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

Xingang Guo, Bin Hu
Department of Electrical and Computer Engineering
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
{xingang2,binhu7}@illinois.edu

Abstract

Direct policy search has been widely applied in modern reinforcement learning and continuous control. However, the theoretical properties of direct policy search on nonsmooth robust control synthesis have not been fully 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 fundamental 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 coerciveness of the closed-loop $\mathcal{H}_\infty$ objective function, and prove that all the sublevel sets of the resultant policy search problem are compact. Based on these properties, we show that Goldstein's subgradient method and its implementable variants can be guaranteed to stay in the nonconvex feasible set and eventually find the global optimal solution of the $\mathcal{H}_\infty$ state-feedback synthesis problem. Our work builds a new connection between nonconvex nonsmooth optimization theory and robust control, leading to an interesting 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 [55, 37], and policy optimization is one of the main workhorses for such applications [17, 61, 54, 56]. 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) [20, 7, 38, 66, 41, 21, 27], stabilization [49, 48], linear robust/risk-sensitive control [69, 68, 25, 70, 71], Markov jump linear quadratic control [30, 29, 31, 51], output feedback control [19, 73, 36, 16, 15, 40, 72], and dynamic filtering [64]. 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 can be non-differentiable over certain points in the policy space [1–3, 26, 9, 12, 45]. Different from LQR which considers stochastic disturbance sequences, $\mathcal{H}_\infty$ control directly addresses the worst-case disturbance, and provides arguably the most fundamental robust control paradigm [74, 18, 58, 4, 14, 22]. Regarding the connection with

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

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 quadratic setup [68, 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. In this work, we study and prove the global convergence properties of direct policy search on the $\mathcal{H}_\infty$ state-feedback 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 [20, 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 [1–3, 26, 9, 12]. However, a satisfying global convergence theory is still missing from 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 policy search problem are also global minimum.
2. We identify the coerciveness of the $\mathcal{H}_\infty$ cost function and use this property to show that Goldstein's subgradient method [24] and its implementable variants [67, 13, 9, 10, 34, 35] 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.

Our work sheds new light on the theoretical properties of policy optimization methods on $\mathcal{H}_\infty$ control problems, and serves as a meaningful initial step towards a general global 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 [23, 33, 32, 44]. 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 [69], 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 [69] cannot be applied to our $\mathcal{H}_\infty$ control setting. More discussions on the connections and differences between these two problems will be given 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 [1–3, 26, 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 [74, 18, 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 to address (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 [1–3, 26, 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 [20, 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)

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

[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 strictly differentiable. For completeness, a simple proof of Proposition 2 based on Clarke's chain rule [11] is included in the supplementary material.

In general, the Clarke directional derivative can be different from the (ordinary) directional derivative. Sometimes the ordinary directional derivative may not even 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.*

See [52, Theorem 10.1] for related proofs and more discussions. 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 good descent direction for nonsmooth optimization is not trivial. Many nonsmooth optimization algorithms are based on the concept of Goldstein subdifferential [24]. 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 $\mathbb{B}_\delta(K) \subset \mathcal{K}$.

Based on the above definition, one can further define the notion of (δ, ϵ)-stationarity. A point K is said to be (δ, ϵ)-stationary if $\operatorname{dist}(0, \partial_\delta J(K)) \leq \epsilon$. It is well-known that the minimal norm element of the Goldstein subdifferential generates a good descent direction. This fact is stated as follows.

Proposition 4 ([24])**.** *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, 26, 9, 12]. We will show that such policy search algorithms can be guaranteed to find the global minimum of (6). It is worth mentioning that there are other notions of enlarged subdifferential [2] which can lead to good descent directions for nonsmooth $\mathcal{H}_\infty$ synthesis. In this paper, we focus on the notion of Goldstein subdifferential and related policy search algorithms.

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

In this section, we investigate the optimization landscape of the $\mathcal{H}_\infty$ state-feedback policy search problem, and show that any Clarke stationary points of (6) are also global minimum. We start by showing the coerciveness 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 coerciveness 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 policy search 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 coerciveness 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 [28]. We can slightly modify the proof in [28] 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 [39, 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 also 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 strong sense [39, 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 policy search 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[2] states that $J(K) \leq \gamma$ if and only if there exists a positive definite matrix P such that the following matrix inequality holds

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

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 $\mathrm{LMI}(Y, L, \gamma) \preceq 0$ and $Y \succ 0$ if and only if (12) is feasible with $K = LY^{-1}$ and $P = \gamma Y^{-1} \succ 0$. 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 a matrix direction $d \neq 0$ such that $J'(K, d) \leq J^* - J(K) < 0$, 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

[2]The difference between the strict and non-strict versions of the bounded real lemma is quite subtle [6, Section 2.7.3]. For completeness, we will provide more explanations for the non-strict version of the bounded real lemma in the supplementary material.

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 $d = \Delta L Y^{-1} - LY^{-1}\Delta Y Y^{-1}$. Then we have

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

A detailed verification of the above inequality is provided in the supplementary material. Notice $d \neq 0$. If $\Delta L Y^{-1} - LY^{-1}\Delta Y Y^{-1} = 0$, 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 subdifferentially regular property, Lemma 3 itself is not sufficient for proving Theorem 1. It is also worth mentioning that Lemma 3 can be viewed as a modification of the convex parameterization/lifting results in [60, 64] for non-differentiable points.

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

In this section, we first show that Goldstein's subgradient method [24] 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 implementable forms of Goldstein's subgradient method.

4.1 Global convergence and complexity of Goldstein's subgradient method

We will investigate the global convergence of Goldstein's subgradient method for direct policy search of the optimal $\mathcal{H}_\infty$ state-feedback policy. Goldstein's subgradient method iterates as follows

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

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 [20]. More recently, some provable guarantees have been obtained for finding such stabilizing policies via direct policy search [49, 48]. 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 policy search problem* (6) *with the cost function $J(K)$ defined in* (5)*. Denote the complement of the feasible set $\mathcal{K}$ as $\mathcal{K}^c$. Suppose Assumption 1 holds and $\gamma \geq J^*$. Then there is a strict separation between the sublevel set $\mathcal{K}_\gamma$ and $\mathcal{K}^c$. 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 policy search problem* (6) *with the cost function $J(K)$ defined in* (5)*. Suppose Assumption 1 holds, and 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 Goldstein's subgradient method* (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. 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 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 will present a more detailed proof in the supplementary material. We believe that this is the first result showing that direct policy search can be guaranteed to converge to the global optimal solution of 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 Assumption 1 holds, and $K^0 \in \mathcal{K}$. 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 Goldstein's subgradient method* (13) *stays in $\mathcal{K}$ and satisfies the following finite-time complexity bound:*

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

In other words, we have $\min_{0\leq n\leq N} \|F^n\|_2 \leq \epsilon$ 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 Goldstein's subgradient 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 $\mathrm{dist}(0, \partial_\delta J(K)) \leq \epsilon$, while ϵ-stationarity means $\mathrm{dist}(0, \partial_C J(K)) \leq \epsilon$. As commented in [57, 67, 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 far from J^* even for small δ and ϵ. Theorem 5 in [67] 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 Goldstein's subgradient method and related guarantees.

Gradient sampling [9, 10, 34]. The gradient sampling (GS) method is the main optimization algorithm used in the robust control package HIFOO [3, 26]. Suppose we can access a first-order oracle which can evaluate ∇J for any differentiable points in the feasible set[3]. Based on Rademacher's

[3]When (A, B) is known, one can calculate the $\mathcal{H}_\infty$ gradient at differential points using the chain rule in [2]. More explanations can be found in the supplementary material.

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 for the Goldstein subdifferential $\partial_{\delta^n} K^n$. 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 on the $\mathcal{H}_\infty$ state-feedback synthesis problem. The following theorem will be treated formally in the supplementary material.

Theorem 4 (Informal statement)**.** *Consider the policy optimization problem* (6) *with the $\mathcal{H}_\infty$ cost function defined in* (5)*. Suppose Assumption 1 holds, and $K^0 \in \mathcal{K}$. The iterations generated from the GS algorithm (described in the supplementary material) is guaranteed to stay in $\mathcal{K}$ for all iterations and achieve $J(K^n) \to J^*$ with probability one.*

Non-derivative sampling (NS) [35]. 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 [35, Theorem 3.8]. We can combine [35, Theorem 3.8] with our results (Theorem 1 and Lemma 4) to prove the global convergence of NS in our setting. A detailed discussion is given in the supplementary material.

Model-free implementation of NS. When the model is unknown, there are various methods available for estimating the $\mathcal{H}_\infty$-norm from data [42, 43, 53, 50, 65, 47, 63, 62]. Based on our own experiences/tests, the multi-input multi-output (MIMO) power iteration method [46] 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 behaviors of its model-based counterpart.

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 [67, 13], the INGD method has been developed as another implementable approximation of Goldstein's subgradient 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[4]. It has been show [67, 13] 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 [13, Theorem 2.6] and Theorem 3 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 Assumption 1 holds, and the initial policy is stabilizing, i.e. $K^0 \in \mathcal{K}$. 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 $\mathcal{O}\left(\frac{\Delta L_0^2}{\epsilon^3 \delta} \log(\frac{1}{p})\right)$, where p is the failure probability.*

[4]This is true for the algorithm developed in [13]. For the original developments in [67], a stronger oracle is required.

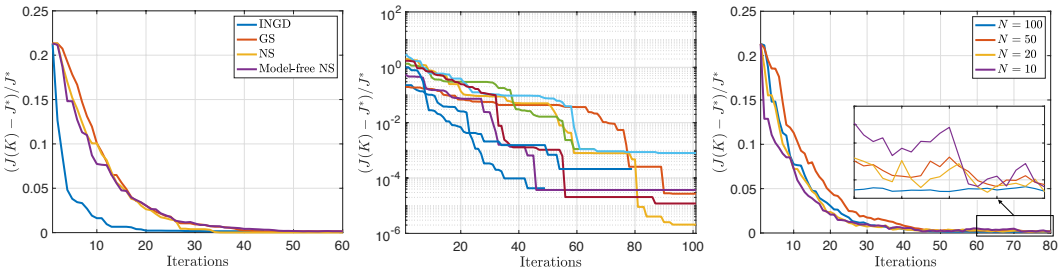

Figure 1: Simulation results. Left: The trajectory of relative error of GS, NS, INGD, and Model-free NS methods on (16). Middle: The trajectory of relative optimality gap of 8 randomly generated cases for NS method. Right: The trajectory of Model-free NS method with more noisy oracle on (16).

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.

6 Conclusions and Future Work

In this paper, we present the global convergence theory for direct policy search on the $\mathcal{H}_\infty$ state-feedback synthesis problem. Although the resultant policy optimization problem is nonconvex and nonsmooth, the concept of Golstein subdifferential can be used to build direct policy search algorithms which are guaranteed to converge to the global $\mathcal{H}_\infty$ optimal policy. The finite-time guarantees in this paper are developed only for finding (δ, ϵ)-stationary point. An important future task is to investigate the finite-time bounds for the optimality gap. In addition, sample complexity of direct policy search in the model-free setting remains open. Finally, it is also interesting to investigate the convergence properties of direct policy search in the nonlinear/output-feedback settings.

