ANNUAL REVIEWS

Annual Review of Control, Robotics, and Autonomous Systems

Toward a Theoretical Foundation of Policy Optimization for Learning Control Policies

Bin Hu,[1] Kaiqing Zhang,[2,3] Na Li,[4] Mehran Mesbahi,[5] Maryam Fazel,[6] and Tamer Başar[1]

[1]Coordinated Science Laboratory and Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Urbana, Illinois, USA; email: binhu7@illinois.edu, basar1@illinois.edu

[2]Laboratory for Information and Decision Systems and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[3]Current affiliation: Department of Electrical and Computer Engineering and Institute for Systems Research, University of Maryland, College Park, Maryland, USA; email: kaiqing@umd.edu

[4]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA; email: nali@seas.harvard.edu

[5]Department of Aeronautics and Astronautics, University of Washington, Seattle, Washington, USA; email: mesbahi@uw.edu

[6]Department of Electrical and Computer Engineering, University of Washington, Seattle, Washington, USA; email: mfazel@uw.edu

Annu. Rev. Control Robot. Auton. Syst. 2023. 6:123–58

The *Annual Review of Control, Robotics, and Autonomous Systems* is online at control.annualreviews.org

https://doi.org/10.1146/annurev-control-042920-020021

Keywords

policy optimization, reinforcement learning, feedback control synthesis

Abstract

Gradient-based methods have been widely used for system design and optimization in diverse application domains. Recently, there has been a renewed interest in studying theoretical properties of these methods in the context of control and reinforcement learning. This article surveys some of the recent developments on policy optimization, a gradient-based iterative approach for feedback control synthesis that has been popularized by successes of reinforcement learning. We take an interdisciplinary perspective in our exposition that connects control theory, reinforcement learning, and large-scale optimization. We review a number of recently developed theoretical results on the optimization landscape, global convergence, and sample complexity

of gradient-based methods for various continuous control problems, such as the linear quadratic regulator (LQR), $\mathcal{H}_\infty$ control, risk-sensitive control, linear quadratic Gaussian (LQG) control, and output feedback synthesis. In conjunction with these optimization results, we also discuss how direct policy optimization handles stability and robustness concerns in learning-based control, two main desiderata in control engineering. We conclude the survey by pointing out several challenges and opportunities at the intersection of learning and control.

1. INTRODUCTION

Reinforcement learning (RL) has recently shown impressive performance in a wide range of applications, from playing Atari (1, 2) and mastering the game of Go (3, 4) to complex robotic manipulations (5–7). Key to RL's success is the algorithmic framework of policy optimization (PO), where the policy, mapping observations to actions, is parameterized and directly optimized upon to improve system-level performance. Mastering Go using PO (combined with techniques such as efficient tree search) is particularly encouraging,[1] as the main idea behind PO is rather straightforward—when learning has been formalized as minimizing a certain cost as a function of the policy, devise an iterative procedure on the policy to improve the objective. For example, in the policy gradient (PG) variant of PO, when learning is represented as minimizing a (differentiable) cost $J(K)$ over the policy K, the policy is improved upon via a gradient update of the form $K^{n+1} = K^n - \alpha \nabla J(K^n)$, for some step size α (also referred to as the learning rate) and data-driven evaluation of the cost gradient ∇J at each iteration n. In fact, PO provides an umbrella formalism for not only PG methods (8) but also actor–critic (9), trust-region (10), and proximal PO (11) methods.

More generally, PO provides a streamlined approach to learning-based system design. For example, it gives a general-purpose paradigm for addressing complex nonlinear dynamics with user-specified cost functions: For tasks involving nonlinear dynamics and complex design objectives, one can parameterize the policy as a neural network to be trained using gradient-based methods to obtain a reasonable solution. The PO perspective can also be adopted for other insufficiently parameterized decision problems, such as end-to-end perception-based control (12–14). In this setting, one may wish to synthesize a policy directly on images. As such, one can envision parameterizing a mapping from pixels (observation) to actions (decisions) as a neural network, and learn the corresponding policy using the PO formalism. Lastly, we mention the use of scalable gradient-based algorithms to efficiently train nonlinear policies on many parameters, making PO suitable for high-dimensional tasks. The computational flexibility and conceptual accessibility of PO have made it a main workhorse for modern RL.

In yet another decision-theoretic science, PO has a long history in control theory (15–20); in fact, it has been popular among control practitioners when the system model is poorly understood or parameterized. Nevertheless, despite its generality and flexibility, the PO formulation of control synthesis is typically nonconvex and, as such, challenging for obtaining strong performance certificates, rendering it unpopular among system theorists. Since the 1980s, convex reformulations or relaxations of control problems have become popular due to the development of convex programming and related global convergence theory (21). It has been realized that many problems in optimal and robust control can be reformulated as convex programs [namely, semidefinite programs (22–24)] or relaxed via sum of squares (25, 26), expressed in terms of certificates (e.g.,

[1] Go is considered a challenging game to master, partly because the number of legal board positions is significantly larger than the number of atoms in the observable universe.

matrix inequalities that represent Lyapunov or dissipativity conditions). However, these formulations have limitations when there is deviation from the canonical synthesis problems (e.g., when there are constraints on the structure of the desired control policy).

When convex reformulations are not available, PO assumes an important role as the main viable option. Examples of such scenarios include static output feedback problems (27), structured $\mathcal{H}_\infty$ synthesis (28–33), and distributed control (34), all of which have significant importance in applications. The PO framework is more flexible, as evidenced by the recent advances in deep RL. PO is also more scalable for high-dimensional problems, as it does not generally introduce extra variables in the optimization problems and enjoys a broader range of optimization methods as compared with the semidefinite-program or sum-of-squares formulations. However, as pointed out above, the nonconvexity of the PO formulation, even on relatively simple linear control problems, has made deriving theoretical guarantees for direct PO challenging, hindering the acceptance of PO as a mainstream control design tool.

In this survey, our aim is to revisit these issues from a modern optimization perspective and provide a unified perspective on the recently developed global convergence/complexity theory for PO in the context of control synthesis. Recent theoretical results on PO for particular classes of control synthesis problems, some of which are discussed in this survey, not only are exciting but also lead to a new research thrust at the interface of control theory and machine learning. This survey includes control synthesis related to linear quadratic regulator (LQR) theory (35–44), stabilization (45–47), linear robust/risk-sensitive control (48–55), Markov jump linear quadratic control (56–59), Lur'e system control (60), output feedback control (61–67), and dynamic filtering (68). Surprisingly, some of these strong global convergence results for PO have been obtained in the absence of convexity in the design objective and/or the underlying feasible set.

These global convergence guarantees have several implications for learning and control. First, these results facilitate examining other classes of synthesis problems in the same general framework. As will be pointed out in this survey, there is an elegant geometry at play between certificates and controllers in the synthesis process, with immediate algorithmic implications. Second, the theoretical developments in PO have generated a renewed interest in the control community in examining synthesis of dynamic systems from a complementary perspective that, in our view, is more integrated with learning in general and RL in particular. This will complement and strengthen the existing connections between RL and control (69–71). Lastly, the geometric analysis of PO-inspired algorithms may shed light on issues in state-of-the-art policy-based RL, critical for deriving guarantees for any subsequent RL-based synthesis procedure for dynamic systems.

This survey is organized to reflect our perspective on—and excitement about—how PO methods (and PG methods in particular) provide a streamlined approach for system synthesis, and to build a bridge between control and learning. First, we provide the PO formulations for various control problems in Section 2. Then we delve into the PO convergence theory on the classical LQR problem in Section 3. As it turns out, a key ingredient for analyzing LQR PO hinges on the coerciveness of the cost function and its gradient dominance property (see Section 3.2). These properties can then be utilized to devise gradient updates ensuring stabilizing feedback policies at each iteration and convergence to the globally optimal policy. In Section 3.3, we highlight some of the challenges in extending the LQR PO theory to other classes of problems, including the role of coerciveness, gradient dominance, smoothness, and the landscape of the optimization problem. In Section 4, we extend the PO perspective to more elaborate synthesis problems, such as linear robust/risk-sensitive control, dynamic games, and nonsmooth $\mathcal{H}_\infty$ state-feedback synthesis. Through these extensions, we highlight how variations on the general theme set by the LQR PO theory can be adopted to address lack of coerciveness or nonsmoothness of the objective in

these problems while ensuring the convergence of the iterations to solutions of interest. This is followed by examining PO for control synthesis with partial observations and, in particular, PO theory for linear quadratic Gaussian (LQG) and output feedback control in Section 5. Our discussion in Section 5 underscores the importance of the underlying geometry of the policy landscape in developing any PO-based algorithms. Fundamental connections between PO theory and convex parameterization in control are then discussed in Section 6. In particular, we demonstrate how the geometry of policies and certificates are intertwined through appropriately constructed maps between nonconvex PO formulation of the synthesis problems and the (convex) semidefinite programming parameterizations. This provides a unified approach for analyzing PO in various control problems that have so far been studied on a case-by-case basis. Finally, in Section 7, we present current challenges and our outlook for a comprehensive PO theory for synthesizing dynamical systems that ensures stability, robustness, safety, and optimality and underscore the challenges in addressing synthesis problems in the face of partial observations and nonlinearities and in multiagent settings. Section 7 also examines further connections between PO theory and machine learning and highlights the possibility of integrating model-based (70) and model-free methods to achieve the best of both worlds, illustrating how the main theme of this survey fits within the big picture of learning-based control.

