Policy Optimization for $\mathcal{H}_2$ Linear Control with $\mathcal{H}_\infty$ Robustness Guarantee: Implicit Regularization and Global Convergence

Kaiqing Zhang KZHANG66@ILLINOIS.EDU
Bin Hu BINHU7@ILLINOIS.EDU
Tamer Başar BASAR1@ILLINOIS.EDU
Department of ECE & Coordinated Science Laboratory
University of Illinois at Urbana-Champaign, Champaign, IL, USA, 61820

Editors: A. Bayen, A. Jadbabaie, G. J. Pappas, P. Parrilo, B. Recht, C. Tomlin, M. Zeilinger

Abstract

Policy optimization (PO) is a key ingredient for modern reinforcement learning (RL). 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 (non-convex) 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 (LQR) 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, as if they are regularized by the algorithms. Furthermore, convergence to the *globally optimal* policies with *globally sublinear* and *locally (super-)linear* rates are provided under certain conditions, despite the nonconvexity of the problem. 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.

Keywords: Reinforcement learning; $\mathcal{H}_\infty$ robust control; policy optimization; implicit regularization; global convergence

1. Introduction

Recent years have seen tremendous success of reinforcement learning (RL) in various sequential decision-making applications (Silver et al., 2016; OpenAI, 2018; Vinyals et al., 2019) and continuous control tasks (Lillicrap et al., 2015; Schulman et al., 2015b; Recht, 2019). Interestingly, most successes hinge on the algorithmic framework of *policy optimization* (PO), umbrellaing policy gradient (PG) methods (Sutton et al., 2000; Kakade, 2002), actor-critic methods (Konda and Tsitsiklis, 2000; Bhatnagar et al., 2009; Zhang et al., 2018), trust-region (Schulman et al., 2015a) and proximal PO (Schulman et al., 2017) methods, etc. This inspires an increasing interest in studying the

This manuscript is a shorter version of the technical report Zhang et al. (2019a), with some of the technical results and simulations simplified/removed. Interested readers are referred to Zhang et al. (2019a) for a more complete treatment.

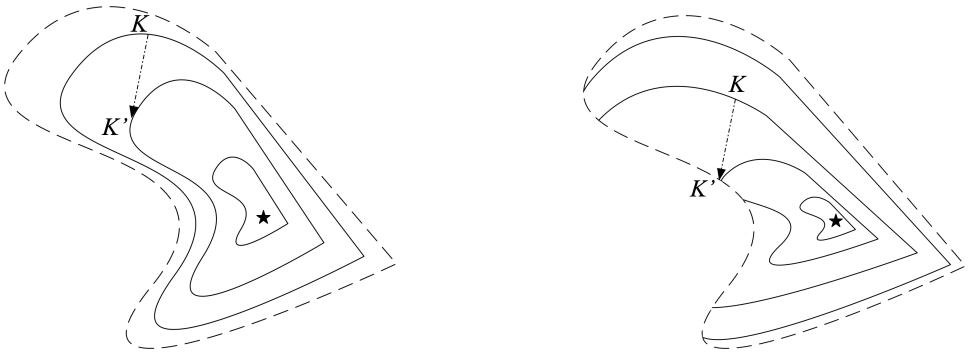
convergence theory, especially global convergence to optimal policies, of PO methods; see recent progresses in both classical RL contexts (Bhandari and Russo, 2019; Zhang et al., 2019b; Wang et al., 2019; Agarwal et al., 2019; Shani et al., 2019), and continuous control benchmarks (Fazel et al., 2018; Bu et al., 2019a; Malik et al., 2019; Tu and Recht, 2018; Zhang et al., 2019c).