Acknowledgments and Disclosure of Funding

This work is generously supported by the NSF award CAREER-2048168 and the 2020 Amazon research award. The authors would like to thank Michael L. Overton, Maryam Fazel, Yang Zheng, Peter Seiler, Geir Dullerud, Aaron Havens, Darioush Kevian, Kaiqing Zhang, Na Li, Mehran Mesbahi, Tamer Başar, Mihailo Jovanovic, and Javad Lavaei for the valuable discussions, as well as the helpful suggestions from the anonymous reviewers of NeurIPS.

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, D. Drusvyatskiy, Y. Tat Lee, S. Padmanabhan, and G. Ye. 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] J. Duan, W. Cao, Y. Zheng, and L. Zhao. On the optimization landscape of dynamical output feedback linear quadratic control. *arXiv preprint arXiv:2201.09598*, 2022.

[16] J. Duan, J. Li, and L. Zhao. Optimization landscape of gradient descent for discrete-time static output feedback. *arXiv preprint arXiv:2109.13132*, 2021.

[17] 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.

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

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

[20] 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.

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

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

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] B. Hu and Y. Zheng. Connectivity of the feasible and sublevel sets of dynamic output feedback control with robustness constraints. *IEEE Control Systems Letters*, 7:442–447, 2023.

[29] 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.

[30] 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.

[31] J. P. Jansch-Porto, B. Hu, and G. E. Dullerud. Policy optimization for Markovian jump linear quadratic control: Gradient method and global convergence. *IEEE Transactions on Automatic Control*, 2022.

[32] 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.

[33] 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.

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

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

[36] 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.

[37] 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.

[38] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *International Conference on Artificial Intelligence and Statistics*, pages 2916–2925, 2019.

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

[40] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović. On the lack of gradient domination for linear quadratic gaussian problems with incomplete state information. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 1120–1124, 2021.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

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

[46] 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.

[47] 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.

[48] I. K. Ozaslan, H. Mohammadi, and M. R. Jovanović. Computing stabilizing feedback gains via a model-free policy gradient method. *IEEE Control Systems Letters*, 7:407–412, 2022.

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

[50] 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.

[51] 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.

[52] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[53] 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.

[54] 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.

[55] 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.

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

[57] O. Shamir. Can we find near-approximately-stationary points of nonsmooth nonconvex functions? *arXiv preprint arXiv:2002.11962*, 2020.

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

[59] S. Subramani. *On sums of singular values*. PhD thesis, The University of New South Wales Sydney, 1993.

[60] Y. Sun and M. Fazel. Learning optimal controllers by policy gradient: Global optimality via convex parameterization. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 4576–4581, 2021.

[61] 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.

[62] 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.

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

[64] 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.

[65] 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.

[66] Z. Yang, Y. Chen, M. Hong, and Z. Wang. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *Advances in neural information processing systems*, 32, 2019.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] Y. Zheng, Y. Sun, M. Fazel, and N. Li. Escaping high-order saddles in policy optimization for linear quadratic Gaussian (LQG) control. *arXiv preprint arXiv:2204.00912*, 2022.

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

[74] 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]

Supplementary Material

A More Details on Problem Formulation and Preliminaries

A.1 Difference between $\mathcal{H}_\infty$ control and mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Design

In this paper, we are interested in $\mathcal{H}_\infty$ control whose objective is to design a state-feedback policy K that stabilizes the closed-loop system and minimizes the closed-loop $\mathcal{H}_\infty$ norm (or equivalently the $\ell_2 \to \ell_2$ induced norm) of G_K at the same time. As discussed in the main paper, $\mathcal{H}_\infty$ control can be formulated as $\min_{K\in\mathcal{K}}\|G_K\|_{2\to 2}$. We have mentioned that such a formulation corresponds to a worst-case assumption on the disturbance $\mathbf{w}$. Since $\|G_K\|_{2\to 2}$ is a nonsmooth function in K, the $\mathcal{H}_\infty$ control leads to a nonsmooth optimization problem. The global optimal $\mathcal{H}_\infty$ norm for this problem is denoted as γ^*.

In contrast, the design objective of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control is to synthesize a a linear state-feedback policy which minimizes an upper bound on the $\mathcal{H}_2$ cost and satisfies an additional $\mathcal{H}_\infty$-robustness requirement in the form of $\|G_K\|_{2\to 2} < \gamma$ with some pre-specified value of γ. Recall that we denote the non-strict γ-admissible sublevel set of the $\mathcal{H}_\infty$ objective function as $\mathcal{K}_\gamma = \{K \in \mathcal{K} : \|G_K\|_{2\to 2} \le \gamma\}$. Similarly, we definie the strict γ-admissible sublevel set of the $\mathcal{H}_\infty$ objective function as $\tilde{\mathcal{K}}_\gamma = \{K \in \mathcal{K} : \|G_K\|_{2\to 2} < \gamma\}$. Then for any $\gamma > \gamma^*$, the $\mathcal{H}_2/\mathcal{H}_\infty$ mixed design problem can be formulated as

$$\min_{K\in\tilde{\mathcal{K}}_\gamma} \operatorname{tr}(P_K)$$

where P_K is the minimal positive definite solution to the following Riccati equation

$$(A-BK)^\mathsf{T}(P_K + P_K(\gamma^2 I - P_K)^{-1}P_K)(A-BK) + Q + K^\mathsf{T}RK - P_K = 0.$$

Notice that as $\gamma \to \infty$, the above mixed design problem reduces to the LQR problem. For finite γ, $\operatorname{tr}(P_K)$ provides an upper bound for the LQR cost. For the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem, the closed-loop $\mathcal{H}_\infty$ norm $\|G_K\|_{2\to 2}$ only appears in the constraint. One can show that cost function $\operatorname{tr}(P_K)$ is still differentiable in the policy space for any $K \in \tilde{\mathcal{K}}$. The main result in [69] states that given an initial policy $K^0 \in \tilde{\mathcal{K}}_\gamma$ for $\gamma > \gamma^*$, then the natural policy gradient method with well-chosen stepsize is guaranteed to stay in $\tilde{\mathcal{K}}_\gamma$ and fine the optimal policy minimizes $\operatorname{tr}(P_K)$. This means that the natural policy gradient method applied to this mixed design problem can improve the average performance (via minimizing $\operatorname{tr}(P_K)$) while maintaining the robustness level γ. A missing step is how to use direct policy search to obtain such an initial policy satisfying $\|G_{K^0}\|_{2\to 2} < \gamma$ in the first place. One by-product of the results in this paper is a direct policy search method for obtaining such initial policy provably.

A.2 Formulating $\mathcal{H}_\infty$ control with arbitrary ℓ_2 bound on the disturbance

To address the worst-case disturbance $\mathbf{w}$, one may be interested in the following problem formulation

$$\min_{\mathbf{u}} \max_{\mathbf{w}:\|\mathbf{w}\|\le\Lambda} \sum_{t=0}^{\infty}(x_t^\mathsf{T}Qx_t + u_t^\mathsf{T}Ru_t) \tag{A.1}$$

where Λ is an arbitrary positive number. An interesting fact is that (A.1) still leads to the same policy optimization problem $\min_{K\in\mathcal{K}}\|G_K\|_{2\to 2}$ regardless of the value of Λ. Now we briefly explain this fact. Again, it is well-know that it suffices to consider linear state-feedback policies for solving (A.1). Notice that G_K is a linear operator for any fixed K. Therefore, it is straightforward to verify

$$\max_{\mathbf{w}:\|\mathbf{w}\|\le\Lambda} \sum_{t=0}^{\infty}(x_t^\mathsf{T}Qx_t + u_t^\mathsf{T}Ru_t) = \Lambda^2\|G_K\|_{2\to 2}^2$$

Consequently, (A.1) can be equivalently formulated as $\min_{K\in\mathcal{K}} \Lambda^2\|G\|_{2\to 2}^2$. The constant Λ^2 can be removed without changing the optimization problem. Therefore, no matter what positive value we use for Λ, (A.1) is always equivalent to $\min_{K\in\mathcal{K}}\|G\|_{2\to 2}$.

A.3 More explanations for Proposition 1

Now we provide some explanations for Proposition 1 which states several facts observed in the existing literature. The continuity of the cost function (5) can be seen from its analytic form. The openness of $\mathcal{K}$ can be seen from the fact that the spectral radius $\rho(A-BK)$ is a continuous function of K. The unboundedness of $\mathcal{K}$ has been proved in [8]. To see that $\mathcal{K}$ can be nonconvex, we provide a simple example here. Suppose $A = B = I$, and we choose two policies as follows

$$K = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}, \quad K' = \begin{bmatrix} 1 & -5 \\ 0 & 1 \end{bmatrix}.$$

Then one can easily check that $K, K' \in \mathcal{K}$, while their convex combination $\frac{1}{2}(K_1 + K_2) \notin \mathcal{K}$.

Finally, we provide an example to show $J(K)$ is nonconvex in K. Suppose $A = B = Q = R = I$ and choose two policies as follows:

$$K = \begin{bmatrix} 0.7 & -3.4 \\ 0.2 & 1 \end{bmatrix}, \quad K' = \begin{bmatrix} 0 & -1.3 \\ 0.6 & 1.2 \end{bmatrix}.$$

Let $K'' = \frac{1}{2}(K + K')$, then we have $J(K'') = 35.99$ and $\frac{1}{2}(J(K) + J(K')) = 13.3$. Hence $J(K'') > 0.5J(K) + 0.5J(K')$, which shows $J(K)$ is nonconvex in K.

A.4 Proof of Proposition 2

Proposition 2 is a consequence of the chain rule stated in [11, Theorem 2.3.10]. As mentioned previously, Proposition 2 should be considered as a well-known fact, and we do not claim any credits in proving it. Although not explicitly stated, the proof of Proposition 2 has been hinted in the discussion of [2, Section III]. We only present the proof for completeness.

To apply [11, Theorem 2.3.10], we can rewrite the closed-loop $\mathcal{H}_\infty$ norm $\|G_K\|_{2\to 2}$ as a composition $g_2 \circ g_1(K)$ where g_1 maps the policy parameter K to a stable transfer function $(Q + K^\mathsf{T} RK)^{\frac{1}{2}}(zI - A + BK)^{-1}$ which lives in the infinite-dimensional $\mathcal{H}_\infty$ Hardy space, and g_2 maps any stable transfer function in that Hardy space to its $\mathcal{H}_\infty$ norm. It is well known that the infinite-dimensional $\mathcal{H}_\infty$ Hardy space consisting of all stable LTI systems is a Banach space [74]. We also know that g_2 is convex on the Hardy space (by triangle inequality of the $\mathcal{H}_\infty$ norm). The convexity of g_2 implies that g_2 is also subdifferentially regular. One can also show that g_1 is a strictly differentiable mapping from K to the Hardy space given $\rho(A - BK) < 1$ and Q being positive definite. Now we can immediately apply [11, Theorem 2.3.10] to obtain the desired conclusion in Proposition 2.

