{K'} - P_K)(I \\
&\quad - \gamma^{-2} D D^\top P_K)^{-1}\big](A - BK') + C^\top C + (K')^\top R K'.
\end{aligned}
$$

Continuing unrolling the RHS of (A.25) and substituting into (A.24), we obtain the desired lower bound in (5.20), which completes the proof. ∎

Appendix B. Simulations. We present some simulation results to illustrate the effectiveness of our PO methods, by comparing them with several existing $\mathcal{H}_2/\mathcal{H}_\infty$ mixed control solvers, including the HIFOO package [3], the `systune` function [2], and the `h2hinfsyn` function [13] in MATLAB's robust control toolbox. We note that HIFOO and `h2hinfsyn` can only handle continuous-time settings, while `systune` and our PO methods can handle both continuous- and discrete-time settings. To make the comparison fair, we compare all four algorithms in the continuous-time setting, even though this paper focuses on the theory for the discrete-time one. Details of the simulation setups and theory for the continuous-time setting can be found in [46], and we expect any comparison in the discrete-time setting to lead to similar conclusions. We summarize the following findings observed from Table 1, based on solving Case 3 in [46, sect. 7.3]:

- *(PO methods return competitive $\mathcal{H}_2$-norm performance, and globally optimal $\mathcal{J}(K)$).* In terms of the $\mathcal{H}_2$-norm, our PO methods can achieve competitive performance. Note that the $\mathcal{H}_2$ performance of our PO methods is almost identical to that of `h2hinfsyn`, since the latter also optimizes $\mathcal{J}(K)$, an upper bound of the actual $\mathcal{H}_2$-norm $\|\mathcal{T}(K)\|_2$, but based on an LMI-based procedure [13]. This in turn verifies the global optimality of our PO methods as proved, as the LMI-based approach can find the global optimum of $\mathcal{J}(K)$ directly. Also, compared to the other two methods, which directly optimize the $\mathcal{H}_2$-norm, the $\mathcal{H}_2$-performance achieved by minimizing $\mathcal{J}(K)$ is reasonably good. Moreover, in contrast to the global convergence we established for PO methods, HIFOO does not have convergence guarantees, and `systune` only has convergence guarantees to local optimum [2]. But still, the local optimum returned by `systune` can have lower $\mathcal{H}_2$-norms, especially when γ is small, e.g., $\gamma = 0.54$. This shows the advantages of `systune` in the setting with stringent $\mathcal{H}_\infty$-norm constraints.
- *(PO methods provably preserve $\mathcal{H}_\infty$-norm constraint).* For all choices of γ, our PO methods consistently preserve the $\|\mathcal{T}(K)\|_\infty < \gamma$ constraint during the optimization process, validating our theoretical findings. However, for the cases of $\gamma = 5, 1, 0.54$, there are $5\%, 22\%, 90\%, 100\%$ of the HIFOO trials that violate the $\|\mathcal{T}(K)\|_\infty$-constraint during optimization. We note that $\|\mathcal{T}(K)\|_\infty$-constraint violation information is not available or not applicable when using `h2hinfsyn` and `systune`. Moreover, a smaller γ, even for $\gamma = 0.54$

TABLE 1

Average statistics over 100 *trials for HIFOO,* `h2hinfsyn`*,* `systune`*, and two proposed PO methods, for solving Case* 3 *in* [46, *sect.* 7]*. An entry for "$\|\mathcal{T}(K)\|_\infty$ violation," e.g., m% (n), represents the $\|\mathcal{T}(K)\|_\infty$ constraint was violated in m% of the trials with an average violation of n. The data in the rows of "$\|\mathcal{T}(K)\|_2$ reached," "$\mathcal{J}(K)$ reached," "$\|\mathcal{T}(K)\|_\infty$ reached" are averaged over trials* with no $\|\mathcal{T}(K)\|_\infty$*-constraint violation. The columns "*`systune` w/ Init*" and "*NPG/GN w/ Init*" display the total runtime used by the optimization processes* after *a feasible K_0 is given. We feed the same K_0 used in our PO methods into the initialization of* `systune`*. In contrast, HIFOO and* `h2hinfsyn` *use in-house methods for algorithm initialization.*