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 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 (Fazel et al., 2018; Bu et al., 2019a). 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 (Fazel et al., 2018; Bu et al., 2019b). 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 $\mathcal{K}$ (Bu et al., 2019a). 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 (Zhou et al., 1996; Skogestad and Postlethwaite, 2007; Dullerud and Paganini, 2013; Apkarian et al., 2008) and risk-sensitive control (Whittle, 1990; Glover and Doyle, 1988). 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 (Glover and Doyle, 1988; Khargonekar and Rotea, 1991; Kaminer et al., 1993; Mustafa and Glover, 1990; Mustafa and Bernstein, 1991; Mustafa, 1989; Apkarian et al., 2008). 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 pre-specified value. This general framework also includes risk-sensitive linear control, modeled as linear exponential quadratic Gaussian (LEQG) (Jacobson, 1973; Whittle, 1990) problems as a special case. In addition, this framework is closely related to dynamic zero-sum LQ games (Başar and Bernhard, 1995).

Two challenges exist in the analysis of PO methods for mixed design problems. First, by definition, $\mathcal{H}_\infty$-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

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 v.s. reward, control v.s. action, etc.



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

Figure 1: Comparison of the landscapes of LQR and mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design that illustrates the hardness 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.

design is no longer *coercive*, as illustrated in Figure 1(b) (and formally established later). Therefore, the decrease of 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 stepsize that induces global convergence to the optimal policy. In this paper, we are able to show that two PO methods can indeed preserve the robustness constraint along the iterations, and enjoy global convergence guarantees.

Contribution. Our key contributions are three-fold: 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 (Fazel et al., 2018). 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. To the best of our knowledge, our work appears to be the first studying the implicit regularization properties of PO methods for *learning-based control* in general.

Due to space limitations, we refer to Zhang et al. (2019a) for a detailed review of the related work on RL with robustness/safety/risk-sensitivity concerns, the overarching theme of this work. Also, we defer the results for continuous-time settings to the longer report Zhang et al. (2019a).

2. Background

We first provide some background on $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantees.

2.1. Motivating Example: LEQG

We start with an example of *risk-sensitive* control, the infinite-horizon state-feedback linear exponential quadratic Gaussian problem[2] (Jacobson, 1973), which is motivating in that: i) the cost is closely related to the well-known linear optimal control problems, e.g., LQR and state-feedback

2. Unless otherwise noted, we will just refer to this problem as LEQG hereafter.

LQ Gaussian (LQG); ii) it illustrates the idea of mixed control design, especially introducing the $\mathcal{H}_\infty$-norm constraint, though implicit, that guarantees robustness.

Specifically, at time $t \geq 0$, the agent takes an action $u_t \in \mathbb{R}^d$ at state $x_t \in \mathbb{R}^m$, which leads the system to a new state x_{t+1} by a linear dynamical system

$$x_{t+1} = Ax_t + Bu_t + w_t, \qquad x_0 \sim \mathcal{N}(\mathbf{0}, X_0), \quad w_t \sim \mathcal{N}(\mathbf{0}, W),$$

where A and B are matrices of proper dimensions, $x_0 \in \mathbb{R}^m$ and $w_t \in \mathbb{R}^m, \forall t \geq 0$ are independent zero-mean Gaussian random variables with positive-definite covariance matrices X_0 and W, respectively. The one-stage cost of applying control u at state x is given by $c(x,u) = x^\top Qx + u^\top Ru$, where Q and R are positive-definite matrices. Then, the long-term cost to minimize is

$$\mathcal{J} := \limsup_{T\to\infty} \ \frac{1}{T}\frac{2}{\beta} \log \mathbb{E} \exp\left[\frac{\beta}{2}\sum_{t=0}^{T-1} c(x_t, u_t)\right], \tag{2.1}$$

where $\beta > 0$ describes the intensity of risk-sensitivity. We have proved in (Zhang et al., 2019a, Sec. 2.1), which is part of our contribution therein, that LEQG can be equivalently written as a constrained optimization problem over *linear time-invariant (LTI)* control gain $K \in \mathbb{R}^{d\times m}$

