Optimization Landscape of Policy Gradient Methods for Discrete-Time Static Output Feedback

Jingliang Duan, *Member, IEEE*, Jie Li, Xuyang Chen, Kai Zhao, *Member, IEEE*, Shengbo Eben Li, *Senior Member, IEEE*, and Lin Zhao, *Member, IEEE*

***Abstract*—In recent times, significant advancements have been made in delving into the optimization landscape of policy gradient methods for achieving optimal control in linear time-invariant (LTI) systems. Compared with state-feedback control, output-feedback control is more prevalent since the underlying state of the system may not be fully observed in many practical settings. This article analyzes the optimization landscape inherent to policy gradient methods when applied to static output feedback (SOF) control in discrete-time LTI systems subject to quadratic cost. We begin by establishing crucial properties of the SOF cost, encompassing coercivity, L-smoothness, and M-Lipschitz continuous Hessian. Despite the absence of convexity, we leverage these properties to derive novel findings regarding convergence (and nearly dimension-free rate) to stationary points for three policy gradient methods, including the vanilla policy gradient method, the natural policy gradient method, and the Gauss–Newton method. Moreover, we provide proof that the vanilla policy gradient method exhibits linear convergence toward local minima when initialized near such minima. This article concludes by presenting numerical examples that validate our theoretical findings. These results not only characterize the performance of gradient descent for optimizing the SOF problem but also provide insights into the effectiveness of general policy gradient methods within the realm of reinforcement learning.**

***Index Terms*—Policy gradient, reinforcement learning (RL), static output feedback (SOF).**

I. Introduction

REINFORCEMENT learning (RL) has showcased remarkable proficiency comparable to human capabilities in a variety of challenging tasks, spanning from games to robot control [1], [2], [3], [4]. RL methods relying on policy gradient, including DDPG [5], SAC [6], and DSAC [7], are commonly employed to identify parameterized optimal control policies for tasks with continuous action space. However, despite these achievements, a complete theoretical grasp of the complexity and performance of such algorithms remains lacking, even in fundamental scenarios like linear time-invariant (LTI) systems.

Optimal control problems have served as a potent tool for exploring various characteristics of RL, including aspects like stability [8], [9]. Within the framework of policy gradient methods, prior investigations have delved into the optimization landscape and the attributes of convergence, particularly within the context of linear quadratic regulator (LQR) problems [10]. It is widely acknowledged that the optimal solution of LQR problems can be derived through the solution of the algebraic Riccati equation (ARE). However, in the pursuit of unveiling the characteristics of policy gradient during the training process, the focus shifts toward the direct optimization of the linear policy using the LQR cost, rather than solving the corresponding ARE. In this context, it is noteworthy that the related optimization problem generally assumes a nonconvex nature since the set of stabilizing state-feedback gains may lack convexity [11]. An influential work by Fazel et al. [11] discovered that the discrete-time LQR objective function exhibits properties of gradient dominance and almost smoothness, enabling policy gradient methods to achieve linear global convergence, despite the nonconvexity of the LQR. Subsequent studies have explored akin attributes, with specific attention to both discrete-time and continuous-time LQR [12], [13], [14], [15], as well as various LQR variations [16], [17], [18], [19], [20], [21].

Compared with state-feedback control, output-feedback control is more common since the underlying state of the system may not be fully observed in practical settings [22], [23], [24], [25]. Most of the existing convergence results of gradient descent are built on full state feedback, whereas the convergence for static output feedback (SOF) LQR has received little attention. As one of the most crucial open topics in LTI systems, SOF can only acquire some linear combinations of states, rather than entire states [26]. Unlike the state feedback LQR, the output feedback gain of SOF may have a disconnected domain, with local minima, saddle points, or even local maxima in each component [27], [28], [29]. Therefore, finding the globally optimal SOF controller using

Manuscript received 31 March 2023; revised 24 June 2023 and 9 September 2023; accepted 24 September 2023. The work of Jingliang Duan was supported in part by NSF, China, under Grant 52202487, and in part by the State Key Laboratory of Automotive Safety and Energy, China, under Project KFY2212. The work of Lin Zhao was supported in part by the Singapore Ministry of Education Tier 1 Academic Research Fund under Grant A0009030-00-00 and Grant 22-5460-A0001. This article was recommended by Associate Editor D. Zhao. *(Jingliang Duan and Jie Li contributed equally to this work.) (Corresponding author: Lin Zhao.)*

Jingliang Duan is with the School of Mechanical Engineering, University of Science and Technology Beijing, Beijing 100083, China, and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: duanjl@ustb.edu.cn).

Jie Li and Shengbo Eben Li are with the School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China (e-mail: jie-li18@mails.tsinghua.edu.cn; lishbo@mail.tsinghua.edu.cn).

Xuyang Chen, Kai Zhao, and Lin Zhao are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: chenxuyang@u.nus.edu; zhaokai@cqu.edu.cn; elezhli@nus.edu.sg).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2023.3323316.

Digital Object Identifier 10.1109/TCYB.2023.3323316

gradient descent is generally intractable. However, it is still of great significance to investigate the optimization landscape of SOF, particularly concerning the convergence toward stationary points, which will bring new insights into the performance of policy gradient methods for partially observed control problems.

Recent efforts have elucidated the optimization landscape pertaining to continuous-time LTI systems in the context of SOF, delineating convergence rates to stationary points [27]. However, these findings are limited to the vanilla policy gradient method, with the convergence behaviors of popular alternatives like the natural policy gradient [30] and the Gauss–Newton method [31] yet to be fully clarified. Given the inherent distinctions between difference equations and differential equations, the analysis of optimization landscapes in discrete-time systems assumes a distinct character. Notably, discrete-time SOF holds practical significance due to its alignment with control frequency limitations, and the utilization of discrete-time data from real-world systems holds the promise of extending convergence insights to model-free contexts.

Despite these considerations, the theoretical properties of policy gradient methods applied to discrete-time SOF scenarios have been overlooked in existing studies. This study takes the initial step toward bridging this gap and offers the following principal contributions.

1) We unveil several crucial properties of the SOF cost function, in spite of its nonconvex nature. Notable among these properties are the compact sublevel set, L-smoothness, and M-Lipschitz continuous Hessian, which are instrumental in the subsequent theoretical analyses. A standout feature of our work is the establishment of Hessian Lipschitz continuity, a property that provides critical insights into the path of convergence toward local minima within SOF problems. This property is typically overlooked in the extant literature on both SOF and state-feedback LQR. Diverging from approaches that establish L-smoothness [19], we prove Hessian Lipschitz continuity through a direct application of its definition, thereby avoiding complex tensor operations.
2) Unlike state-feedback LQR, where theories of convergence often hinge on the concept of gradient dominance [11], [12], [13], [14], [16], [18], the landscape of SOF problems presents greater complexities. This complexity arises from the disconnected nature of the stabilizing SOF domain and the potential multiplicity of stationary points. Leveraging the compactness and L-smoothness of the SOF cost function, we show that three different policy gradient methods (the vanilla policy gradient, the natural policy gradient, and the Gauss–Newton method) can converge to stationary points at a (nearly) dimension-free rate, given an initial stabilizing policy.
3) Furthermore, when the initial point is proximate to a local minimum, we demonstrate that the vanilla policy gradient method converges linearly toward it, predicated on the Lipschitz continuity of the Hessian.

It is worth noting that the primary goal of this study is not the introduction of a new control algorithm for specific control problems. Rather, we focus on the SOF problem as a fertile ground for investigating the convergence, complexity, and optimality of policy gradient-based RL algorithms. Our findings offer new perspectives on the effectiveness of policy gradient methods in SOF problems and illuminate the efficacy of employing general policy gradient methods when learning SOF policies with unknown systems.

Notation: $\|M\|$, $\|M\|_F$, and $\rho(M)$ denote the induced 2-norm, Frobenius norm, and spectral radius of a matrix M; for square matrices, $\mathrm{Tr}(M)$, $\lambda_{\min}(M)$, and $\sigma_{\min}(M)$ represent the trace, minimal eigenvalue, and minimal singular value; $\mathrm{vec}(M)$ indicates the vectorized form; $\partial\mathbb{M}$ signifies the boundary of the set $\mathbb{M}$; $M \succ N$ ($M \succeq N$) implies that $M-N$ is positive definite (semidefinite); $\mathbb{S}_{+}^{n}$ ($\mathbb{S}_{++}^{n}$) refers to the set of symmetric $n \times n$ positive semidefinite (definite) matrices; $\mathbb{N}$ stands for the set of natural numbers; $\mathbb{E}_x$ denotes taking expectation over x; and I_n denotes the identity matrix.

II. Problem Statement

Consider the discrete-time LTI dynamic model

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t \\ y_t &= Cx_t \end{aligned} \tag{1}$$

with x denoting the state, y representing the output, and matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{d\times n}$ describing the system dynamics. The LQR problem aims to find a control policy to minimize the accumulated linear quadratic cost

$$\mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\left(x_t^\top Q x_t + u_t^\top R u_t\right)\right] \tag{2}$$

where it is assumed that $\mathbb{E}_{x_0\sim\mathcal{D}}[x_0x_0^\top] \succ 0$, and $Q \in \mathbb{S}_{++}^{n}$ and $R \in \mathbb{R}_{++}^{m}$ are performance weights. The assumption on $\mathbb{E}_{x_0\sim\mathcal{D}}[x_0x_0^\top] \succ 0$ is quite standard in learning-based control [11], [12] and can be somehow informally thought as the persistent excitation condition in data-driven control.

The SOF is defined as

$$u_t = -Ky_t \tag{3}$$

with $K \in \mathbb{R}^{m\times d}$. Substituting the SOF controller into the dynamic model (1) yields

$$x_{t+1} = \mathcal{A}_K x_t \tag{4}$$

where $\mathcal{A}_K := A - BKC$. We can further reformulate the linear quadratic cost (2) as

$$J(K) = \mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^{\infty} x_t^\top\left(Q + C^\top K^\top RKC\right)x_t\right]. \tag{5}$$

This study assumes that a stabilizing controller is present. We refer to the set of all stabilizing control gain K as the feasible set, that is

$$\mathbb{K} := \left\{K \in \mathbb{R}^{m\times d} : \rho(\mathcal{A}_K) < 1\right\}. \tag{6}$$

For the LTI systems (1), the value function of state x takes the quadratic closed-loop form

$$V_K(x_t) := x_t^\top P_K x_t \tag{7}$$

where $P_K \in \mathbb{S}_{+}^{n}$.

We define the accumulated state correlation matrix as

$$\Sigma_K := \mathbb{E}_{x_0\sim\mathcal{D}} \sum_{t=0}^{\infty} x_t x_t^\top. \tag{8}$$

If the initial state correlation matrix is positive definite, that is

$$X_0 := \mathbb{E}_{x_0\sim\mathcal{D}}\left[x_0 x_0^\top\right] \succ 0 \tag{9}$$

one has that the minimal singular value of X_0

$$\mu := \sigma_{\min}(X_0) \ > 0. \tag{10}$$

Since $\Sigma_K \succeq X_0$, it is straightforward that

$$\sigma_{\min}(\Sigma_K) \geq \mu. \tag{11}$$

With P_K and Σ_K, it is well known in [11] and [32] that SOF control of discrete-time LTI systems with quadratic cost can be formulated into the following problem.

Problem 1 (Policy Optimization for SOF):

$$\min_{K\in\mathbb{K}} \ J(K) = \mathrm{Tr}(P_K X_0) = \mathrm{Tr}\Big(\big(Q + C^\top K^\top RKC\big)\Sigma_K\Big) \tag{12}$$

where P_K and Σ_K satisfy the following Lyapunov equations:

$$P_K = Q + C^\top K^\top RKC + \mathcal{A}_K^\top P_K \mathcal{A}_K \tag{13a}$$

$$\Sigma_K = X_0 + \mathcal{A}_K \Sigma_K \mathcal{A}_K^\top. \tag{13b}$$

The unique positive-definite solution of (13a) can be expressed as

$$P_K = \sum_{j=0}^{\infty} \mathcal{A}_K^{\top j}\Big(Q + C^\top K^\top RKC\Big)\mathcal{A}_K{}^j. \tag{14}$$

The formulation of Problem 1 enables us to derive the analytical policy gradients to analyze the optimization landscape. For this problem, we make the following standard assumption.

Assumption 1: (A, B) is controllable, (C, A) is observable, and C has independent rows.

Note that the feasible set $\mathbb{K}$ of Problem 1 can possess a disconnected domain, replete with local minima, saddle points, or even local maxima at the stationary points of each component [27], [28], [29]; therefore, Problem 1 is generally nonconvex, making the convergence analysis far from straightforward.

III. Gradients and Hessian

In this section, we give the analytical expression for both the gradient and Hessian. The derivations follow similar lines as the state-feedback LQR case [11], [19].

