Global Convergence of Direct Policy Search for State-Feedback $\mathcal{H}_\infty$ Robust Control: A Revisit of Nonsmooth Synthesis with Goldstein Subdifferential

Anonymous Author(s)
Affiliation
Address
email

Abstract

Direct policy search has been widely applied in modern reinforcement learning and continuous control. However, the performance of direct policy search on nonsmooth robust control synthesis has not been well understood. The optimal $\mathcal{H}_\infty$ control framework aims at designing a policy to minimize the closed-loop $\mathcal{H}_\infty$ norm, and is arguably the most important robust control paradigm. In this work, we show that direct policy search is guaranteed to find the global solution of the robust $\mathcal{H}_\infty$ state-feedback control design problem. Notice that policy search for optimal $\mathcal{H}_\infty$ control leads to a constrained nonconvex nonsmooth optimization problem where the nonconvex feasible set consists of all the policies stabilizing the closed-loop dynamics. We show that for this nonsmooth optimization problem, all Clarke stationary points are global minimum. Next, we identify the coercivity of the closed-loop $\mathcal{H}_\infty$ objective function, and prove that the sublevel sets of the resultant policy search problem are compact. Based on these properties, we show that the Goldstein subdifferential method and its various implementable variants can be guaranteed to stay in the nonconvex feasible set and eventually find the global optimal solution for the $\mathcal{H}_\infty$ state-feedback synthesis problem. Our work builds a new connection between nonconvex nonsmooth optimization theory and robust control, leading to the first global convergence result for direct policy search on optimal $\mathcal{H}_\infty$ synthesis.

1 Introduction

Reinforcement learning (RL) has achieved impressive performance on many continuous control tasks [49, 34], and policy optimization is one of the main workhorses for such applications [15, 52, 48, 50]. Recently, there have been extensive research efforts studying the global convergence properties of policy optimization methods on benchmark control problems including linear quadratic regulator (LQR) [18, 7, 35, 57, 37, 19, 33, 25], stabilization [44], linear robust/risk-sensitive control [60, 59, 23, 61, 62], Markov jump linear quadratic control [28, 27, 46], output feedback control [17, 63], and dynamic filtering [55]. For all these benchmark problems, the objective function in the policy optimization formulation is always differentiable over the entire feasible set, and the existing convergence theory heavily relies on this fact. Consequently, an important open question remains whether direct policy search can enjoy similar global convergence properties when applied to the famous $\mathcal{H}_\infty$ control problem whose objective function is not always differentiable in the policy space [1, 2, 3, 24, 9, 12, 41]. Different from LQR which considers stochastic disturbance, $\mathcal{H}_\infty$ control directly addresses the worst-case disturbance, and provides arguably the most fundamental robust control paradigm [64, 16, 51, 4, 14, 20]. Regarding the connection with RL, it has also been shown that $\mathcal{H}_\infty$ control can be applied to stabilize the training of adversarial RL schemes in the linear

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

quadratic setup [59, Section 5]. Given the fundamental importance of $\mathcal{H}_\infty$ control, we view it as an important benchmark for understanding the theoretical properties of direct policy search in the context of robust control and adversarial RL. The main contribution of our paper is the first global convergence guarantee of direct policy search for the $\mathcal{H}_\infty$ state-feedback control synthesis problem.

The objective of the $\mathcal{H}_\infty$ state-feedback synthesis is to design a linear state-feedback policy that stabilizes the closed-loop system and minimizes the $\mathcal{H}_\infty$ norm from the disturbance to a performance signal at the same time. The design goal is also equivalent to synthesizing a state-feedback policy that minimizes a quadratic cost subject to the worst-case disturbance. We will present the problem formulation for the $\mathcal{H}_\infty$ state-feedback synthesis and discuss such connections in Section 2. Essentially, $\mathcal{H}_\infty$ state-feedback synthesis can be formulated as a constrained policy optimization problem $\min_{K\in\mathcal{K}} J(K)$, where the decision variable K is a matrix parameterizing the linear state-feedback policy, the objective function $J(K)$ is the closed-loop $\mathcal{H}_\infty$-norm for given K, and the feasible set $\mathcal{K}$ consists of all the linear state-feedback policies stabilizing the closed-loop dynamics. Notice that the feasible set for the $\mathcal{H}_\infty$ state-feedback control problem is the same as the nonconvex feasible set for the LQR policy search problem [18, 7]. However, the objective function $J(K)$ for the $\mathcal{H}_\infty$ control problem can be non-differential over certain feasible points, introducing new difficulty to direct policy search. There has been a large family of nonsmooth $\mathcal{H}_\infty$ policy search algorithms developed based on the concept of Clarke subdifferential [1, 2, 3, 24, 9, 12]. However, no global convergence guarantee has been reported in the literature. Our paper bridges this gap by making the following two contributions.

1. We show that all Clarke stationary points for the $\mathcal{H}_\infty$ state-feedback control problem are also global minimum.
2. We identify the coercivity property of the $\mathcal{H}_\infty$ cost function and use this property to show that the Goldstein subdifferential method [22] and its implementable variants [58, 13, 9, 10, 31, 32] can be guaranteed to stay in the nonconvex feasible set of stabilizing policies during the optimization process and eventually find the global optimal solution of the $\mathcal{H}_\infty$ state-feedback control problem. Finite-time complexity bounds for finding (δ, ϵ)-stationary points are also provided.

To the best of our knowledge, our work offers the first global convergence guarantee of direct policy search on $\mathcal{H}_\infty$ control problems. We believe that our work also serves as a meaningful initial step towards a general convergence theory of direct policy search on nonsmooth robust control synthesis.

Finally, it is worth clarifying the differences between $\mathcal{H}_\infty$ control and mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design. For mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control, the objective is to design a stabilizing policy that minimizes an $\mathcal{H}_2$ performance bound and satisfies an $\mathcal{H}_\infty$ constraint at the same time [21, 30, 29, 40]. In other words, mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control aims at improving the average $\mathcal{H}_2$ performance while "maintaining" a certain level of robustness by keeping the closed-loop $\mathcal{H}_\infty$ norm to be smaller than a pre-specified number. In contrast, $\mathcal{H}_\infty$ control aims at "improving" the system robustness and the worst-case performance via achieving the smallest closed-loop $\mathcal{H}_\infty$ norm. In [60], it has been shown that the natural policy gradient method initialized from a policy satisfying the $\mathcal{H}_\infty$ constraint can be guaranteed to maintain the $\mathcal{H}_\infty$ requirement during the optimization process and eventually converge to the optimal solution of the mixed design problem. However, notice that the objective function for the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem is still differentiable over all the feasible points, and hence the analysis technique in [60] cannot be applied to our $\mathcal{H}_\infty$ control setting. More discussion on the connections and differences between these two problems will be presented in the supplementary material.

.