Case 3 with $\gamma = 5$	*HIFOO*	`h2hinfsyn`	`systune` *w/ Init*	*NPG/GN w/ Init*
Runtime (s)	0.4569	0.0305	0.0857	0.0084/0.0095
$\|\mathcal{T}(K)\|_2$ *reached*	0.9810	0.9811	0.9810	0.9811
$\mathcal{J}(K)$ *reached*	0.9699	0.9699	0.9699	0.9699
$\|\mathcal{T}(K)\|_\infty$ *reached*	1.0229	1.0126	1.0229	1.0124
$\|\mathcal{T}(K)\|_\infty$ *violation*	5% (2.3092)	*n.a.*	*n.a.*	0% (0)

Case 3 with $\gamma = 1$	*HIFOO*	`h2hinfsyn`	`systune` *w/ Init*	*NPG/GN w/ Init*
Runtime (s)	1.8370	0.0351	0.1959	0.0064/0.0071
$\|\mathcal{T}(K)\|_2$ *reached*	1.0876	1.0037	0.9812	1.0038
$\mathcal{J}(K)$ *reached*	1.8506	1.2082	1.4962	1.2082
$\|\mathcal{T}(K)\|_\infty$ *reached*	1.0000	0.8145	0.9929	0.8143
$\|\mathcal{T}(K)\|_\infty$ *violation*	90% (2.0772)	*n.a.*	*n.a.*	0% (0)

Case 3 with $\gamma = 0.54$	*HIFOO*	`h2hinfsyn`	`systune` *w/ Init*	*NPG/GN w/ Init*
Runtime (s)	1.1163	0.0363	0.3959	0.0050/0.0051
$\|\mathcal{T}(K)\|_2$ *reached*	*NaN*	2.2174	2.0915	2.2174
$\mathcal{J}(K)$ *reached*	*NaN*	9.3070	10.7198	9.3070
$\|\mathcal{T}(K)\|_\infty$ *reached*	*NaN*	0.5397	0.5400	0.5397
$\|\mathcal{T}(K)\|_\infty$ *violation*	100% (2.3708)	*n.a.*	*n.a.*	0% (0)

(very close to $\gamma^* = 0.53$ in Case 3 in [46, sect. 7]), does not prevent `systune` from returning a solution that has no constraint violation. However, the $\mathcal{H}_\infty$-norms of the returned controllers, which are almost binding for small γ (e.g., $\gamma = 1$ or 0.54), are consistently larger than those returned by `h2hinfsyn` and our PO methods. In other words, the controller returned by `h2hinfsyn` and our PO methods can be more robust.

- *(PO methods have competitive, if not much faster, runtimes).* In terms of computation runtime, our PO methods are competitive, and even much faster than some existing methods, when a feasible initial point is available.

More numerical results for large-scale systems can be found in the extended version [46, sect. 7.3]. The overall observation is that, given the same feasible initialization, the higher the system dimension is, the faster our PO methods are, compared to the other solvers. That being said, the advantage does not come for free, as for high-dimensional systems, finding a feasible (robustly stable) initialization becomes more challenging. This is one limitation of our policy search methods. We have left it as our future work to find a robustly stable initialization efficiently.

REFERENCES

[1] A. AGARWAL, S. M. KAKADE, J. D. LEE, AND G. MAHAJAN, *Optimality and approximation with policy gradient methods in Markov decision processes*, J. Mach. Learn. Res., 125 (2020), pp. 64–66.

[2] P. APKARIAN, D. NOLL, AND A. RONDEPIERRE, *Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control via nonsmooth optimization*, SIAM J. Control Optim., 47 (2008), pp. 1516–1546.