Lemma 1 (Policy Gradient Formula): For any control gain K in the feasible set $\mathbb{K}$, we have

$$\nabla J(K) = 2E_K \Sigma_K C^\top \tag{15}$$

with $E_K := (R + B^\top P_K B)KC - B^\top P_K A$.

Proof: From (7) and (13a), it follows that:

$$\begin{aligned} V_K(x_0) &= x_0^\top\Big(Q + C^\top K^\top RKC\Big)x_0 + x_0^\top \mathcal{A}_K^\top P_K \mathcal{A}_K x_0 \\ &= x_0^\top\Big(Q + C^\top K^\top RKC\Big)x_0 + V_K(\mathcal{A}_K x_0). \end{aligned} \tag{16}$$

Taking the gradient of $V_K(x_0)$ w.r.t. K, one has

$$\begin{aligned} \nabla V_K(x_0) &= 2E_K x_0 x_0^\top C^\top + x_1^\top \nabla P_K x_1\big|_{x_1=\mathcal{A}_K x_0} \\ &= 2E_K x_0 x_0^\top C^\top + \nabla V_K(x_1)\big|_{x_1=\mathcal{A}_K x_0} \\ &= 2E_K \sum_{t=0}^{\infty}\Big(x_t x_t^\top\Big)C^\top. \end{aligned} \tag{17}$$

By taking expectation w.r.t. $\mathcal{D}$, the expression of policy gradient is obtained

$$\nabla J(K) = \mathbb{E}_{x_0\sim\mathcal{D}} \nabla V_K(x_0) = 2E_K \Sigma_K C^\top \tag{18}$$

which completes the proof. ■

Note that the objective function $J(K)$ is twice differentiable. Thus, the analytical form of the Hessian of the objective function can be derived. To simplify our analysis without delving into tensors, we analyze the Hessian along a certain matrix $Z \in \mathbb{R}^{m\times d}$, whose expression is as follows:

$$\begin{aligned} \nabla^2 J(K)[Z, Z]: &= \frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\bigg|_{\lambda=0} J(K + \lambda Z) \\ &= \mathrm{Tr}\left(\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\bigg|_{\lambda=0} P_{K+\lambda Z} X_0\right). \end{aligned} \tag{19}$$

Lemma 2: For any control gain K in the feasible set $\mathbb{K}$, the Hessian of the objective function $J(K)$ along a certain matrix $Z \in \mathbb{R}^{m\times d}$ is

$$\begin{aligned} \nabla^2 J(K)[Z, Z] = {} & \mathrm{Tr}\Big(2(ZC)^\top\Big(B^\top P_K B + R\Big)ZC\Sigma_K\Big) \\ & - \mathrm{Tr}\Big(4(BZC)^\top P_K'[Z]\mathcal{A}_K\Sigma_K\Big) \end{aligned} \tag{20}$$

where

$$P_K'[Z] = \sum_{j=0}^{\infty} \mathcal{A}_K^{\top j}\Big(C^\top Z^\top E_K + E_K^\top ZC\Big)\mathcal{A}_K{}^j. \tag{21}$$

Proof: Denote $P_K'[Z] := (\mathrm{d}/\mathrm{d}\lambda)|_{\lambda=0} P_{K+\lambda Z}$. From (13a), we have

$$\begin{aligned} P_K'[Z] &= C^\top Z^\top E_K + E_K^\top ZC + \mathcal{A}_K^\top P_K'[Z]\mathcal{A}_K \\ &= \sum_{j=0}^{\infty} \mathcal{A}_K^{\top j}\Big(C^\top Z^\top E_K + E_K^\top ZC\Big)\mathcal{A}_K{}^j. \end{aligned} \tag{22}$$

Then, its second derivative $P_K''[Z] := (\mathrm{d}^2/\mathrm{d}\lambda^2)|_{\lambda=0} P_{K+\lambda Z}$ can be derived as

$$P_K''[Z] = S_1 + \mathcal{A}_K^\top P_K''[Z]\mathcal{A}_K = \sum_{j=0}^{\infty} \mathcal{A}_K^{\top j} S_1 \mathcal{A}_K{}^j \tag{23}$$

where

$$\begin{aligned} S_1 := 2\Big(& C^\top Z^\top\Big(R + B^\top P_K B\Big)ZC \\ & - (BZC)^\top P_K'[Z]\mathcal{A}_K - \mathcal{A}_K^\top P_K'[Z]BZC\Big). \end{aligned} \tag{24}$$

Furthermore, from (19) and (13b), we can show that

$$\begin{aligned}\nabla^2 J(K)[Z,Z] &= \mathrm{Tr}\left(\sum_{j=0}^{\infty} \mathcal{A}_K^{\top j} S_1 \mathcal{A}_K{}^j X_0\right)\\ &= \mathrm{Tr}\left(S_1 \sum_{j=0}^{\infty} \mathcal{A}_K{}^j X_0 \mathcal{A}_K^{\top j}\right)\\ &= \mathrm{Tr}(S_1 \Sigma_K)\\ &= \mathrm{Tr}\Big(2(ZC)^\top \Big(B^\top P_K B + R\Big) ZC\Sigma_K\Big)\\ &\quad - \mathrm{Tr}\Big(4(BZC)^\top P_K'[Z]\mathcal{A}_K \Sigma_K\Big). \qquad (25)\end{aligned}$$

■

IV. Cost Function Properties

Building upon the derived explicit formulas for the gradient and Hessian, we are now ready to discuss the optimization landscape for the SOF problem. This section develops some essential properties of the cost function, which will play an important role in the final convergence analysis. The intermediate lemmas required by the property analysis are provided in Appendix A.

Lemma 3 (Coercive Property): The SOF cost (12) is coercive, that is, for all sequence $\{K_i\}_{i=1}^{\infty} \subseteq \mathbb{K}$, we have

$$J(K_i) \to +\infty, \quad \text{if } K_i \to K \in \partial\mathbb{K} \text{ or } \|K_i\| \to +\infty.$$

See Appendix B for detailed proof. Based on the coercivity nature, we can obtain the compactness of the sublevel set.

Lemma 4 (Compactness of Sublevel Set): Given a scalar $\alpha \geq J(K^\star)$ with the globally optimal SOF gain $K^\star$, the sublevel set $\mathbb{K}_\alpha := \{K | J(K) \leq \alpha\} \subseteq \mathbb{K}$ is compact.

Proof: Upon the coercivity proven in Lemma 3, and referring to [33, Proposition 11.12], it becomes evident that the set $\mathbb{K}_\alpha$ is bounded. Given the continuity of $J(K)$ over $\mathbb{K}$, it follows that $\mathbb{K}_\alpha$ is also closed, which completes the proof. ■

With the compactness property in place, it becomes possible to demonstrate that the monotonicity of the objective function guarantees that the line segment between two neighboring iterations remains within $\mathbb{K}_\alpha$.

Lemma 5 (Smoothness on Sublevel Set): For all control gain K in the sublevel set $\mathbb{K}_\alpha$, the norm of the Hessian of the cost function is bounded by a constant, that is, $\|\nabla^2 J(\mathrm{vec}(K))\| \leq L$, where

$$L = \frac{2\alpha}{\sigma_{\min}(Q)}\left(\|R\| + \frac{\alpha}{\mu}\left(1 + \frac{2\zeta_1}{\|B\|\|C\|}\right)\|B\|^2\right)\|C\|^2$$

with

$$\zeta_1 = \frac{1}{\sigma_{\min}(Q)}\left(\frac{\alpha}{\mu}\left(1 + \|B\|^2\|C\|^2\right) + \|R\|\|C\|^2\right) - 1. \quad (26)$$

Proof: From (19), applying the Taylor series expansion about direction Z, we can show that

$$\nabla^2 J(K)[Z,Z] = \mathrm{vec}(Z)^\top \nabla^2 J(\mathrm{vec(K)})\mathrm{vec}(Z). \quad (27)$$

Since $\nabla^2 J(\mathrm{vec}(K))$ is symmetric, one has

$$\begin{aligned}\|\nabla^2 J(\mathrm{vec}(K))\| &= \sup_{\|Z\|_F=1} |\mathrm{vec}(Z)^\top \nabla^2 J(\mathrm{vec(K)})\mathrm{vec}(Z)|\\ &= \sup_{\|Z\|_F=1} |\nabla^2 J(K)[Z,Z]|. \qquad (28)\end{aligned}$$

Based on (20), we further have

$$\begin{aligned}&\|\nabla^2 J(\mathrm{vec}(K))\|\\ &\quad \leq 2 \sup_{\|Z\|_F=1}\left|\mathrm{Tr}\Big(C^\top Z^\top \Big(R + B^\top P_K B\Big) ZC\Sigma_K\Big)\right|\\ &\qquad + 4 \sup_{\|Z\|_F=1}\left|\mathrm{Tr}\Big((BZC)^\top P_K'[Z]\mathcal{A}_K\Sigma_K\Big)\right|\\ &\quad =: 2q_1 + 4q_2. \qquad (29)\end{aligned}$$

Actually, q_1 and q_2 are bounded above by

$$q_1 \leq \frac{J(K)}{\sigma_{\min}(Q)}\left(\|R\| + \frac{J(K)}{\mu}\|B\|^2\right)\|C\|^2 \quad (30a)$$

$$q_2 \leq \frac{\zeta_1 J(K)^2}{\mu\sigma_{\min}(Q)}\|B\|\|C\|. \quad (30b)$$

The detailed derivation of (30) is referred to Appendix C.

Plugging (30) into (29), we finally complete the proof. ■

In light of Lemma 5, consider any scalar $\delta \in [0, 1]$ and any control gains K and K' residing in the sublevel set $\mathbb{K}_\alpha$. If any point along the segment defined by $(1-\delta)K + \delta K'$ remains within this sublevel set, then the cost function has

$$J\big(K'\big) \leq J(K) + \mathrm{Tr}\Big(\nabla J(K)^\top \Big(K' - K\Big)\Big) + \frac{L}{2}\|K - K'\|_F^2. \ (31)$$

Moreover, if the cost function exhibits global L-smoothness, it is widely acknowledged that the gradient descent method can attain a stationary point with a gradient step complexity that is independent of the dimension [34], [35]. However, the L-smooth property (31) and its derived conclusions are not applicable to all control gains $K, K' \in \mathbb{K}_\alpha$ because the domain can be nonconvex or even disconnected [27].

Denote the output correlation matrix as

$$\mathcal{L}_K := C\Sigma_K C^\top = \mathbb{E}_{x_0\sim\mathcal{D}} \sum_{t=0}^{\infty} y_t y_t^\top. \quad (32)$$

Next, we will give the gradient domination condition for the fully observed case. These results are already established in [11]; we provide a short proof in Appendix D for completeness.

Lemma 6 (Gradient Domination): Denote $\mathbb{C} := \{C \in \mathbb{R}^{n\times n} : \mathrm{rank}(C) = n\}$. The globally optimal gain of the SOF problem and the globally optimal performance of the corresponding LQR problem with the state-feedback controller are denoted as $K^\star$ and $J_s^\star$, respectively. Assuming $X_0 \succ 0$ and that the control gain K attains a finite performance, we can express the upper bound of the cost function for K as

$$J(K) - J_s^\star \leq \frac{\|\Sigma_{K^\star}\|\|\nabla J(K)\|_F^2}{4\mu^2\sigma_{\min}(C)^2\sigma_{\min}(R)} \quad \forall C \in \mathbb{C}. \quad (33)$$

Additionally, we have the following lower bound:

$$J(K) - J_s^\star \geq \frac{\mu\mathrm{Tr}\big(E_K^\top E_K\big)}{\|R + B^\top P_K B\|} \quad \forall C \in \mathbb{C}. \quad (34)$$

When $1 \leq \mathrm{rank}(C) < n$, one only has

$$J(K) - J\big(K^\star\big) \leq \|\Sigma_{K^\star}\|\mathrm{Tr}\Big(E_K^\top \Big(R + B^\top P_K B\Big)^{-1} E_K\Big). \quad (35)$$

The concept of gradient dominance is crucial for achieving global convergence in gradient descent algorithms, as it signifies that no stationary points exist aside from the global minimum [34], [36]. Nevertheless, when C is not a full-rank square matrix, this property ceases to be valid (see [27, Example 3.4]), making it challenging to achieve results beyond convergence toward a stationary point. Such limitations on gradient dominance extend to dynamic output-feedback controllers as well, as the set of stabilizing controllers contains at most two disconnected components [37], [38].