2 Problem Formulation and Preliminaries

2.1 Notation

The set of p-dimensional real vectors is denoted as $\mathbb{R}^p$. For a matrix A, we use the notation A^T, $\|A\|$, $\operatorname{tr} A$, $\sigma_{\min}(A)$, $\|A\|_2$, and $\rho(A)$ to denote its transpose, largest singular value, trace, smallest singular value, Frobenius norm, and spectral radius, respectively. When a matrix P is negative semidefinite (definite), we will use the notation $P \preceq (\prec) 0$. When P is positive semidefinite (definite), we use the notation $P \succeq (\succ) 0$. Consider a (real) sequence $\mathbf{u} := \{u_0, u_1, \cdots\}$ where $u_t \in \mathbb{R}^{n_u}$ for all t. This sequence is said to be in $\ell_2^{n_u}$ if $\sum_{t=0}^{\infty} \|u_t\|^2 < \infty$ where $\|u_t\|$ denotes the standard (vector) 2-norm of u_t. In addition, the 2-norm for $\mathbf{u} \in \ell_2^{n_u}$ is defined as $\|\mathbf{u}\|^2 := \sum_{t=0}^{\infty} \|u_t\|^2$.

2.2 Problem statement: $\mathcal{H}_\infty$ state-feedback synthesis and a policy optimization formulation

We consider the following linear time-invariant (LTI) system

$$x_{t+1} = Ax_t + Bu_t + w_t,\ x_0 = 0 \tag{1}$$

where $x_t \in \mathbb{R}^{n_x}$ is the state, $u_t \in \mathbb{R}^{n_u}$ is the control action, and $w_t \in \mathbb{R}^{n_w}$ is the disturbance. We have $A \in \mathbb{R}^{n_x \times n_x}$, $B \in \mathbb{R}^{n_x \times n_u}$, and $n_w = n_x$. We denote $\mathbf{x} := \{x_0, x_1, \cdots\}$, $\mathbf{u} := \{u_0, u_1, \cdots\}$, and $\mathbf{w} := \{w_0, w_1, \cdots\}$. The initial condition is fixed as $x_0 = 0$. The objective of $\mathcal{H}_\infty$ control is to choose $\{u_t\}$ to minimize the quadratic cost $\sum_{t=0}^{\infty}(x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t)$ in the presence of the worst-case ℓ_2 disturbance satisfying $\|\mathbf{w}\| \leq 1$. In this paper, the following assumption is adopted.

Assumption 1. *The matrices Q and R are positive definite. The matrix pair (A, B) is stabilizable.*

In $\mathcal{H}_\infty$ control, $\{w_t\}$ is considered to be the worst-case disturbance satisfying the ℓ_2 norm bound $\|\mathbf{w}\| \leq 1$, and can be chosen in an adversarial manner. This is different from LQR which makes stochastic assumptions on $\{w_t\}$. Without loss of generality, we have chosen the ℓ_2 upper bound on $\mathbf{w}$ to be 1. In principle, we can formulate the $\mathcal{H}_\infty$ control problem with any arbitrary ℓ_2 upper bound on $\mathbf{w}$, and there is no technical difference. We will provide more explanations on this fact in the supplementary material. Therefore, $\mathcal{H}_\infty$ control can be formulated as the following minimax problem

$$\min_{\mathbf{u}} \max_{\mathbf{w}:\|\mathbf{w}\|\leq 1} \sum_{t=0}^{\infty}(x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t) \tag{2}$$

Under Assumption 1, it is well known that the optimal solution for (2) can be achieved using a linear state-feedback policy $u_t = -Kx_t$ (see [4]). Given any K, the LTI system (1) can be rewritten as

$$x_{t+1} = (A - BK)x_t + w_t,\ x_0 = 0. \tag{3}$$

Now we define $z_t = (Q + K^\mathsf{T} R K)^{\frac{1}{2}} x_t$. We have $\|z_t\|^2 = x_t^\mathsf{T}(Q + K^\mathsf{T} R K)x_t = x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t$. We denote $\mathbf{z} := \{z_0, z_1, \cdots\}$. If $\mathbf{x} \in \ell_2^{n_x}$, then we have $\|\mathbf{z}\|^2 = \sum_{t=0}^{\infty}(x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t) < +\infty$. Therefore, the closed-loop LTI system (3) can be viewed as a linear operator mapping any disturbance sequence $\{w_t\}$ to another sequence $\{z_t\}$. We denote this operator as G_K, where the subscript highlights the dependence of this operator on K. If K is stabilizing, i.e. $\rho(A - BK) < 1$, then G_K is bounded in the sense that it maps any ℓ_2 sequence $\mathbf{w}$ to another sequence $\mathbf{z}$ in $\ell_2^{n_x}$. For any stabilizing K, the $\ell_2 \to \ell_2$ induced norm of G_K can be defined as:

$$\|G_K\|_{2\to 2} := \sup_{0 \neq \|\mathbf{w}\| \leq 1} \frac{\|\mathbf{z}\|}{\|\mathbf{w}\|} \tag{4}$$

Since G_K is a linear operator, it is straightforward to show

$$\|G_K\|_{2\to 2}^2 := \max_{\mathbf{w}:\|\mathbf{w}\|\leq 1} \sum_{t=0}^{\infty} x_t^\mathsf{T}(Q + K^\mathsf{T} R K)x_t = \max_{\mathbf{w}:\|\mathbf{w}\|\leq 1} \sum_{t=0}^{\infty}(x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t).$$

Therefore, the minimax optimization problem (2) can be rewritten as the policy optimization problem: $\min_{K\in\mathcal{K}} \|G_K\|_{2\to 2}^2$, where $\mathcal{K}$ is the set of all linear state-feedback stabilizing policies, i.e. $\mathcal{K} = \{K \in \mathbb{R}^{n_x \times n_u} : \rho(A - BK) < 1\}$. In the robust control literature [1, 2, 3, 24, 9, 12], it is standard to drop the square in the cost function and just reformulate (2) as $\min_{K\in\mathcal{K}} \|G_K\|_{2\to 2}$. This is exactly the policy optimization formulation for $\mathcal{H}_\infty$ state-feedback control. The main reason why this problem is termed as $\mathcal{H}_\infty$ state-feedback control is that in the frequency domain, G_K can be viewed as a transfer function which lives in the Hardy $\mathcal{H}_\infty$ space and has an $\mathcal{H}_\infty$ norm being exactly equal to $\|G_K\|_{2\to 2}$. Applying the frequency-domain formula for the $\mathcal{H}_\infty$ norm, we can calculate $\|G_K\|_{2\to 2}$ as

$$\|G_K\|_{2\to 2} = \sup_{\omega\in[0,2\pi]} \lambda_{\max}^{1/2}\big((e^{-j\omega}I - A + BK)^{-\mathsf{T}}(Q + K^\mathsf{T} R K)(e^{j\omega}I - A + BK)^{-1}\big), \tag{5}$$

where I is the identity matrix, and $\lambda_{\max}$ denotes the largest eigenvalue of a given symmetric matrix. Therefore, eventually the $\mathcal{H}_\infty$ state-feedback control problem can be formulated as

$$\min_{K\in\mathcal{K}} J(K), \tag{6}$$

where $J(K)$ is equal to the $\mathcal{H}_\infty$ norm specified by (5). Classical $\mathcal{H}_\infty$ control theory typically solves (6) via introducing extra Lyapunov variables and reparameterizing the problem into a higher-dimensional convex domain over which convex optimization algorithms can be applied [64, 16, 6]. In this paper, we revisit (6) as a benchmark for direct policy search, and discuss how to search the optimal solution of (6) in the policy space directly. Applying direct policy search for (6) leads to a nonconvex nonsmooth optimization problem. A main technical challenge is that the objective function (5) can be non-differentiable over some important feasible points [1, 2, 3, 24, 9, 12].