2. POLICY OPTIMIZATION FOR LINEAR CONTROL: FORMULATION

Control design can generally be formulated as a PO problem of the form

$$\min_{K \in \mathcal{K}} \quad J(K), \tag{1.}$$

where the decision variable K is determined by the controller parameterization (linear mapping, polynomials, kernels, neural networks, etc.), the cost function $J(K)$ is some task-dependent control performance measure (tracking errors, closed-loop $\mathcal{H}_2$ or $\mathcal{H}_\infty$ norm, etc.), and the feasible set $\mathcal{K}$ represents the class of controllers of interest, for example, ensuring closed-loop stability/robustness requirements. Such a PO formulation is general and enables flexible policy parameterizations. For example, consider a modern deep RL setting where one wants to design a policy maximizing some task-dependent reward function for a complicated nonlinear system $x_{t+1} = f(x_t, u_t, w_t)$, with (x_t, u_t, w_t) being the state, action, and disturbance triplet. PO has served as the main workhorse for addressing such tasks. Specifically, one just needs to parameterize the policy as a (deep) neural network and then apply iterative PO algorithms such as trust-region PO (10) and proximal PO (11) to learn the optimal weights.

The focus of this article is the recently developed (global) convergence, complexity, and landscape theory of PO on classical control tasks, including LQR, risk-sensitive/robust control, and output feedback control. In this section, we formulate these linear control problems as PO via proper selection of K, J, and $\mathcal{K}$ in Equation 1.

2.1. Case I: The Linear Quadratic Regulator

There are several ways to formulate the LQR problem. For simplicity, we start by considering a discrete-time linear time-invariant (LTI) system $x_{t+1} = Ax_t + Bu_t$, where x_t is the state and u_t is the control action. The design objective is to choose the control actions $\{u_t\}$ to minimize a quadratic cost function $J := \mathbb{E}_{x_0 \sim \mathcal{D}} \sum_{t=0}^{\infty} (x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t)$, with $Q \succeq 0$ and $R \succ 0$ being preselected cost-weighting matrices. In this setting, the only randomness stems from the initial condition x_0, which is sampled from a certain distribution $\mathcal{D}$ with a full rank covariance matrix.

It is well known that under some standard stabilizability and detectability assumptions, the optimal cost is finite and can be achieved by a linear state-feedback controller of the form

$u_t = -Kx_t$. Therefore, we can formulate the LQR problem as a special case of the PO problem in Equation 1. Specifically, the decision variable K is simply the feedback gain matrix. Under a fixed policy K, we have $u_t = -Kx_t$ for all t, and the LQR cost can be rewritten as $J(K) = \mathbb{E}_{x_0\sim\mathcal{D}}[\sum_{t=0}^{\infty} x_0^\top((A-BK)^\top)^t(Q+K^\top RK)(A-BK)^t x_0]$, which is a function of K. This cost can also be computed as $J(K) = \mathrm{Tr}(P_K\Sigma_0)$, where $\Sigma_0 = \mathbb{E}x_0x_0^\top$ is the (full-rank) covariance matrix of x_0, and P_K is the solution of the following Lyapunov equation:

$$(A-BK)^\top P_K(A-BK) + Q + K^\top RK = P_K. \tag{2.}$$

The above cost $J(K)$ is only well defined when the closed-loop system matrix $(A-BK)$ is Schur stable, i.e., when the spectral radius satisfies $\rho(A-BK) < 1$. Therefore, one can define the feasible set $\mathcal{K}$ as

$$\mathcal{K} = \{K : \rho(A-BK) < 1\}. \tag{3.}$$

Now we can see that the LQR problem is a special case of the PO problem in Equation 1.

There are also several slightly different ways to formulate the LQR problem. In an alternative formulation, we can add stochastic process noise and consider the following LTI system:

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{4.}$$

where the disturbance $\{w_t\}$ is a zero-mean independent and identically distributed (i.i.d.) process with a full rank covariance matrix W. The design objective is then to choose $\{u_t\}$ to minimize the time-average cost

$$J := \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\left(x_t^\top Qx_t + u_t^\top Ru_t\right)\right], \tag{5.}$$

where $Q \succeq 0$ and $R \succ 0$ are preselected weighting matrices. Again, it suffices to parameterize the policy as $u_t = -Kx_t$. For a fixed policy K, the cost in Equation 5 can be computed as $J(K) = \mathrm{Tr}(P_KW)$, where P_K is the solution for Equation 2. Again, the cost is well defined only for K satisfying $\rho(A-BK) < 1$. This setting leads to almost the same PO formulation as before. Similarly, the discounted LQR can be formulated as PO.

2.2. Case II: Linear Risk-Sensitive/Robust Control

One can enforce risk sensitivity and robustness via the formulation of linear exponential quadratic Gaussian (LEQG) (72) and $\mathcal{H}_\infty$ control (73), respectively. For linear risk-sensitive control, we still consider the LTI system in Equation 4, with $w_t \sim \mathcal{N}(0, W)$ being an i.i.d. Gaussian noise, and the design objective is to choose control actions $\{u_t\}$ to minimize an exponentiated quadratic cost,

$$J := \limsup_{T\to\infty}\frac{1}{T}\frac{2}{\beta}\log\mathbb{E}\exp\left[\frac{\beta}{2}\sum_{t=0}^{T-1}\left(x_t^\top Qx_t + u_t^\top Ru_t\right)\right], \tag{6.}$$

where β is the parameter quantifying the intensity of risk sensitivity, and the expectation is taken over the distributions for x_0 and w_t for all $t \geq 0$. One typically chooses $\beta > 0$ to make the control risk averse. As $\beta \to 0$, the objective in Equation 6 reduces to the LQR cost.

The above LEQG problem is also a special case of the PO problem in Equation 1. It is known that the optimal cost can be achieved by a linear state-feedback controller. Again, one can just parameterize the controller as $u_t = -Kx_t$, where the gain matrix K is the decision variable. Then the cost function can be specified as $J(K) = -\frac{1}{\beta}\log\det(I-\beta P_KW)$, where P_K is the unique stabilizing

solution to the following algebraic Riccati equation:[2]

$$P_K = Q + K^\mathsf{T} R K + (A - BK)^\mathsf{T}\big[P_K - P_K W^{\frac{1}{2}}(-\beta^{-1} I + W^{\frac{1}{2}} P_K W^{\frac{1}{2}})^{-1} W^{\frac{1}{2}} P_K\big](A - BK).$$

Notice that, in this case, J is well defined only when K is in the following feasible set:

$$\mathcal{K} = \left\{K : \rho(A - BK) < 1, \text{ and } \|(Q + K^\mathsf{T} R K)^{\frac{1}{2}}(zI - A + BK) W^{\frac{1}{2}}\|_\infty < \frac{1}{\sqrt{\beta}}\right\}, \tag{7.}$$

where $\|\cdot\|_\infty$ denotes the $\mathcal{H}_\infty$ norm of a given discrete-time transfer function. Hence, the LEQG problem is a special case of the PO problem with J and $\mathcal{K}$ as defined above.

For the LEQG problem, the $\mathcal{H}_\infty$ constraint $\|(Q + K^\mathsf{T} R K)^{\frac{1}{2}}(zI - A + BK) W^{\frac{1}{2}}\|_\infty < \frac{1}{\sqrt{\beta}}$ is implicitly required by the problem formulation. Importantly, the LEQG problem can be viewed as a special case of the more general mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem studied in robust control. In this article, we will cover two important robust control settings, namely, the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design and the $\mathcal{H}_\infty$ state-feedback synthesis.

For mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design, consider the following system, where w_t is the disturbance and z_t is the controlled output:

$$x_{t+1} = Ax_t + Bu_t + Dw_t, \quad z_t = Cx_t + Eu_t. \tag{8.}$$

It is standard to assume $E^\mathsf{T}[C\ E] = [0\ R]$ for some $R \succ 0$. The mixed design objective is to synthesize a linear state-feedback controller that minimizes an upper bound on the $\mathcal{H}_2$ cost and satisfies an additional $\mathcal{H}_\infty$ robustness requirement on the channel from w_t to z_t. The $\mathcal{H}_\infty$ constraint is posed explicitly and is powerful in guaranteeing robust stability in the presence of any small gain type of uncertainty, including being time varying, dynamic, or nonlinear. For the mixed design problem, the robustness constraint is directly enforced on K, and hence the feasible set $\mathcal{K}$ is modified as

$$\mathcal{K} = \big\{K : \rho(A - BK) < 1, \text{ and } \|(C - EK)(zI - A + BK)D\|_\infty < \gamma\big\}, \tag{9.}$$

where γ quantifies the robustness level. The smaller γ is, the more robust the system is in the $\mathcal{H}_\infty$ sense (since it can tolerate the small gain uncertainty at the level $1/\gamma$ by the small gain theorem). There exist several objective functions that upper bound the $\mathcal{H}_2$ cost (74, 75); a common one is $J(K) = \mathrm{Tr}(P_K D D^\mathsf{T})$, where P_K is the solution to the algebraic Riccati equation introduced earlier with $Q = C^\mathsf{T} C$, $\gamma = 1/\sqrt{\beta}$, and $W = DD^\mathsf{T}$. Notice that the 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 smaller than a prespecified number.

By contrast, the $\mathcal{H}_\infty$ state-feedback synthesis aims at improving the system robustness and the worst-case performance by achieving the smallest closed-loop $\mathcal{H}_\infty$ norm. For simplicity, consider the LTI system $x_{t+1} = Ax_t + Bu_t + w_t$ initialized at $x_0 = 0$. The design objective of $\mathcal{H}_\infty$ control is to choose $\{u_t\}$ to minimize the quadratic cost $J := \sum_{t=0}^\infty (x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t)$ in the presence of the worst-case ℓ_2 disturbance satisfying $\sum_{t=0}^\infty \|w_t\|^2 \le 1$. This problem can be reformulated as the PO problem with the cost $J(K)$ being defined as the following closed-loop $\mathcal{H}_\infty$ norm:

$$J(K) = \sup_{\omega \in [0, 2\pi]} \lambda_{\max}^{\frac{1}{2}}\big((e^{-j\omega} I - A + BK)^{-\mathsf{T}}(Q + K^\mathsf{T} R K)(e^{j\omega} I - A + BK)^{-1}\big). \tag{10.}$$

The reason is that the above cost actually satisfies

$$J^2(K) = \max_{\sum_{t=0}^\infty \|w_t\|^2 \le 1} \sum_{t=0}^\infty x_t^\mathsf{T}(Q + K^\mathsf{T} R K) x_t = \max_{\sum_{t=0}^\infty \|w_t\|^2 \le 1} \sum_{t=0}^\infty (x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t).$$

[2]The solution $P_K \succeq 0$ satisfies $\rho((A - BK)^\mathsf{T}(I - \beta P_K W)^{-1}) < 1$, and $W^{-1} - \beta P_K \succ 0$.

The above cost is well defined only for K satisfying $\rho(A - BK) < 1$. Therefore, minimizing the $\mathcal{H}_\infty$ cost function defined by Equation 10 over $\mathcal{K}$ given by Equation 3 leads to a policy that minimizes the quadratic cost under the worst-case ℓ_2 disturbance.

2.3. Case III: Linear Quadratic Gaussian and Output Feedback Control

Consider the following LTI system that can only be partially observed:

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{11a.}$$

$$y_t = Cx_t + v_t. \tag{11b.}$$

Here, w_t and v_t are zero-mean white Gaussian noises with covariance matrices $W \succeq 0$ and $V \succ 0$. At step t, one can only observe y_t, and the state x_t is not directly measured. The design objective is to choose actions $\{u_t\}$ to minimize the time-averaged cost defined in Equation 5 given such partial observation information. Again, $Q \succeq 0$ and $R \succ 0$ are preselected weighting matrices. It is assumed that the pairs (A, B) and $(A, W^{\frac{1}{2}})$ are controllable and that the pairs (C, A) and $(Q^{\frac{1}{2}}, A)$ are observable.

This problem can also be formulated as a special case of the PO formulation given by Equation 1. Under our assumptions, it suffices to consider (full-order) dynamic controllers of the form

$$\xi_{t+1} = A_K \xi_t + B_K y_t, \quad u_t = C_K \xi_t, \tag{12.}$$

where ξ_t is the internal state of the controller and has the same dimension as x_t. For convenience, we encode the dynamic controller as

$$K := \begin{bmatrix} 0 & C_K \\ B_K & A_K \end{bmatrix}. \tag{13.}$$

The cost function $J(K)$ is well defined when the closed-loop system is stable, and hence, the feasible set should be specified as

$$\mathcal{K} = \left\{ K : \begin{bmatrix} A & BC_K \\ B_K C & A_K \end{bmatrix} \text{ is Schur stable} \right\}. \tag{14.}$$

For any $K \in \mathcal{K}$, the cost $J(K)$ can be represented as

$$J(K) = \mathrm{Tr}\left(\begin{bmatrix} Q & 0 \\ 0 & C_K^\mathsf{T} R C_K \end{bmatrix} X_K \right) = \mathrm{Tr}\left(\begin{bmatrix} W & 0 \\ 0 & B_K V B_K^\mathsf{T} \end{bmatrix} Y_K \right), \tag{15.}$$

where X_K and Y_K are the unique positive semidefinite solutions to the following Lyapunov equations:

$$X_K = \begin{bmatrix} A & BC_K \\ B_K C & A_K \end{bmatrix} X_K \begin{bmatrix} A & BC_K \\ B_K C & A_K \end{bmatrix}^\mathsf{T} + \begin{bmatrix} W & 0 \\ 0 & B_K V B_K^\mathsf{T} \end{bmatrix}, \tag{16a.}$$

$$Y_K = \begin{bmatrix} A & BC_K \\ B_K C & A_K \end{bmatrix}^\mathsf{T} Y_K \begin{bmatrix} A & BC_K \\ B_K C & A_K \end{bmatrix} + \begin{bmatrix} Q & 0 \\ 0 & C_K^\mathsf{T} R C_K \end{bmatrix}. \tag{16b.}$$

Thereby, the LQG design problem can be formulated as a special case of PO.

It is possible to use other control parameterizations and enforce more structures on K. This will lead to PO formulations for general output feedback control. Such formulations are particularly useful for decentralized control.

For all three cases, the PO formulation is nonconvex in the policy space (35, 76). This is in contrast to convex reformulations of these problems. Next, we review the recently developed PO theory for Cases I–III in Sections 3–5, respectively.

3. CASE I: GLOBAL CONVERGENCE AND COMPLEXITY OF POLICY OPTIMIZATION FOR THE LINEAR QUADRATIC REGULATOR

LQR provides arguably the most fundamental optimal control formulation. A main challenge for the PO formulation of LQR is that the stability constraints are nonconvex in the policy space. The global convergence and complexity of PO methods for LQR have not been established until very recently. We review such results in this section.

3.1. Background: Optimization and Complexity

Consider the constrained optimization problem $\min_{K\in\mathcal{K}} J(K)$, with $\mathcal{K}$ being nonconvex. If the feasible set $\mathcal{K}$ is open and the optimal value of J is achieved by some interior point K^* in $\mathcal{K}$, then we have $\nabla J(K^*) = 0$ (in this case, the Karush–Kuhn–Tucker condition reduces to the first-order optimality condition for unconstrained problems), and it is possible to solve for K^* by applying an iterative gradient-based algorithm with the update rule $K^{n+1} = K^n - \alpha F^n$, where K^n denotes the controller parameter at iteration n, and F^n is some descent direction of the cost J. The most common example of F^n is the gradient direction $\nabla J(K^n)$ at K^n. Some other examples include the natural gradient direction and Gauss–Newton (or other quasi-Newton) directions (for more details, see Section 3.2).

It is important to know whether and how fast $\{K^n\}$ converges to K^*. We will introduce one important optimization result for coercive and/or gradient-dominant function $J(K)$.

Definition 1 (coercive and gradient dominant properties). We call a function $J(K)$ coercive on $\mathcal{K}$ if 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 on the boundary $\partial\mathcal{K}$. We call the function μ-gradient dominant of degree p if it is continuously differentiable and satisfies

$$J(K) - J(K^*) \le \frac{1}{2\mu}\|\nabla J(K)\|_F^p, \quad \forall K \in \mathcal{K}, \tag{17.}$$

where μ is some positive constant, and K^* is an optimal solution of $J(K)$ over $\mathcal{K}$.[3]

If $J(K)$ is coercive, then it serves as a barrier function over the feasibility set $\mathcal{K}$, and hence projection is not needed for maintaining feasibility. In addition, gradient dominance is useful for establishing global convergence. The following optimization result is fundamental and useful.