Remark 1. *For readability, we also provide some brief explanations for the fact that the mapping g_1 is strictly differentiable. For each $K \in \mathcal{K}$, g_1 maps K to an infinite-dimensional bounded operator which maps from $\ell_2^{n_w}$ to $\ell_2^{n_x}$ and has a Toeplitz structure. The Toeplitz structure can be combined with Fubini's theorem to show simple upper bounds for the operator norm of g_1. One such upper bound is provided by $\|(Q + K^\mathsf{T} RK)^{\frac{1}{2}}\| \sum_{k=0}^{\infty} \|(A - BK)^k\|$, which is obviously finite for $K \in \mathcal{K}$. Such a bound allows us to study the operator g_1 via treating it as an infinite-dimensional Toeplitz matrix whose properties are completely determined by its first block column. By definition, g_1 admits a strict derivative at K if the following operator norm convergence result holds for any V:*

$$\lim_{K' \to K,\, t \downarrow 0} \|(g_1(K' + tV) - g_1(K'))/t - \langle D_s g_1(K), V\rangle\| = 0, \tag{A.2}$$

where $\|\cdot\|$ denotes the operator norm. Since we know $\rho(A - BK) < 1$, the Toeplitz structure of g_1 can be combined with Fubini's theorem again to provide simple upper bounds for the operator norm on the left side of (A.2)*, leading to the desired operator norm convergence result in* (A.2)*.*

A.5 A non-strict version of the bounded real lemma

The bounded real lemma (also referred to as the Kalman–Yakubovich–Popov (KYP) lemma) is an important tool for characterizing the sublevel sets of the closed-loop $\mathcal{H}_\infty$ norm. There is some subtlety in the assumptions needed for different versions of the bounded real lemma. We briefly clarify this subtlety here. Consider the following closed-loop linear system:

$$\begin{aligned} x_{t+1} &= A_{\mathrm{cl}} x_t + B_{\mathrm{cl}} w_t \\ z_t &= C_{\mathrm{cl}} x_t, \end{aligned} \tag{A.3}$$

where $\rho(A_{\mathrm{cl}}) < 1$. Recall that the closed-loop $\mathcal{H}_\infty$ norm of the above system (from $\{w_t\}$ to $\{z_t\}$) is defined as

$$H_{cl} = \sup_{\omega \in [0, 2\pi]} \lambda_{\max}^{1/2}\big(B_{\mathrm{cl}}^\mathsf{T}(e^{-j\omega} I - A_{\mathrm{cl}})^{-\mathsf{T}} C_{\mathrm{cl}}^\mathsf{T} C_{\mathrm{cl}} (e^{j\omega} I - A_{\mathrm{cl}})^{-1} B_{\mathrm{cl}}\big).$$

The strict version of the bounded real lemma [74, Theorem 21.12] states that $H_{\mathrm{cl}} < \gamma$ if and only if there exists a positive definite matrix P such that

$$\begin{bmatrix} A_{\mathrm{cl}}^\mathsf{T} P A_{\mathrm{cl}} - P & A_{\mathrm{cl}}^\mathsf{T} P B_{\mathrm{cl}} \\ B_{\mathrm{cl}}^\mathsf{T} P A_{\mathrm{cl}} & B_{\mathrm{cl}}^\mathsf{T} P B_{\mathrm{cl}} \end{bmatrix} + \begin{bmatrix} C_{\mathrm{cl}}^\mathsf{T} C_{\mathrm{cl}} & 0 \\ 0 & -\gamma^2 I \end{bmatrix} \prec 0.$$

The above strict-version version of the bounded real lemma does not require any assumptions on $(A_{\mathrm{cl}}, B_{\mathrm{cl}}, C_{\mathrm{cl}})$. There is also a non-strict version of the bounded real lemma which requires the extra assumption that $(A_{\mathrm{cl}}, B_{\mathrm{cl}}, C_{\mathrm{cl}})$ is minimal (i.e. $(A_{\mathrm{cl}}, B_{\mathrm{cl}})$ is controllable and $(A_{\mathrm{cl}}, C_{\mathrm{cl}})$ is observable). Consider (A.3) with $(A_{\mathrm{cl}}, B_{\mathrm{cl}}, C_{\mathrm{cl}})$ being a minimal realization. Then the non-strict version of the bounded real lemma[5] states that

[5] See [6, Section 2.7.3] for the continuous-time counterpart.

$H_{\rm cl} \leq \gamma$ if and only if there exists a positive definite matrix P such that

$$\begin{bmatrix} A_{\rm cl}^{\mathsf T} P A_{\rm cl} - P & A_{\rm cl}^{\mathsf T} P B_{\rm cl} \\ B_{\rm cl}^{\mathsf T} P A_{\rm cl} & B_{\rm cl}^{\mathsf T} P B_{\rm cl} \end{bmatrix} + \begin{bmatrix} C_{\rm cl}^{\mathsf T} C_{\rm cl} & 0 \\ 0 & -\gamma^2 I \end{bmatrix} \preceq 0.$$

For the main result in this paper, the proof require the above non-strict version of the bounded real lemma. Specifically, in our setting, we have $\|G_K\|_{2\to 2} = H_{\rm cl}$ if we choose $A_{\rm cl} = A - BK$, $B_{\rm cl} = I$, and $C_{\rm cl} = (Q + K^{\mathsf T} R K)^{\frac{1}{2}}$. Since Q is positive definite, such choice of $(A_{\rm cl}, B_{\rm cl}, C_{\rm cl})$ leads to a minimal realization, and hence it is valid to apply the non-strict version of the bounded real lemma to characterize the non-strict sublevel sets of $\|G_K\|_{2\to 2}$.

B Detailed Proofs of Our Main Results in Sections 3 & 4

In the main paper, we have only provided the proof sketches for Lemma 1, Lemma 2, Theorem 1, Lemma 3, and Theorem 2. In this section, we provide detailed proofs for these results.

B.1 Proof for Lemma 1

Suppose we have a sequence $\{K^l\}$ satisfying $\|K^l\|_2 \to +\infty$. We can choose $\mathbf{w} = \{w_0, 0, 0, \cdots\}$ with $\|w_0\| = 1$. Then we have:

$$\begin{aligned} J(K^l) &= \max_{\mathbf{w}:\|\mathbf{w}\|\leq 1} \sum_{t=0}^{\infty} x_t^{\mathsf T}(Q + (K^l)^{\mathsf T} R K^l) x_t \\ &\geq w_0^{\mathsf T}(Q + (K^l)^{\mathsf T} R K^l) w_0 \\ &= w_0^{\mathsf T} Q w_0 + (K^l w_0)^{\mathsf T} R (K^l w_0) \\ &\geq \lambda_{\min}(R) \|K^l w_0\|^2, \end{aligned}$$

where the first inequality holds since we plugged into a specific $\mathbf{w}$ over the $\max$ operation and the matrix $Q + (K^l)^{\mathsf T} R K^l$ is positive definite. The second inequality uses the fact that $R \geq \lambda_{\min}(R) I$. Then by carefully choosing w_0, we can ensure $J(K^l) \to +\infty$ as $\|K^l\| \to +\infty$.

Next, we assume $K^l \to K$ where K is on the boundary $\partial\mathcal{K}$. Clearly we have $\rho(A - BK) = 1$. We will use a frequency-domain argument to prove $J(K^l) \to +\infty$. Notice $(e^{-j\omega} I - A + BK^l)^{-\mathsf T}(Q + (K^l)^{\mathsf T} R K^l)(e^{j\omega} I - A + BK^l)^{-1}$ is positive semidefinite for all l and $\omega \in [0, 2\pi]$. Since $\rho(A - BK) = 1$, there exists some ω_0 such that $(e^{\pm j\omega_0} I - A + BK)$ become singular. Therefore, we have:

$$\begin{aligned} J(K^l) &= \sup_{\omega\in[0,2\pi]} \lambda_{\max}^{1/2}\big((e^{-j\omega} I - A + BK^l)^{-\mathsf T}(Q + (K^l)^{\mathsf T} R K^l)(e^{j\omega} I - A + BK^l)^{-1}\big) \\ &= \sup_{\omega\in[0,2\pi]} \|(e^{-j\omega} I - A + BK^l)^{-\mathsf T}(Q + (K^l)^{\mathsf T} R K^l)(e^{j\omega} I - A + BK^l)^{-1}\|^{1/2} \\ &\geq \lambda_{\min}^{1/2}(Q) \sup_{\omega\in[0,2\pi]} (\|(e^{-j\omega} I - A + BK^l)^{-1}\| \|(e^{j\omega} I - A + BK^l)^{-1}\|)^{1/2} \\ &\geq \lambda_{\min}^{1/2}(Q) (\|(e^{-j\omega_0} I - A + BK^l)^{-1}\| \|(e^{j\omega_0} I - A + BK^l)^{-1}\|)^{1/2}. \end{aligned}$$

Clearly the above argument relies on the fact that Q is positive definite. Notice that we have $\rho(A - BK^l) < 1$ for each l , and hence we have $\sigma_{\min}(e^{\pm j\omega_0} I - A + BK^l) > 0$, i.e. the smallest singular value of matrix $(e^{\pm j\omega_0} I - A + BK^l)$ is strictly positive for all l. As $l \to \infty$, one can show that the matrix $(e^{\pm j\omega_0} I - A + BK^l)$ converges to the singular matrix $(e^{\pm j\omega_0} I - A + BK)$ with $K = \lim_{l\to\infty} K^l \in \partial\mathcal{K}$. Hence we have $\sigma_{\min}(e^{\pm j\omega_0} I - A + BK^l) \to 0$ as $l \to \infty$, which implies $\|(e^{\pm j\omega_0} I - A + BK^l)^{-1}\| \to +\infty$ as $l \to \infty$. Therefore, we have $J(K^l) \to +\infty$ as $K^l \to K \in \partial\mathcal{K}$. This completes the proof.

B.2 Proof for Lemma 2

We first show the compactness of $\mathcal{K}_\gamma$. Since J is continuous, we know $\mathcal{K}_\gamma = \{K \in \mathcal{K} : J(K) \leq \gamma\}$ is a closed set. It remains to show $\mathcal{K}_\gamma$ is bounded. Suppose there exist $\gamma > 0$ such that $\mathcal{K}_\gamma$ is unbounded. Then there exists a sequence $\{K^l\}_{l=1}^{\infty} \subset \mathcal{K}_\gamma$ such that $\|K^l\|_2 \to +\infty$ as $l \to \infty$. But by coerciveness of $J(K)$, we must have $J(K^l) \to +\infty$ as well, which contradicts that $J(K^l) \leq \gamma$ for all l. Hence $\mathcal{K}_\gamma$ is bounded. Therefore, $\mathcal{K}_\gamma$ is compact. The path-connectedness of the strict sublevel sets for the continuous-time $\mathcal{H}_\infty$ control problem has been proved in [28]. We can slightly modify the proof in [28] 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, we can apply [39, Theorem 5.2] to show $\mathcal{K}_\gamma$ is also path-connected.