2.3 Direct policy search: A nonsmooth optimization perspective

Now we briefly review several key facts known for the $\mathcal{H}_\infty$ policy optimization problem (6).

Proposition 1. *The set $\mathcal{K} = \{K : \rho(A - BK) < 1\}$ is open. In general, it can be unbounded and nonconvex. The cost function* (5) *is continuous and nonconvex in K.*

See [18, 8] for some related proofs. We have also included more explanations in the supplementary material. An immediate consequence is that (6) becomes a nonconvex optimization problem. Another important fact is that the objective function (5) is also nonsmooth. As a matter of fact, (5) is subject to two sources of nonsmoothness. Based on (5), we can see that the largest eigenvalue for a fixed frequency ω is nonsmooth, and the optimization step over $\omega \in [0, 2\pi]$ is also nonsmooth. As a matter of fact, the $\mathcal{H}_\infty$ objective function (5) can be non-differentiable over important feasible points, e.g. optimal points. Fortunately, it is well known[1] that the $\mathcal{H}_\infty$ objective function (5) has the following desired property so it is Clarke subdifferentiable.

Proposition 2. *The $\mathcal{H}_\infty$ objective function* (5) *is locally Lipschitz and subdifferentially regular over the stabilizing feasible set $\mathcal{K}$.*

Recall that $J : \mathcal{K} \to \mathbb{R}$ is locally Lipschitz if for any bounded $S \subset \mathcal{K}$, there exists a constant $L > 0$ such that $|J(K) - J(K')| \leq L\|K - K'\|_2$ for all $K, K' \in S$. Based on Rademacher's theorem, a locally Lipschitz function is differentiable almost everywhere, and the Clarke subdifferential is well defined for all feasible points. Formally, the Clarke subdifferential is defined as

$$\partial_C J(K) := \mathrm{conv}\{\lim_{i\to\infty} \nabla J(K_i) : K_i \to K,\ K_i \in \mathrm{dom}(\nabla J) \subset \mathcal{K}\} \tag{7}$$

where conv denotes the convex hull. Then we know that the Clarke subdifferential for the $\mathcal{H}_\infty$ objective function (5) is well defined for all $K \in \mathcal{K}$. We say that K is a Clarke stationary point if $0 \in \partial_C J(K)$. The following fact is also well known.

Proposition 3. *If K is a local min of J, then $0 \in \partial_C J(K)$ and K is a Clarke stationary point.*

Under Assumption 1, it is well known that there exists $K^* \in \mathcal{K}$ achieving the minimum of (6). Since $\mathcal{K}$ is an open set, K^* has to be an interior point of $\mathcal{K}$ and hence K^* has to be a Clarke stationary point. In Section 3, we will prove that any Clarke stationary points for (6) are actually global minimum.

Now we briefly elaborate on the subdifferentially regular property stated in Proposition 2. For any given direction d (which has the same dimension as K), the generalized Clarke directional derivative of J is defined as

$$J^\circ(K, d) := \lim_{K' \to K} \sup_{t \searrow 0} \frac{J(K' + td) - J(K')}{t}. \tag{8}$$

In contrast, the (ordinary) directional derivative is defined as follows (when existing)

$$J'(K, d) := \lim_{t \searrow 0} \frac{J(K + td) - J(K)}{t}. \tag{9}$$

In general, the Clarke directional derivative can be different from the directional derivative, and the directional derivative does not always exist. The objective function $J(K)$ is subdifferentially regular if for every $K \in \mathcal{K}$, the ordinary directional derivative always exists and coincides with the generalized one for every direction, i.e. $J'(K, d) = J^\circ(K, d)$. The most important consequence of the subdifferentially regular property is given as follows.

Corollary 1. *Suppose $K^\dagger \in \mathcal{K}$ is a Clarke stationary point for J. If J is subdifferentially regular, then the directional derivatives $J'(K^\dagger, d)$ are non-negative for all d.*

[1] We cannot find a formal statement of Proposition 2 in the literature. However, based on our discussion with other researchers who have worked on nonsmooth $\mathcal{H}_\infty$ synthesis for long time, this fact is well known and hence we do not claim any credits in deriving this result. As a matter of fact, although not explicitly stated, the proof of Proposition 2 is hinted in the last paragraph of [2, Section III] given the facts that the $\mathcal{H}_\infty$ norm is a convex function over the Hardy $\mathcal{H}_\infty$ space (which is a Banach space) and the mapping from $K \in \mathcal{K}$ to the (infinite-dimensional) Hardy $\mathcal{H}_\infty$ space is smooth. For completeness, a simple proof of Proposition 2 based on Clarke's chain rule is included in the supplementary material.

See [11] for related proofs. Notice that having non-negative directional derivatives does not mean that the point $K^\dagger$ is a local minimum. Nevertheless, the above fact will be used in our main theoretical developments. Now we briefly summarize two key difficulties in establishing a global convergence theory for direct policy search on the $\mathcal{H}_\infty$ state-feedback control problem (6). First, it is unclear whether the direct policy search method will get stuck at some local minimum. Second, it is challenging to guarantee the direct policy search method to stay in the nonconvex feasible set $\mathcal{K}$ during the optimization process. Since $\mathcal{K}$ is nonconvex, we cannot use a projection step to maintain feasibility. Our main results will address these two issues.

2.4 Goldstein subdifferential

Generating a descent direction for nonsmooth optimization is not trivial. Many nonsmooth optimization algorithms are based on the concept of Goldstein subdifferential [22]. Before proceeding to our main result, we briefly review this concept here.

Definition 1 (Goldstein subdifferential)**.** *Suppose J is locally Lipschitz. Given a point $K \in \mathcal{K}$ and a parameter $\delta > 0$, the Goldstein subdifferential of J at K is defined to be the following set*

$$\partial_\delta J(K) := \operatorname{conv}\left\{\cup_{K' \in \mathbb{B}_\delta(K)} \partial_C J(K')\right\}, \tag{10}$$

where $\mathbb{B}_\delta(K)$ denotes the δ-ball around K. The above definition implicitly requires $\mathbb{B}_\delta(K) \subset \mathcal{K}$.