Lemma 7 (M-Lipschitz Continuous Hessian): For any control gain K in the sublevel set $\mathbb{K}_\alpha$, define $\gamma := \max_{K\in\mathbb{K}_\alpha} \|\mathcal{A}_K\|$ and denote the upper bound of $\|KC\|$ as ψ, whose explicit form is given Lemma 9. Given any scalar $\delta \in [0, 1]$ and any control gains K and K' in the sublevel set, if any point along the segment between these two gains, represented as $(1-\delta)K+\delta K'$, remains within the sublevel set, then the Hessian of the cost function satisfies

$$\|\nabla^2 J\big(\mathrm{vec}\big(K'\big)\big) - \nabla^2 J(\mathrm{vec}(K))\|_F \le M\|K' - K\|_F \quad (36)$$

where

$$M = \frac{4\alpha^2\sqrt{md}}{\mu\sigma_{\min}(Q)}\left(\Big(\zeta_1 + \frac{\zeta_2}{2}\Big)\|B\|\|C\| + \zeta_3 + \frac{\zeta_4}{2}\right)\|B\|\|C\|$$

ζ_1 is defined in (26), and other intermediate parameters are

$$\zeta_2 = \frac{2\|C\|}{\sigma_{\min}(Q)}\left(\frac{\alpha\gamma}{\mu}\|B\| + \psi\|R\|\right)$$
$$\zeta_3 = \frac{2\|C\|}{\sigma_{\min}(Q)}\left(\frac{\alpha}{\mu}(\zeta_1\gamma + \zeta_2\gamma + \|B\|\|C\|)\|B\| + \|R\|\|C\|\right)$$
$$\zeta_4 = \frac{2\|C\|}{\sigma_{\min}(Q)}\left(\frac{\alpha}{\mu}(\zeta_1\gamma + \|B\|\|C\|)\|B\| + \|R\|\|C\|\right).$$

Proof: Similar to (28), since $\nabla^2 J(\mathrm{vec}(K)) - \nabla^2 J(\mathrm{vec}(K'))$ is symmetric, we have

$$\begin{aligned}&\|\nabla^2 J(\mathrm{vec}(K)) - \nabla^2 J\big(\mathrm{vec}\big(K'\big)\big)\| \\ &\quad= \sup_{\|Z\|_F=1} |\nabla^2 J(K)[Z,Z] - \nabla^2 J\big(K'\big)[Z,Z]|. \end{aligned}\quad (37)$$

By (19), we define

$$g(\delta) := \nabla^2 J\big((1-\delta)K + \delta K'\big)[Z,Z] \quad (38)$$

and denote $\bar{K} := K+\delta(K'-K)$, $\Delta K := K'-K$. Then, from (19), one has

$$g'(\delta) = \mathrm{Tr}\left(\frac{\partial^3}{\partial\lambda^2\partial\delta}\Big|_{\lambda=0} P_{K+\delta(K'-K)+\lambda Z} X_0\right). \quad (39)$$

By the fundamental theorem of calculus, it follows that:

$$\begin{aligned}\|\nabla^2 J(\mathrm{vec}(K)) - \nabla^2 J\big(\mathrm{vec}\big(K'\big)\big)\| &= \sup_{\|Z\|_F=1} |g(0) - g(1)| \\ &= \sup_{\|Z\|_F=1} \left|\int_0^1 g'(\delta)\mathrm{d}\delta\right| \\ &\le \int_0^1 \sup_{\|Z\|_F=1} |g'(\delta)|\mathrm{d}\delta. \end{aligned}\quad (40)$$

Based on (23), we can observe that

$$\frac{\partial^3}{\partial\lambda^2\partial\delta}\Big|_{\lambda=0} P_{K+\delta(K'-K)+\lambda Z} = \sum_{j=0}^{\infty} \mathcal{A}_{\bar{K}}^{\top j} S_2 \mathcal{A}_{\bar{K}}^{j} \quad (41)$$

where

$$\begin{aligned} S_2 := {}& 2C^\top Z^\top B^\top \frac{\partial P_{\bar{K}}}{\partial\delta} BZC + 2C^\top Z^\top B^\top P'_{\bar{K}}[Z]B\Delta KC \\ &+ 2C^\top \Delta K^\top B^\top P'_{\bar{K}}[Z]BZC - 2(BZC)^\top \frac{\partial P'_{\bar{K}}[Z]}{\partial\delta}\mathcal{A}_{\bar{K}} \\ &- 2\mathcal{A}_{\bar{K}}^\top \frac{\partial P'_{\bar{K}}[Z]}{\partial\delta} BZC - (B\Delta KC)^\top P''_{\bar{K}}[Z]\mathcal{A}_{\bar{K}} \\ &- \mathcal{A}_{\bar{K}}^\top P''_{\bar{K}}[Z]B\Delta KC. \end{aligned}\quad (42)$$

According to (39), it follows that:

$$\begin{aligned} g'(\delta) = {}& \mathrm{Tr}\left(2(BZC)^\top \frac{\partial P_{\bar{K}}}{\partial\delta} BZC\Sigma_{\bar{K}}\right) \\ &+ \mathrm{Tr}\Big(4(BZC)^\top P'_{\bar{K}}[Z]B\Delta KC\Sigma_{\bar{K}}\Big) \\ &- \mathrm{Tr}\left(4(BZC)^\top \frac{\partial P'_{\bar{K}}[Z]}{\partial\delta}\mathcal{A}_{\bar{K}}\Sigma_{\bar{K}}\right) \\ &- \mathrm{Tr}\Big(2(B\Delta KC)^\top P''_{\bar{K}}[Z]\mathcal{A}_{\bar{K}}\Sigma_{\bar{K}}\Big). \end{aligned}\quad (43)$$

Similar to the derivation of (30), we can further show that

$$\begin{aligned} \sup_{\|Z\|_F=1} |g'(\delta)| \le {}& 2\|B\|^2\|C\|^2\left\|\frac{\partial P_{\bar{K}}}{\partial\delta}\right\|\mathrm{Tr}\big(\Sigma_{\bar{K}}\big) \\ &+ 4\|B\|^2\|C\|^2\|P'_{\bar{K}}[Z]\|\mathrm{Tr}\big(\Sigma_{\bar{K}}\big)\|\Delta K\| \\ &+ 4\|B\|\|C\|\left\|\frac{\partial P'_{\bar{K}}[Z]}{\partial\delta}\right\|\mathrm{Tr}\big(\Sigma_{\bar{K}}\big) \\ &+ 2\|B\|\|C\|\|P''_{\bar{K}}[Z]\|\mathrm{Tr}\big(\Sigma_{\bar{K}}\big)\|\Delta K\|. \end{aligned}\quad (44)$$

According to Lemma 5, we know that $P'_{\bar{K}}[Z] \preceq \zeta_1 P_{\bar{K}}$. As a matter of fact, we can also show that $(\partial P_{\bar{K}}/\partial\delta) \preceq \zeta_2\|\Delta K\|P_{\bar{K}}$, $(\partial P'_{\bar{K}}[Z]/\partial\delta) \preceq \zeta_3\|\Delta K\|P_{\bar{K}}$, and $P''_{\bar{K}}[Z] \preceq \zeta_4 P_{\bar{K}}$ (see Appendix E for detailed derivations). Utilizing the results of Lemma 8, we can further show that

$$\begin{aligned} &\sup_{\|Z\|_F=1} |g'(\delta)| \\ &\quad\le 2\|B\|\|C\|\big((2\zeta_1 + \zeta_2)\|B\|\|C\| + 2\zeta_3 + \zeta_4\big)\frac{\alpha^2\|\Delta K\|}{\mu\sigma_{\min}(Q)}. \end{aligned}\quad (45)$$

Plugging (45) into (40) and remembering $\|X\| \le \|X\|_F \le \sqrt{\mathrm{rank}(X)}\|X\|$, we finally complete the proof. ■

To the best of our knowledge, the Lipschitz continuity of the Hessian for the SOF cost function has not been previously examined. Nonetheless, this discovery is notable for enhancing the convergence toward a local minimum in nonconvex optimization scenarios, under relatively mild conditions [35]. Moreover, recent studies [39], [40], [41] indicate that the Hessian Lipschitz property facilitates efficient navigation away from strict saddle points in general gradient-based nonconvex optimization problems.

Remark 1: The coercive property, compactness of the sublevel set, and L-smoothness of the cost function in the SOF problem, can be deemed as partially observed counterparts to the properties of the state-feedback LQR cost. The associated proofs follow similar lines as the state-feedback LQR case [12], [19]. Different from these properties, to the best of our knowledge, we are the first to establish the M-Lipschitz continuous Hessian in both SOF and state-feedback LQR problems. Notably, this property cannot be straightforwardly derived using methods akin to those employed for establishing L-smoothness [19]. This is because the analysis of $\nabla^3 J(\mathrm{vec}(K))$ necessitates complicated tensor operations. To circumvent these tensor-related complexities, we directly establish the M-Lipschitz continuous Hessian by adhering to the Lipschitz continuity definition (36).

V. Convergence

This section presents new convergence findings for three variants of policy gradient methods applied to SOF: 1) the vanilla policy gradient; 2) the natural policy gradient; and 3) the Gauss–Newton method. These three methods have been extensively analyzed in related studies [11], [12], [19]. Given that the cost function $J(K)$ is nonconvex, the properties outlined in Section IV play a crucial role in facilitating the convergence analysis for these policy gradient methods.

A. *Vanilla Policy Gradient*

The vanilla policy gradient method iterates as follows:

$$K_{i+1} = K_i - \eta \nabla J(K_i) \tag{46}$$

with the initial gain $K_0 \in \mathbb{K}$ and the step size η. If the line segment $[K_i, K_{i+1}]$ is verified to lie within a sublevel set, then the convergence of the iteration gain obtained by (46) can be directly inferred by leveraging the L-smoothness property specified in Lemma 5 [35]. Before we proceed to the main proofs, let us give the following definition.

Definition 1: Given a differentiable function $J(\cdot)$, if $\|\nabla J(K)\|_F \leq \epsilon$, K is an ϵ-stationary point.

Theorem 1: Assume that $J(K_0) = \alpha$ and $X_0 \succ 0$. If we run the vanilla policy gradient method (46) with any step size $\eta \in (0,\ 1/L]$, $J(K_i)$ is monotonically diminishing (which indicates $K_i \in \mathbb{K}_\alpha \subseteq \mathbb{K}$, that is, K_i is stabilizing), and an ϵ-stationary point will be obtained in

$$\frac{2\alpha}{\eta\epsilon^2} \tag{47}$$

iterations. Additionally, for each iteration i, the line segment $[K_i, K_{i+1}] \subseteq \mathbb{K}_\alpha$. If C is full rank, an ϵ_J-optimal gain K_N, that is, $J(K_N) - J_s^\star \leq \epsilon_J$, is obtained when the iteration step

$$N \geq \frac{\|\Sigma_{K^\star}\|}{2\eta\mu^2\sigma_{\min}(C)^2\sigma_{\min}(R)} \log\left(\frac{J(K_0) - J_s^\star}{\epsilon_J}\right). \tag{48}$$

Proof: We first define an open set $\mathbb{K}_\alpha^o := \{K | J(K) < \alpha\} \subseteq \mathbb{K}$, whose complement $(\mathbb{K}_\alpha^o)^c$ is a closed set. By invoking Lemma 5, for a given $\phi \in (0, 1)$, there is a positive number ς so that $\|\nabla^2 J(\mathrm{vec}(K_i))\| \leq L < L + \phi L$ holds for all $K_i \in \mathbb{K}_\alpha \subset \mathbb{K}_{\alpha+\varsigma}$.

Due to the compactness of $\mathbb{K}_\alpha$ established by Lemma 4, the distance between $\mathbb{K}_\alpha$ and $(\mathbb{K}_{\alpha+\varsigma}^o)^c$, represented by $d = \inf\{\|K_i - K_j\|\ \forall K_i \in \mathbb{K}_\alpha\ \forall K_j \in (\mathbb{K}_{\alpha+\varsigma}^o)^c\}$, is guaranteed to be positive. Now, choose a step size t so that $t \leq \min\{2/(L + \phi L), d/\|\nabla J(K_i)\|\}$. This ensures that the segment $[K_i, K_i - t\nabla J(K_i)] \subseteq \mathbb{K}_{\alpha+\varsigma}$. According to the L-smoothness result (31), one has

$$J(K_i) \geq J(K_i - t\nabla J(K_i)) + t\left(1 - \frac{(L + \phi L)t}{2}\right)\|\nabla J(K_i)\|_F^2. \tag{49}$$

Given the range of step size t, we confirm $J(K_i - t\nabla J(K_i)) \leq J(K_i) < \alpha$, ensuring the iteration point $K_i - t\nabla J(K_i) \in \mathbb{K}_\alpha$ and the segment $[K_i, K_i - t\nabla J(K_i)] \subseteq \mathbb{K}_\alpha$. By applying similar reasoning through (31), we can demonstrate that $[K, K - 2t\nabla J(K)] \in \mathbb{K}_\alpha$ when $2t \leq 2/((1+\phi)L)$. Furthermore, using induction, we generalize this result for $T \in \mathbb{N}^+$ steps, establishing that $[K, K - Tt\nabla J(K)] \in \mathbb{K}_\alpha$ if $Tt \leq 2/((1+\phi)L)$.