[3] D. ARZELIER, D. GEORGIA, S. GUMUSSOY, AND D. HENRION, *H2 for HIFOO*, in International Conference on Control and Optimization with Industrial Applications, 2011.

[4] T. BAŞAR AND P. BERNHARD, *H-infinity Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, Birkhäuser, Boston, 1995.

[5] D. S. Bernstein and W. M. Haddad, *LQG control with an $\mathcal{H}_\infty$ performance bound: A Riccati equation approach*, IEEE Trans. Automat. Control, 34 (1989), pp. 293–305.
[6] J. Bhandari and D. Russo, *Global Optimality Guarantees for Policy Gradient Methods*, preprint, arXiv:1906.01786, 2019.
[7] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, Stud. Appl. Numer. Math., 15, SIAM, Philadelphia, 1994.
[8] S. P. Boyd, V. Balakrishnan, C. H. Barratt, N. M. Khraishi, X. Li, D. G. Meyer, and S. A. Norman, *A new CAD method and associated architectures for linear controllers*, IEEE Trans. Automat. Control, 33 (1988), pp. 268–283.
[9] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, *Adaptive linear quadratic control using policy iteration*, in IEEE American Control Conference, IEEE, Piscataway, NJ, 1994, pp. 3475–3479.
[10] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, *LQR Through the Lens of First Order Methods: Discrete-time case*, preprint, arXiv:1907.08921, (2019).
[11] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems*, SIAM J. Optim., 20 (2010), pp. 2833–2852.
[12] X. Chen and J. T. Wen, *A linear matrix inequality approach to the general mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem*, in IEEE American Control Conference, IEEE, Piscataway, NJ, 1995, pp. 1443–1447.
[13] M. Chilali and P. Gahinet, *$\mathcal{H}_\infty$ design with pole placement constraints: An LMI approach*, IEEE Trans. Automat. Control, 41 (1996), pp. 358–367.
[14] G. E. Dullerud and F. Paganini, *A Course in Robust Control Theory: A Convex Approach*, Texts Appl. Math. 36, Springer, New York, 2000.
[15] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, *Global convergence of policy gradient methods for the linear quadratic regulator*, in International Conference on Machine Learning, Curran Associates, Red Hook, NY, 2018, 380.
[16] K. Glover and J. C. Doyle, *State-space formulae for all stabilizing controllers that satisfy an $\mathcal{H}_\infty$-norm bound and relations to relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.
[17] B. Gravell, P. M. Esfahani, and T. Summers, *Learning Robust Controllers for Linear Quadratic Systems with Multiplicative Noise Via Policy Gradient*, preprint, arXiv:1907.03680, 2019.
[18] W. M. Haddad, D. S. Bernstein, and D. Mustafa, *Mixed-norm $\mathcal{H}_2/\mathcal{H}_\infty$ regulation and estimation: The discrete-time case*, Systems Control Lett., 16 (1991), pp. 235–247.
[19] G. Hewer, *An iterative technique for the computation of the steady state gains for the discrete optimal regulator*, IEEE Trans. Automat. Control, 16 (1971), pp. 382–384.
[20] M. Hintermuller and J. Von Neumann Haus, *Nonlinear Optimization*, manuscript.
[21] D. Jacobson, *Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games*, IEEE Trans. Automat. Control, 18 (1973), pp. 124–131.
[22] I. Kaminer, P. P. Khargonekar, and M. A. Rotea, *Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control for discrete-time systems via convex optimization*, Automatica, J. IFAC, 29 (1993), pp. 57–70.
[23] P. P. Khargonekar and M. A. Rotea, *Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control: A convex optimization approach*, IEEE Trans. Automat. Control, 36 (1991), pp. 824–837.
[24] V. R. Konda and J. N. Tsitsiklis, *Actor-critic algorithms*, in Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2000, pp. 1008–1014.
[25] S. G. Krantz and H. R. Parks, *The Implicit Function Theorem: History, Theory, and Applications*, Springer, New York, 2012.
[26] T. Lillicrap, J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, *Continuous Control with Deep Reinforcement Learning*, in International Conference on Learning Representations, 2016.
[27] J. R. Magnus and H. Neudecker, *Matrix differential calculus with applications to simple, Hadamard, and Kronecker products*, J. Math. Psych., 29 (1985), pp. 474–492.
[28] P. Makila and H. Toivonen, *Computational methods for parametric LQ problems—A survey*, IEEE Trans. Automat. Control, 32 (1987), pp. 658–671.
[29] 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 International Conference on Artificial Intelligence and Statistics, 2019, pp. 2916–2925.
[30] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, *Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem*, IEEE Trans. Automat. Control, to appear.