$$\min_K \ \mathcal{J}(K) := -\frac{1}{\beta}\log\det(I - \beta P_K W), \quad s.t. \ \rho(A - BK) < 1, \ \|\mathcal{T}(K)\|_\infty < 1/\sqrt{\beta}, \tag{2.2}$$

where $\mathcal{T}(K)$ is the closed-loop transfer function from the noise $\{w_t\}$ to the output $\{z_t\}$ with $z_t = Q^{1/2}x_t + R^{1/2}u_t$ under stabilizing controller $u_t = -Kx_t$, and P_K is the solution to some algebraic Riccati equation.

Note that the feasible set for LEQG above is implicit in the original formulation (2.1), which, though quite concise to characterize, is hard to enforce directly onto the control gain K, since it is a frequency-domain characterization using the $\mathcal{H}_\infty$-norm. In fact, this reformulation of LEQG belongs to a general class problems, named *mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design* with state-feedback.

2.2. Bigger Picture: Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Control Synthesis

Consider the following discrete-time linear dynamical system

$$x_{t+1} = Ax_t + Bu_t + Dw_t, \quad z_t = Cx_t + Eu_t, \tag{2.3}$$

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 Kaminer et al. (1993) that *LTI* state-feedback controller (without memory) suffices to achieve the optimal performance of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design under the *state-feedback* information structure. Hence, it suffices to consider only state-feedback controller parametrized as $u_t = -Kx_t$.

In accordance with Glover and Doyle (1988); Khargonekar and Rotea (1991); Başar and Bernhard (1995), we make the following assumption on the matrices A, B, C, D and E.

Assumption 2.1 *The matrices A, B, C, D, E in (2.3) satisfy $E^\top [C \ \ E] = [\mathbf{0} \ \ R]$ for some $R > 0$.*

Hence, the transfer function from the disturbance w_t to the output z_t can be represented as

$$\mathcal{T}(K) := \left[\begin{array}{c|c} A - BK & D \\ \hline (C^\top C + K^\top RK)^{1/2} & \mathbf{0} \end{array}\right]. \tag{2.4}$$

Then, robustness of the controller can be ensured by the constraint on the $\mathcal{H}_\infty$-norm, i.e., $\|\mathcal{T}(K)\|_\infty < \gamma$ for some $\gamma > 0$. The intuition, which follows from small gain theorem (Zames, 1966), 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 $\|\Delta\|_\infty < 1/\gamma$ may be connected from z_t back to w_t without destablizing the system. For convenience, we define the feasible set of the mixed-design problem as

$$\mathcal{K} := \big\{K \,\big|\, \rho(A - BK) < 1, \text{ and } \|\mathcal{T}(K)\|_\infty < \gamma\big\}. \tag{2.5}$$

The objective of the mixed design problem is usually an upper bound of the $\mathcal{H}_2$ norm of the closed-loop system. By a slight abuse of notation, let $\mathcal{J}(K)$ be the cost function. Several common forms of $\mathcal{J}(K)$ in the literature can be found in (Zhang et al., 2019a, Sec. 2.2). Here we choose

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

where P_K is the solution to the following Riccati equation

$$(A - BK)^\top \widetilde{P}_K (A - BK) + C^\top C + K^\top R K - P_K = 0, \tag{2.7}$$

with $\widetilde{P}_K$ defined as

$$\widetilde{P}_K := P_K + P_K D(\gamma^2 I - D^\top P_K D)^{-1} D^\top P_K. \tag{2.8}$$

The cost in (2.6) is closely related to maximum entropy $\mathcal{H}_\infty$-control (Mustafa and Glover, 1990), which, interestingly, also coincides with the closed-form cost of LEQG we have derived in (2.2) (up to some changes of variables). Thus, studying (2.6) solves LEQG as a by-product. In sum, the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design problem can be formulated as

$$\min_K \quad \mathcal{J}(K), \qquad s.t. \quad K \in \mathcal{K}, \tag{2.9}$$

with $\mathcal{J}(K)$ and $\mathcal{K}$ defined in (2.6) and (2.5), respectively. Next, we develop policy optimization methods for solving the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem in (2.9).