A point K is (δ, ϵ)-stationary if $\operatorname{dist}(0, \partial_\delta J(K)) \leq \epsilon$. It is well-known that the minimal norm 183 element of the Goldstein subdifferential generates a descent direction. This fact is stated as follows.

Proposition 4 ([22])**.** *Let F be the minimal norm element in $\partial_\delta J(K)$. Suppose $K - \alpha F/\|F\|_2 \in \mathcal{K}$ for any $0 \leq \alpha \leq \delta$. Then we have*

$$J(K - \delta F/\|F\|_2) \leq J(K) - \delta\|F\|_2. \tag{11}$$

The idea of Goldstein subdifferential has been used in designing algorithms for nonsmooth $\mathcal{H}_\infty$ control [3, 24, 9, 12]. We will show that the Goldstein subdifferential method can be guaranteed to find the global minimum of (6).

3 Optimization Landscape for $\mathcal{H}_\infty$ State-Feedback Control

In this section, we investigate the optimization landscape of $\mathcal{H}_\infty$ state-feedback control, and show that any Clarke stationary points of (6) are also global minimum. We start by showing the coercivity of the $\mathcal{H}_\infty$ objective function (5).

Lemma 1. *The $\mathcal{H}_\infty$ objective function $J(K)$ defined by* (5) *is coercive over the set $\mathcal{K}$ in the sense that for any sequence $\{K^l\}_{l=1}^\infty \subset \mathcal{K}$ we have $J(K^l) \to +\infty$, if either $\|K^l\|_2 \to +\infty$, or K^l converges to an element in the boundary $\partial\mathcal{K}$.*

Proof. We will only provide a proof sketch here. A detailed proof is presented in the supplementary material. Suppose we have a sequence $\{K^l\}$ satisfying $\|K^l\|_2 \to +\infty$. We can choose $\mathbf{w} = \{w_0, 0, 0, \cdots\}$ with $\|w_0\| = 1$ and show $J(K^l) \geq w_0^\mathsf{T}(Q + (K^l)^\mathsf{T} R K^l) w_0 \geq \lambda_{\min}(R)\|K^l w_0\|^2$. Clearly, we have used the positive definiteness of R in the above derivation. Then by carefully choosing w_0, we can ensure $J(K^l) \to +\infty$ as $\|K^l\|_2 \to +\infty$. Next, we assume $K^l \to K \in \partial\mathcal{K}$. We have $\rho(A - BK) = 1$, and hence there exists some ω_0 such that $(e^{j\omega_0} I - A + BK)$ becomes singular. Then we can use the positive definiteness of Q to show $J(K^l) \geq \lambda_{\min}^{1/2}(Q)\|(e^{j\omega_0} I - A + BK^l)^{-1}\|$. Notice $\sigma_{\min}(e^{j\omega_0} I - A + BK^l) \to 0$ as $l \to \infty$, which implies $\|(e^{j\omega_0} I - A + BK^l)^{-1}\| \to +\infty$ as $l \to \infty$. Therefore, we have $J(K^l) \to +\infty$ as $K^l \to K \in \partial\mathcal{K}$. More details for the proof can be found in the supplementary material. □

We want to emphasize that the positive definiteness of (Q, R) are crucial for proving the coercivity of the cost function (5). Built upon Lemma 1, we can obtain the following nice properties of the sublevel sets of (6).

Lemma 2. *Consider the $\mathcal{H}_\infty$ state-feedback control problem* (6) *with the objective function $J(K)$ defined in* (5)*. Under Assumption 1, the sublevel set defined as $\mathcal{K}_\gamma = \{K \in \mathcal{K} : J(K) \leq \gamma\}$ is compact and path-connected for every $\gamma \geq J(K^*)$ where K^* is the global minimum of* (6)*.*

Proof. The compactness of $\mathcal{K}_\gamma$ directly follows from the continuity and coercivity of $J(K)$, and is actually a consequence of [5, Proposition 11.12]. The path-connectedness of the strict sublevel sets for the continuous-time $\mathcal{H}_\infty$ control problem has been proved in [26]. We can slightly modify the proof in [26] to show that the strict sublevel set $\{K \in \mathcal{K} : J(K) < \gamma\}$ is path-connected. Based on the fact that every non-strict sublevel sets are compact, now we can apply [36, Theorem 5.2] to show $\mathcal{K}_\gamma$ is also path-connected. An independent proof based on the non-strict version of the bounded real lemma is provided in the supplementary material. □

The path-connectedness of $\mathcal{K}_\gamma$ for every γ actually implies the uniqueness of the minimizing set in a certain sense [36, Sections 2&3]. Due to the space limit, we will defer the discussion on the uniqueness of the minimizing set to the supplementary material. Here, we present a stronger result which is one of the main contributions of our paper.

Theorem 1. *Consider the $\mathcal{H}_\infty$ state-feedback control problem* (6)*. Under Assumption 1, any Clarke stationary point of $J(K)$ is a global minimum.*

A detailed proof is presented in the supplementary material. Here we provide a proof sketch. Since Q and R are positive definite, the non-strict version of the bounded real lemma [6, Section 2.7.3] states that $J(K) \leq \gamma$ if and only if there exists a positive definite matrix P such that the following matrix inequality holds

$$\begin{bmatrix} (A-BK)^\mathsf{T} P(A-BK) - P & (A-BK)^\mathsf{T} P \\ P(A-BK) & P \end{bmatrix} + \begin{bmatrix} Q + K^\mathsf{T} R K & 0 \\ 0 & -\gamma^2 I \end{bmatrix} \preceq 0. \tag{12}$$

The above matrix inequality is linear in P but not linear in K. A standard trick from the control theory can be combined with the Schur complement lemma to convert the above matrix inequality condition to another condition which is linear in all the decision variables [6]. Specifically, there exists a matrix function $\mathrm{LMI}(Y, L, \gamma)$ which is linear in (Y, L, γ) such that $\mathrm{LMI}(Y, L, \gamma) \preceq 0$ if and only if (12) is feasible with $K = LY^{-1}$ and $P = \gamma Y^{-1}$. The matrix function $\mathrm{LMI}(Y, L, \gamma)$ involves a larger matrix. Hence we present the analytical formula of $\mathrm{LMI}(Y, L, \gamma)$ in the supplementary material and skip it here. Since $\mathrm{LMI}(Y, L, \gamma)$ is linear in (Y, L, γ), we know $\mathrm{LMI}(Y, L, \gamma) \preceq 0$ is just a convex semidefinite programming condition. Based on this convex necessary and sufficient condition for $J(K) \leq \gamma$, we can prove the following important lemma.