[31] D. Mustafa and D. S. Bernstein, *LQG cost bounds in discrete-time $\mathcal{H}_2/\mathcal{H}_\infty$ control*, Trans. Inst. Measure. Control, 13 (1991), pp. 269–275.
[32] D. Mustafa and K. Glover, *Minimum entropy $\mathcal{H}_\infty$ control*, Lect. Notes Control Inf. Sci. 146, Springer, Berlin (1990).
[33] Y. Nesterov and B. T. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Program., 108 (2006), pp. 177–205.
[34] A. Ran and R. Vreugdenhil, *Existence and comparison theorems for algebraic Riccati equations for continuous- and discrete-time systems*, Linear Algebra Appl., 99 (1988), pp. 63–83.
[35] A. Rantzer, *On the Kalman-Yakubovich-Popov lemma*, Systems Control Lett., 28 (1996), pp. 7–10.
[36] H. Rotstein and M. Sznaier, *An exact solution to general four-block discrete-time mixed $\mathcal{H}_2/\mathcal{H}_\infty$ problems via convex optimization*, IEEE Trans. Automat. Control, 43 (1998), pp. 1475–1480.
[37] V. Roulet, M. Fazel, S. Srinivasa, and Z. Harchaoui, *On the convergence of the iterative linear exponential quadratic Gaussian algorithm to stationary points*, in American Control Conference, IEEE, Piscataway, NJ, 2020, pp. 132–137.
[38] C. W. Scherer, *Multiobjective $\mathcal{H}_2/\mathcal{H}_\infty$ control*, IEEE Trans. Automat. Control, 40 (1995), pp. 1054–1062.
[39] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, *Trust region policy optimization*, in International Conference on Machine Learning, 2015, Curran Associates, Red Hook, NY, pp. 1889–1897.
[40] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., *Mastering the game of Go with deep neural networks and tree search*, Nature, 529 (2016), pp. 484–489.
[41] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, *Policy gradient methods for reinforcement learning with function approximation*, in Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2000, pp. 1057–1063.
[42] E. E. Tyrtyshnikov, *A Brief Introduction to Numerical Analysis*, Birkhäuser, Boston, 1997.
[43] H. K. Venkataraman and P. J. Seiler, *Recovering robustness in model-free reinforcement learning*, in IEEE American Control Conference, IEEE, Piscataway, NJ, 2019, pp. 4210–4216.
[44] P. Whittle, *Risk-Sensitive Optimal Control*, Wiley, Chichester, England, 1990.
[45] G. Zames, *On the input-output stability of time-varying nonlinear feedback systems Part one: Conditions derived using concepts of loop gain, conicity, and positivity*, IEEE Trans. Automat. Control, 11 (1966), pp. 228–238.
[46] 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*, Proc. Mach. Learn. Res. (PMLR), 120 (2020), pp. 179–190.
[47] K. Zhang, Z. Yang, and T. Başar, *Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games*, in Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2019, pp. 11602–11614.
[48] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice Hall, Upper Saddle River, New Jersey, 1996.