Next, we consider a step size $\eta \leq 1/L$. We can then choose a positive $t > 0$ and a positive integer T so that $Tt \in [\eta, 2/(L + \phi L)]$. Then, the segment $[K_i, K_i - \eta\nabla J(K_i)] \subseteq \mathbb{K}_\alpha$. Following a parallel argument to that for (49), we get:

$$J(K_i - \eta\nabla J(K_i)) \leq J(K_i) - \frac{\eta}{2}\|\nabla J(K_i)\|_F^2 \tag{50}$$

where the inequality takes into account that $\eta \leq 1/L$, with the boundary $1/L$ selected to achieve the fastest convergence rate.

Given $J(K_0) = \alpha$, (50) indicates that $K_1 \in \mathbb{K}_\alpha$. Then, for any iteration i, we can use mathematical induction to arrive at

$$J(K_{i+1}) \leq J(K_i) - \frac{\eta}{2}\|\nabla J(K_i)\|_F^2. \tag{51}$$

Also, the line segment $[K_i, K_{i+1}] \subseteq \mathbb{K}_\alpha$. Summing up the above inequality yields

$$\frac{\eta}{2}\sum_{i=0}^{N}\|\nabla J(K_i)\|_F^2 \leq J(K_0) - J(K_{N+1}) \leq J(K_0) - J\big(K^\star\big). \tag{52}$$

It then follows that $\lim_{i\to\infty}\|\nabla J(K_i)\|_F^2 = 0$ and:

$$\min_{0\leq i\leq N}\|\nabla J(K_i)\|_F^2 \leq \frac{2(J(K_0) - J(K^\star))}{\eta N} \leq \frac{2\alpha}{\eta N} \tag{53}$$

which shows the vanilla policy gradient can reach an ϵ-stationary point within $(2\alpha/\eta\epsilon^2)$ iterations.

Furthermore, when $C \in \mathbb{C}$, combining (51) and (33) yields

$$J(K_{i+1}) - J(K_i) \leq -\frac{2\eta\mu^2\sigma_{\min}(C)^2\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|}\big(J(K_i) - J_s^\star\big). \tag{54}$$

This subsequently results in

$$J(K_i) - J_s^\star \leq \left(1 - \frac{2\eta\mu^2\sigma_{\min}(C)^2\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|}\right)^i\big(J(K_0) - J_s^\star\big). \tag{55}$$

This proves the second claim of this theorem. ■

Theorem 1 establishes that, starting with an initial stabilizing control gain, the vanilla policy gradient method for the SOF problem ensures both the recursive stability of the control policy and a monotonically decreasing cost function. Moreover, the convergence rate to a stationary point is dimension-independent. To offer a unified view that encompasses both SOF and state-feedback LQR, our findings also reveal that the vanilla policy gradient method globally converges to a unique minimum at a linear rate when the state is fully observed. In this context, the convergence rate outlined in (48) aligns with the conclusions in [11, Th. 7]. Notably, in contrast to [11, Th. 7], we provide an explicit upper bound of the step size η such that (51) is satisfied.

Although the convergence to stationary points of the vanilla policy gradient for SOF has been established, it is important to note that these stationary points can be local minima, saddle points, or even local maxima. Next, we will proceed to demonstrate that under mild assumptions, the vanilla method can indeed converge to a local minimum.

Theorem 2: Suppose all the conditions in Lemmas 5 and 7 hold. Assume that $\mathbb{K}_\beta \subset \mathbb{K}_\alpha$, where $\beta < \alpha$. So, the distance $d = \inf\{\|K_i - K_j\| \forall K_i \in \mathbb{K}_\beta \forall K_j \in (\mathbb{K}_\alpha^o)^c\}$ between two compact sets is positive. For the sublevel set $\mathbb{K}_\beta$, assume that there is a local minimum $K^\#$ with $l = \lambda_{\min}(\nabla^2 J(\mathrm{vec}(K^\#)) > 0$. Given that the initial gain K_0 is sufficiently close to this local minimum $K^\#$, denoted by an initial error $r_0 = \|K_0 - K^\#\|_F < \bar{r} = 2l/M$, and fulfilling the condition $\bar{r}r_0/(\bar{r} - r_0) \le d$, the vanilla policy gradient using a constant step size $\eta \le 1/L$ has an upper error bound

$$\|K_i - K^\#\|_F \le \frac{\bar{r}r_0}{\bar{r} - r_0}\left(\frac{1}{1+\eta l}\right)^i. \tag{56}$$

Proof: Denote the set of gains around the minimum $K^\#$ as $\mathbb{K}^\# := \{K \in \mathbb{R}^{m\times d} : \|K - K^\#\|_F \le \bar{r}r_0/(\bar{r} - r_0) \le d\}$. Then, we have $\mathbb{K}^\# \subset \mathbb{K}_\alpha$. For any scalar $\delta \in [0, 1]$ and any control gains $K, K' \in \mathbb{K}^\#$, it follows that $(1-\delta)K + \delta K' \in \mathbb{K}^\# \subset \mathbb{K}_\alpha$. Therefore, the conclusions of Lemmas 5 and 7 can be applied directly. Finally, the upper error bound of iterative gain can be immediately derived from [35, Th. 1.2.4]. ■

When initialized near local minima, Theorem 2 assures that vanilla policy gradient will exhibit linear convergence concerning the control gain. Although the aforementioned theoretical analysis relies on full awareness of model parameters and cost function details, it is worth noting the applicability of this analysis in model-free environments. In such settings, data-driven approaches like zeroth-order optimization techniques can be employed to offer an unbiased estimation of $\nabla J(K)$ [11], [42], [43]. Hence, our findings suggest that data-driven methods can also effectively handle discrete-time SOF problems, provided the gradient is approximated with reasonable precision.

B. Natural Policy Gradient

Besides the vanilla policy gradient method, the natural policy gradient method is also widely used in RL research [11], [19], [30]. The natural gradient method iterates as follows:

$$K' = K - \eta \nabla^{\mathrm{NA}} J(K) \tag{57}$$

where

$$\nabla^{\mathrm{NA}} J(K) = \nabla J(K)\mathcal{L}_K^{-1}$$

is the natural policy gradient. More explanations for this update rule can be found in [11].

Theorem 3: Suppose $J(K_0) = \alpha$ and $X_0 \succ 0$. The cost $J(K_i)$ of natural gradient descent (57) is monotonically diminishing (which indicates $K_i \in \mathbb{K}_\alpha \subseteq \mathbb{K}$, that is, K_i is stabilizing), and an ϵ-stationary point, that is, $\|\nabla^{\mathrm{NA}} J(K_i)\|_F \le \epsilon$, can be reached in

$$\frac{2\alpha}{\eta\mu\sigma_{\min}(C)^2\epsilon^2} \tag{58}$$

iterations, where the step size $\eta \le \mu\sigma_{\min}(C)^2/L$. If C is full rank, an ϵ_J-optimal control gain K_N, satisfying $J(K_N) - J_s^\star \le \epsilon_J$, is achieved when the iteration step

$$N \ge \frac{\|\Sigma_{K^\star}\|}{2\eta\mu\sigma_{\min}(R)}\log\left(\frac{J(K_0) - J_s^\star}{\epsilon_J}\right). \tag{59}$$

The proof of Theorem 3 is provided in Appendix F, which is similar to that of Theorem 1. Theorem 3 illustrates that the natural policy gradient technique also converges to a stationary point in SOF problems at a nearly dimension-free rate. The term "nearly dimension-free rate" suggests that the convergence does not explicitly depend on the system dimension. Besides, the explicit form of the convergence rate (59) for the fully observed case ($C \in \mathbb{C}$) is also provided for completeness, which is consistent with the result given in [11, Th. 7]. Similar to the vanilla policy gradient method, the natural policy gradient method can also be implemented in a model-free manner. Since $\mathcal{L}_K = \mathbb{E}_{x_0\sim\mathcal{D}}\sum_{t=0}^{\infty} y_t y_t^\top$, one can just estimate $\nabla^{\mathrm{NA}} J(K)$ from cost and output information trajectories. The numerical evidence given in existing studies [11], [19] shows that the natural policy gradient method usually leads to a faster convergence speed than the vanilla policy gradient method.

C. Gauss–Newton Policy Gradient

Next, we consider the Gauss–Newton policy gradient method, which iterates as follows:

$$K' = K - \eta \nabla^{GN} J(K) \tag{60}$$

where

$$\nabla^{GN} J(K) = \left(R + B^\top P_K B\right)^{-1} \nabla J(K)\mathcal{L}_K^{-1}$$

is the Gauss–Newton policy gradient. More explanations for this update rule can be found in [11].

Theorem 4: Suppose $J(K_0) = \alpha$ and $X_0 \succ 0$. If we run Gauss–Newton natural gradient descent (60) with any step size $\eta \le \mu\sigma_{\min}(R)\sigma_{\min}(C)^2/L$, $J(K_i)$ is monotonically diminishing (which indicates $K_i \in \mathbb{K}_\alpha \subseteq \mathbb{K}$, that is, K_i is stabilizing), and an ϵ-stationary point, that is, $\|\nabla^{\mathrm{GN}} J(K_i)\|_F \le \epsilon$, will be reached in

$$\frac{2\alpha}{\eta\mu\sigma_{\min}(R)\sigma_{\min}(C)^2\epsilon^2} \tag{61}$$

iterations. If $C \in \mathbb{C}$, an ϵ_J-optimal control gain K_N, satisfying $J(K_N) - J_s^\star \le \epsilon_J$, is achieved when the iteration step

$$N \ge \frac{\|\Sigma_{K^\star}\|}{2\eta\mu}\log\left(\frac{J(K_0) - J_s^\star}{\epsilon_J}\right). \tag{62}$$

See Appendix G for details on deriving Theorem 4. Theorem 4 establishes the result of nearly dimension-free convergence to stationary points of the Gauss–Newton method. The explicit form of the convergence rate (62) for the fully observed case ($C \in \mathbb{C}$) is consistent with the result given in [11, Th. 7]. Different from the vanilla policy gradient and the natural policy gradient methods, the Gauss–Newton method is not suitable for model-free settings since it requires the knowledge of matrices B and P_K.

Remark 2: The Gauss–Newton method and the natural policy gradient method generally converge faster than the vanilla policy gradient method in terms of iteration number [11], [19]. As a tradeoff, these two methods need more information to calculate the update gradients, taking up more computational resources. Notably, the Gauss–Newton method is ill-suited for model-free settings, as its gradient estimation relies on matrices B and P_K.

D. Impact of Initial Distribution

Up to this point, we have demonstrated that all three policy gradient methods are capable of converging to stationary points at a nearly dimension-free rate. When implementing these policy gradient algorithms in practice, it is crucial to recognize that these stationary points are not fixed; they are influenced by the initial state distribution.

Proposition 1: Let $K^\ddagger$ represent the stationary point of the SOF problem. When C lacks full rank and $K^\ddagger C \neq K_s^\star$, $K^\ddagger$ is influenced by the initial state distribution. Here, $K_s^\star$ represents the optimal solution for state feedback LQR.

Proof: For the case where $C \in \mathbb{C}$, Lemmas 1 and (33) in Lemma 6 establish that

$$\|E_{K_s^\star}\|_F = 0. \tag{63}$$

From the definition of the SOF controller (3), one has

$$u_t = -KCx_t$$

where KC effectively serves as a state-feedback gain.

Lemma 6 asserts that $K_s^\star$ is unique, which means

$$\|E_K\|_F = 0 \iff KC = K_s^\star.$$

However, when C lacks full rank, it is possible that no gain K will satisfy $KC = K_s^\star$.

If $K \in \mathbb{K}$ and $X_0 \succ 0$, theory [32, Lemma 2] suggests that the Lyapunov equation (13b) admits a unique positive-definite solution Σ_K. For different initial distributions $\mathcal{D}$ and $\mathcal{D}'$, (13b) indicates that

$$\Sigma_{K^\ddagger} \neq \Sigma'_{K^\ddagger} \text{ if } X_0 \neq X_0' = \mathbb{E}_{x_0\sim\mathcal{D}'}x_0x_0^\top.$$

According to (15), a stationary point $K^\ddagger$ meets the condition

$$\|\nabla J\left(K^\ddagger\right)\|_F = 2\|E_{K^\ddagger}\Sigma_{K^\ddagger}C^\top\|_F = 0.$$

However, if $\|E_{K^\ddagger}\|_F \neq 0$ (that is, $K^\ddagger C \neq K_s^\star$), we cannot guarantee that $\|\nabla J(K^\ddagger)\|_F = \|E_{K^\dagger}\Sigma'_{K^\ddagger}C^\top\|_F$ will be zero for all possible distribution $\mathcal{D}'$, due to its influence on $\Sigma'_{K^\ddagger}$.

Nevertheless, when $C \notin \mathbb{C}$, the stationary point $K^\ddagger$ in SOF is influenced by the initial state distribution $\mathcal{D}$. In other words, different initial distributions could yield distinct stationary points unless $K^\ddagger C = K_s^\star$. ■

The foregoing theoretical discussion suggests that to achieve an effective SOF policy, the initial state distribution should be carefully selected to match the practical application conditions.