Lemma 3. *For any $K \in \mathcal{K}$ satisfying $J(K) > J^*$, there exists $d \neq 0$ such that $J'(K, d) < 0$, where $J^* = J(K^*)$ and K^* is the global minimum of* (6)*.*

Proof. Suppose we have $K = LY^{-1}$ where $(Y, L, J(K))$ is a feasible point for the convex regime $\mathrm{LMI}(Y, L, J(K)) \preceq 0$. In addition, we have $K^* = L^*(Y^*)^{-1}$ where $(Y^*, L^*, J(K^*))$ is a point satisfying $\mathrm{LMI}(Y^*, L^*, J(K^*)) \preceq 0$. Since the LMI condition is convex, the line segment between $(Y, L, J(K))$ and $(Y^*, Q^*, J(K^*))$ is also in this convex set. For any $t > 0$, we know $(Y + t\Delta Y, L + t\Delta L, J(K) + t(J(K^*) - J(K)))$ also satisfies $\mathrm{LMI}(Y + t\Delta Y, L + t\Delta L, J(K) + t(J(K^*) - J(K))) \preceq 0$, where $\Delta L = L^* - L$, and $\Delta Y = Y^* - Y$. Therefore, based on the bounded real lemma, we know $J((L + t\Delta L)(Y + t\Delta Y)^{-1}) \leq J(K) + t(J(K^*) - J(K))$. Let's choose $d = \Delta L Y^{-1} + L Y^{-1} \Delta Y Y^{-1}$. Then we have

$$J'(K, d) = \lim_{t \searrow 0} \left(\frac{J((L + t\Delta L)(Y + t\Delta Y)^{-1}) - J(K)}{t} + o(t) \right) < 0.$$

A detailed verification of the above inequality is provided in the supplementary material. Notice $d \neq 0$. If $\Delta L Y^{-1} + L Y^{-1} \Delta Y Y^{-1} = 0$, the above argument still works and we reach to the conclusion $J'(K, 0) < 0$. But this is impossible since we always have $J'(K, 0) = 0$. Hence we have $d \neq 0$. This completes the proof for this lemma. □

Now we are ready to provide the proof for Theorem 1. Based on Lemma 3 and the fact that $J(\cdot)$ is subdifferentially regular, the proof can be done by contradiction. Suppose K^* is the global minimum, and $K^\dagger$ is a Clarke stationary point. If $K^\dagger$ is not a global minimum. Then by Lemma 3, there exists $d \neq 0$ such that $J'(K^\dagger, d) < 0$, this contradicts the fact that $J'(K^\dagger, d) \geq 0$ for all d by Corollary 1. Therefore, $K^\dagger$ has to be the global minimum of (6).

The above proof relies on Lemma 3 and the fact that J is subdifferentially regular. Without using the subdifferential regular property, Lemma 3 itself is not sufficient for proving Theorem 1. It is also worth mentioning that Lemma 3 can be viewed as an extension of the lifting technique in [55, Theorem 4] to non-differentiable points.

4 Global Convergence of Direct Policy Search on $\mathcal{H}_\infty$ State-Feedback Control

In this section, we first show that the Goldstein subdifferential method can be guaranteed to stay in the nonconvex feasible regime $\mathcal{K}$ during the optimization process and eventually converge to the global minimum of (6). The complexity of finding (δ, ϵ)-stationary points of (6) is also presented. Then we further discuss the convergence guarantees for various more implementable forms of the Goldstein subdifferential method.

4.1 Global convergence and complexity of the Goldstein subdifferential method

We will investigate the global convergence of the following Goldstein subdifferential method for direct policy search of the $\mathcal{H}_\infty$ optimal state-feedback policy:

$$K^{n+1} = K^n - \delta^n F^n / \|F^n\|_2 \tag{13}$$

where F^n is the minimum norm element of the Goldstein subdifferential $\partial_{\delta^n} J(K^n)$. We assume that an initial stabilizing policy is available, i.e. $K^0 \in \mathcal{K}$. The same initial policy assumption has also been made in the global convergence theory for direct policy search on LQR [18]. More recently, some provable guarantees have been obtained for finding such stabilizing policies via direct policy search [44]. Hence such an assumption on the initial policy K^0 is reasonable. Our global convergence result relies on the fact that there is a strict separation between any sublevel set of (6) and the boundary of $\mathcal{K}$. This fact is formalized as follows.

Lemma 4. *Consider the $\mathcal{H}_\infty$ state-feedback control problem* (6) *with the cost function $J(K)$ defined in* (5)*. Suppose $\gamma \geq J^*$. Under Assumption 1, there is a strict separation between the sublevel set $\mathcal{K}_\gamma$ and $\mathcal{K}^c$ which is the complement of the feasible set $\mathcal{K}$. In other words, we have $dist(\mathcal{K}_\gamma, \mathcal{K}^c) > 0$.*

Proof. Obviously, the set $\mathcal{K}_\gamma \cap \mathcal{K}^c$ is empty (since we know $\mathcal{K}_\gamma \subset \mathcal{K}$). Based on Lemma 2, we know $\mathcal{K}_\gamma$ is compact. Since $\mathcal{K}$ is open, we know $\mathcal{K}^c$ is closed. Therefore, there is a strict separation between $\mathcal{K}_\gamma$ and $\mathcal{K}^c$, and we have $dist(\mathcal{K}_\gamma, \mathcal{K}^c) > 0$. □

Now we are ready to present our main convergence result.

Theorem 2. *Consider the $\mathcal{H}_\infty$ state-feedback control problem* (6) *with the cost function $J(K)$ defined in* (5)*. Suppose an initial stabilizing policy is given, i.e. $K^0 \in \mathcal{K}$. Denote $\Delta_0 := \mathrm{dist}(\mathcal{K}_{J(K^0)}, \mathcal{K}^c) > 0$. Choose $\delta^n = \frac{c\Delta_0}{n+1}$ for all n with c being a fixed number in $(0, 1)$. Then the Goldstein subdifferential algorithm* (13) *is guaranteed to stay in $\mathcal{K}$ for all n. In addition, we have $J(K^n) \to J^*$ as $n \to \infty$.*

Proof. We have $\delta^n \leq c\Delta_0 < \Delta_0$ for all n. Now we use an induction proof to show $K^n \in \mathcal{K}_{J(K^0)}$ for all $n = 0, 1, \cdots$. For $n = 0$, we know $K^0 - c\Delta_0 F^0/\|F^0\|_2$ has to be within the Δ_0 ball around K^0 since we know the norm of $F^0/\|F^0\|_2$ is exactly equal to 1. Since $\Delta_0 := \mathrm{dist}(\mathcal{K}_{J(K^0)}, \mathcal{K}^c) > 0$, we know $K^0 - \delta^0 F^0/\|F^0\|_2 \in \mathcal{K}$. As a matter of fact, we know $\mathbb{B}_{\delta^0}(K^0)$ has to be a subset of $\mathcal{K}$. Hence we can apply (11) to show that K^1 exists and is also in $\mathcal{K}_{J(K^0)}$. Similarly, we can repeat this argument to show $K^n \in \mathcal{K}_{J(K^0)}$ for all n. Next, we can apply (11) to every step and then sum the inequalities over all n. Then the following inequality holds for all N:

$$\sum_{n=0}^{N} \delta_n \|F^n\|_2 \leq J(K^0) - J^* \tag{14}$$

Since we have $\sum_{n=0}^{\infty} \delta^n = \infty$, we know $\liminf_{n\to\infty} \|F^n\|_2 = 0$. There exists one subsequence $\{i_n\}$ such that $\|F^{i_n}\|_2 \to 0$. For this subsequence, the resultant policy sequence $\{K^{i_n}\}$ is also bounded (notice that the policy parameter sequence stays in the compact set $\mathcal{K}_{J(K^0)}$ for all n) and has a convergent subsequence. We can show that the limit of this subsequence is a Clarke stationary point. Hence the function value $\{J(K^{i_n})\}$ associated with this subsequence converges to J^*. Notice that $J(K^n)$ is monotonically decreasing for the entire sequence $\{n\}$. Hence we have $J(K^n) \to J^*$. □