Theorem 1. Suppose $J(K)$ is coercive. Assume further that J is twice continuously differentiable over $\mathcal{K}$. Then the following statements hold:

1. The sublevel set $\mathcal{K}_\gamma := \{K \in \mathcal{K} : J(K) \le \gamma\}$ is compact.

[3]This property (also referred to as the Polyak–Łojasiewicz condition) appears commonly in the optimization literature (77–79) but is often used only locally. Here, we are interested in special problems where this property holds globally.

2. The function $J(K)$ is L-smooth on $\mathcal{K}_\gamma$, and the constant L depends on γ and the problem parameters. Specifically, for any (K, K') satisfying $tK + (1-t)K' \in \mathcal{K}_\gamma \ \forall t \in [0, 1]$, the following inequality holds:

$$J(K') \leq J(K) + \langle \nabla J(K), (K' - K) \rangle + \frac{L}{2} \|K' - K\|_F^2. \tag{18.}$$

3. Consider the gradient descent method

$$K^{n+1} = K^n - \alpha \nabla J(K^n). \tag{19.}$$

Suppose $K^0 \in \mathcal{K}$. Let $\gamma_0 = J(K^0)$. Suppose L is the smoothness constant of $J(K)$ on $\mathcal{K}_{\gamma_0}$. Then, for any $0 < \alpha < \frac{2}{L}$, we have $K^n \in \mathcal{K}$ for all n. In addition, we have $\nabla J(K^n) \to 0$, and the following convergence rate bound holds with $C = \alpha - \frac{L\alpha^2}{2} > 0$:

$$\min_{0 \leq l \leq k} \|\nabla J(K^l)\|_F^2 \leq \frac{\gamma_0}{C(k+1)}. \tag{20.}$$

4. If the function J also satisfies the gradient dominance property with degree 2, then we have the linear convergence

$$J(K^n) - J(K^*) \leq (1 - 2\mu\alpha + \mu L\alpha^2)^n (J(K^0) - J(K^*)). \tag{21.}$$

Proof. This result is important, so a proof is included for illustrative purposes. Statement 1 can be proved using the continuity and coerciveness of $J(K)$ and is actually a direct consequence of proposition 11.12 in Reference 80.

Since J is twice continuously differentiable, we know that the function $\|\nabla^2 J(K)\|$ (with $\|\cdot\|$ being the operator norm) is continuous. By the Weierstrass theorem, we know that $\|\nabla^2 J(K)\|$ has to be bounded on the compact set $\mathcal{K}_\gamma$. We denote this uniform upper bound as L, and hence J is L-smooth on $\mathcal{K}_\gamma$. By the mean value theorem, Equation 18 holds as desired. This proves Statement 2.

The proof of Statement 3 is based on smoothness, and we will use standard arguments. Suppose we have chosen $\alpha = \frac{2}{L+\omega}$ for some positive constant $\omega > 0$. First, we need to show that given $K \in \mathcal{K}_{\gamma_0}$, the line segment connecting K and $K' = K - \alpha \nabla J(K)$ is also in $\mathcal{K}_{\gamma_0}$. By continuity of $\|\nabla^2 J(K)\|$ and $J(K)$, there exists a small constant $c > 0$ such that $\|\nabla^2 J(K)\| \leq L + \omega$ for all $K \in \mathcal{K}_{\gamma_0 + c}$. Denote the closure of the complement of $\mathcal{K}_{\gamma_0+c}$ as S_1. Obviously, $\mathcal{K}_{\gamma_0} \cap S_1$ is empty. Since $\mathcal{K}_{\gamma_0}$ is compact, we know that the distance between $\mathcal{K}_{\gamma_0}$ and S_1 is strictly positive. We denote this distance as δ. Let us choose $\tau = \min\{0.9\delta / \|\nabla J(K)\|_F, 2/(L+\omega)\}$. Clearly, the line segment between K and $(K - \tau \nabla J(K))$ is in $\mathcal{K}_{\gamma_0+c}$. Notice that $\|\nabla^2 J(K)\| \leq L + \omega$ for all $K \in \mathcal{K}_{\gamma_0+c}$, and hence we have

$$J(K - \tau \nabla J(K)) \leq J(K) + \langle \nabla J(K), K - \tau \nabla J(K) - K \rangle + \frac{L+\omega}{2} \|K - \tau \nabla J(K) - K\|_F^2,$$

which leads to $J(K - \tau \nabla J(K)) \leq J(K) + (-\tau + \frac{(L+\omega)\tau^2}{2}) \|\nabla J(K)\|_F^2$. As long as $\tau \leq 2/(L+\omega)$, we have $-\tau + \frac{(L+\omega)\tau^2}{2} \leq 0$ and $J(K - \tau \nabla J(K)) \leq J(K) \leq \gamma_0$. Hence, we have $K - \tau \nabla J(K) \in \mathcal{K}_{\gamma_0}$. Actually, it is straightforward to see that the line segment between K and $(K - \tau \nabla J(K))$ is in $\mathcal{K}_{\gamma_0}$ by varying τ.

The rest of the proof follows from induction. We can apply the same argument to show that the line segment between $(K - \tau \nabla J(K))$ and $(K - 2\tau \nabla J(K))$ is also in $\mathcal{K}_{\gamma_0}$. This means that the line segment between K and $(K - 2\tau \nabla J(K))$ is in $\mathcal{K}_{\gamma_0}$. Since $\tau > 0$, we only need to apply the above argument for finite times, and then will be able to show that the line segment between K and $(K - \alpha \nabla J(K))$ is in $\mathcal{K}_{\gamma_0}$ for $\alpha = \frac{2}{L+\omega}$. Now we can apply Equation 18 to show

the convergence result. Since $\|\nabla^2 J(K)\| \le L$ for all $K \in \mathcal{K}_{\gamma_0}$, we can use the mean value theorem to show

$$J(K') \le J(K) + \langle \nabla J(K), K' - K \rangle + \frac{L}{2}\|K' - K\|_F^2 = J(K) + \left(-\alpha + \frac{L\alpha^2}{2}\right)\|\nabla J(K)\|_F^2,$$

which can be summed over a finite window to get the desired convergence result.

Finally, we can combine the gradient dominance inequality with the above smoothness inequality to show $J(K') - J(K) \le -(2\mu\alpha - \mu L\alpha^2)(J(K) - J(K^*))$. This immediately leads to Equation 21, thus completing the proof. □

Next, we show that the convergence/complexity of the gradient descent method for LQR follows as a consequence of the above result.

3.2. Policy Optimization Theory for the Linear Quadratic Regulator

There are multiple ways to show the global convergence/complexity of the gradient descent method for the LQR problem (35, 36, 62, 81). In this section, we review one proof that is based on Theorem 1. Interestingly, the LQR cost is coercive, real analytic, and gradient dominant, so that Theorem 1 can be directly applied, though the problem is nonconvex in the parameter K (35, 76). Recall that the LQR cost can be computed as $J(K) = \mathrm{Tr}(P_K \Sigma_0)$, where $\Sigma_0 = \mathbb{E}x_0 x_0^\mathsf{T}$ and P_K satisfies $(A - BK)^\mathsf{T} P_K (A - BK) + Q + K^\mathsf{T} R K = P_K$. For simplicity, we assume here that $Q \succ 0$, following Reference 35.[4] The following result holds.

Lemma 1. The LQR cost satisfies the following properties: The cost function J is

1. real analytical and hence twice continuously differentiable;
2. coercive over the feasible set $\mathcal{K}$; and
3. μ-gradient dominant, with $\mu = \frac{2(\sigma_{\min}(\mathbb{E}x_0 x_0^\mathsf{T}))^2 \sigma_{\min}(R)}{\|\Sigma_{K^*}\|}$, where $\sigma_{\min}$ and $\|\cdot\|$ denote the smallest and largest singular values, respectively.

Proof. To prove Statement 1, notice that the analytical solution of the Lyapunov equation can be calculated as $\mathrm{vec}(P_K) = \left(I - (A - BK)^\mathsf{T} \otimes (A - BK)^\mathsf{T}\right)^{-1} \mathrm{vec}(Q + K^\mathsf{T} R K)$. Hence, P_K is a rational function of the elements of K. Then we know that J is a rational function of the elements of K. Therefore, J is real analytical and twice continuously differentiable.

To prove Statement 2, one can apply the contradiction argument in Reference 36. We refer readers to that work for details of this argument.