We can also prove the path-connectedness of $\mathcal{K}_\gamma$ by using the non-strict version of the bounded real lemma reviewed in Section A.5. This proof is more self-contained, and hence also included here. 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 non-strict matrix inequality holds

$$\begin{bmatrix} (A-BK)^\mathsf{T} P (A-BK) - P & (A-BK)^\mathsf{T} P \\ P(A-BK) & P \end{bmatrix} + \begin{bmatrix} Q + K^\mathsf{T} R K & 0 \\ 0 & -\gamma^2 I \end{bmatrix} \preceq 0. \tag{B.1}$$

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 variable [6]. Specifically, there exists a matrix function $\mathrm{LMI}(Y, L, \gamma)$ which is linear in (Y, L, γ) such that $Y \succ 0$ and $\mathrm{LMI}(Y, L, \gamma) \preceq 0$ if and only if (B.1) is feasible with $K = LY^{-1}$, and $P = \gamma Y^{-1}$. For completeness, we provide the detailed derivation of $\mathrm{LMI}(Y, L, \gamma)$ as follows.

Step 1: Let $\tilde{P} = \frac{1}{\gamma} P$, dividing both sides of (B.1) by γ. Then by Schur complement lemma, (B.1) can be rewritten as:

$$\begin{bmatrix} (A-BK)^\mathsf{T} \tilde{P} (A-BK) - \tilde{P} + \frac{1}{\gamma} K^\mathsf{T} R K & (A-BK)^\mathsf{T} \tilde{P} & I \\ \tilde{P}(A-BK) & \tilde{P} - \gamma I & 0 \\ I & 0 & -\gamma Q^{-1} \end{bmatrix} \preceq 0, \tag{B.2}$$

To see this, noticing that $-\gamma Q^{-1}$ is negative definite and (B.1) can be rewritten as:

$$\begin{bmatrix} (A-BK)^\mathsf{T} \tilde{P} (A-BK) - \tilde{P} + \frac{1}{\gamma} K^\mathsf{T} R K & (A-BK)^\mathsf{T} \tilde{P} \\ \tilde{P}(A-BK) & \tilde{P} - \gamma I \end{bmatrix} - \begin{bmatrix} I \\ 0 \end{bmatrix} (-\frac{1}{\gamma} Q) \begin{bmatrix} I & 0 \end{bmatrix} \preceq 0. \tag{B.3}$$

Step 2: Now we can apply Schur complement lemma to (B.2) again to show its equivalence to:

$$\begin{bmatrix} -\tilde{P} + \frac{1}{\gamma} K^\mathsf{T} R K & 0 & I & (A-BK)^\mathsf{T} \tilde{P} \\ 0 & -\gamma I & 0 & \tilde{P} \\ I & 0 & -\gamma Q^{-1} & 0 \\ \tilde{P}(A-BK) & \tilde{P} & 0 & -\tilde{P} \end{bmatrix} \preceq 0. \tag{B.4}$$

To see this, noticing that $-\tilde{P}$ is negative definite and the LHS of (B.2) can be rewritten as:

$$\begin{bmatrix} -\tilde{P} + \frac{1}{\gamma} K^\mathsf{T} R K & 0 & I \\ 0 & -\gamma I & 0 \\ I & 0 & -\gamma Q^{-1} \end{bmatrix} - \begin{bmatrix} (A-BK)^\mathsf{T} \tilde{P} \\ \tilde{P} \\ 0 \end{bmatrix} (-\tilde{P}^{-1}) \begin{bmatrix} \tilde{P}(A-BK) & \tilde{P} & 0 \end{bmatrix}. \tag{B.5}$$

Step 3: Then we left and right multiply $\mathrm{diag}(\begin{bmatrix} \tilde{P}^{-1} & I & I & \tilde{P}^{-1} \end{bmatrix})$ on both sides of (B.4), this will not change the definiteness of (B.4) and leads the LHS of (B.4) to:

$$\begin{bmatrix} \tilde{P}^{-1} & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & \tilde{P}^{-1} \end{bmatrix} \begin{bmatrix} -\tilde{P} + \frac{1}{\gamma} K^\mathsf{T} R K & 0 & I & (A-BK)^\mathsf{T} \tilde{P} \\ 0 & -\gamma I & 0 & \tilde{P} \\ I & 0 & -\gamma Q^{-1} & 0 \\ \tilde{P}(A-BK) & \tilde{P} & 0 & -\tilde{P} \end{bmatrix} \begin{bmatrix} \tilde{P}^{-1} & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & \tilde{P}^{-1} \end{bmatrix} \tag{B.6}$$

$$= \begin{bmatrix} -\tilde{P}^{-1} + \frac{1}{\gamma} \tilde{P}^{-1} K^\mathsf{T} R K \tilde{P}^{-1} & 0 & \tilde{P}^{-1} & \tilde{P}^{-1}(A-BK)^\mathsf{T} \\ 0 & -\gamma I & 0 & I \\ \tilde{P}^{-1} & 0 & -\gamma Q^{-1} & 0 \\ (A-BK)\tilde{P}^{-1} & I & 0 & -\tilde{P}^{-1} \end{bmatrix} \tag{B.7}$$

Step 4: Substituting $Y = \tilde{P}^{-1}$ and $KY = L$ into the above matrix leads (B.4) to:

$$\begin{bmatrix} -Y + \frac{1}{\gamma} L^\mathsf{T} R L & 0 & Y & (AY-BL)^\mathsf{T} \\ 0 & -\gamma I & 0 & I \\ Y & 0 & -\gamma Q^{-1} & 0 \\ AY-BL & I & 0 & -Y \end{bmatrix} \preceq 0 \tag{B.8}$$

Step 5: Furthermore, (B.8) is equivalent to:

$$\begin{bmatrix} -Y & 0 & Y & (AY-BL)^\mathsf{T} \\ 0 & -\gamma I & 0 & I \\ Y & 0 & -\gamma Q^{-1} & 0 \\ AY-BL & I & 0 & -Y \end{bmatrix} - \begin{bmatrix} L^\mathsf{T} \\ 0 \\ 0 \\ 0 \end{bmatrix} (-\frac{1}{\gamma} R) \begin{bmatrix} L & 0 & 0 & 0 \end{bmatrix} \preceq 0 \tag{B.9}$$

Applying Schur complement lemma to (B.9) again leads to:

$$\mathrm{LMI}(Y, L, \gamma) := \begin{bmatrix} -Y & 0 & Y & (AY - BL)^\mathsf{T} & L^\mathsf{T} \\ 0 & -\gamma I & 0 & I & 0 \\ Y & 0 & -\gamma Q^{-1} & 0 & 0 \\ AY - BL & I & 0 & -Y & 0 \\ L & 0 & 0 & 0 & -\gamma R^{-1} \end{bmatrix} \preceq 0. \tag{B.10}$$

This proves the equivalence between (B.10) and (B.1). Therefore, we have:

$$\begin{aligned} &\{K \in \mathcal{K} : J(K) \leq \gamma\} \\ \iff &\{K : \text{(B.1) is feasible, } P \succ 0\} \\ \iff &\{K = LY^{-1} : \text{(B.10) is feasible, } Y \succ 0\}. \end{aligned} \tag{B.11}$$

Noticing that the set of (Y, L) satisfying (B.10) and $Y \succ 0$ is convex and hence path-connected. In addition, the map $K = LY^{-1}$ is continuous for positive definite Y. We can conclude that $\mathcal{K}_\gamma = \{K \in \mathcal{K} : J(K) \leq \gamma\}$ is path-connected. Such a proof is actually quite similar to the proof presented in [28]. The main difference is that the assumptions in this paper allow us to directly use the non-strict version of the bounded real lemma.

B.3 Proof for Lemma 3

The proof of this lemma depends on the above convex parameterization (Y, L). Since $\mathrm{LMI}(Y, L, \gamma)$ is linear in (Y, L, γ), the condition (B.10) is convex. Suppose we have $K = LY^{-1}$ where $(Y, L, J(K))$ is a feasible point for the convex regime $\mathrm{LMI}(Y, L, J(K)) \preceq 0$ and $Y \succ 0$. In addition, by the non-strict version of the bounded real lemma, we know there exists a pair (Y^*, L^*) such that $K^* = L^*(Y^*)^{-1}$, $\mathrm{LMI}(Y^*, L^*, J(K^*)) \preceq 0$ and $Y^* \succ 0$. By convexity of the LMI condition $\mathrm{LMI}(Y, L, \gamma) \preceq 0$ and $Y \succ 0$, we know that the line segment between $(Y, L, J(K))$ and $(Y^*, L^*, J(K^*))$ is also in this convex set. Therefore, for any $0 \leq t \leq 1$, we know the point $(Y + t\Delta Y, L + t\Delta L, J(K) + t(J(K^*) - J(K)))$ also satisfies

$$\mathrm{LMI}(Y + t\Delta Y, L + t\Delta L, J(K) + t(J(K^*) - J(K))) \preceq 0, \ Y + t\Delta Y \succ 0$$

where $\Delta L = L^* - L$, and $\Delta Y = Y^* - Y$. Since $Y \succ 0$ and $Y^* \succ 0$, we automatically have $Y + t\Delta Y \succ 0$. Based on (B.11), we can construct a policy $(L + t\Delta L)(Y + t\Delta Y)^{-1}$ and the resultant closed-loop $\mathcal{H}_\infty$ norm must be smaller than or equal to $J(K) + t(J(K^*) - J(K))$. Formally, we have

$$J((L + t\Delta L)(Y + t\Delta Y)^{-1}) \leq J(K) + t(J(K^*) - J(K)).$$

Based on the fact $J(K^*) < J(K)$, we can use the above inequality to construct a direction d such that $J'(K, d) < 0$. Specifically, let's choose $d = \Delta L Y^{-1} - LY^{-1}\Delta Y Y^{-1}$. Then we have

$$\begin{aligned} J'(K, d) &= \lim_{t \searrow 0} \frac{J(K + t(\Delta L Y^{-1} - LY^{-1}\Delta Y Y^{-1})) - J(K)}{t} \\ &= \lim_{t \searrow 0} \frac{J(LY^{-1} + t(\Delta L Y^{-1} - LY^{-1}\Delta Y Y^{-1})) - J(K)}{t} \\ &=_{(a)} \lim_{t \searrow 0} \frac{J((L + t\Delta L)(Y + t\Delta Y)^{-1}) - J(K)}{t} \\ &\quad + \lim_{t \searrow 0} \frac{J((L + t\Delta L)(Y + t\Delta Y)^{-1} + O(t^2)) - J((L + t\Delta L)(Y + t\Delta Y)^{-1})}{t} \\ &\leq_{(b)} \lim_{t \searrow 0} \left(\frac{J((L + t\Delta L)(Y + t\Delta Y)^{-1}) - J(K)}{t} + O(t) \right) \\ &\leq \lim_{t \searrow 0} \left(\frac{J(K) + t(J(K^*) - J(K)) - J(K)}{t} + O(t) \right) \\ &= J(K^*) - J(K) < 0, \end{aligned}$$

where the step (a) relies on the fact that $(Y + t\Delta Y)^{-1} = Y^{-1} - tY^{-1}\Delta Y Y^{-1} + O(t^2)$ and the step (b) uses the fact that $J(\cdot)$ is locally Lipschitz (Proposition 2). Finally, the last inequality holds since we know $J(K) > J(K^*)$. Notice $d \neq 0$. Otherwise the above argument still works and we have $J'(K, 0) < 0$. This is impossible since we know $J'(K, 0) = 0$. This leads to a contradiction, and we must have $d \neq 0$. This completes the proof for this lemma.