We have tried to be brief in giving the above proof. We present a more detailed proof in the supplementary material. We believe that this is the first global convergence result for direct policy search on the $\mathcal{H}_\infty$ state-feedback control problem. The above result only provides an asymptotic convergence guarantee to ensure $J(K^n) \to J^*$. One can use a similar argument to establish a finite time complexity bound for finding the (δ, ϵ)-stationary points of (6). Such a result is given as follows.

Theorem 3. *Consider the $\mathcal{H}_\infty$ problem* (6) *with the cost function* (5)*. Suppose $K^0 \in \mathcal{K}$. Denote $\Delta_0 := \mathrm{dist}(\mathcal{K}_{J(K^0)}, \mathcal{K}^c) > 0$. For any $\delta < \Delta_0$, we can choose $\delta^n = \delta$ for all n to ensure that the Goldstein subdifferential algorithm* (13) *to stay in $\mathcal{K}$ and satisfies the following finite-time bound:*

$$\min_{n:0\le n\le N} \|F^n\|_2 \le \frac{J(K^0) - J^*}{(N+1)\delta} \tag{15}$$

In other words, we have $\min_{0\le n\le N}\|F^n\|_2 \le \epsilon$ after $N = \mathcal{O}\left(\frac{\Delta}{\delta\epsilon}\right)$ where $\Delta := J(K^0) - J^*$. For any $\delta < \Delta_0$ and $\epsilon > 0$, the complexity of finding a (δ, ϵ)-stationary point is $\mathcal{O}\left(\frac{\Delta}{\delta\epsilon}\right)$.

The above result can be proved using a similar argument from Theorem 2. We can use the same induction argument to show $K^n \in \mathcal{K}_{J(K^0)}$ for all n, and (14) holds with $\delta^n = \delta$. Then the desired conclusion directly follows. The complexity for nonsmooth optimization of Lipschitz functions is quite subtle. While the above result gives a reasonable characterization of the finite time performance of the Goldstein subdifferential method on the $\mathcal{H}_\infty$ state-feedback control problem, it does not quantify how fast $J(K^n)$ converges to J^*. Recall that (δ, ϵ)-stationarity means $\mathrm{dist}(0, \partial_\delta J(K)) \le \epsilon$, while ϵ-stationarity means $\mathrm{dist}(0, \partial_C J(K)) \le \epsilon$. As commented in [58, 13], (δ, ϵ)-stationarity does not imply being δ-close to an ϵ-stationary point of J. Consequently, the function value of a (δ, ϵ)-stationary point can be very far from J^* even for very small δ and ϵ. Theorem 5 in [58] shows that there is no finite time algorithm that can find ϵ-stationary points provably for all Lipschitz functions. It is still possible that one can develop some finite time bounds for $(J(K^n) - J^*)$ via exploiting other advanced properties of the $\mathcal{H}_\infty$ cost function (5). This is an important future task.

4.2 Implementable variants and related convergence results

In practice, it can be difficult to evaluate the minimum norm element of the Goldstein subdifferential. Now we discuss implementable variants of the Goldstein subdifferential algorithm and the guarantees.

Gradient sampling [9, 10, 31]. The gradient sampling (GS) method is the main optimization algorithm used in the robust control package HIFOO [3, 24]. Suppose we can access a first-order oracle which can evaluate ∇J for any differentiable points in $\mathcal{K}$[2]. Based on Rademacher's theorem, a locally Lipschitz function is differentiable almost everywhere. Therefore, for any $K^n \in \mathcal{K}$, we can randomly sample policy parameters over $\mathbb{B}_{\delta^n}(K^n)$ and obtain differentiable points with probability one. For all these sampled differentiable points, the Clarke subdifferential at each point is just the gradient. Then the convex hull of these sampled gradients can be used as an approximation of $\partial_{\delta^n} K^n$. The minimum norm element from the convex hull of the sampled gradients can be solved via a simple convex quadratic program, and is sufficient for generating a reasonably good descent direction for updating K^{n+1} as long as we sample at least $(n_x n_u + 1)$ differentiable points for each n [9]. In the unconstrained setup, the cluster points of the GS algorithm can be guaranteed to be Clarke stationary [9, Theorem 3.1]. This result can be combined with Theorem 1 and Lemma 4 to show the global convergence of the GS method. The following theorem will be treated formally in the supplementary.

Theorem 4 (Informal statement)**.** *Consider the policy optimization problem* (6) *with the $\mathcal{H}_\infty$ cost function defined in* (5)*. Suppose $K^0 \in \mathcal{K}$. The iterations generated from the GS algorithm is guaranteed to stay in $\mathcal{K}$ for all iterations and achieve $J(K^n) \to J^*$ with probability one.*

Non-derivative sampling (NS) [32]. The NS method can be viewed as the derivative-free version of the GS algorithm. Suppose we only have the zeroth-order oracle which can evaluate the function value $J(K)$ for $K \in \mathcal{K}$. The main difference between NS and GS is that the NS algorithm relies on estimating the gradient from function values via Gupal's estimation method. In the unconstrained setting, the cluster points of the NS method can be guaranteed to be Clarke stationary with probability one [32, Theorem 3.8]. We can combine [32, Theorem 3.8] with our results (Theorem 1 and Lemma 4) to prove the global convergence of NS in our setting. The details are given in the supplementary.

Model-free implementation of NS. When the model is unknown, there are various methods available for estimating the $\mathcal{H}_\infty$-norm from data [38, 39, 47, 45, 56, 43, 54, 53]. Based on our experience, the multi-input multi-output (MIMO) power iteration method [42] works quite well as a stochastic zeroth-order oracle for the purpose of implementing NS in the model-free setting. While the sample complexity for model-free NS is unknown, we will provide some numerical justifications to show that such a model-free implementation closely tracks the convergence of its model-based counterpart.