3. Landscape and Algorithms

In this section, we investigate the optimization landscape of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design, and develop policy optimization algorithms with convergence guarantees.

3.1. Optimization Landscape

We start by showing that the mixed-design problem in (2.9) is a *nonconvex* optimization problem with a *non-coercive* cost.

Lemma 3.1 (Nonconvexity and No Coercivity of Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Design) *The discrete-time mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem (2.9) is nonconvex. Moreover, the cost function (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.*

Lemma 3.1, which is the basis of the illustration in Figure 1, is a combination of Lemmas 3.1 and 3.2 in Zhang et al. (2019a), whose proofs are deferred to §B therein. In particular, the proofs follow by constructing examples showing that the constraint set $\mathcal{K}$ is nonconvex, and $\mathcal{J}(K)$ approaches to a finite value as $K \to \partial\mathcal{K}$. Due to the nonconvexity, finding the global optimum using policy gradient methods is NP-hard in general. The lack of coercivity further complicates the analysis for the *stability/feasibility* of the iterates as the methods proceed, in contrast to that for LQR problems.

For algorithm design, we then derive the form of the policy gradient of $\mathcal{J}(K)$ within $\mathcal{K}$.

Lemma 3.2 (Policy Gradient for Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Design) *The objective $\mathcal{J}(K)$ is differentiable with respect to K for any $K \in \mathcal{K}$. The policy gradient has the following form:*

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

with $\Delta_K \in \mathbb{R}^{m\times m}$ being the unique solution to the Lyapunov equation

$$\Delta_K = D(I - \gamma^{-2} D^\top P_K D)^{-1} D^\top + A_K \Delta_K A_K^\top, \tag{3.1}$$

where $A_K := (I - \gamma^{-2} P_K D D^\top)^{-\top}(A - BK)$, and $\widetilde{P}_K$ is defined in (2.8).

Lemma 3.2 corresponds to Lemmas 3.3 and 3.4 in Zhang et al. (2019a), whose proofs are provided in §B.4 and §B.5 therein. Note that in the proof, by the Bounded Real Lemma, see Lemma 2.7 in Zhang et al. (2019a), any $K \in \mathcal{K}$ ensures that A_K is stabilizing and $I - \gamma^{-2} D^\top P_K D > 0$, making $\Delta_K \geq 0$ well defined as (3.1) admits a non-negative definite and unique solution in this case. Lemma 3.2 also implies that if Δ_K is full-rank, then $\nabla \mathcal{J}(K) = 0$ admits the unique solution $K = (R + B^\top \widetilde{P}_K B)^{-1} B^\top \widetilde{P}_K A$. This unique stationary point thus becomes the global optimum in $\mathcal{K}$. We formally establish it in the following corollary proved in §B.6 of Zhang et al. (2019a).

Corollary 3.3 *Suppose that the discrete-time mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design admits a control gain solution $K^* \in \mathcal{K}$, and for any stationary point $K \in \mathcal{K}$ such that $\nabla \mathcal{J}(K) = 0$, the pair $\big((I - \gamma^{-2} P_K D D^\top)^{-\top}(A - BK), D\big)$ is controllable. Then, such a solution is unique, and has the form of $K^* = (R + B^\top \widetilde{P}_{K^*} B)^{-1} B^\top \widetilde{P}_{K^*} A$.*

Note that for LEQG with $D = W^{1/2} > 0$, the controllability condition in Corollary 3.3 is satisfied for any $K \in \mathcal{K}$. Therefore, the argument that *stationary point implies global optimum* holds for LEQG naturally. Also, the controllability assumption above is standard for mixed design, and has also been made in Mustafa and Bernstein (1991).