Finally, Statement 3 can be proved using the cost difference lemma (lemma 10 in Reference 35). Lemma 11 in Reference 35 provides one such argument.[5] □

It is worth mentioning that we can explicitly bound the smoothness constant L over any sublevel set of $J(K)$ in terms of problem parameters, which helps in establishing refined convergence rates for the gradient descent method. We are now ready to state the global convergence result for the gradient descent method for LQR.

Theorem 2. Consider the LQR PO problem with (Q, R) being positive definite, and apply the gradient method with the update rule $K^{n+1} = K^n - \alpha \nabla J(K^n)$. Suppose $K^0 \in \mathcal{K}$ is stabilizing. Then, with a step size satisfying $0 < \alpha \le 1/L_{K^0}$, where L_{K^0} denotes the smoothness

[4] The results can be generalized to the cases where $Q \succeq 0$ and even where Q is indefinite (82).

[5] The constant coefficient used in lemma 11 of Reference 35 can be slightly tightened to match the exact value of μ given in Lemma 1.

constant of $J(K)$ over the sublevel set of level $J(K^0)$, we have that (*a*) for all $n \geq 1$, K^n stabilizes the system [i.e., $\rho(A - BK^n) < 1$], and (*b*) the sequence $\{K^n\}$ converges to the global optimum of LQR at a linear rate as

$$J(K^n) - J(K^*) \leq \Big(1 - \mu\alpha\Big)^n \big(J(K^0) - J(K^*)\big),$$

where μ is the gradient dominance coefficient given in Lemma 1.

The proof of Theorem 2 follows Theorem 1 and Lemma 1. The above result requires an initial stabilizing controller $K^0 \in \mathcal{K}$. This is not an issue, since it is also known that one can obtain such stabilizing policies using PO methods (45–47, 83).

3.2.1. Zeroth-order optimization. In many cases, the exact gradient $\nabla J(K)$ is not available, especially when the dynamical system model is unknown. In the optimization and learning community, one method that has been actively studied is to estimate the gradient through the cost value, $J(K)$. For example, we can estimate the gradient using the following single-point zeroth-order gradient estimator:

$$\mathsf{G}_J(K; r, z) = \frac{d}{r} J(K + rz)\, z, \qquad z \sim \mathcal{Z}. \tag{22.}$$

Here, $r > 0$ is a positive parameter called the smoothing radius. We slightly abuse the notation by letting z denote the random perturbation, which is a d-dimensional random vector following the probability distribution $\mathcal{Z}$. Usually, $\mathcal{Z}$ is chosen to be either (*a*) the Gaussian distribution $\mathcal{N}(0, d^{-1}I)$ or (*b*) the uniform distribution on the unit sphere $\mathbb{S}_{d-1} := \{z \in \mathbb{R}^d : \|z\| = 1\}$, which we denote by $\mathrm{Unif}(\mathbb{S}_{d-1})$. For this single-point estimator, it can be shown that $\mathbb{E}_{z\sim\mathcal{Z}}[\mathsf{G}_J(K; r, z)] = \nabla J_r(K)$, where J_r is a smoothed version of J and the radius r controls the approximation accuracy. Besides the single-point estimator given in Equation 22, multipoint gradient estimators can be used to improve the convergence rate (84, 85).

The optimization and learning literature (e.g., 84, 86–89) has studied the properties and complexity of the zeroth-order methods under different settings [convex or nonconvex $J(\cdot)$, stochastic or deterministic optimization, etc.]. When using the zeroth-order methods for LQR, we need to pay extra attention to the following issues:

1. Feasible initial K^0: As in the exact gradient case, the initial K^0 should be feasible, i.e., stabilizing the system. If a system is unknown, this is challenging. Recent works have provided the convergence/complexity theory for using PO-based discount annealing methods to obtain initial stabilizing policies (45–47); the related issue of online regularizability from streaming control/state pairs has also been examined (90).
2. Impact of bias and variance of zeroth-order estimation on the feasibility and convergence of K^n: Though Theorem 2 ensures feasibility for gradient descent iterations, zeroth-order estimation introduces both bias and variance in the gradient evaluation. To address this, one can tune the parameter r and use the average of several single-point estimators.
3. Feasibility of the perturbed controller $K^n + rz$: We need to ensure that the iteration $K^n + rz$ is in $\mathcal{K}$. This limits the choices of random exploration rz (e.g., K^n should be at least strictly feasible to allow perturbations).
4. Evaluation of $J(K)$: When the cost is defined on the infinite time horizon (e.g., in the LQR problem), it is challenging to evaluate in practice; often, one can obtain only a finite-time truncated estimate for $J(K)$. Handling the truncation requires care, as it also introduces bias and variance in the estimator.

5. Dependence of the sample complexity on system parameters: Fazel et al. (35) and Malik et al. (37) showed that the sample complexity of LQR zeroth-order methods depends on the system parameters A, B, Q, and R and the initial controller K^0. If the system is ill conditioned, then the number of samples can potentially be quite large, as discussed in a recent work by Ziemann et al. (91). A concurrent survey by Tsiamis et al. (92) provided more general discussions on the interplay between statistical learning theory and control.

Due to space limitations, we refer readers to References 35, 37, and 40 for details on how the above issues were handled when implementing zeroth-order methods for LQR. There are also other data-driven PG estimation methods, such as the policy gradient theorem (7, 8) and iterative feedback tuning (93, 94); sample complexity for these methods is less understood.

3.2.2. Additional remarks. We end this section with a few remarks on other aspects of the LQR PO theory:

1. LQR with stochastic noise: The above convergence result extends to more general forms of LQR, e.g., with process noise as in Equation 4. The major change in the derivations is to replace the matrix $\mathbb{E}(x_0 x_0^\top)$ with the covariance of the process noise.
2. Natural policy gradient (NPG): NPG is a standard RL algorithm that enforces a Kullback–Leibler divergence constraint on the updated and the old policies (95). In the LQR setting, the deterministic counterpart of NPG is given as follows:

$$K^{n+1} = K^n - \alpha \nabla J(K^n) \Sigma_{K^n}^{-1}, \tag{23.}$$

   where Σ_{K^n} is the state correlation matrix for K^n (for details, see 35). For a discounted LQR problem with a stochastic Gaussian policy, the NPG update (which inverts the Fisher information matrix) exactly reduces to the above iterative scheme. The NPG method in Equation 23 also converges at a linear rate. If we denote $F^n = \nabla J(K^n)\Sigma_{K^n}^{-1}$, then we have $J(K^n) - J(K^*) \leq \frac{\|\Sigma_{K^*}\|}{\sigma_{\min}(R)} \operatorname{Tr}((F^n)^\top F^n)$, which can be combined with the cost difference lemma (see lemma 10 in Reference 35) to show the linear convergence of NPG.
3. Policy iteration and Kleinman's algorithm: An important variant of the gradient descent method is the Gauss–Newton method with the following iterations:

$$K^{n+1} = K^n - \alpha (B^\top P_{K^n} B + R)^{-1} \nabla J(K^n) \Sigma_{K^n}^{-1}, \tag{24.}$$

   which is equivalent to $K^{n+1} = K^n - 2\alpha \left(K^n - (B^\top P_{K^n} B + R)^{-1} B^\top P_{K^n} A\right)$. If $\alpha = \frac{1}{2}$, then this algorithm reduces to the policy iteration algorithm in the RL literature (96) or, equivalently, Kleinman's algorithm in the control literature (97, 98). We can show that the Gauss–Newton method has a global linear convergence rate for any $\alpha \leq \frac{1}{2}$. In addition, when $\alpha = \frac{1}{2}$, the above method has a superlinear local rate. This explains why policy iteration is typically fast on the LQR problem. Policy iteration can be implemented in a model-free manner via least-squares techniques (99). There are also sample complexity results for approximate policy iteration on the LQR problem (100).
4. Further extensions: The above LQR PO theory can also be extended to cover (*a*) time-varying/nonlinear systems, such as Markovian jump linear systems (56–59) and Lur'e systems (60); (*b*) more complicated RL algorithms, such as actor–critic (42, 43); and (*c*) different settings, including the continuous-time setting (38, 44, 81, 101), the multiplicative noise setting (50), and the finite-horizon setting (41).

3.3. Technical Challenges for Settings Beyond the Linear Quadratic Regulator

The global convergence of LQR PO relies heavily on several important properties of the cost function and feasible set. Here, we summarize the importance of four properties: (*a*) coerciveness of cost, (*b*) gradient dominance (Polyak–Łojasiewicz property) of cost, (*c*) smoothness of cost, and (*d*) the connectivity of the feasible set.