B.4 Proof for Theorem 2

We first show that $K^n \in \mathcal{K}_{J(K^0)}$ for all n by induction. By choice, we have $\delta^n = \frac{c\Delta_0}{n+1} \leq c\Delta_0$ for all n with a fixed $c \in (0, 1)$. For $n = 1$, we have $K^1 = K^0 - c\Delta_0 F^0/\|F^0\|_2$. Since the norm of $F^0/\|F^0\|_2$

equals to one and $\delta^0 = c\Delta_0 > 0$, we have $K^1 \in \mathbb{B}_{\delta^0}(K^0)$, where $\mathbb{B}_{\delta^0}(K^0)$ is the ball centered at K^0 with radius δ^0. By definition of Δ_0, we have $\mathbb{B}_{\delta^0}(K^0) \subseteq \mathcal{K}$, hence $K^1 \in \mathcal{K}$. In addition, (13) implies $J(K^1) \leq J(K^0) - \delta^0 \|F^0\|_2$. Hence we have $\mathcal{K}_{J(K^1)} \subseteq \mathcal{K}_{J(K^0)}$, which implies $K^1 \in \mathcal{K}_{J(K^0)}$. 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:

$$\sum_{n=0}^{N} \delta^n \|F^n\|_2 \leq J(K^0) - J(K^{N+1}) \leq J(K^0) - J^*, \tag{B.12}$$

where the second inequality holds since $J(K^{N+1}) \geq J^*$. Since we have $\sum_{n=0}^{\infty} \delta^n = c\Delta_0 \sum_{n=1}^{\infty} \frac{1}{n} = \infty$, we know $\liminf_{n\to\infty} \|F^n\|_2 = 0$. There exists one subsequence $\{n_i\}$ such that $\|F^{n_i}\|_2 \to 0$. For this subsequence, since the policy parameter sequence stays in the compact set $\mathcal{K}_{J(K^0)}$ for all n, the resultant policy sequence $\{K^{n_i}\}$ is also bounded and has a convergent subsequence $\{K^{n_{i(l)}}\}$ which converges to some limit point K^∞. In addition, we have $\|F^{n_{i(l)}}\|_2 \to 0$ as well.

For any fixed $\delta > 0$, there exists N_δ such that $\mathbb{B}_{\delta^{n_{i(l)}}}(K^{n_{i(l)}}) \subset \mathbb{B}_\delta(K^\infty)$ for all $n_{i(l)} \geq N_\delta$. Hence we have $F^{n_{i(l)}} \in \partial_{\delta^{n_{i(l)}}} J(K^{n_{i(l)}}) \subset \partial_\delta J(K^\infty)$ for all $n_{i(l)} \geq N_\delta$ by definition. Noticing that $F^{n_{i(l)}} \to 0$ and $\partial_\cdot J(\cdot)$ is closed, we have $0 \in \partial_\delta J(K^\infty)$ for any $\delta > 0$. Since $\bigcap_{n_{i(l)}} \partial_{\delta^{n_{i(l)}}} J(K^\infty) = \partial_C J(K^\infty)$ by Remark 2, we have $0 \in \partial_C(K^\infty)$ as well. Hence K^∞ is a Clarke stationary point by definition. Therefore, the function value $\{J(K^{n_{i(l)}})\}$ converges to J^*. Notice that $J(K^n)$ is monotonically decreasing for the entire sequence $\{n\}$ (not just the subsequence $\{n_i\}$ and $\{n_{i(l)}\}$). Hence the sequence $\{J(K^n)\}$ has a limit, and this limit has to be $J(K^\infty) = J^*$. This completes the proof.

Remark 2. *For completeness, we also include a proof for $\bigcap_n \partial_{\delta^n} J(K) = \partial_C J(K)$. For small enough $\delta > 0$ we have $\partial_C J(K) \subset \partial_\delta J(K)$. If $\delta^1 > \delta^2$, then $\partial_{\delta^2} J(K) \subset \partial_{\delta^1} J(K)$. Therefore, we have*

$$\partial_C J(K) \subseteq \lim_{\delta^n \searrow 0} \partial_{\delta^n} J(K) = \bigcap_n \partial_{\delta^n} J(K).$$

To show the reverse, we assume on the contrary that there exists $L \in \bigcap_n \partial_{\delta^n} J(K) \setminus \partial_C J(K)$. Denote $S_n := \{\cup_{K' \in \mathbb{B}_{\delta^n}(K)} \partial_C J(K')\}$, we know that S_n is compact and nested. By [24, Lemma 2.6] we have

$$L \in \bigcap_n \partial_{\delta^n} J(K) = \operatorname{conv} \bigcap_n S_n.$$

Hence $L = \sum_j t_j L_j$ with $L_j \in \bigcap_n S_n$, $t_j \geq 0$, and $\sum_j t_j = 1$. Therefore, we have $L_j \in S_n$ for each n. For each n, there exist a sequence $\{K_j^{n,r}\}$ such that $K_j^{n,r} \to K_j^n$, $\nabla J(K_j^{n,r}) \to L_j$ as $r \to \infty$, and $\|K_j^n - K\| \leq \delta^n$. Then there exists a large enough $r(n,j)$ such that $\|L_j - \nabla J(K_j^{n,r(n,j)})\| \leq \delta^n$, and $\|K_j^{n,r(n,j)} - K\| \leq 2\delta^n$. Since $\delta^n \to 0$ as $n \to \infty$, we have found a sequence $\{K_j^{n,r(n,j)}\}_{n=1}^\infty$, such that $K_j^{n,r(n,j)} \to K$ and $\nabla J(K_j^{n,r(n,j)}) \to L_j$ as $n \to \infty$. Therefore, we have $L_j \in \partial_C J(K)$. Finally, by convexity of $\partial_C J(K)$ [24, Proposition 2.3], we have $L = \sum_j t_j L_j \in \partial_C J(K)$, contradiction. Therefore, we have $\bigcap_n \partial_{\delta^n} J(K) = \partial_C J(K)$.

C Discussions on Implementable Variants and Related Convergence Results

In this section, we discuss several implementable variants of Goldstein's subgradient method and the related convergence guarantees.

C.1 Gradient Sampling

The gradient sampling (GS) method can be viewed as an approximation of Goldstein's subgradient method. The idea of GS has been explained in Section 4.2. The details of the GS method are presented in Algorithm 1. For Algorithm 1, we can show that if $\epsilon^n \downarrow 0$ and $\delta^n \downarrow 0$, every cluster point of $\{K^n\}$ is Clark stationary for $J(K)$ [9, Theorem 3.1]. It is worth to mention that Algorithm 1 keeps the iterates to stay in the feasible region $\mathcal{K}$ due to a slightly modification of the line search step (C.1) to (C.4) (See [34, Section 4.2] for more details about such modification). Such a modification relies on the fact that there is a strict separation between $\mathcal{K}_{J(K^0)}$ and $\mathcal{K}^c$. Then it is straightforward to combine Theorem 1 and the proof of [9, Theorem 3.1] to show the global convergence of the gradient sampling method under the same assumption stated in [9, Theorem 3.1]. We formalize Theorem 4 as the following global convergence result of GS as below.

Theorem C.1. *Consider the policy optimization problem* (6) *with the $\mathcal{H}_\infty$ cost function defined in* (5)*. Suppose $K^0 \in \mathcal{K}$. Let $\{K^n\}$ be a sequence generated by Algorithm 1 with $\delta_{opt} = \epsilon_{opt} = 0$ and $\mu_\delta, \mu_\epsilon < 1$. With probability one the algorithm does not stop such that $\delta^n \downarrow 0$ and $\epsilon^n \downarrow 0$. In addition, with probability one, every*

Algorithm 1: Gradient Sampling (GS)

Require: $K^0 \in \mathcal{K}$ at which J is differentiable, optimality tolerances $(\delta_{opt}, \epsilon_{opt})$, initial sampling radius $\delta^0 = c\Delta_0$ with $c \in (0,1)$, a stationary target $\epsilon^0 \in [0, \infty)$, reduction factors $\mu_\delta, \mu_\epsilon \in (0,1]$, sample size $m \geq n_x n_u + 1$, line search parameters $(\beta, \theta) \in (0,1] \times (0,1]$

for $n \in \mathbb{N} \cup \{0\}$ **do**

Independently sample $\{K_{n,1}, \cdots K_{n,m}\}$ uniformly from $\mathbb{B}_{\delta^n}(K^n)$

Compute F^n as the solution of

$$\min \ \frac{1}{2}\|F\|_2^2$$
$$\text{subject to } F \in \mathcal{F}^n = \text{conv}\{\nabla J(K^n), \nabla J(K_{n,1}), \cdots, \nabla J(K_{n,m})\}$$

if $\|F^n\|_2 \leq \epsilon_{opt}$ and $\delta^n \leq \delta_{opt}$ **then** terminate

if $\|F^n\|_2 \leq \epsilon^n$ **then** set $\delta^{n+1} \leftarrow \mu_\delta \delta^n$, $\epsilon^{n+1} \leftarrow \mu_\epsilon \epsilon^n$, and $t^n \leftarrow 0$

else set $\delta^{n+1} \leftarrow \delta^n$, $\epsilon^{n+1} \leftarrow \epsilon^n$, define:

$$\hat{F}^n = \delta^n F^n / \|F^n\|_2, \tag{C.1}$$

and compute t^n such that

$$t^n \leftarrow \max\{t \in \{1, \theta, \theta^2, \cdots\} : J(K^n - t\hat{F}^n) < J(K^n) - \beta t \delta^n \|F^n\|_2\} \tag{C.2}$$

if J is differentiable at $K^n - t^n \hat{F}^n$

then set $K^{n+1} \leftarrow K^n - t^n \hat{F}^n$

else set K^{n+1} randomly as any point where J is differentiable such that

$$J(K^{n+1}) < J(K^n) - \beta t^n \delta^n \|F^n\|_2 \tag{C.3}$$

$$\|K^n - t^n \hat{F}^n - K^{n+1}\|_2 \leq \frac{1}{2} \min\{t^n, \delta^n\} \|\hat{F}^n\|_2 \tag{C.4}$$

end for

cluster point of $\{K^n\}$ is a Clark stationary point of $J(K)$. The sequence $\{J(K^n)\}$ is monotonically decreasing almost surely, and we have $J(K^n) \to J^*$ as $n \to \infty$ with probability one.