In order to find the global optimum under the conditions of Corollary 3.3, it suffices to find the first-order stationary point, which can be obtained using first-order policy optimization methods.

3.2. Policy Optimization Algorithms

Consider three policy-gradient methods as follows. For simplicity, we define $E_K := (R + B^\top \widetilde{P}_K B)K - B^\top \widetilde{P}_K A$, and use K and K' to denote the control gain before and after one-step of the update.

Policy Gradient:
$$K' = K - \eta \nabla \mathcal{J}(K) = K - 2\eta E_K \Delta_K \tag{3.2}$$

Natural Policy Gradient:
$$K' = K - \eta \nabla \mathcal{J}(K) \Delta_K^{-1} = K - 2\eta E_K \tag{3.3}$$

Gauss-Newton:
$$\begin{aligned} K' &= K - \eta (R + B^\top \widetilde{P}_K B)^{-1} \nabla \mathcal{J}(K) \Delta_K^{-1} \\ &= K - 2\eta (R + B^\top \widetilde{P}_K B)^{-1} E_K \end{aligned} \tag{3.4}$$

where $\eta > 0$ is the stepsize. The updates are motivated by and resemble the policy optimization updates for LQR (Fazel et al., 2018; Bu et al., 2019a), but with P_K therein replaced by $\widetilde{P}_K$. The natural PG update is related to gradient over a Riemannian manifold; while the Gauss-Newton update is one type of quasi-Newton update. In particular, with $\eta = 1/2$, the Gauss-Newton update (3.4) can be viewed as the *policy iteration* update for infinite-horizon mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design.

4. Theoretical Results

In this section, we investigate the convergence of the PO methods proposed in §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 important 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 (2.9), 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.

The concept of *(implicit) regularization* has been adopted in many recent studies on nonconvex optimization, including training neural networks (Allen-Zhu et al., 2018; Kubo et al., 2019), phase retrieval (Chen and Candes, 2015; Ma et al., 2017), etc., referring to any scheme that biases the search direction of gradient-based algorithms. *Implicit* here means that the iterates are regularized/biased as if an explicit regularization (projection) is imposed. In fact, by Lemma 3.1, projection onto $\mathcal{K}$ is in general intractable, as $\mathcal{K}$ is a nonconvex set. See more discussions on the use of this terminology in (Zhang et al., 2019a, Remark 4.2).

As mentioned earlier, the iterates of PG for LQR are implicitly regularized, due to the *coercivity* of the cost: any decrease of the cost ensures stay in the feasibility set. This holds for *any* descent direction of the cost. Though coercivity does not hold in mixed design, interestingly, we show in the following theorem that the updates following certain directions, the natural PG and Gauss-Newton updates in (3.3)-(3.4), still enjoy the implicit regularization feature, with *constant* stepsize.

Theorem 4.2 (Implicit Regularization for Mixed Design) *For any control gain $K \in \mathcal{K}$, i.e., $\rho(A - BK) < 1$ and $\|\mathcal{T}(K)\|_\infty < \gamma$, with $\|K\| < \infty$, suppose that the stepsize η satisfies: for Natural policy gradient (3.3) $\eta \leq 1/(2\|R + B^\top \widetilde{P}_K B\|)$ and for Gauss-Newton (3.4): $\eta \leq 1/2$. Then the K' obtained from (3.3)-(3.4) also lies in $\mathcal{K}$.*

4.2. Global Convergence

We now focus on the convergence property of the two methods with implicit regularization. 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 in the following theorem.