VI. Numerical Results

In this section, we will present some numerical simulations to verify the performance of the above gradient descent methods in optimizing SOF problems. Since the vanilla policy gradient method and the natural policy gradient method can be implemented in a model-free manner, their model-free versions are also developed and tested.

Algorithm 1 Model-Free Vanilla and Natural Policy Gradient

```
Input: stabilizing policy gain $K_0$, number of trajectories $z$, roll-out length $l$, perturbation amplitude $r$, step size $\eta$
repeat
  Gradient Estimation:
  for $i = 1, \ldots, z$ do
    Sample $x_0$ from $\mathcal{D}$
    Simulate $K_j$ for $l$ steps starting from $x_0$ and observe $y_0, \cdots, y_{l-1}$ and $c_0, \cdots, c_{l-1}$.
    Draw $U_i$ uniformly at random over matrices such that $\|U_i\|_F = 1$, and generate a policy $K_{j,U_i} = K_j + rU_i$.
    Simulate $K_{j,U_i}$ for $l$ steps starting from $x_0$ and observe $c'_0, \cdots, c'_{l-1}$.
    Calculate empirical estimates:
      $\widehat{J^i_{K_j}} = \sum_{t=0}^{l-1} c_t,\ \widehat{\mathcal{L}^i_{K_j}} = \sum_{t=0}^{l-1} y_t y_t^\top,\ \widehat{J_{K_{j,U_i}}} = \sum_{t=0}^{l-1} c'_t.$
  end for
  Return estimates:
    $\widehat{\nabla J(K_j)} = \frac{1}{z}\sum_{i=1}^{z} \frac{\widehat{J_{K_{j,U_i}}} - \widehat{J^i_{K_j}}}{r} U_i,\ \widehat{\mathcal{L}_{K_j}} = \frac{1}{z}\sum_{i=1}^{z} \widehat{\mathcal{L}^i_{K_j}}.$
  Policy Update:
  Vanilla policy gradient $K_{j+1} = K_j - \eta\widehat{\nabla J(K_j)}$.
  Natural policy gradient $K_{j+1} = K_j - \eta\widehat{\nabla J(K_j)}\widehat{\mathcal{L}_{K_j}}^{-1}$.
  $j = j + 1$.
until $\|\widehat{\nabla J(K_{j-1})}\|_F \leq \epsilon$
```

A. Model-Free Optimization

In the model-free setting, the model parameters, A, B, C, Q, R, are unknown. In keeping with other work in [11], we assume the algorithm has access to the observation y_t and running cost c_t at each time step, where $c_t := x_t^\top Q x_t + u_t^\top R u_t$. Using the zeroth-order optimization approach [11], [42], [43], Algorithm 1 provides a data-driven procedure to estimate the gradients of both vanilla and natural policy gradient methods.

B. Example I: Open-Loop Unstable Linear System

Consider an internally unstable linear system

$$A = \begin{bmatrix} 1.1 & 0.1 \\ 0 & 1.1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0.1 \end{bmatrix}, C = \begin{bmatrix} 1.0 & 1.0 \end{bmatrix} \tag{64}$$

which is a discrete version of the famous Doyle's LQG example. Let $Q = 0.25I_2$, $R = 0.2$, and $X_0 = 0.1I_2$. We employ all three policy gradient methods in model-based settings and Algorithm 1 in model-free settings to learn a suboptimal SOF policy. The initial controller is set as $K_0 = 9$. The optimal gain $K^\star = 4.0637$ can be found by solving several Lyapunov equations given in [44, Th. 1]. The step size of all methods is set as $\eta = 0.2$. Besides, other hyperparameters of Algorithm 1 are set as: $r = 0.001$, $z = 2^{14}$, and $l = 100$.[1]

The relative errors of both the control gain and the cost function are presented in Fig. 1, which are computed as $\|K - K^\star\|_F/\|K^\star\|_F$ and $|J(K) - J(K^\star)|/|J(K^\star)|$, respectively. We can easily observe that all model-based policy gradient

[1] Our code is available athttps://github.com/jieli18/sof.

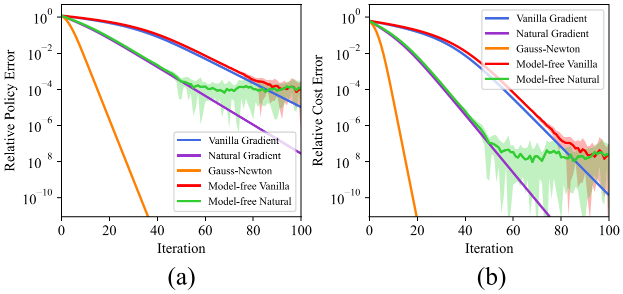

Fig. 1. Learning curves of different methods for Example I. The solid lines correspond to the mean and the shaded regions correspond to an interval between maximum and minimum values over ten runs. (a) Policy error. (b) Cost error.

methods converge to the optimal solution within 100 iterations. As expected, the two model-free methods, especially the model-free natural policy gradient method, converge more slowly and unsteadily than their model-based counterparts due to gradient estimation errors. These results provide numerical evidence for our theoretical convergence analysis.

For the internally unstable system, such as (64), the stability of the controller can be assessed by evaluating the spectral radius of the closed-loop system matrix. This process, however, requires the knowledge of model dynamics. As a result, finding stabilizing controllers can be relatively complex when using model-free methods. In such instances, the trial and error approach could provide a practical strategy for obtaining an initial stabilizing controller. In terms of applying a controller to the dynamic system, the convergence or divergence of the observation output provides a useful criterion for determining the stability of the closed-loop system. Through the application of this manner, we are able to establish the set of stabilizing controllers for the internally unstable system (64), which is $\mathbb{K} = \{K{:}K \in (2.1, 22.05)\}$.

We run all three policy gradient methods with 10 randomly generated initial stabilizing controllers. The relative errors of control gains are shown in Fig. 2, where the curves of the same color start from the same initial point. It can be seen that all methods converge within 100 iterations under different initial controllers. This further confirms our theoretical convergence results within the context of an internally unstable system with randomly chosen initial controllers.

C. Example II: Four-Dimensional Open-Loop Stable System

Consider a circuit system given in [45] with

$$A = \begin{bmatrix} 0.90031 & -0.00015 & 0.09048 & -0.00452 \\ -0.00015 & 0.90031 & 0.00452 & -0.09048 \\ -0.09048 & -0.00452 & 0.90483 & -0.09033 \\ 0.00452 & 0.09048 & -0.09033 & 0.90483 \end{bmatrix}$$

$$B = \begin{bmatrix} 0.00468 & -0.00015 \\ 0.00015 & -0.00468 \\ 0.09516 & -0.00467 \\ -0.00467 & 0.09516 \end{bmatrix}, C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

where $Q = \mathrm{diag}([0.1, 0.2, 0, 0])$, $R = \mathrm{diag}([10^{-6}, 10^{-4}])$, and $X_0 = I_4$. According to [44, Th. 1], the optimal gain is

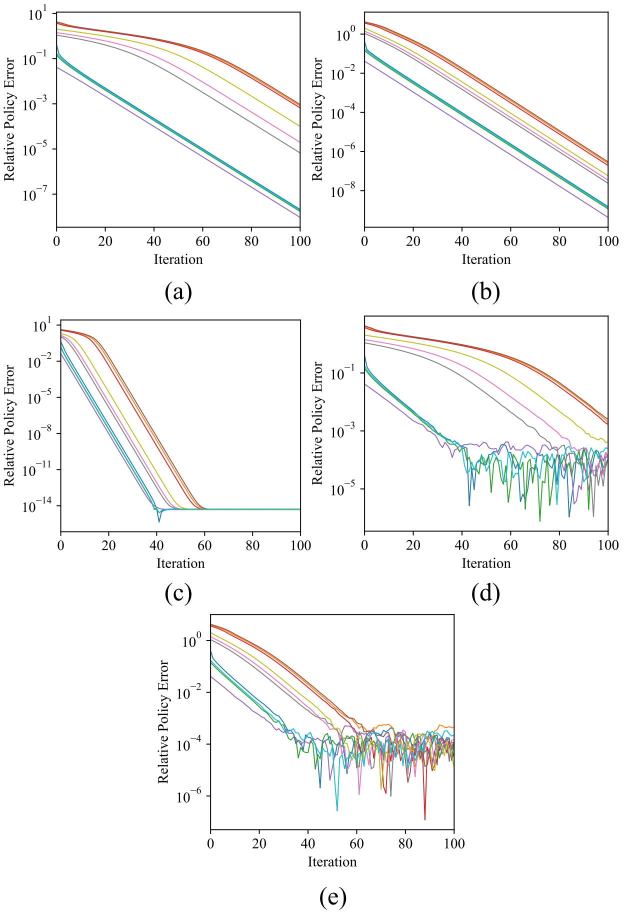

Fig. 2. Learning curves of different methods with ten different random initializations (corresponding to curves with different colors). (a) Vanilla gradient. (b) Natural gradient. (c) Gauss–Newton. (d) Model-free vanilla. (e) Model-free natural.

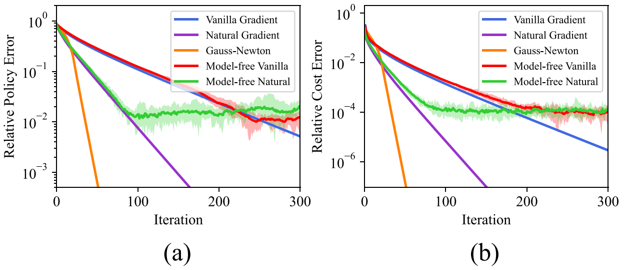

Fig. 3. Learning curves of different methods for Example II. The solid lines correspond to the mean and the shaded regions correspond to the interval between maximum and minimum values over ten runs. (a) Policy error. (b) Cost error.

$$K^* = \begin{bmatrix} 2.9738 & -7.2907 \\ 2.1067 & -12.5384 \end{bmatrix}.$$

We set $K_0 = \begin{bmatrix} 0 & -1 \\ 0 & -2 \end{bmatrix}$ for all methods and adopt the same hyperparameters as outlined in Section VI-B. The relative errors in control gain and cost function for various methods are shown in Fig. 3. The observed trend of this example is quite similar to the example given in Section VI-B. Overall, these numerical findings corroborate our theoretical analysis on convergence.

VII. Conclusion

In this work, we have investigated the optimization landscape of three distinct policy gradient algorithms for SOF problems. Initially, we demonstrated various crucial properties of the SOF cost function, including coercivity, L-smoothness, and M-Lipschitz continuity of its Hessian. Utilizing these foundational properties, we unearthed new understandings about the convergence behaviors and rates at which all three policy gradient algorithms arrive at stationary points. These stationary points are generally influenced by the initial state distribution. Moreover, provided that the initial gain is around a local minimum, we demonstrated that the vanilla policy gradient exhibits linear convergence toward that minimum. Our numerical experiments suggest that both the vanilla policy gradient method and the natural policy gradient method can be implemented in a model-free manner, as long as the gradient estimations are sufficiently accurate. Additionally, recent literature highlights the potential of model-free SOF in H_∞ control [46]. Our next steps will involve expanding the convergence analysis in this specialized domain.