Proof. We first show that the iterates $\{K^n\}$ stays inside the feasible set $\mathcal{K}$. It is possible that there is no update at step n when $\|F^n\|_2 \leq \epsilon^n$. Furthermore, K is updated when a descent condition is found (C.2). There are two cases: if J is differentiable at $K^n - t^n \hat{F}^n$, then we have $K^{n+1} = K^n - t^n \hat{F}^n \in \mathbb{B}_{\delta^n}(K^n)$ as $\|\hat{F}^n\|_2 = \delta^n$ and $t^n < 1$. If J is not differentiable at $K^n - t^n \hat{F}^n$, then K^{n+1} is randomly sampled such that (C.3) and (C.4) hold. From (C.4), we have $\|K^{n+1} - K^n\| \leq t^n \|\hat{F}^m\| = t^n \delta^n \leq \delta^n$. Hence we have $K^{n+1} \in \mathbb{B}_{\delta^n}(K^n)$ as well. Since we have set $\delta^0 = c\Delta_0$ for some $c \in (0,1)$, we have $K^1 \in \mathbb{B}_{\delta^0}(K^0) \in \mathcal{K}$. In addition, since $J(K^1) \leq J(K^0)$, we further have $K^1 \in \mathcal{K}_{J(K^1)} \subseteq \mathcal{K}_{J(K^0)}$. Repeating such argument shows that $K^n \in \mathcal{K}_{J(K^0)}$ and hence $K^n \in \mathcal{K}$ for all n. From [9, Theorem 3.1], each cluster point of $\{K^n\}$ is a Clark stationary point of $J(K)$[6].

Since $\{K^n\}$ stays in the compact set $\mathcal{K}_{J(K^0)}$ for all n, there exists a subsequence $\{n_i\}$ such that $K^{n_i} \to K^\infty$ as $n_i \to \infty$. In addition, K^∞ has to be one of the cluster point of $\{K^n\}$. Therefore, K^∞ is a global minimum and the function value $\{J(K^{i_n})\}$ associated with the subsequence $\{K^{n_i}\}$ converges to J^*. Notice that $J(K^n)$ is monotonically decreasing for the entire sequence $\{n\}$ (not only the subsequence $\{n_i\}$). Hence we have $J(K_n) \to J^*$. This completes the proof. □

Remark 3. *The implementation of* (C.3) *and* (C.4) *can be done as follows: From* (C.2)*, we know that there exists t^n such that $J(K^n - t^n \hat{F}^n) < J(K^n) - \beta t^n \delta^n \|F^n\|_2$. Since J is continuous, there exist ε such that $J(K) < J(K^n) - \beta t^n \delta^n \|F^n\|_2 \ \forall K \in \mathbb{B}_\varepsilon(K^n - t^n \hat{F}^n)$.*

Now we can let $r = \min\{t^n, \delta^n\} \|\hat{F}^n\|_2 / N$ and N large enough so that $r \leq \varepsilon$. Then we can sample K^{n+1} uniformly from $\mathbb{B}_r(K^n - t^n \hat{F}^n)$ until K^{n+1} is differentiable. Since J is differentiable almost everywhere, this process terminates with probability one.

[6] Based on the specific form of the $\mathcal{H}_\infty$ cost (5) and the chain rule in [2], it is obvious that the $\mathcal{H}_\infty$ cost is continuously differentiable on a set of full measure.

C.2 Non-derivative Sampling

In [35], it has been shown that the gradient oracle in the gradient sampling method can be replaced with the function value oracle and similar convergence guarantees can still be obtained. This leads to the non-derivative sampling (NS) method. To avoid the gradient computation, one can use the Gupal estimation to approximate the derivatives. For completeness, we briefly review such approximation below, more details can be found in [35].

For $\alpha > 0$, let $n = n_x n_u$, then the Steklov averaged function J_α is defined as follows:

$$J_\alpha(K) := \int_{\mathbb{R}^n} J(K - Y)\psi_\alpha(Y)dY, \tag{C.5}$$

where $\psi_\alpha : \mathbb{R}^n \to \mathbb{R}_+$ is the Steklov mollifier defined by

$$\psi_\alpha := \begin{cases} 1/\alpha^n & \text{if } Y \in [-\alpha/2, \alpha/2]^n \\ 0 & \text{otherwise.} \end{cases} \tag{C.6}$$

Then the partial derivatives of J_α are given by

$$\frac{\partial J_\alpha}{\partial K_i}(K) = \int_{\mathbb{B}_\infty} \chi_i(K, \alpha, \zeta)d\zeta$$

for $i = 1, \cdots, n$, where K_i refers to the i-th element of vectorized K, $\mathbb{B}_\infty := [-1/2, 1/2]^n$ is the unit cube centered at 0 and

$$\begin{aligned}\chi_i(K, \alpha, \zeta) := \frac{1}{\alpha}[&J(K_1 + \alpha\zeta_1, \cdots, K_{i-1} + \alpha\zeta_{i-1}, K_i + \frac{1}{2}\alpha, K_{i+1} + \alpha\zeta_{i+1}, \cdots, K_n + \alpha\zeta_n) \\ &-J(K_1 + \alpha\zeta_1, \cdots, K_{i-1} + \alpha\zeta_{i-1}, K_i - \frac{1}{2}\alpha, K_{i+1} + \alpha\zeta_{i+1}, \cdots, K_n + \alpha\zeta_n)].\end{aligned}$$

Therefore, for the unit cube $Z := \prod_{i=1}^n \mathbb{B}_\infty$ in $\mathbb{R}^{n\times n}$, we can use the Gupal estimate of $\nabla J_\alpha(K)$:

$$\chi(K, \alpha, z) := (\chi_1(K, \alpha, \zeta_1), \cdots, \chi_n(K, \alpha, \zeta_n)) \text{ for } z := (\zeta_1, \cdots, \zeta_n) \in Z.$$

It is worth to mention that such approximation needs $2n$ evaluations of J compared to the n evaluations for gradient sampling Algorithm 1. The pseudo code for the NS method is presented in Algorithm 2 [35].

Algorithm 2: Nonderivative Gradient Sampling (NS)

Require: $K^0 \in \mathcal{K}$, optimality tolerances $(\delta_{opt}, \epsilon_{opt})$, initial sampling radius $\delta^0 = c\Delta_0$ with $c \in (0, 1)$, a stationary target $\epsilon^0 \in [0, \infty)$, reduction factors $\mu_\delta, \mu_\epsilon \in (0, 1]$, sample size $m \geq n_x n_u + 1$, line search parameters $(\beta, \theta) \in (0, 1)^2$. A sequence of mollifier parameters $\alpha_n \downarrow 0$ in $(0, 1]$.

for $n \in \mathbb{N} \cup \{0\}$ **do**

Independently sample $\{K_{n,1}, \cdots K_{n,m}\}$ uniformly from $\mathbb{B}_\delta(K^n)$

Independently sample $\{\zeta_{n,1}, \cdots \zeta_{n,m}\}$ uniformly from Z

Compute F^n as the solution of

$$\begin{aligned}&\min \ \frac{1}{2}\|F\|_2^2 \\ &\text{subject to } F \in \mathcal{F}^n = \text{conv}\{\chi(K_{n,1}, \alpha_n, \zeta_{n,1}), \cdots, \chi(K_{n,m}, \alpha_n, \zeta_{n,m})\}\end{aligned}$$

if $\|F^n\|_2 \leq \epsilon_{opt}$ and $\delta^n \leq \delta_{opt}$ **then** terminate

if $\|F^n\| \leq \epsilon^n$, set $\epsilon^{n+1} \leftarrow \mu_\epsilon \epsilon^n$, $\delta^{n+1} \leftarrow \mu_\delta \delta^n$, and $t^n \leftarrow 0$.

else set $\delta^{n+1} \leftarrow \delta^n$, $\epsilon^{n+1} \leftarrow \epsilon^n$, $\hat{F}^n = F^n/\|F^n\|_2$, and compute t^n as follows:

(i) Choose an initial step size $t = t^n_{ini} = \delta^n$

(ii) If $J(K^n - t\hat{F}^n) \leq J(K^n) - \beta t\|F^n\|$, return $t^n := t$

(iii) If $\theta t < t^n_{\min}$, return $t^n := 0$

(iv) Set $t := \theta t$, and go to (ii).

Update $K^{n+1} = K^n - t^n \hat{F}^n$

end for

Similar to the GS algorithm, the diminishing δ^n guarantees the iterations of NS Algorithm stay inside the feasible set $\mathcal{K}$. Furthermore, [35, Theorem 3.8] shows that every cluster point of $\{K^n\}$ is a Clark stationary for $J(K)$. Therefore, we have the global convergence based on Theorem 1.

Theorem C.2. *Consider the policy optimization problem* (6) *with the $\mathcal{H}_\infty$ cost function defined in* (5)*. Suppose $K^0 \in \mathcal{K}$. Let $\{K^n\}$ be a sequence generated by Algorithm 2 with $\delta_{opt} = \epsilon_{opt} = 0$ and $\mu_\delta, \mu_\epsilon < 1$. With probability one the algorithm does not stop such that $\delta^n \downarrow 0$ and $\epsilon^n \downarrow 0$. In addition, with probability one, every cluster point of $\{K^n\}$ is a Clark stationary point. The sequence of $\{J(K^n\}$ is monotonically decreasing almost surely, and we have $J(K^n) \to J^*$ as $n \to \infty$ with probability one.*

Proof. We first show that the iterates $\{K^n\}$ stays inside the feasible set $\mathcal{K}$. Noticing that K is updated only when a descent condition is found. In addition, at the line search step, we can choose the initial step size $t^n_{ini} = \delta^n$. Therefore, when $n = 1$, we have $K^1 = K^0$ or $K^1 = K^0 - t^0 F^0/\|F^0\|_2 \in \mathcal{K}$ since $t^n \leq \delta^0$. In addition, since $J(K^1) \leq J(K^0)$, we further have $K^1 \in \mathcal{K}_{J(K^1)} \subseteq \mathcal{K}_{J(K^0)}$. Repeating such argument shows that $K^n \in \mathcal{K}_{J(K^0)}$ and hence $K^n \in \mathcal{K}$ for all n. From Theorem 3.8 in [35], each cluster point of $\{K^n\}$ is a Clark stationary point of $J(K)$.