Theorem 4.3 (Global Convergence for Discrete-Time Mixed Design) *Suppose that $K_0 \in \mathcal{K}$ and $\|K_0\| < \infty$. Then, under the stepsize choices[3] in Theorem 4.2, both updates (3.3) and (3.4) converge to the stationary points K where $E_K = \mathbf{0}$, such that the average of $\{\|E_{K_n}\|_F^2\}$ over iterations converges to 0 with $O(1/N)$ rate. Also, if $\big((I - \gamma^{-2} P_K D D^\top)^{-\top}(A - BK), D\big)$ is controllable at the stationary point K, then such a convergence is towards the unique* global *optimal policy.*

The proof of Theorem 4.3 is detailed in §5.2 in Zhang et al. (2019a). The controllability assumption has been made in Corollary 3.3, and is satisfied automatically for LEQG problems with $D = W^{1/2} > 0$. Moreover, in contrast to the results for LQR (Fazel et al., 2018), only globally

3. In fact, for natural PG (3.3), it suffices to require the stepsize $\eta \leq 1/(2\|R + B^\top \widetilde{P}_{K_0} B\|)$ for the initial K_0.

sublinear $O(1/N)$, instead of *linear*, convergence rate 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 general nonconvex optimization (Cartis et al., 2010, 2017; Khamaru and Wainwright, 2018).

Though sublinear globally, much faster rates of (super-)linear can be shown locally around the optimum as below. Proof of the following theorem can be found in §5.3 in Zhang et al. (2019a). The intuition of locally linear rates is that the property of *gradient dominance* (Polyak, 1963; Nesterov and Polyak, 2006) holds locally around the optimum for mixed design problems. Such a property has been shown to hold globally for LQR problems (Fazel et al., 2018), and also hold locally for zero-sum LQ games (Zhang et al., 2019c). The Q-quadratic rate also echoes back the rate of Gauss-Newton with $\eta = 1/2$ for LQR problems (Hewer, 1971; Bu et al., 2019a).

Theorem 4.4 (Local (Super-)Linear Convergence for Mixed Design) *Suppose that the conditions in Theorem 4.3 hold, and additionally* $DD^\top > 0$. *Then, under the stepsize choices as in Theorem 4.3, both updates* (3.3) *and* (3.4) *converge to the optimal control gain* K^* *with* locally linear *rate, such that the objective* $\{\mathcal{J}(K_n)\}$ *converges to* $\mathcal{J}(K^*)$ *with a linear rate. In addition, if* $\eta = 1/2$, *the Gauss-Newton update* (3.4) *converges to* K^* *with a locally* Q-quadratic *rate.*

Remark 4.5 (Comparison to Zhang et al. (2019c)) *Due to the close relationship between mixed design and zero-sum LQ games, see* §6 *in Zhang et al. (2019a), one may compare the convergence results and find the rates here (*globally sublinear and locally linear*) not improved over Zhang et al. (2019c). However, one key difference is that an extra* projection *step is required to guarantee the* stability *of the system in Zhang et al. (2019c), which is essentially to* regularize *the iterates* explicitly. *More importantly, such a projection can only be calculated under more restrictive assumptions (see Assumption 2.1 therein), which, though covering a class of LQ games, are not standard in robust control. Here, similar convergence results are established, without projections or non-standard assumptions in robust control, thanks to* implicit regularization. *Moreover, we have established the local "superlinear" rate for the Gauss-Newton update.*

5. Concluding Remarks

In this paper, we 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 PO methods with provable convergence to the global optimal policy. More importantly, we showed that the proposed PO methods enjoy the implicit regularization property, despite the lack of coercivity of the cost function. We note that due to the equivalence of mixed design and zero-sum linear quadratic games, see §6 in Zhang et al. (2019a) and the references therein, our results can be applied to study PO methods for LQ games as well. Moreover, our results, together with this connection to LQ games, pave the way for developing model-free versions of our algorithms, which is left as our future work.

Acknowledgments

K. Zhang and T. Başar were supported by the US Army Research Laboratory (ARL) Cooperative Agreement W911NF-17-2-0196.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018.

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

Tamer Başar and Pierre Bernhard. *H-infinity Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Birkhäuser, Boston., 1995.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Jingjing Bu, Afshin Mesbahi, Maryam Fazel, and Mehran Mesbahi. LQR through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019a.