First, a key factor in LQR results is the coerciveness of the objective, as stated in Lemma 1. Coerciveness ensures that as long as the cost function value decreases, the controller K^n remains stabilizing, i.e., feasible. Furthermore, coerciveness ensures that the sublevel sets of the cost function are compact, which, together with the real analytical property of the cost, implies that the gradient of the cost is globally Lipschitz over any finite level set (i.e., the cost is globally smooth over its sublevel sets). This smoothness property serves as one of the pillars in nonconvex optimization analysis in determining the step size that sufficiently decreases the cost (see, e.g., 102). The coercive property makes the cost function a valid barrier function that explicitly regularizes the iterations to be feasible during the optimization processes. However, the cost is not necessarily coercive for other control problems, and only decreasing the cost value thus may no longer ensure the feasibility of the iterations.

Second, convergence to the global minimum of PG methods for LQR, especially with a linear convergence rate, relies heavily on the benign landscape property of gradient dominance (see Lemma 1). Together with the smoothness of the objective, the gradient dominance property (of degree 2; see Definition 1) naturally leads to a global convergence rate that can be linear (78). This benign property is a blessing for certain control problems (see Section 6) and does not necessarily hold in general.

Third, sometimes the cost may not be differentiable over the entire feasible set. For example, for optimal $\mathcal{H}_\infty$ control, the cost function can be nondifferentiable at stationary points (28, 103). The lack of smoothness causes difficulty for such PO problems.

Finally, another key to the success of LQR PO, as a local search approach, is that the feasible set, though nonconvex in general, is connected. This is important since local search algorithms typically cannot jump between connected components, and a single connected component must include the global optimum. Unfortunately, such connectivity is lost when extending PO to partially observable control systems, creating additional challenges toward establishing global convergence.

Next, we study several control problems where some (if not all) of these desired properties are lacking, and more careful and advanced analyses are needed in order to obtain global convergence guarantees for PO methods.

4. CASE II: POLICY OPTIMIZATION FOR RISK-SENSITIVE AND ROBUST CONTROL

In this section, we review the global convergence results of PO methods for the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem and the state-feedback $\mathcal{H}_\infty$ optimal control problem. For the mixed design problem, the main issue is the lack of coerciveness—i.e., the cost function close to the boundary of the feasible set may not approach infinity. For the $\mathcal{H}_\infty$ synthesis problem, the main difficulty is the lack of smoothness—i.e., the cost function may be nondifferentiable over some important points in the feasible set. We discuss how to modify PO algorithms to mitigate these issues and provably achieve global convergence. It is worth emphasizing that the idea of applying RL methods to solve $\mathcal{H}_\infty$ control is not new (104–106). This section focuses mainly on the recently developed global convergence theory for PO methods for such robust control tasks (48, 49, 54).

4.1. Policy Optimization for Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Design: Implicit Regularization

Recall the formulation of linear risk-sensitive and mixed design problems in Case II in Section 2. For simplicity, we use one common objective of the problem, which we restate as follows:

$$\min_K \quad J(K) := \mathrm{Tr}(P_K DD^\mathsf{T}) \qquad 25.$$

$$\text{subject to} \quad K \in \mathcal{K} \text{ in Equation 9 and}$$

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

The above cost function $J(K)$ is known to be differentiable over the feasible set. One may wonder whether the analysis for the LQR case can be tailored to establish the global convergence of the gradient descent method on the above mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem. However, the following lemma reveals the less desired landscape properties of the cost function.

Lemma 2 (nonconvexity and no coerciveness). The feasible set for the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem in Equation 25 is nonconvex. Moreover, the cost function given in Equation 25 is not coercive. In particular, as $K \to \partial\mathcal{K}$, where $\partial\mathcal{K}$ is the boundary of the constraint set $\mathcal{K}$, the cost $J(K)$ does not necessarily approach infinity.

The difference between the landscapes of LQR and mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control is illustrated in **Figure 1**. The cost function for mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control is not necessarily coercive and hence cannot serve as a barrier function over the feasible set by itself (for further discussion, see 48). The lack of coerciveness for the mixed design problem calls for a more careful analysis on maintaining the feasibility of the iterations during optimization. Zhang et al. (48) adopted the concept of implicit regularization to address this issue. Specifically, a PO algorithm is referred to as being implicitly regularized if the iterations $\{K_n\}$ generated by the algorithm remain in $\mathcal{K}$ without using projection. The implicit regularization property has been investigated in nonconvex optimization and machine learning, including training neural networks (107), phase retrieval (108), and matrix

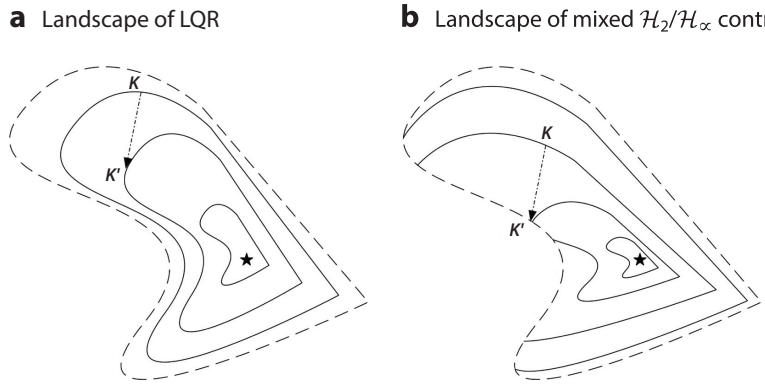


Figure 1

Comparison of the landscapes of LQR and mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design. The dashed lines represent the boundaries of the constraint sets $\mathcal{K}$. For (*a*) LQR, $\mathcal{K}$ is the set of all linear stabilizing state-feedback controllers; for (*b*) mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control, $\mathcal{K}$ is the set of all linear stabilizing state-feedback controllers satisfying an extra $\mathcal{H}_\infty$ constraint. The solid lines are the contour lines of the cost $J(K)$. K and K' denote the control gains of two consecutive iterations; $\star$ denotes the global optimizer. Abbreviation: LQR, linear quadratic regulator. Figure adapted from Reference 48 with permission from SIAM.

completion (109, 110). We emphasize that implicit regularization is a feature of both the problem and the algorithm. Next, we discuss two PO algorithms that are guaranteed to stay in the feasible set and achieve global convergence, thanks to the implicit regularization property.

4.1.1. Algorithms. For ease of exposition, we introduce the following notation:

$$\tilde{P}_K := P_K + P_K D(\gamma^2 I - D^\mathsf{T} P_K D)^{-1} D^\mathsf{T} P_K, \quad E_K := (R + B^\mathsf{T}\tilde{P}_K B)K - B^\mathsf{T}\tilde{P}_K A, \tag{26.}$$

$$\Delta_K := \sum_{t=0}^{\infty} \left[(I - \gamma^{-2} P_K DD^\mathsf{T})^{-\mathsf{T}}(A - BK)\right]^t D(I - \gamma^{-2} D^\mathsf{T} P_K D)^{-1} D^\mathsf{T} \left[(A - BK)^\mathsf{T}(I - \gamma^{-2} P_K DD^\mathsf{T})^{-1}\right]^t. \tag{27.}$$

Then we have the explicit gradient formula $\nabla J(K) = 2((R + B^\mathsf{T}\tilde{P}_K B)K - B^\mathsf{T}\tilde{P}_K A)\Delta_K$, and we can show that the following two PO methods enjoy the implicit regularization property (48):

$$\text{NPG:} \quad K^{n+1} = K^n - \eta \nabla J(K^n)\Delta_{K^n}^{-1} = K^n - 2\eta E_{K^n}, \tag{28.}$$

$$\text{Gauss–Newton:} \quad K^{n+1} = K^n - \eta(R + B^\mathsf{T}\tilde{P}_{K^n}B)^{-1}\nabla J(K^n)\Delta_{K^n}^{-1} = {}'K^n - 2\eta(R + B^\mathsf{T}\tilde{P}_{K^n}B)^{-1}E_{K^n}, \tag{29.}$$

where $\eta > 0$ is the step size. The updates resemble the PO updates for LQR, as discussed in Section 3.2, but with P_K replaced by $\tilde{P}_K$. The natural PG update is related to the gradient over a Riemannian manifold, while the Gauss–Newton update can be viewed as a special case of the quasi-Newton update.

4.1.2. Global convergence guarantees. The natural PG and Gauss–Newton updates in Equations 28 and 29 enjoy the implicit regularization property, formalized as below.