[2] When (A, B) is known, one can calculate the $\mathcal{H}_\infty$ gradient at differential points using the chain rule in [2]

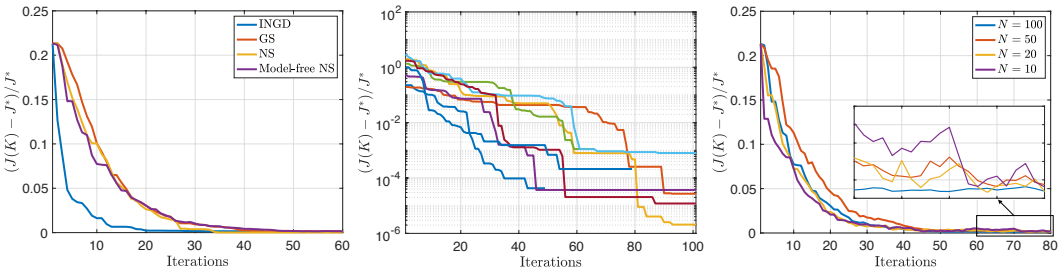

Figure 1: Simulation results. Left: The trajectory of relative error of GS, NS, INGD, and Model-free NS methods on (16). Middle: The trajectory of relative optimality gap of 8 randomly generated cases for NS method. Right: The trajectory of Model-free NS method with more noisy oracle on (16).

Interpolated normalized gradient descent (INGD) with finite-time complexity. No finite-time guarantees for finding (δ, ϵ)-stationary points have been reported for the GS/NS methods. In [58, 13], the INGD method has been developed as another implementable approximation of the Goldstein subdifferential method, and is proved to satisfy high-probability finite-time complexity bounds for finding (δ, ϵ)-stationary points of Lipschitz functions. INGD uses an iterative sampling strategy to generate a descent direction which serves a role similar to the minimal norm element of the Goldstein subdifferential. A first-order oracle for differentiable points is needed for implementing INGD. It has been show [58, Theorem 8] that for unconstrained nonsmooth optimization of L-Lipschitz functions, the above INGD algorithm can be guaranteed to find the (δ, ϵ)-stationary point with the high-probability iteration complexity $\mathcal{O}\left(\frac{\Delta L^2}{\epsilon^3 \delta} \log(\frac{1}{p})\right)$, where $\Delta := J(K^0) - J^*$ is the initial function value gap, and p is the failure probability (i.e. the optimization succeeds with the probability $(1-p)$). We can combine the proofs for Theorem 3 and [58, Theorem 8] to obtain the following complexity result for our $\mathcal{H}_\infty$ setting. A formal treatment is given in the supplementary material.

Theorem 5 (Informal statement)**.** *Consider the policy optimization problem* (6) *with the $\mathcal{H}_\infty$ cost function defined in* (5)*. Suppose $K^0 \in \mathcal{K}$. Denote $\Delta_0 := \mathrm{dist}(\mathcal{K}_{J(K^0)}, \mathcal{K}^c) > 0$, and let L_0 be the Lipschitz constant of $J(K)$ over the set $\mathcal{K}_{J(K^0)}$. For any $\delta < \Delta_0$, we can choose $\delta^n = \delta$ for all n to ensure that the iterations of the INGD algorithm stay in $\mathcal{K}$ almost surely, and find a (δ, ϵ)-stationary point with the high-probability iteration complexity $\mathcal{O}\left(\frac{\Delta L_0^2}{\epsilon^3 \delta} \log(\frac{1}{p})\right)$.*

5 Numerical Simulations

We provide some numerical simulations in this section as well as the supplementary material. The left plot in Figure 1 shows that GS, NS, INGD, and model-free NS work well for the following example:

$$A = \begin{bmatrix} 1 & 0 & -5 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \; B = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \; Q = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}, \; R = 1. \tag{16}$$

For this example, we have $J^* = 7.3475$. We initialize from $K^0 = [0.4931 \quad -0.1368 \quad -2.2654]$, which satisfies $\rho(A - BK^0) = 0.5756 < 1$. The hyperparameter choices are detailed in the supplementary material. We can see that model-free NS closely tracks the trajectory of NS and works well. In the middle plot of Figure 1, we test the NS method on randomly generated cases. We set $A \in \mathbb{R}^{3\times 3}$ to be $I + \xi$, where each element of $\xi \in \mathbb{R}^{3\times 3}$ is sampled uniformly from $[0, 1]$. For $B \in \mathbb{R}^{3\times 1}$, each element is uniformly sampled from $[0, 1]$. We have $Q = I + \zeta I \in \mathbb{R}^{3\times 3}$ with ζ uniformly sampled from $[0, 0.1]$, and $R \in \mathbb{R}$ uniformly sampled from $[1, 1.5]$. For each experiment, the initial condition $K^0 \in \mathbb{R}^{1\times 3}$ is also randomly sampled such that $\rho(A - BK^0) < 1$. The NS method converges globally for all the cases. In the right plot, we focus on the model-free setting for (16). We decrease the samples used in the $\mathcal{H}_\infty$ estimation and show how this increases the noise in the zeroth-order oracle and worsens the convergence behaviors of the model-free NS method. Nevertheless, the model-free NS method tracks its model-based counterpart with enough samples.

References

[1] P. Apkarian and D. Noll. Controller design via nonsmooth multidirectional search. *SIAM Journal on Control and Optimization*, 44(6):1923–1949, 2006.

[2] P. Apkarian and D. Noll. Nonsmooth $\mathcal{H}_\infty$ synthesis. *IEEE Transactions on Automatic Control*, 51(1):71–86, 2006.

[3] D. Arzelier, G. Deaconu, S. Gumussoy, and D. Henrion. H2 for HIFOO. In *International Conference on Control and Optimization With Industrial Applications (COIA 2011)*, 2011.

[4] T. Başar and P. Bernhard. *H^∞-Optimal Control and Related Minimax Design Problems*. Birkhäuser, 2nd edition, 1995.

[5] H. H. Bauschke, P. L. Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[6] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*, volume 15 of *Studies in Applied Mathematics*. SIAM, 1994.

[7] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi. LQR through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019.

[8] J. Bu, A. Mesbahi, and M. Mesbahi. On topological and metrical properties of stabilizing feedback gains: the MIMO case. *arXiv preprint arXiv:1904.02737*, 2019.

[9] J. V. Burke, F. E. Curtis, A. S. Lewis, M. L. Overton, and L. E. Simões. Gradient sampling methods for nonsmooth optimization. *Numerical Nonsmooth Optimization*, pages 201–225, 2020.

[10] J. V. Burke, A. S. Lewis, and M. L. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005.

[11] F. H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[12] F. E. Curtis, T. Mitchell, and M. L. Overton. A BFGS-SQP method for nonsmooth, nonconvex, constrained optimization and its evaluation using relative minimization profiles. *Optimization Methods and Software*, 32(1):148–181, 2017.

[13] D. Davis and D. Drusvyatskiy. A gradient sampling method with complexity guarantees for general lipschitz functions. *arXiv preprint arXiv:2112.06969*, 2021.

[14] J. Doyle, K. Glover, P. Khargonekar, and B. Francis. State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems. *IEEE Transactions on Automatic Control*, 34:831–847, 1989.

[15] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.

[16] G. Dullerud and F. Paganini. *A Course in Robust Control Theory: A Convex Approach*. Springer, 1999.