Appendix A
Intermediate Lemmas

Lemma 8: The upper bound of $\|P_K\|$ and $\|\Sigma_K\|$ are given by

$$\|P_K\| \le \frac{J(K)}{\mu}, \quad \|\Sigma_K\| \le \mathrm{Tr}(\Sigma_K) \le \frac{J(K)}{\sigma_{\min}(Q)}.$$

Proof: From (12), one has

$$J(K) = \mathrm{Tr}(P_K X_0) \ge \mu\|P_K\|.$$

Then, the first claim can be directly obtained. Similarly, $J(K)$ can also be lower bounded by

$$\begin{aligned} J(K) &= \mathrm{Tr}\Big(\Big(Q + C^\top K^\top RKC\Big)\Sigma_K\Big) \ge \sigma_{\min}(Q)\mathrm{Tr}(\Sigma_K) \\ &\ge \sigma_{\min}(Q)\|\Sigma_K\| \end{aligned}$$

which leads to the second claim. ■

Lemma 9: For any $K \in \mathbb{K}_\alpha$, it holds that

$$\|KC\| \le \psi := \frac{\sqrt{\|R\|\alpha + \|B\|^2\alpha^2/\mu}}{\sqrt{\mu}\sigma_{\min}(R)} + \frac{\|B\|\|A\|\alpha}{\mu\sigma_{\min}(R)}.$$

Proof: First, we can observe that

$$\begin{aligned} \|KC\| &= \left\|\Big(R + B^\top P_K B\Big)^{-1}\Big(R + B^\top P_K B\Big)KC\right\| \\ &\le \left\|\Big(R + B^\top P_K B\Big)^{-1}\right\|\left\|\Big(R + B^\top P_K B\Big)KC\right\| \\ &\le \frac{\left\|\big(R + B^\top P_K B\big)KC - B^\top P_K A + B^\top P_K A\right\|}{\sigma_{\min}(R)} \\ &\le \frac{\|E_K\| + \|B^\top P_K A\|}{\sigma_{\min}(R)} \\ &\le \frac{\sqrt{\mathrm{Tr}\big(E_K^\top E_K\big)} + \|B^\top P_K A\|}{\sigma_{\min}(R)}. \end{aligned}$$

From (34), we know that

$$\mathrm{Tr}\Big(E_K^\top E_K\Big) \le \frac{\|R + B^\top P_K B\|}{\mu} J(K).$$

Thereby, we finally have

$$\begin{aligned} \|KC\| &\le \frac{\sqrt{\|R + B^\top P_K B\| J(K)}}{\sqrt{\mu}\sigma_{\min}(R)} + \frac{\|B^\top P_K A\|}{\sigma_{\min}(R)} \\ &\le \frac{\sqrt{\|R\|\alpha + \|B\|^2\alpha^2/\mu}}{\sqrt{\mu}\sigma_{\min}(R)} + \frac{\|B\|\|A\|\alpha}{\mu\sigma_{\min}(R)} \end{aligned}$$

where the last step follows from Lemma 8. ■

Appendix B
Proof of Lemma 3

Proof: From (12), we can show that

$$\begin{aligned} J(K_i) &= \mathrm{Tr}\Big(\Big(Q + C^\top K_i^\top RK_iC\Big)\Sigma_{K_i}\Big) \\ &\ge \mu\sigma_{\min}(R)\sigma_{\min}(C)^2\|K_i\|^2 \end{aligned}$$

which directly leads to that $J(K_i) \to +\infty$ as $\|K_i\| \to +\infty$.

By (14), we also have

$$\begin{aligned} J(K_i) &= \mathrm{Tr}\left(\sum_{j=0}^{\infty} \mathcal{A}_{K_i}^{\top j}\Big(Q + C^\top K_i^\top RK_iC\Big)\mathcal{A}_{K_i}{}^j X_0\right) \\ &\ge \mu\sigma_{\min}(Q)\sum_{j=0}^{\infty}\|\mathcal{A}_{K_i}{}^j\|_F^2 \ge \mu\sigma_{\min}(Q)\sum_{j=0}^{\infty}\rho\big(\mathcal{A}_{K_i}\big)^{2j} \\ &= \mu\sigma_{\min}(Q)\frac{1 - \rho\big(\mathcal{A}_{K_i}\big)^\infty}{1 - \rho\big(\mathcal{A}_{K_i}\big)^2}. \end{aligned}$$

Since $\rho(\mathcal{A}_K) = 1$ when $K \in \partial\mathbb{K}$, by continuity of the $\rho(\mathcal{A}_{K_i})$, we have $\rho(\mathcal{A}_{K_i}) \to 1$ as $K_i \to K \in \partial\mathbb{K}$. Therefore, for every $\epsilon > 0$, there exists some $N(\epsilon) \in \mathbb{N}$ such that $1 - \rho(\mathcal{A}_{K_i}) < \epsilon$ for all $i \ge N(\epsilon)$. That is, $1 > \rho(\mathcal{A}_{K_i}) > 1 - \epsilon$ for $i \ge N(\epsilon)$. Hence, $J(K_i)$ is bounded below by

$$J(K_i) \ge \mu\sigma_{\min}(Q)\frac{1}{1 - (1 - \epsilon)^2}.$$

It thus follows that $J(K_i) \to +\infty$ as $K_i \to \partial\mathbb{K}$. This completes the proof of Lemma 3. ■

Appendix C
Derivations of Bounds q_1 and q_2 in Lemma 5

First, for the bound q_1, we can easily observe that

$$\begin{aligned} q_1 &\le \sup_{\|Z\|_F=1}\Big(\|C^\top Z^\top\Big(R + B^\top P_K B\Big)ZC\|\mathrm{Tr}(\Sigma_K)\Big) \\ &\le \sup_{\|Z\|_F=1}\Big(\|C\|^2\|Z\|_F^2\Big(\|R\| + \|B\|^2\|P_K\|\Big)\mathrm{Tr}(\Sigma_K)\Big) \\ &\le \|C\|^2\left(\|R\| + \|B\|^2\frac{J(K)}{\mu}\right)\frac{J(K)}{\sigma_{\min}(Q)} \end{aligned} \tag{65}$$

where the last step follows from Lemma 8.

Next, we focus on the upper bound of q_2. Using the Cauchy–Schwarz inequality, we can show that

$$\begin{aligned} q_2 &\le \sup_{\|Z\|_F=1}\left(\|(BZC)^\top P_K'[Z]\mathcal{A}_K\Sigma_K^{1/2}\|_F\left\|\Sigma_K^{1/2}\right\|_F\right) \\ &\le \sup_{\|Z\|_F=1}\left(\|C\|\|Z\|\|B\|\|P_K'[Z]\|\|\mathcal{A}_K\Sigma_K^{1/2}\|_F\sqrt{\mathrm{Tr}(\Sigma_K)}\right) \\ &\le \|C\|\|B\|\sup_{\|Z\|_F=1}\Big(\left\|P_K'[Z]\right\|\Big)\sqrt{\mathrm{Tr}\Big(\mathcal{A}_K\Sigma_K\mathcal{A}_K^\top\Big)}\sqrt{\mathrm{Tr}(\Sigma_K)}. \end{aligned}$$

By (13b), it is not hard to see that $\Sigma_K \succ \mathcal{A}_K \Sigma_K \mathcal{A}_K^\top$. Therefore, we further have

$$
\begin{aligned}
q_2 &\le \|C\|\|B\|\mathrm{Tr}(\Sigma_K) \sup_{\|Z\|_F=1} \|P_K'[Z]\| \\
&\le \|C\|\|B\| \frac{J(K)}{\sigma_{\min}(Q)} \sup_{\|Z\|_F=1} \|P_K'[Z]\| \qquad (66)
\end{aligned}
$$

where the last step follows from Lemma 8. Then, the only thing left is to show the following bound holds:

$$
\sup_{\|Z\|_F=1} \|P_K'[Z]\| \le \zeta_1 \|P_K\|
$$

where ζ_1 is as given by (26).

We will prove the above inequality by showing that $P_K'[Z] \preceq \zeta_1 P_K$. Based on (14) and (21), $(C^\top Z^\top E_K + E_K^\top Z C) \preceq \zeta_1(Q + C^\top K^\top R K C)$ will directly lead to $P_K'[Z] \preceq \zeta_1 P_K$ for a given $\zeta_1 \in \mathbb{R}^+$. Now the remaining task is to find such ζ_1. From (13a), we have

$$
\begin{aligned}
&C^\top Z^\top E_K + E_K^\top Z C \\
&\quad = C^\top Z^\top R K C + C^\top K^\top R Z C \\
&\qquad - C^\top Z^\top B^\top P_K \mathcal{A}_K - \mathcal{A}_K^\top P_K B Z C \\
&\quad \prec eq C^\top Z^\top R Z C + C^\top K^\top R K C \\
&\qquad + \mathcal{A}_K^\top P_K \mathcal{A}_K + (BZC)^\top P_K BZC \\
&\quad = P_K - Q + C^\top Z^\top R Z C + (BZC)^\top P_K BZC \\
&\quad \prec eq \|P_K + C^\top Z^\top R Z C + (BZC)^\top P_K BZC\| I - Q \\
&\quad \prec eq \frac{Q}{\sigma_{\min}(Q)} \left(\frac{\alpha}{\mu} \left(1 + \|B\|^2 \|C\|^2\right) + \|R\|\|C\|^2 \right) - Q. \quad (67)
\end{aligned}
$$

Therefore, we prove that $P_K'[Z] \preceq \zeta_1 P_K$. According to (66) and Lemma 8, this directly leads to (30b).

Appendix D
Proof of Lemma 6

The performance difference lemma, also referred to as almost smoothness, serves as the foundational element for establishing the gradient domination condition.

Lemma 10 (Performance Difference Lemma): Let $K, K' \in \mathbb{K}$. Then, the following relationship exists:

$$
\begin{aligned}
J(K') - J(K) = {}& 2\mathrm{Tr}\left(\Sigma_{K'} (K'C - KC)^\top E_K\right) + \\
&+ \mathrm{Tr}\left(\Sigma_{K'} (K'C - KC)^\top \left(R + B^\top P_K B\right)(K'C - KC)\right).
\end{aligned}
$$

Proof: Consider state and action sequences x_t' and u_t' generated by K', and let $c_t' = x_t'^\top Q x_t' + u_t'^\top R u_t'$. Then, one has

$$
\begin{aligned}
J(K') - J(K) &= \mathbb{E}_{x_0 \sim \mathcal{D}} \left[\sum_{t=0}^{\infty} c_t' - V_K(x_0) \right] \\
&= \mathbb{E}_{x_0 \sim \mathcal{D}} \left[\sum_{t=0}^{\infty} \left(c_t' + V_K(x_t') - V_K(x_t')\right) - V_K(x_0) \right] \\
&= \mathbb{E}_{x_0 \sim \mathcal{D}} \left[\sum_{t=0}^{\infty} \left(c_t' + V_K(x_{t+1}') - V_K(x_t')\right) \right]
\end{aligned}
$$

where the last step takes advantage of the fact that $x_0 = x_0'$.

Let $A_K(x_t, K') = c_t + V_K(x_{t+1}) - V_K(x_t)|_{u_t = -K'Cx_t}$, which can be expanded as

$$
\begin{aligned}
A_K(x_t, K') &= x_t^\top \left(Q + C^\top K'^\top R K' C\right) x_t \\
&\quad + x_t^\top \mathcal{A}_{K'}^\top P_K \mathcal{A}_{K'} x_t - V_K(x_t) \\
&= x_t^\top \left(Q + \left(K'C - KC + KC\right)^\top R \left(K'C - KC + KC\right)\right) x_t \\
&\quad + x_t^\top \left(A - B\left(K'C - KC + KC\right)\right)^\top \\
&\qquad \times P_K \left(A - B\left(K'C - KC + KC\right)\right) x_t - V_K(x_t) \\
&= 2x_t^\top \left(K'C - KC\right)^\top \left(\left(R + B^\top P_K B\right) KC - B^\top P_K A\right) x_t \\
&\quad + x_t^\top \left(K'C - KC\right)^\top \left(R + B^\top P_K B\right) (K'C - KC) x_t \\
&= 2x_t^\top \left(K'C - KC\right)^\top E_K x_t \\
&\quad + x_t^\top \left(K'C - KC\right)^\top \left(R + B^\top P_K B\right) (K'C - KC) x_t.
\end{aligned}
$$

Then, we get that

$$
\begin{aligned}
&J(K') - J(K) \\
&\quad = \mathbb{E}_{x_0 \sim \mathcal{D}} \left[\sum_{t=0}^{\infty} A_K(x_t', K') \right] \\
&\quad = \mathbb{E}_{x_0 \sim \mathcal{D}} \Bigg[\sum_{t=0}^{\infty} \Big(2\mathrm{Tr}\left(x_t' x_t'^\top (K'C - KC)^\top E_K\right) \\
&\qquad + \mathrm{Tr}\left(x_t' x_t'^\top (K'C - KC)^\top \left(R + B^\top P_K B\right)(K'C - KC)\right)\Big)\Bigg] \\
&\quad = 2\mathrm{Tr}\left(\Sigma_{K'} (K'C - KC)^\top E_K\right) \\
&\qquad + \mathrm{Tr}\left(\Sigma_{K'} (K'C - KC)^\top \left(R + B^\top P_K B\right)(K'C - KC)\right).
\end{aligned}
$$

■

Next, we show the main proof of Lemma 6.

Proof: Let $X = (R + B^\top P_K B)^{-1} E_K \Sigma_{K'} C^\top \mathcal{L}_{K'}^{-1}$. From Lemma 10, we find that

$$
\begin{aligned}
&J(K') - J(K) \\
&\quad = 2\mathrm{Tr}\left(\Sigma_{K'} (K'C - KC)^\top E_K\right) \\
&\qquad + \mathrm{Tr}\left(\Sigma_{K'} (K'C - KC)^\top \left(R + B^\top P_K B\right)(K'C - KC)\right) \\
&\quad = \mathrm{Tr}\left(\Sigma_{K'} C^\top (\Delta K + X)^\top \left(R + B^\top P_K B\right)(\Delta K + X) C\right) \\
&\qquad - \mathrm{Tr}\Big(\Sigma_{K'} C^\top \mathcal{L}_{K'}^{-1} C \Sigma_{K'} E_K^\top \\
&\qquad\qquad \left(R + B^\top P_K B\right)^{-1} E_K \Sigma_{K'} C^\top \mathcal{L}_{K'}^{-1} C\Big) \\
&\quad \ge -\mathrm{Tr}\left(\mathcal{L}_{K'}^{-1} C \Sigma_{K'} E_K^\top \left(R + B^\top P_K B\right)^{-1} E_K \Sigma_{K'} C^\top\right) \qquad (68)
\end{aligned}
$$

where $\Delta K = K' - K$ and the equality holds when $K' = K - X$.