Theorem 3 (implicit regularization). For any iteration $K = K^n \in \mathcal{K}$ [i.e., $\rho(A - BK) < 1$ and $\|(C - EK)(zI - A + BK)D\|_\infty < \gamma$], suppose that the step size η satisfies (*a*) $\eta \le 1/(2\|R + B^\mathsf{T}\tilde{P}_K B\|)$ for NPG in Equation 28 and (*b*) $\eta \le 1/2$ for Gauss–Newton in Equation 29. Then the next iteration $K' = K^{n+1}$ obtained from Equations 28 and 29 also lies in $\mathcal{K}$.

The proof of Theorem 3 can be found in section 5 of Reference 48, which has a deep connection with the bounded real lemma (73, 111, 112). Theorem 3 shows that the robustness of the controller is preserved when certain policy search directions are used. Intuitively, the natural PG and Gauss–Newton methods somehow exploit the information of P_K to avoid the directions that may lead to infeasibility. This implicit regularization property is due to a combination of a certain nonconvex objective (mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control) and certain algorithms (natural PG and Gauss–Newton). With this property in hand, we are ready to state the global convergence result.

Theorem 4 (global convergence and local faster rates). Suppose that $K^0 \in \mathcal{K}$ and $\|K^0\| < \infty$. Then, under the step-size choices[6] as in Theorem 3, updates in Equations 28 and 29 both converge to the global optimum $K^* = (R + B^\mathsf{T}\tilde{P}_{K^*}B)^{-1}B^\mathsf{T}\tilde{P}_{K^*}A$, in the sense that $\sum_{n=1}^{N} \|E_{K^n}\|_F^2/N = O(1/N)$. Moreover, if $DD^\mathsf{T} > 0$, then under the same step-size choices, both updates converge to the optimal K^* with a locally linear rate—i.e., the objective $\{J(K^n)\}$

[6]For natural PG, it suffices to require the step size $\eta \le 1/(2\|R + B^\mathsf{T}\tilde{P}_{K_0}B\|)$ for the initial K^0.

converges to $J(K^*)$ with a linear rate. In addition, if $\eta = \frac{1}{2}$, then the Gauss–Newton update in Equation 29 converges to K^* with a locally Q-quadratic rate.

The proof of Theorem 4 can be found in section 5 of Reference 48. The theorem shows that although the problem is nonconvex and noncoercive, certain PO methods can still find the globally optimal solution of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control at globally sublinear and locally superlinear rates. The global rate is sublinear, in contrast to the linear one for LQR, as the global gradient dominance property does not necessarily hold here.

Finally, we remark that, interestingly, Zhang et al. (48) also numerically compared the computation efficiency of PO methods and existing solvers for mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control (see, e.g., 33). It has been shown that PO methods can indeed be much faster than these existing solvers (though note that these solvers can handle more general cases, such as the output feedback case), especially for large-scale dynamical systems. This justifies the desired scalability of PO methods for control synthesis (for more numerical examples, see 48).

4.1.3. Model-free implementations: connections to dynamic games and adversarial reinforcement learning. There is a fundamental connection between mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control and linear quadratic dynamic games (111). This connection will not only allow us to implement the PO algorithms using model-free adversarial RL techniques, but also enable the development of PO methods for solving these dynamic games.

As LQR can be viewed as the benchmark for single-agent RL in continuous space, linear quadratic dynamic games serve as the benchmark for studying multiagent RL. Indeed, zero-sum linear quadratic games have been investigated as fundamental settings in multiagent RL (82, 113–117). Specifically, consider a zero-sum dynamic game with linear dynamics $x_{t+1} = Ax_t + Bu_t + Dw_t$. The objective of player 1 (player 2) is to minimize (maximize) the value function $\mathcal{C} := \mathbb{E}_{x_0\sim\mathcal{D}}\big[\sum_{t=0}^{\infty}(x_t^\top Q x_t + u_t^\top R^u u_t - w_t^\top R^w w_t)\big]$, where $x_0 \sim \mathcal{D}$ for some distribution $\mathcal{D}$, and (Q, R^u, R^w) are positive definite matrices. It is known that the Nash equilibrium—the solution concept for the problem—can be achieved with state-feedback policy classes; that is, there exists a pair (K^*, L^*) such that the Nash equilibrium satisfies $u_t^* = -K^* x_t$ and $v_t^* = -L^* x_t$ (111, 118). Hence, one can parameterize the controllers using matrices (K, L) and solve for $\min_K \max_L \mathcal{C}(K, L)$, where $\mathcal{C}$ is the accumulated cost under this pair (K, L). This leads to a multiagent PO problem.

Intriguingly, the Nash equilibrium to the game is provided by the solution to a specific mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem (111). With this connection, the natural PG and Gauss–Newton methods in Equations 28 and 29 can be equivalently transformed into provably convergent double-loop PO algorithms for the above linear quadratic game. Related algorithmic developments have been documented by Zhang et al. (114) and Bu et al. (119). The game formulation for mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control is powerful in that these double-loop variants of the natural PG and Gauss–Newton methods can be implemented in a model-free manner. Zhang et al. (51, 114) and Keivan et al. (55) provided detailed discussions of model-free implementations and related sample complexity results.

An important issue in RL is the simulation-to-real gap. One common remedy is to use robust adversarial RL algorithms that jointly learn a protagonist and an adversary, where the former learns to robustly perform the control tasks under the disturbances created by the latter (120, 121). Policy-based robust adversarial RL methods can be viewed as model-free variants of multiagent PO methods for dynamic games. Therefore, the PO theory for mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control can also be applied to study the properties of robust adversarial RL algorithms in the linear quadratic setting. Zhang et al. (49) provided more details about this connection and results.

Remark 1. The inner-loop subroutine in both zero-sum dynamic games and robust adversarial RL reduces to a generalized LQR problem whose state cost matrix Q is not positive

semidefinite. The formulation of the indefinite LQR is similar to that in Section 2, except that the Q and R matrices are symmetric but indefinite. Then the cost may not be coercive, and the descent of the cost does not ensure the stability of the iterations (82). Nevertheless, global convergence of PO methods can be established (see 82).

4.2. Policy Optimization for $\mathcal{H}_\infty$ State-Feedback Synthesis: Nonsmoothness and Convergence

In this section, we consider the state-feedback $\mathcal{H}_\infty$ optimal control problem. Classical convex approaches for this task require reparameterizing the problem into a higher-dimensional convex domain (22, 73, 122). By contrast, we view this problem as a benchmark of PO for robust control. Here, we discuss how to provably find the optimal $\mathcal{H}_\infty$ controller in the policy space directly.

Recall that for the PO formulation of $\mathcal{H}_\infty$ state-feedback synthesis, the cost function $J(K)$ is given by Equation 10, and the feasible set $\mathcal{K}$ is specified by Equation 3. A main technical challenge here is that this $\mathcal{H}_\infty$ cost can be nondifferentiable at some important feasible points, e.g., the optimal points (28, 32, 33, 103). From Equation 10, we can see that this cost function is subject to two sources of nonsmoothness: The largest eigenvalue for a fixed frequency ω is nonsmooth, and the optimization step over $\omega \in [0, 2\pi]$ is also nonsmooth. We also know that the feasible set from Equation 3 is nonconvex. Hence, the resultant PO problem for $\mathcal{H}_\infty$ state-feedback synthesis is nonconvex and nonsmooth. A large family of nonsmooth $\mathcal{H}_\infty$ policy search algorithms have been developed based on the concept of the Clarke subdifferential (28, 32, 33, 103). However, the global convergence theory of PO methods for the $\mathcal{H}_\infty$ state-feedback synthesis was not established until very recently. Next, we review such global convergence results from Guo & Hu (54).

First, we introduce a few concepts related to subdifferential of nonconvex functions. A function $J : \mathcal{K} \to \mathbb{R}$ is locally Lipschitz if for any bounded $S \subset \mathcal{K}$ there exists a constant $L > 0$ such that $|J(K) - J(K')| \le L\|K - K'\|_F$ 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. We define the Clarke subdifferential as $\partial_C J(K) := \mathrm{conv}\{\lim_{i\to\infty} \nabla J(K_i) : K_i \to K,\ K_i \in \mathrm{dom}(\nabla J) \subset \mathcal{K}\}$, where conv denotes the convex hull. For any given direction V (which has the same dimension as K), the generalized Clarke directional derivative of J is defined as

$$J^\circ(K, V) := \lim_{K' \to K} \sup_{t \searrow 0} \frac{J(K' + tV) - J(K')}{t}. \tag{30.}$$

By contrast, the (ordinary) directional derivative is defined as follows (when it exists):

$$J'(K, V) := \lim_{t \searrow 0} \frac{J(K + tV) - J(K)}{t}. \tag{31.}$$

In general, the Clarke directional derivative can be different from the (ordinary) directional derivative, which may not even exist for some feasible points. 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., $J'(K, V) = J^\circ(K, V)$. The following result holds for the $\mathcal{H}_\infty$ objective function.