[17] I. Fatkhullin and B. Polyak. Optimizing static linear feedback: Gradient method. *SIAM Journal on Control and Optimization*, 59(5):3887–3911, 2021.

[18] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1467–1476, 2018.

[19] L. Furieri, Y. Zheng, and M. Kamgarpour. Learning the globally optimal distributed LQ regulator. In *Learning for Dynamics and Control*, pages 287–297, 2020.

[20] P. Gahinet and P. Apkarian. A linear matrix inequality approach to H_∞ control. *International Journal of Robust and Nonlinear Control*, 4:421–448, 1994.

[21] 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 Letters*, 11(3):167–172, 1988.

[22] A. Goldstein. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977.

[23] B. Gravell, P. M. Esfahani, and T. Summers. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, 66(11):5283–5298, 2020.

[24] S. Gumussoy, D. Henrion, M. Millstone, and M. L. Overton. Multiobjective robust control with HIFOO 2.0. *IFAC Proceedings Volumes*, 42(6):144–149, 2009.

[25] B. Hambly, R. Xu, and H. Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59(5):3359–3391, 2021.

[26] B. Hu and Y. Zheng. Connectivity of the feasible and sublevel sets of dynamic output feedback control with robustness constraints. *arXiv preprint arXiv:2203.11177*, 2022.

[27] J. P. Jansch-Porto, B. Hu, and G. Dullerud. Policy learning of MDPs with mixed continuous/discrete variables: A case study on model-free control of Markovian jump systems. In *Learning for Dynamics and Control*, pages 947–957, 2020.

[28] J. P. Jansch-Porto, B. Hu, and G. E. Dullerud. Convergence guarantees of policy optimization methods for Markovian jump linear systems. In *American Control Conference*, pages 2882–2887, 2020.

[29] 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*, 29(1):57–70, 1993.

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

[31] K. C. Kiwiel. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 18(2):379–388, 2007.

[32] K. C. Kiwiel. A nonderivative version of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 20(4):1983–1994, 2010.

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

[34] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[35] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *arXiv:1812.08305*, 2018.

[36] D. H. Martin. Connected level sets, minimizing sets, and uniqueness in optimization. *Journal of Optimization Theory and Applications*, 36(1):71–91, 1982.

[37] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanovic. Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem. *IEEE Transactions on Automatic Control*, 2021.

[38] M. Müller and C. R. Rojas. Gain estimation of linear dynamical systems using thompson sampling. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pages 1535–1543, 2019.

[39] M. I. Müller, P. E. Valenzuela, A. Proutiere, and C. R. Rojas. A stochastic multi-armed bandit approach to nonparametric $\mathcal{H}_\infty$-norm estimation. In *IEEE 56th Annual Conference on Decision and Control*, pages 4632–4637, 2017.

[40] D. Mustafa and D. 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.

[41] D. Noll and P. Apkarian. Spectral bundle methods for non-convex maximum eigenvalue functions: second-order methods. *Mathematical Programming*, 104(2):729–747, 2005.

[42] T. Oomen, R. van der Maas, C. R. Rojas, and H. Hjalmarsson. Iteratively learning the $\mathcal{H}_\infty$-norm of multivariable systems applied to model-error-modeling of a vibration isolation system. In *2013 American Control Conference*, pages 6703–6708. IEEE, 2013.

[43] T. Oomen, R. van der Maas, C. R. Rojas, and H. Hjalmarsson. Iterative data-driven $\mathcal{H}_\infty$ norm estimation of multivariable systems with application to robust active vibration isolation. *IEEE Transactions on Control Systems Technology*, 22(6):2247–2260, 2014.

[44] J. Perdomo, J. Umenberger, and M. Simchowitz. Stabilizing dynamical systems via policy gradient methods. *Advances in Neural Information Processing Systems*, 34, 2021.

[45] G. Rallo, S. Formentin, C. R. Rojas, T. Oomen, and S. M. Savaresi. Data-driven $\mathcal{H}_\infty$-norm estimation via expert advice. In *IEEE 56th Annual Conference on Decision and Control*, pages 1560–1565, 2017.

[46] S. Rathod, M. Bhadu, and A. De. Global convergence using policy gradient methods for model-free markovian jump linear quadratic control. *arXiv preprint arXiv:2111.15228*, 2021.

[47] C. R. Rojas, T. Oomen, H. Hjalmarsson, and B. Wahlberg. Analyzing iterations in identification with application to nonparametric $\mathcal{H}_\infty$-norm estimation. *Automatica*, 48(11):2776–2790, 2012.

[48] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

[49] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[50] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[51] S. Skogestad and I. Postlethwaite. *Multivariable Feedback Control: Analysis and Design*, volume 2. Wiley New York, 2007.

[52] 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*, pages 1057–1063, 2000.

[53] S. Tu, R. Boczar, and B. Recht. On the approximation of toeplitz operators for nonparametric $\mathcal{H}_\infty$-norm estimation. In *American Control Conference*, pages 1867–1872, 2018.

[54] S. Tu, R. Boczar, and B. Recht. Minimax lower bounds for $\mathcal{H}_\infty$-norm estimation. In *American Control Conference*, pages 3538–3543, 2019.

[55] J. Umenberger, M. Simchowitz, J. C. Perdomo, K. Zhang, and R. Tedrake. Globally convergent policy search over dynamic filters for output estimation. *arXiv preprint arXiv:2202.11659*, 2022.

[56] B. Wahlberg, M. B. Syberg, and H. Hjalmarsson. Non-parametric methods for l_2-gain estimation using iterative experiments. *Automatica*, 46(8):1376–1381, 2010.

[57] Z. Yang, Y. Chen, M. Hong, and Z. Wang. On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *arXiv preprint arXiv:1907.06246*, 2019.

[58] J. Zhang, H. Lin, S. Jegelka, S. Sra, and A. Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11173–11182. PMLR, 2020.

[59] K. Zhang, B. Hu, and T. Başar. On the stability and convergence of robust adversarial reinforcement learning: A case study on linear quadratic systems. *Advances in Neural Information Processing Systems*, 33, 2020.

[60] K. Zhang, B. Hu, and T. Basar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. *SIAM Journal on Control and Optimization*, 59(6):4081–4109, 2021.

[61] K. Zhang, X. Zhang, B. Hu, and T. Başar. Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[62] F. Zhao and K. You. Primal-dual learning for the model-free risk-constrained linear quadratic regulator. In *Learning for Dynamics and Control*, pages 702–714, 2021.

[63] Y. Zheng, Y. Tang, and N. Li. Analysis of the optimization landscape of linear quadratic gaussian (LQG) control. *arXiv preprint arXiv:2102.04393*, 2021.

[64] K. Zhou, J. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice-Hall, 1996.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Our works only addresses the $\mathcal{H}_\infty$ state-feedback control problem. More study is needed for the partial observation case.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We do not believe that our current work will create any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix in the supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] The nature of our work is theoretical.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]