Then, one has

$$
\begin{aligned}
&J(K) - J(K^*) \\
&\quad \le \mathrm{Tr}\left(\mathcal{L}_{K^*}^{-1} C \Sigma_{K^*} E_K^\top \left(R + B^\top P_K B\right)^{-1} E_K \Sigma_{K^*} C^\top\right) \\
&\quad \le \|\Sigma_{K^*} C^\top \mathcal{L}_{K^*}^{-1} C \Sigma_{K^*}\| \mathrm{Tr}\left(E_K^\top \left(R + B^\top P_K B\right)^{-1} E_K\right) \\
&\quad \le \|\Sigma_{K^*} C^\top \mathcal{L}_{K^*}^{-1} C\| \|\Sigma_{K^*}\| \mathrm{Tr}\left(E_K^\top \left(R + B^\top P_K B\right)^{-1} E_K\right) \\
&\quad \le \|\Sigma_{K^*}\| \mathrm{Tr}\left(E_K^\top \left(R + B^\top P_K B\right)^{-1} E_K\right) \\
&\quad \le \frac{\|\Sigma_{K^*}\| \mathrm{Tr}\left(E_K^\top E_K\right)}{\sigma_{\min}(R)}. \qquad (69)
\end{aligned}
$$

From (15), it follows that:

$$
\begin{aligned}
\|\nabla J(K))\|_F^2 &= 4\mathrm{Tr}\left(C \Sigma_K E_K^\top E_K \Sigma_K C^\top\right) \\
&\ge 4\mu^2 \sigma_{\min}(C)^2 \mathrm{Tr}\left(E_K^\top E_K\right) \ \forall C \in \mathbb{C}. \qquad (70)
\end{aligned}
$$

By (69) and (70), one has

$$J(K) - J(K^*) \leq \frac{\|\Sigma_{K^*}\|\|\nabla J(K))\|_F^2}{4\mu^2\sigma_{\min}(C)^2\sigma_{\min}(R)} \quad \forall C \in \mathbb{C}. \tag{71}$$

Suppose K' satisfies that $K' = K - X$. According to (68), we get

$$\begin{aligned} &J(K) - J(K^*) \\ &\geq J(K) - J(K') \\ &= \mathrm{Tr}\left(\mathcal{L}_{K'}^{-1} C \Sigma_{K'} E_K^\top \left(R + B^\top P_K B\right)^{-1} E_K \Sigma_{K'} C^\top\right) \\ &\geq \frac{\mu \mathrm{Tr}\left(E_K^\top E_K\right)}{\|R + B^\top P_K B\|} \quad \forall C \in \mathbb{C}. \end{aligned} \tag{72}$$

In addition, when $C \in \mathbb{C}$, since we can always identity the state x by $x = C^{-1}y$, it is clear that $J(K^*) = J_s^*$ for every $C \in \mathbb{C}$. By replacing $J(K^*)$ in (71) and (72) with J_s^*, we finally complete the proof. ■

Appendix E
Derivations of ζ_2, ζ_3, and ζ_4 in Lemma 7

From (21), it is clear that

$$\frac{\partial P_{\bar{K}}}{\partial \delta} = \sum_{j=0}^{\infty} \mathcal{A}_{\bar{K}}^{\top j}\left(C^\top \Delta K^\top E_{\bar{K}} + E_{\bar{K}}^\top \Delta K C\right)\mathcal{A}_{\bar{K}}^{j}. \tag{73}$$

Then, we can observe that

$$\begin{aligned} &C^\top \Delta K^\top E_{\bar{K}} + E_{\bar{K}}^\top \Delta K C \\ &= C^\top \Delta K^\top R \bar{K} C + C^\top \bar{K}^\top R \Delta K C \\ &\quad - C^\top \Delta K^\top B^\top P_{\bar{K}} \mathcal{A}_{\bar{K}} - \mathcal{A}_{\bar{K}}^\top P_{\bar{K}} B \Delta K C \\ &\leq 2\|C\|\left(\|R\|\|\bar{K}C\| + \|B\|\|P_{\bar{K}}\|\|\mathcal{A}_{\bar{K}}\|\right)\|\Delta K\| I \\ &\leq \frac{2\|C\|Q}{\sigma_{\min}(Q)}\left(\|R\|\psi + \gamma\|B\|\frac{\alpha}{\mu}\right)\|\Delta K\| \end{aligned} \tag{74}$$

where the last step follows from Lemma 9. Therefore, according to (14), we have $(\partial P_{\bar{K}}/\partial\delta) \preceq \zeta_2\|\Delta K\|P_{\bar{K}}$.

Next, we will prove that $(\partial P'_{\bar{K}}[Z]/\partial\delta) \preceq \zeta_3\|\Delta K\|P_{\bar{K}}$. Based on (22), we get

$$\frac{\partial P'_{\bar{K}}[Z]}{\partial \delta} = \sum_{j=0}^{\infty} \mathcal{A}_{\bar{K}}^{\top j} S_3 \mathcal{A}_{\bar{K}}^{j} \tag{75}$$

where

$$\begin{aligned} S_3 := &\, C^\top Z^\top \left(R + B^\top P_{\bar{K}} B\right)\Delta K C - (BZC)^\top \frac{\partial P_{\bar{K}}}{\partial\delta}\mathcal{A}_{\bar{K}} \\ &+ C^\top \Delta K^\top \left(R + B^\top P_{\bar{K}} B\right) Z C - \mathcal{A}_{\bar{K}}^\top \frac{\partial P_{\bar{K}}}{\partial\delta} BZC \\ &- (B\Delta K C)^\top P'_{\bar{K}}[Z]\mathcal{A}_{\bar{K}} - \mathcal{A}_{\bar{K}}^\top P'_{\bar{K}}[Z] B \Delta K C. \end{aligned}$$

Recalling that $P'_{\bar{K}}[Z] \preceq \zeta_1 P_{\bar{K}}$ and $(\partial P_{\bar{K}}/\partial\delta) \preceq \zeta_2\|\Delta K\|P_{\bar{K}}$, we can also show that

$$\begin{aligned} S_3 \leq &\, 2\Big(\|C\|^2\|R\| + \|C\|^2\|B\|^2\|P_{\bar{K}}\| + \zeta_2\gamma\|B\|\|C\|\|P_{\bar{K}}\| \\ &\quad + \zeta_1\gamma\|B\|\|C\|\|P_{\bar{K}}\|\Big)\|\Delta K\| I \\ \leq &\, \frac{2\|C\|Q}{\sigma_{\min}(Q)}\Big(\|C\|\|R\| + \|B\|(\|C\|\|B\| \\ &\qquad\qquad + \zeta_1\gamma + \zeta_2\gamma)\frac{\alpha}{\mu}\Big)\|\Delta K\|. \end{aligned}$$

Therefore, we get $(\partial P'_{\bar{K}}[Z]/\partial\delta) \preceq \zeta_3\|\Delta K\|P_{\bar{K}}$.

Similarly, for $P''_{\bar{K}}[Z]$, from (23), we can show that

$$\begin{aligned} S_1 &\leq 2\left(\|C\|^2\|R\| + \|C\|^2\|B\|^2\|P_K\| + \zeta_1\gamma\|B\|\|C\|\|P_K\|\right) I \\ &\leq \frac{2\|C\|Q}{\sigma_{\min}(Q)}\left(\|C\|\|R\| + \|B\|(\|C\|\|B\| + \zeta_1\gamma)\frac{\alpha}{\mu}\right). \end{aligned}$$

So, it is clear $P''_{\bar{K}}[Z] \preceq \zeta_4 P_{\bar{K}}$, which completes the derivations.

Appendix F
Proof of Theorem 3

Proof: We can easily modify the proof of Theorem 1 to show that for every $K \in \mathbb{K}_\alpha$, if $\eta \leq \mu\sigma_{\min}(C)^2/L$, the line segment $[K, K - \eta\nabla^{\mathrm{NA}}J(K)] \subseteq \mathbb{K}_\alpha$. Then from (31), one has

$$\begin{aligned} J(K_{i+1}) &\leq J(K_i) - \eta\mathrm{Tr}\left(\nabla J(K_i)^\top \nabla^{\mathrm{NA}} J(K_i)\right) \\ &\quad + \frac{\eta^2 L}{2}\|\nabla^{\mathrm{NA}} J(K_i)\|_F^2 \\ &\leq J(K_i) - \eta\left(\sigma_{\min}\left(\mathcal{L}_{K_i}\right) - \frac{\eta L}{2}\right)\left\|\nabla^{\mathrm{NA}} J(K_i)\right\|_F^2 \\ &\leq J(K_i) - \eta\left(\mu\sigma_{\min}(C)^2 - \frac{\eta L}{2}\right)\left\|\nabla^{\mathrm{NA}} J(K_i)\right\|_F^2 \\ &\leq J(K_i) - \frac{\eta\mu\sigma_{\min}(C)^2}{2}\left\|\nabla^{\mathrm{NA}} J(K_i)\right\|_F^2 \end{aligned}$$

where the last inequality takes into account that $\eta \leq \mu\sigma_{\min}(C)^2/L$. Note that the boundary $\mu\sigma_{\min}(C)^2/L$ is selected for achieving the fastest convergence rate. By summing up the above inequality, one has

$$\frac{\mu\sigma_{\min}(C)^2\eta}{2}\sum_{i=0}^{N}\left\|\nabla^{\mathrm{NA}} J(K_i)\right\|_F^2 \leq J(K_0) - J(K^\star).$$

Consequently, it follows that:

$$\min_{0\leq i\leq N}\left\|\nabla^{\mathrm{NA}} J(K_i)\right\|_F^2 \leq \frac{2\alpha}{\eta\mu\sigma_{\min}(C)^2 N}.$$

Thus, the natural policy gradient method can attain an ϵ-stationary point in $(2\alpha/[\eta\mu\sigma_{\min}(C)^2\epsilon^2])$ iterations.

When $C \in \mathbb{C}$, by (15) and (31), one has

$$\begin{aligned} J(K_{i+1}) &\leq J(K_i) - 4\eta\mathrm{Tr}\left(\Sigma_{K_i} E_{K_i}^\top E_{K_i}\right) + 2\eta^2 L\left\|E_{K_i} C^{-1}\right\|_F^2 \\ &\leq J(K_i) - 4\eta\left(\mu - \frac{L\eta}{2\sigma_{\min}(C)^2}\right)\left\|E_{K_i}\right\|_F^2 \\ &\leq J(K_i) - 2\mu\eta\left\|E_{K_i}\right\|_F^2 \\ &\leq J(K_i) - \frac{2\eta\mu\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|}\left(J(K_i) - J(K^\star)\right) \end{aligned}$$

where the last step follows from (35). It directly follows that:

$$J(K_i) - J_s^\star \leq \left(1 - \frac{2\eta\mu\sigma_{\min}(R)}{\|\Sigma_{K^\star}\|}\right)^i\left(J(K_0) - J_s^\star\right)$$

which completes the proof of the second claim. ■

Appendix G
Proof of Theorem 4

Proof: We can easily modify the proof of Theorem 1 to show that for every $K \in \mathbb{K}_\alpha$, if $\eta \leq \mu\sigma_{\min}(R)\sigma_{\min}(C)^2/L$, the segment $[K, K - \eta\nabla^{GN}J(K)] \subseteq \mathbb{K}_\alpha$. Then from (31), one has

$$\begin{aligned} J(K_{i+1}) &\leq J(K_i) - \eta\mathrm{Tr}\Big(\nabla J(K_i)^\top \nabla^{\mathrm{GN}} J(K_i)\Big) \\ &\quad + \frac{\eta^2 L}{2}\|\nabla^{\mathrm{GN}} J(K_i)\|_F^2 \\ &\leq J(K_i) - \eta\left(\mu\sigma_{\min}(R)\sigma_{\min}(C)^2 - \frac{\eta L}{2}\right)\left\|\nabla^{\mathrm{GN}} J(K_i)\right\|_F^2 \\ &\leq J(K_i) - \frac{\eta\mu\sigma_{\min}(R)\sigma_{\min}(C)^2}{2}\left\|\nabla^{\mathrm{GN}} J(K_i)\right\|_F^2 \end{aligned}$$

where the last inequality considers the boundary of step size, that is, $\eta \leq \mu\sigma_{\min}(R)\sigma_{\min}(C)^2/L$. The boundary $\mu\sigma_{\min}(R)\sigma_{\min}(C)^2/L$ is selected for achieving the fastest convergence rate. By summing up the above inequality, one has

$$\frac{\eta\mu\sigma_{\min}(R)\sigma_{\min}(C)^2}{2}\sum_{i=0}^{N}\left\|\nabla^{\mathrm{NA}} J(K_i)\right\|_F^2 \leq J(K_0) - J\big(K^\star\big).$$

Consequently, it follows that:

$$\min_{0\leq i\leq N}\|\nabla^{\mathrm{NA}} J(K_i)\|_F^2 \leq \frac{2\alpha}{\eta\mu\sigma_{\min}(R)\sigma_{\min}(C)^2 N}.$$

Thus, the Gauss–Newton method can attain an ϵ-stationary point in $(2\alpha/[\eta\mu\sigma_{\min}(R)\sigma_{\min}(C)^2\epsilon^2])$ iterations.