Jingjing Bu, Afshin Mesbahi, and Mehran Mesbahi. On topological and metrical properties of stabilizing feedback gains: the MIMO case. *arXiv preprint arXiv:1904.02737*, 2019b.

Coralia Cartis, Nicholas IM Gould, and Ph L Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010.

Coralia Cartis, Nick IM Gould, and Philippe L Toint. Worst-case evaluation complexity and optimality of second-order methods for nonconvex smooth optimization. *arXiv preprint arXiv:1709.07180*, 2017.

Yuxin Chen and Emmanuel Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747, 2015.

Geir E Dullerud and Fernando Paganini. *A Course in Robust Control Theory: A Convex Approach*, volume 36. Springer Science & Business Media, 2013.

Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, 2018.

Keith Glover and John 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 Letters*, 11(3): 167–172, 1988.

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

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

Sham M Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, pages 1531–1538, 2002.

Isaac Kaminer, Pramod P Khargonekar, and Mario A Rotea. Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control for discrete-time systems via convex optimization. *Automatica*, 29(1):57–70, 1993.

Koulik Khamaru and Martin J Wainwright. Convergence guarantees for a class of non-convex and non-smooth optimization problems. *arXiv preprint arXiv:1804.09629*, 2018.

Pramod P Khargonekar and Mario A Rotea. Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control: A convex optimization approach. *IEEE Transactions on Automatic Control*, 36(7):824–837, 1991.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014, 2000.

Masayoshi Kubo, Ryotaro Banno, Hidetaka Manabe, and Masataka Minoji. Implicit regularization in over-parameterized neural networks. *arXiv preprint arXiv:1903.01997*, 2019.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *International Conference on Artificial Intelligence and Statistics*, pages 2916–2925, 2019.

Denis Mustafa. Relations between maximum-entropy/$\mathcal{H}_\infty$ control and combined $\mathcal{H}_\infty$/LQG control. *Systems & Control Letters*, 12(3):193–203, 1989.

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

Denis Mustafa and Keith Glover. Minimum entropy $\mathcal{H}_\infty$ control. *Lecture Notes in Control and Information Sciences*, 1990.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

OpenAI. Openai five. https://blog.openai.com/openai-five/, 2018.

Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):14–29, 1963.

Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015a.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. *arXiv preprint arXiv:1909.02769*, 2019.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Sigurd Skogestad and Ian Postlethwaite. *Multivariable Feedback Control: Analysis and Design*, volume 2. Wiley New York, 2007.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. *arXiv preprint arXiv:1812.03565*, 2018.

Oriol Vinyals, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojciech M. Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, Timo Ewalds, Dan Horgan, Manuel Kroiss, Ivo Danihelka, John Agapiou, Junhyuk Oh, Valentin Dalibard, David Choi, Laurent Sifre, Yury Sulsky, Sasha Vezhnevets, James Molloy, Trevor Cai, David Budden, Tom Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Toby Pohlen, Yuhuai Wu, Dani Yogatama, Julia Cohen, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Chris Apps, Koray Kavukcuoglu, Demis Hassabis, and David Silver. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/, 2019.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Peter Whittle. *Risk-sensitive Optimal Control*. Wiley Chichester, 1990.

George Zames. On the input-output stability of time-varying nonlinear feedback systems part one: Conditions derived using concepts of loop gain, conicity, and positivity. *IEEE Transactions on Automatic Control*, 11(2):228–238, 1966.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, pages 5872–5881, 2018.

Kaiqing Zhang, Bin Hu, and Tamer Başar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. *arXiv preprint arXiv:1910.09496*, 2019a.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383*, 2019b.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, 2019c.

Kemin Zhou, John Comstock Doyle, and Keith Glover. *Robust and Optimal Control*, volume 40. Prentice Hall New Jersey, 1996.