The proof for the function value convergence is identical to Theorem C.1. We omit it here. □

C.3 Model-Free NS

When the model is unknown, there are various methods available for estimating the $\mathcal{H}_\infty$-norm from data [42, 43, 53, 50, 65, 47, 63, 62]. The model-free NS method just replaces the exact oracle for the $\mathcal{H}_\infty$ cost (5) with some noisy estimates. Based on our experience, the multi-input multi-output (MIMO) power iteration method [46] works quite well as a stochastic zeroth-order oracle for the purpose of implementing NS in the model-free setting. The main idea of the MIMO power iteration method is that the $\mathcal{H}_\infty$ norm of an LTI MIMO system can be approximated from its finite-time approximated representation. Given a black-box simulator for a stable system G_K, the power iteration method provides an efficient oracle for estimating the $\mathcal{H}_\infty$ of G_K from trajectories of G_K. The trajectory length is denoted as N, which quantifies the sample size. The larger N is, the better the $\mathcal{H}_\infty$-norm estimation is. We refer the readers to [46] for implementation details of the power iteration method.

The sample complexity for model-free NS is unknown. However, our numerical results show that such such a model-free method tracks the convergence of its model-based counterpart given reasonable amount of data.

C.4 Interpolated Normalized Gradient Descent (INGD) with Finite-Time Complexity

The INGD method was originally proposed in [67], and can be viewed as an implementable approximation of Goldstein's subgradient method. The main advantage of INGD is that it yields finite-time iteration complexity for finding (δ, ϵ)-stationary points. The original method in [67] works for any directionally differentiable functions. Specially, at an iterate K^n, [67] proposes Algorithm 3 to compute the update direction F^n. In the unconstrained setup, it has been shown that Algorithm 3 terminates with high probability [67, Lemma 13]. In addition, one can combine Algorithm 3 and (13) with a fixed δ yields the INGD method (See Algorithm 4). A non-asymptotic analysis for Algorithm 4 has been provided. In particular, it has been show that Algorithm 4 terminates with high probability as well [67, Theorem 8].

Algorithm 3: MiniNorm1

Input: K, $\delta > 0$, and $\epsilon > 0$
Let $n = 0$, call the generalized gradient oracle $F^0 = \mathbb{O}(K, 0)$
while $\|F^n\| > \epsilon$ **and** $\frac{\delta}{4}\|F^n\| \geq J(K) - J(K - \delta\frac{F^n}{\|F^n\|})$ **do**
  Sample Y^n uniformly from $[K^n, K - \delta\frac{F^n}{\|F^n\|}]$
  Call the generalized gradient oracle $H^n = \mathbb{O}(Y^n, -F^n)$
  Update $F^n = \beta^n F^n + (1 - \beta^n)H^n$ with $\beta^n = \frac{4L^2 - \|F^n\|^2}{4L^2 + 2\|F^n\|^2}$
  $n \leftarrow n + 1$
end while
Return F^n

However, Algorithm 3 (the original algorithm in [67]) requires a special generalized gradient oracle $\mathbb{O}$ defined in [67, Assumption 1]. To relax this requirement, a variant of Algorithm 3 has been proposed in [13]. The version of INGD in [13] relies on a variant of Algorithm 3 which only requires standard gradient oracle for any differentiable points. This variant of the MiniNorm1 sub-routine is documented in Algorithm 5. Therefore, the generalized gradient oracle $\mathbb{O}$ is no longer required.

Similar to the Algorithm 3, MiniNorm2 also terminates with high probability. Now we discuss how to modify the analysis for our $\mathcal{H}_\infty$ control problem. It has been show in [67, Theorem 8] that for unconstrained nonsmooth

Algorithm 4: Interpolated Normalized Gradient Descent (INGD)

Initial: K^0, L
Input: δ, ϵ
for $l = 0, \cdots, L$ **do**
  $F^l = \text{MiniNorm1}(K^l, \delta, \epsilon)$
  $K^{l+1} = K^l - \delta F^l / \|F^l\|_2$
end
Return K^L

Algorithm 5: MiniNorm2

Input: K, $\delta > 0$, and $\epsilon > 0$
Let $n = 0$, $F^0 = \nabla J(Y^0)$ where $Y^0 \sim \mathbb{B}_\delta(K)$
while $\|F^n\|_2 > \epsilon$ **and** $\frac{\delta}{4}\|F^n\|_2 \geq J(K) - J(K - \delta F^n/\|F^n\|_2)$ **do**
  Choose any r satisfying $0 < r < \|F^n\|_2 \sqrt{1 - (1 - \frac{\|F^n\|_2^2}{128L^2})^2}$
  Sample H^n uniformly from $\mathbb{B}_r(F^n)$
  Sample Y^n uniformly from $[K, K - \delta \frac{H^n}{\|H^n\|_2}]$
  $F^{n+1} = \arg\min_{Z \in [F^n, \nabla J(Y^n)]} \|Z\|_2$
  $n \leftarrow n + 1$
end while
Return F^n

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 Δ is the initial function value gap, and p is the failure probability (i.e. the optimization succeeds with the probability $(1-p)$). We can choose $\delta < \Delta_0$ to show that the resultant algorithm can find the (δ, ϵ)-stationary point for $\mathcal{H}_\infty$ state-feedback control problem with the high-probability iteration complexity $\mathcal{O}\left(\frac{\Delta L^2}{\epsilon^3 \delta} \log(\frac{1}{p})\right)$. Now we formalize Theorem 5 as follows.

Theorem C.3. *Consider the $\mathcal{H}_\infty$ state-feedback control problem* (6) *with the objective function $J(K)$ defined in* (5)*. Suppose $K^0 \in \mathcal{K}$ and denote $\Delta_0 := \text{dist}(\mathcal{K}_{J(K^0)}, \mathcal{K}^c)$. Let $\Delta = J(K^0) - J^*$. Denote the Lipschitz constant of $J(K)$ over the sublevel set $\mathcal{K}_{J(K^0)}$ as L_0. For any $\delta < \Delta_0$, the iterations generated by the INGD algorithm will stay in $\mathcal{K}$ almost surely. In addition, the INGD method finds a (δ, ϵ)-stationary point with a high-probability iteration complexity $\mathcal{O}\left(\frac{\Delta L_0^2}{\epsilon^3 \delta} \log(\frac{4\Delta}{p\delta\epsilon})\right)$.*

Proof. We first show that the iterates of INGD stay inside the feasible set $\mathcal{K}$ when $\delta < \Delta_0$. At step n, the Algorithm 3 returns an F^n either $\|F^n\| \leq \epsilon$ or $J(K^n - \delta F^n/\|F^n\|_2) \leq J(K^n) - \frac{\delta}{4}\|F^n\|_2$. If $\|F^n\| \leq \epsilon$, the INGD algorithm will terminate as we have found a (δ, ϵ)-stationary point. For the latter case, we update K^{n+1} as $K^{n+1} = K^n - \delta F^n/\|F^n\|_2$. Therefore, we only update K when a descent condition is satisfied. Since $K^0 \in \mathcal{K}$ and $\delta < \Delta_0$, we have $K^1 = K^0 - \delta F^0/\|F^0\|_2 \in \mathcal{K}$. In addition, since $J(K^1) \leq J(K^0) - \frac{\delta}{4}\|F^0\|_2$, we further have $K^1 \in \mathcal{K}_{J(K^0)}$ as $\mathcal{K}_{J(K^1)} \subseteq \mathcal{K}_{J(K^0)}$. Repeating such argument shows that $K^n \in \mathcal{K}_{J(K^0)}$ and hence $K^n \in \mathcal{K}$ for all n. The high probability complexity bound is a direct result of [67, Theorem 8]. We omit the details here. □

D Numerical experiments

In this section, we provide simulation results to show the effeteness of the aforementioned implantable algorithms. Computations are executed on a desktop computer with a 3.3 GHz Intel Xeon W-1350 processor with MATLAB R2021b.

We implemented the Interpolated Normalized Gradient Descent (INGD) algorithm with with decreasing δ, Gradient Sampling (GS), and Nonderivative Sampling (NS). In addition, model-free NS has also been implemented. Subsequently, we provide the simulation result for one-time INGD method, which aims to illustrate that (δ, ϵ)-stationarity does not imply being δ-close to an ϵ-stationary point of J. Finally, we implement the NS and GS methods in higher dimensional cases.

Table 1: Algorithm parameters for GS and NS

Parameter	(n_x, n_u, m)	δ^0	ϵ^0	μ_δ	μ_ϵ	δ_{opt}	ϵ_{opt}	β	θ	α_n
GS	$(3, 1, 4)$	0.01	100	0.5	0.5	0	0	0.5	0.9	N/A
NS	$(3, 1, 4)$	0.01	100	0.5	0.5	0	0	0.5	0.9	$0.1/(n+1)$

D.1 More details on the simulation results in Figure 1

Implementation of INGD, GS, NS, and model-free NS methods: For the problem matrices given in (16), we have $J^* = 7.3475$. In order to obtain the Clark stationary point by using INGD method. We need to perform the repeated INGD algorithm with δ decreases to 0. To this end, we set $\epsilon = 1 \times 10^{-5}$ and $\delta^{n+1} = 0.7\delta^n$ with $\delta^0 = 0.01$ for $n = 0, 1, \cdots$. In other words, whenever the INGD method successfully finds a (δ^n, ϵ)-stationary point, we reduce δ^n by a constant factor (i.e., $\delta^{n+1} = 0.7\delta^n$), then the INGD method tries to find a (δ^{n+1}, ϵ)-stationary point. As δ^n tends to 0, one should expect that the INGD algorithm returns a Clark stationary point. In addition, we also implement the Gradient Sampling (GS) and Non-derivative sampling (NS) methods, where the algorithm parameters are set as in Table 1. Finally, for the model-free NS method for the given problem matrices (16), we choose the sample size in the $\mathcal{H}_\infty$-norm estimation as $N = 100$.

Implementation of randomly generated cases: The middle plot of Figure 1 shows the performance of NS algorithm on the randomly generated cases. In particular, 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 $[0, 1]$; $B \in \mathbb{R}^{3\times 1}$ with each element uniformly sampled from $[0, 1]$; $Q = I + \zeta I \in \mathbb{R}^{3\times 3}$ with ζ uniformly sampled from $[0, 0.1]$; $R \in \mathbb{R}$ 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 $\rho(A - BK^0) < 1$.

Implementation of Model-free NS method with decreasing sample size N: The right plot of Figure 1 provides the simulation result of model-free NS method with decreasing sample size N used in the $\mathcal{H}_\infty$-norm estimation. The hyperparameters are the same as the ones used in NS. The only difference is that only we need to specify a sample size in the MIMO power iteration method for evaluating $J(K)$ from sampled trajectories. Specifically, we choose $N = [100\ 50\ 20\ 10]$. As we can see from the right plot in Figure 1, as we decrease the sample size, the $\mathcal{H}_\infty$-norm oracle become more noisy, and the algorithm performance has been degraded with more oscillations.