When $C \in \mathbb{C}$, by (15) and (31), one has

$$\begin{aligned} J(K_{i+1}) &\leq J(K_i) - 4\eta\mathrm{Tr}\Big(\Sigma_{K_i}E_{K_i}^\top\Big(R + B^\top P_{K_i}B\Big)^{-1}E_{K_i}\Big) \\ &\qquad + 2\eta^2 L\left\|\Big(R + B^\top P_{K_i}B\Big)^{-1}E_{K_i}C^{-1}\right\|_F^2 \\ &\leq J(K_i) - 4\eta\left(\mu - \frac{\eta L}{2\sigma_{\min}(R)\sigma_{\min}(C)^2}\right) \\ &\quad \times \mathrm{Tr}\Big(E_{K_i}^\top\Big(R + B^\top P_{K_i}B\Big)^{-1}E_{K_i}\Big) \\ &\leq J(K_i) - 2\eta\mu\mathrm{Tr}\Big(E_{K_i}^\top\Big(R + B^\top P_{K_i}B\Big)^{-1}E_{K_i}\Big) \\ &\leq J(K_i) - \frac{2\eta\mu}{\|\Sigma_{K^\star}\|}\big(J(K_i) - J\big(K^\star\big)\big) \end{aligned}$$

where the last step follows from (35). It directly follows that:

$$J(K_i) - J_s^\star \leq \left(1 - \frac{2\eta\mu}{\|\Sigma_{K^\star}\|}\right)^i\big(J(K_0) - J_s^\star\big)$$

which completes the proof of the second claim. ■

Acknowledgment

The authors extend their gratitude to Yang Zheng for his valuable suggestions.

References

[1] D. Silver et al., “Mastering the game of go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[2] S. E. Li, *Reinforcement Learning for Sequential Decision and Optimal Control*. Singapore: Springer, 2022.

[3] J. Duan et al., “Relaxed actor-critic with convergence guarantees for continuous-time optimal control of nonlinear systems,” *IEEE Trans. Intell. Veh.*, vol. 8, no. 5, pp. 3299–3311, May 2023.

[4] Y. Guan et al., “Integrated decision and control: Toward interpretable and computationally efficient driving intelligence,” *IEEE Trans. Cybern.*, vol. 53, no. 2, pp. 859–873, Feb. 2023.

[5] T. P. Lillicrap et al., “Continuous control with deep reinforcement learning,” in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–101.

[6] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 1861–1870.

[7] J. Duan, Y. Guan, S. E. Li, Y. Ren, Q. Sun, and B. Cheng, “Distributional soft actor-critic: Off-policy reinforcement learning for addressing value estimation errors,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6584–6598, Nov. 2022.

[8] D. Wang, M. Ha, and M. Zhao, “The intelligent critic framework for advanced optimal control,” *Artif. Intell. Rev.*, vol. 55, pp. 1–22, Jan. 2022.

[9] D. Wang, J. Ren, M. Ha, and J. Qiao, “System stability of learning-based linear optimal control with general discounted value iteration,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6504–6514, Sep. 2023.

[10] K. Zhang, A. Koppel, H. Zhu, and T. Basar, “Global convergence of policy gradient methods to (almost) locally optimal policies,” *SIAM J. Control Optim.*, vol. 58, no. 6, pp. 3586–3612, 2020.

[11] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, “Global convergence of policy gradient methods for the linear quadratic regulator,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1467–1476.

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

[13] J. Bhandari and D. Russo, “Global optimality guarantees for policy gradient methods,” 2019, *arXiv:1906.01786*.

[14] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, “Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator,” in *Proc. IEEE 58th Conf. Decis. Control (CDC)*, 2019, pp. 7474–7479.

[15] 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 *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 2916–2925.

[16] B. Hambly, R. Xu, and H. Yang, “Policy gradient methods for the noisy linear quadratic regulator over a finite horizon,” *SIAM J. Control Optim.*, vol. 59, no. 5, pp. 3359–3391, 2021.

[17] Y. Zheng, L. Furieri, M. Kamgarpour, and N. Li, “Sample complexity of linear quadratic Gaussian (LQG) control for output feedback systems,” in *Proc. Learn. Dyn. Control*, 2021, pp. 559–570.

[18] Z. Ren, A. Zhong, and N. Li, “LQR with tracking: A zeroth-order approach and its global convergence,” in *Proc. Amer. Control Conf. (ACC)*, 2021, pp. 2562–2568.

[19] J. P. Jansch-Porto, B. Hu, and G. E. Dullerud, “Policy optimization for Markovian jump linear quadratic control: Gradient method and global convergence,” *IEEE Trans. Autom. Control*, vol. 68, no. 4, pp. 2475–2482, Apr. 2023.

[20] L. Furieri, Y. Zheng, and M. Kamgarpour, “Learning the globally optimal distributed LQ regulator,” in *Proc. Learn. Dyn. Control*, 2020, pp. 287–297.

[21] K. Zhang, B. Hu, and T. Basar, “Policy optimization for H_2 linear control with H_∞ robustness guarantee: Implicit regularization and global convergence,” in *Proc. Learn. Dyn. Control*, 2020, pp. 179–190.

[22] H. Wang, P. X. Liu, and P. Shi, “Observer-based fuzzy adaptive output-feedback control of stochastic nonlinear multiple time-delay systems,” *IEEE Trans. Cybern.*, vol. 47, no. 9, pp. 2568–2578, Sep. 2017.

[23] J. Yu, S. Cheng, P. Shi, and C. Lin, “Command-filtered neuroadaptive output-feedback control for stochastic nonlinear systems with input constraint,” *IEEE Trans. Cybern.*, vol. 53, no. 4, pp. 2301–2310, Apr. 2023.

[24] S. Zhao, J. Wang, H. Xu, and B. Wang, “Composite observer-based optimal attitude-tracking control with reinforcement learning for hypersonic vehicles,” *IEEE Trans. Cybern.*, vol. 53, no. 2, pp. 913–926, Feb. 2023.

[25] G. Zong, Q. Xu, X. Zhao, S.-F. Su, and L. Song, “Output-feedback adaptive neural network control for uncertain nonsmooth nonlinear systems with input Deadzone and saturation,” *IEEE Trans. Cybern.*, vol. 53, no. 9, pp. 5957–5969, Sep. 2023.

[26] V. L. Syrmos, C. T. Abdallah, P. Dorato, and K. Grigoriadis, "Static output feedback—A survey," *Automatica*, vol. 33, no. 2, pp. 125–137, 1997.

[27] I. Fatkhullin and B. Polyak, "Optimizing static linear feedback: Gradient method," *SIAM J. Control Optim.*, vol. 59, no. 5, pp. 3887–3911, 2021.

[28] H. Feng and J. Lavaei, "Connectivity properties of the set of stabilizing static decentralized controllers," *SIAM J. Control Optim.*, vol. 58, no. 5, pp. 2790–2820, 2020.

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

[30] S. M. Kakade, "A natural policy gradient," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14, 2001, pp. 1531–1538.

[31] F. D. Foresee and M. T. Hagan, "Gauss-Newton approximation to Bayesian learning," in *Proc. Int. Conf. Neural Netw. (ICNN)*, vol. 3, 1997, pp. 1930–1935.

[32] D. Lee and J. Hu, "Primal-dual Q-learning framework for LQR design," *IEEE Trans. Autom. Control*, vol. 64, no. 9, pp. 3756–3763, Sep. 2019.

[33] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, vol. 408. New York, NY, USA: Springer, 2011.

[34] B. T. Polyak, "Gradient methods for minimizing functionals," *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, vol. 3, no. 4, pp. 643–653, 1963.

[35] Y. Nesterov, "Introductory lectures on convex programming volume I: Basic course," *Lecture Notes*, vol. 3, no. 4, p. 5, 1998.

[36] S. Lojasiewicz, "A topological property of real analytic subsets," *Coll. du CNRS, Les Équ. aux dérivées partielles*, vol. 117, nos. 87–89, p. 2, 1963.

[37] Y. Tang, Y. Zheng, and N. Li, "Analysis of the optimization landscape of linear quadratic Gaussian (LQG) control," in *Proc. Learn. Dyn. Control*, 2021, pp. 599–610.

[38] J. Duan, W. Cao, Y. Zheng, and L. Zhao, "On the optimization landscape of dynamic output feedback linear quadratic control," *IEEE Trans. Autom. Control*, early access, May 12, 2023, doi: 10.1109/TAC.2023.3275732.

[39] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1724–1732.

[40] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—Online stochastic gradient for tensor decomposition," in *Proc. Conf. Learn. Theory*, 2015, pp. 797–842.

[41] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Accelerated methods for nonconvex optimization," *SIAM J. Optim.*, vol. 28, no. 2, pp. 1751–1772, 2018.

[42] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to Derivative-Free Optimization*. Philadelphia, PA, USA: SIAM, 2009.

[43] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Found. Comput. Math.*, vol. 17, no. 2, pp. 527–566, 2017.

[44] J. Yu, "An equivalent discrete-time output feedback linear quadratic regulator theory," in *Proc. 7th Int. Conf. Control Decis. Inf. Technol. (CoDIT)*, vol. 1, 2020, pp. 868–873.

[45] F. Lewis, *Applied Optimal Control & Estimation: Digital Design & Implementation*. Hoboken, NJ, USA: Prentice Hall, 1992. [Online]. Available: https://books.google.fi/books?id=SqkeAQAAIAAJ

[46] S. A. Arogeti and F. L. Lewis, "Static output-feedback H_∞ control design procedures for continuous-time systems with different levels of model knowledge," *IEEE Trans. Cybern.*, vol. 53, no. 3, pp. 1432–1446, Mar. 2023.

Jingliang Duan (Member, IEEE) received the Doctoral degree in mechanical engineering from the School of Vehicle and Mobility, Tsinghua University, Beijing, China, in 2021.

In 2019, he was as a Visiting Student Researcher with the Department of Mechanical Engineering, The University of California at Berkeley, Berkeley, CA, USA. Following his Ph.D. degree, he served as a Research Fellow with the Department of Electrical and Computer Engineering, The National University of Singapore, Singapore, from 2021 to 2022. He is currently a tenured Associate Professor with the School of Mechanical Engineering, University of Science and Technology Beijing, Beijing. His research interests include reinforcement learning, optimal control, and self-driving decision making.

Jie Li received the B.S. degree in automotive engineering from Tsinghua University, Beijing, China, in 2018, where he is currently pursuing the Ph.D. degree with the School of Vehicle and Mobility.

His current research interests include model predictive control, adaptive dynamic programming, and robust reinforcement learning.

Xuyang Chen received the B.S. degree from the Honors College, Beihang University, Beijing, China, in 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore.

His research interests include reinforcement learning, Markov decision process, and policy gradient methods.

Kai Zhao (Member, IEEE) received the Ph.D. degree in control theory and control engineering from Chongqing University, Chongqing, China, in 2019.

He was a Postdoctoral Fellow with the Department of Computer and Information Science, University of Macau, Macau, China, from 2019 to 2021. He is currently a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests include adaptive control and prescribed performance control.

Shengbo Eben Li (Senior Member, IEEE) received the M.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 2006 and 2009, respectively.

He was with Stanford University, Stanford, CA, USA; University of Michigan at Ann Arbor, Ann Arbor, MI, USA; and The University of California at Berkeley, Berkeley, CA, USA. He is currently a tenured Professor with Tsinghua University. His active research interests include intelligent vehicles and driver assistance, reinforcement learning and distributed control, and optimal control and estimation.

Lin Zhao (Member, IEEE) received the B.S. and M.S. degrees in automatic control from the Harbin Institute of Technology, Harbin, China, in 2010 and 2012, respectively, and the M.S. degree in mathematics and the Ph.D. degree in electrical and computer engineering from The Ohio State University, Columbus, OH, USA, in 2017.

From 2018 to early 2020, he was a Research Scientist with Aptiv Pittsburgh Technology Center (currently Motional), Pittsburgh, PA, USA. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His current research focuses on control and reinforcement learning with applications in robotics.

Dr. Zhao serves on the Young Editorial Board for the *Journal of Systems Science and Complexity* (Springer), served as the Program Co-Chair for the 17th IEEE International Conference on Control and Automation (ICCA 2022), and as a Publicity Co-Chair for the 62nd IEEE Conference on Decision and Control (CDC 2023).