Two gradient oracles for differentiable points. To implement INGD and GS methods, we need to compute the gradients of $J(K)$ at differentiable points. We provide two different oracles for computing the gradients of $J(K)$ at differentiable points. To this end, let us rewrite $J(K)$ as follows:

$$J(K) = \sup_{\omega \in [0, 2\pi]} \sigma_{\max}(H(K, \omega)), \tag{D.1}$$

where $H(K, \omega) = (Q + K^\mathsf{T} RK)^{\frac{1}{2}} (e^{j\omega} I - A + BK)^{-1}$ and $\sigma_{\max}$ denotes the maximum singular value of a given matrix. It is obvious that (D.1) is equivalent to the cost function as in (5). It is well known that $J(K)$ is differentiable at K if the $\mathcal{H}_\infty$ norm of $H(K, \omega)$ is achieved at one frequency ω_0 and the the largest singular value of $H(K, \omega_0)$ has multiplicity one [59].

For the first oracle $\mathbb{O}_1(K)$, given a differentiable point $K \in \mathbb{R}^{n_u \times n_x}$, $\mathbb{O}_1(K)$ returns a finite time difference estimation of the gradient $J(K)$ at K defined as follows:

$$\nabla J(K) \approx [\tilde{\nabla} J(K)_{ij}]_{ij}, \qquad \tilde{\nabla} J(K)_{ij} = \frac{J(K + hE_{ij}) - J(K - hE_{ij})}{2h}, \tag{D.2}$$

where E_{ij} has the same dimension as K whose (i, j)-th entry is 1 and other entries are 0, and h is a small positive scalar ($h = 0.0001$ for example).

While the second oracle $\mathbb{O}_2(K)$ returns $\nabla J(K)$ via explicit analytic formula. For completeness, we restate the explicit formula of the gradients of $J(K)$ at differentiable points.

Let $H(K, \omega_0) = U\Sigma V^H$ be a singular value decomposition, where $U, V \in \mathbb{C}^{n_x \times n_x}$ are complex unitary matrices consists of the left singular vectors and right singular vectors of $H(K, \omega_0)$, respectively. Let u_1 be the first column of U and let v_1 be the first column of V. In addition, denote $H_1(K) = (Q + K^\mathsf{T} RK)^{\frac{1}{2}}$, $H_2(K) = (e^{j\omega_0} I - A + BK)^{-1}$, and define $\mathbf{I} = \int_0^\infty e^{-\tau H_1(K)} H_2(K) v_1 u_1^* e^{-\tau H_1(K)} d\tau$. Then the gradient of $J(K)$ at differentiable point K can be calculated as:

$$\nabla J(K) = \mathrm{Re}(RK(\mathbf{I} + \mathbf{I})^\mathsf{T} - (H_2(K) v_1 u_1^* H(K, \omega_0) B)^\mathsf{T}). \tag{D.3}$$

The derivation of the above formula uses the chain rule of the total derivative and $dJ_k = \mathrm{Tr}(\nabla J(K) dK^\mathsf{T})$. We omit the details here.

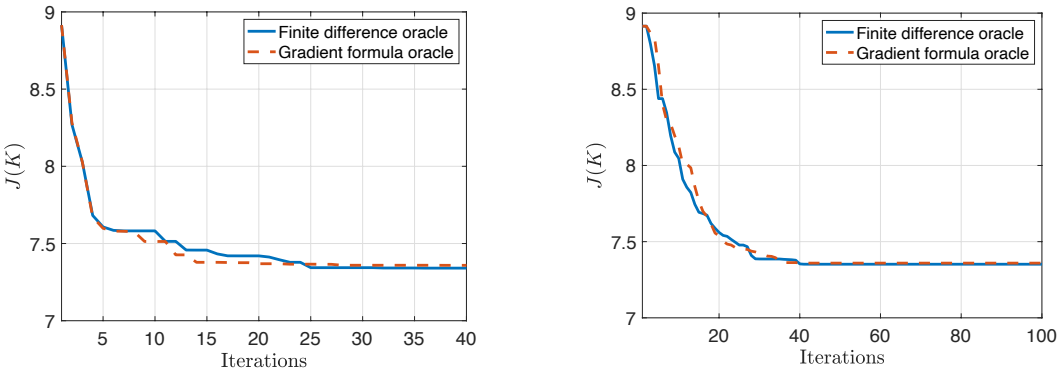

Figure 2: Simulation result for finite-difference oracle and gradient formula oracle. Left: INGD method. Right: GS method.

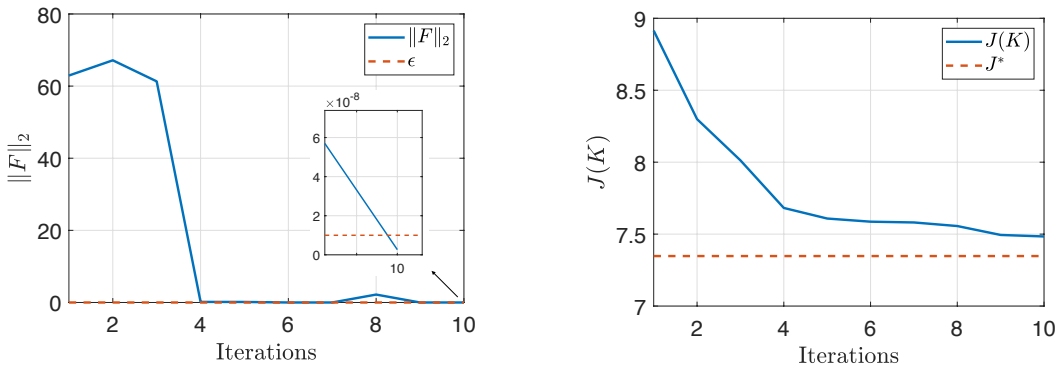

Figure 3: Simulation result for one-time INGD methods with $\delta = 0.01$ and $\epsilon = 1 \times 10^{-8}$. Left: the trajectory of the $\|F\|_2$. Right: the trajectory of function value $J(K)$.

Figure 2 shows the simulation results for the problem matrices given in (16) by using the above two different gradient oracles for INGD and GS, respectively. It can be seen that both oracles work well and the trajectories of two different oracles are consistent to each other.

D.2 Numerical experiment on INGD method with constants (δ, ϵ)

The simulation results for INGD with Mininorm2 and constants (δ, ϵ) are shown in Figure 3, where we set $\delta = 0.01$ and $\epsilon = 1 \times 10^{-8}$. From the left figure of Figure 3, it can be seen that it takes 10 steps for INGD to find a (δ, ϵ)-stationary point. At step 10, we have $\|F\|_2 < \epsilon$. However, $J(K)$ does not converge to the optimal value as shown in the right plot of Figure 3. This result confirms that (δ, ϵ)-stationarity does not imply being δ-close to an ϵ-stationary point of J as commented in [67, 13].

D.3 Numerical experiments on higher-dimension cases

In this section, we test the effeteness of NS and GS methods on higher-dimensional cases. In particular, we first consider the following set of problem matrices:

$$A = \begin{bmatrix} 1.7865 & 0.3912 & 0.8758 & 0.5996 \\ 0.2756 & 1.3175 & 0.7692 & 0.4848 \\ 0.4764 & 0.9786 & 1.0618 & 0.7591 \\ 0.4489 & 0.7918 & 0.6014 & 1.7520 \end{bmatrix}, \quad B = \begin{bmatrix} 0.1303 & 0.0312 \\ 0.1309 & 0.0528 \\ 0.7452 & 0.6727 \\ 0.2460 & 0.0743 \end{bmatrix}, \tag{D.4}$$

$$Q = 1.0613 I_4, \quad R = 1.1315 I_2.$$

For the above example, we have $J^* = 43.26$ and we select the initialization K^0 to be:

$$K^0 = \begin{bmatrix} 2.4364 & 2.2337 & 2.4867 & 1.5551 \\ 12.1213 & -4.6823 & 2.1718 & -2.5906 \end{bmatrix},$$

which satisfies $\rho(A - BK^0) = 0.9567 < 1$.

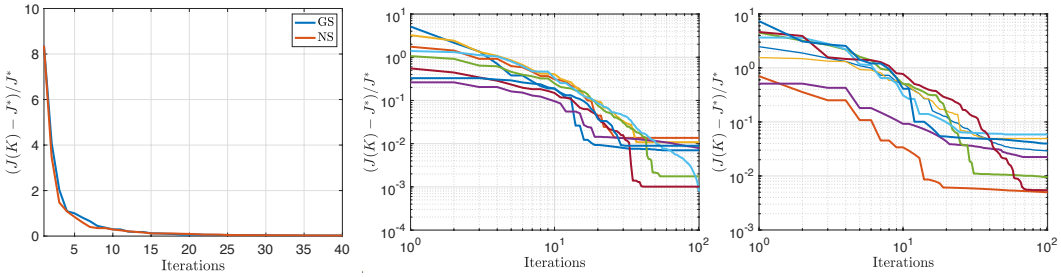

Figure 4: Simulation results for higher dimensional cases. Left: The trajectory of relative error of NS with problem matrices (D.4). Middle: The trajectory of relative error of 8 randomly generated cases for NS method. Right: The trajectory of relative error of 8 randomly generated cases for GS method.

In addition, we also perform the NS and GS algorithm on the randomly generated cases with higher dimensions. We set $A \in \mathbb{R}^{4\times 4}$ to be $I + \xi$, where each element of $\xi \in \mathbb{R}^{4\times 4}$ is sampled uniformly from $[0, 1]$; $B \in \mathbb{R}^{4\times 2}$ with each element uniformly sampled from $[0, 1]$; $Q = I + \zeta I \in \mathbb{R}^{4\times 4}$ with ζ uniformly sampled from $[0, 0.1]$; $R = I + \upsilon I \in \mathbb{R}^{2\times 2}$ with υ uniformly sampled from $[0, 0.5]$. For each experiment, the initial condition $K^0 \in \mathbb{R}^{2\times 4}$ is also chosen such that $\rho(A - BK^0) < 1$.

Figure 4 shows the simulation results on the higher dimensional cases. The left one shows the convergence of NS and GS method with the problem matrices in (D.4). The middle figure and right figure show the performance of NS and GS with the randomly generated cases, respectively. It can be seen that both GS and NS works well with higher dimensional systems.

E Further Discussions

E.1 Uniqueness of the Minimizing set

Our theory does not answer whether the global minimum for the $\mathcal{H}_\infty$ state-feedback synthesis problem is unique or not. As commented in Section 3, the path-connectedness of $\mathcal{K}_\gamma$ for every γ can be used to show that there exists a unique global minimizing set in a certain sense [39, Sections 2&3]. However, we are not able to rule out the possibility that the uniuqe global minimizing set actually consists of multiple points. Whether the global minimum of the $\mathcal{H}_\infty$ state-feedback control problem is unique or not is an interesting open question.

E.2 Possible generalizations for nonlinear systems

For nonlinear systems, it is possible to generalize Theorems 3 and 5. The proofs for the finite-time complexity of finding (δ, ϵ)-stationary points can be generalized to the constrained policy optimization setting, as long as the cost function J is coercive over the nonconvex feasible set $\mathcal{K}$. In contrast, the convergence to global minimum may be too much to ask for general nonlinear robust control problems. If the cost function J is not coercive, one may just add regularization to induce coerciveness such that one can still find some approximated (δ, ϵ)-stationary points provably. Such developments are worth more investigations in the future.