Proposition 1. Let $\mathcal{K}$ be nonempty. Then the $\mathcal{H}_\infty$ objective function defined by Equation 10 is locally Lipschitz and subdifferentially regular over the stabilizing feasible set $\mathcal{K}$.

The above result is well known (for further explanation, see 54). Consequently, the Clarke subdifferential for the $\mathcal{H}_\infty$ objective function is well defined for all $K \in \mathcal{K}$. We say that $K^\dagger$ is a Clarke stationary point if $0 \in \partial_C J(K^\dagger)$. The subdifferentially regular property guarantees that the

directional derivatives at any Clarke stationary points $J'(K^\dagger, V)$ are always nonnegative. Since $\mathcal{K}$ is open, the global minimum has to be a Clarke stationary point. Searching Clarke stationary points provably requires advanced subgradient algorithms, since generating a good descent direction for nonsmooth optimization is nontrivial. The concept of the Goldstein subdifferential (123) is relevant and stated below.

Definition 2 (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:

$$\partial_\delta J(K) := \mathrm{conv}\left\{\cup_{K' \in \mathbb{B}_\delta(K)} \partial_C J(K')\right\}, \tag{32.}$$

where $\mathbb{B}_\delta(K)$ denotes the δ-ball around K. It is implicitly assumed that $\mathbb{B}_\delta(K) \subset \mathcal{K}$.

Importantly, the minimal norm element of the Goldstein subdifferential generates a good descent direction. The minimal norm element in $\partial_\delta J(K)$, denoted as F, will satisfy $J(K - \delta F/\|F\|_F) \leq J(K) - \delta\|F\|_F$, if we have $\partial_\delta J(K) \subset \mathcal{K}$. This fact has inspired the developments of Goldstein's subgradient method (123) and related variants for nonsmooth $\mathcal{H}_\infty$ control (32, 33, 103). Recently, Guo & Hu (54) proved that Goldstein's subgradient method can be guaranteed to find the global minimum of the $\mathcal{H}_\infty$ state-feedback synthesis problem despite the nonconvexity of the feasible set. We summarize this result as follows.

Theorem 5. Suppose that (Q, R) are positive definite and the pair (A, B) is stabilizable. Let $J^* = \min_{K \in \mathcal{K}} J(K)$. For $\mathcal{H}_\infty$ state-feedback synthesis, the following statements hold:

1. The $\mathcal{H}_\infty$ objective function defined by Equation 10 is coercive over $\mathcal{K}$.
2. For any $K \in \mathcal{K}$ satisfying $J(K) > J^*$, there exists $V \neq 0$ such that $J'(K, V) < 0$.
3. Any Clarke stationary points of the $\mathcal{H}_\infty$ objective function are global minima.
4. The sublevel set $\mathcal{K}_\gamma$ is always compact. There is a strict separation between $\mathcal{K}_\gamma$ and $\mathcal{K}^c$ (which is the complement of the feasible set $\mathcal{K}$). In other words, we have $\mathrm{dist}(\mathcal{K}_\gamma, \mathcal{K}^c) > 0$.
5. Suppose $K^0 \in \mathcal{K}$. Denote $\Delta_0 := \mathrm{dist}(\mathcal{K}_{J(K^0)}, \mathcal{K}^c) > 0$. Choose $\delta^n = \frac{0.99\Delta_0}{n+1}$ for all n. Then Goldstein's subgradient method $K^{n+1} = K^n - \delta^n F^n/\|F^n\|_F$, with F^n being the minimum norm element of $\partial_{\delta^n} J(K^n)$, is guaranteed to stay in $\mathcal{K}$ for all n. In addition, we have $J(K^n) \to J^*$ as $n \to \infty$.

The coerciveness can be proved using the positive definiteness of (Q, R). Statement 2 follows from the convex parameterization for $\mathcal{H}_\infty$ state-feedback synthesis, and we further discuss this point in Section 6. Statement 3 can be proved by combining Statement 2 and the subdifferential regular property. Statement 4 is a consequence of Statement 1. Then one can combine the descent property of Goldstein's subgradient method and Statement 4 to prove the convergence result in Statement 5. Due to some subtlety of nonsmooth nonconvex optimization, the sample complexity of PO on nonsmooth $\mathcal{H}_\infty$ synthesis remains unknown. Guo & Hu (54) have also discussed model-free implementations and related issues.

Remark 2. It is worth mentioning that PO has also been investigated in control problems with robustness and risk-sensitivity concerns other than the LEQG/$\mathcal{H}_\infty$ settings discussed in this survey (see 50, 52, 55, 115, 124–129).

5. CASE III: POLICY OPTIMIZATION WITH PARTIAL OBSERVATIONS

In this section, we examine the more challenging case of control with partial observation. When the system's state is not directly measured, there is an intricate balance between the achievable

control performance and the class of controllers used in PO. Depending on the policy parameterization, the optimization landscape can become quite different. We first survey several recent results on the optimization landscape of LQG PO, and then point out some of the subtle aspects for more general output feedback and structured synthesis problems.

5.1. Policy Optimization for Linear Quadratic Gaussian Control: The Optimization Landscape

To characterize the performance of PO algorithms such as PG methods for LQG control, it is necessary to understand the landscape of the associated PO formulation in Equation 1, with the cost function given in Equation 15 and the feasible set given in Equation 14. Following the standard setup in the literature, we assume that the pairs (A, B) and $(A, W^{\frac{1}{2}})$ are controllable and that the pairs (C, A) and $(Q^{\frac{1}{2}}, A)$ are observable. It has been shown that with these assumptions, the set of stabilizing controllers $\mathcal{K}$ is nonempty, open, and unbounded and can be nonconvex. Moreover, the cost function $J(K)$ is real analytic on the underlying set $\mathcal{K}$.

However, beyond these properties, until recently little was known about the geometric and analytical properties of the PO formulation of LQG control. We will mainly summarize results on the optimization landscape of LQG control from Zheng et al. (63), especially with respect to the connectivity of the stabilizing set $\mathcal{K}$ and the structure of the stationary points; several related extensions can be found in other works (64–67). Before introducing the results, we discuss a special structure for LQG control with the state-space dynamical controller parameterization in Equation 11.

It is known that the optimal feedback controller is unique in the frequency domain (73, theorem 14.7). However, in the time domain, this controller is not unique: Consider the similarity transformation for the state-space form of the controller, and note that the two controller parameterizations,

$$K = \begin{bmatrix} 0 & C_K \\ B_K & A_K \end{bmatrix} \quad \text{and} \quad \mathcal{T}(T, K) := \begin{bmatrix} 0 & C_K T^{-1} \\ T B_K & T A_K T^{-1} \end{bmatrix},$$

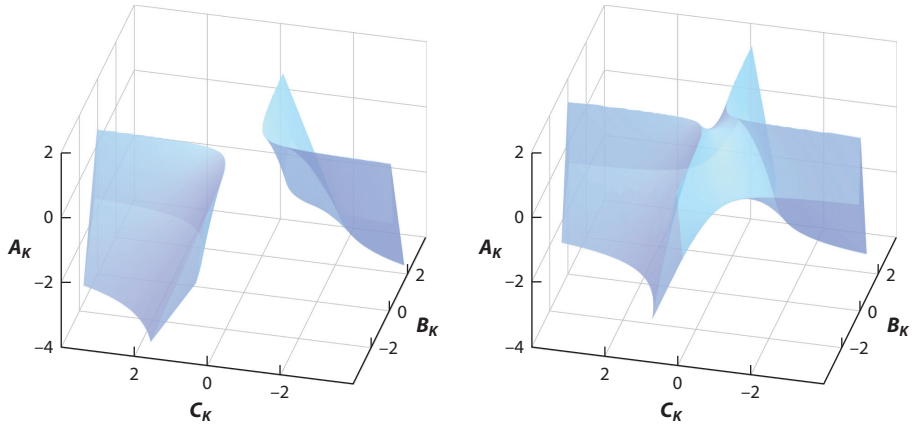
where T is an invertible matrix, have identical input–output behavior regardless of the choice of T. Thus, the cost is invariant with respect to this similarity transformation. Besides this invariance, when a controller K is nonminimal [i.e., (A_K, B_K) is not controllable or (A_K, C_K) is not observable], one can use model reduction to remove the uncontrollable/unobservable modes while keeping the cost the same.

We will now summarize the main results from Zheng et al. (63). First, we have the following theorem on the connectivity of $\mathcal{K}$.

> **Theorem 6.** The set $\mathcal{K}$ has at most two path-connected components. When $\mathcal{K}$ has two connected components, these components are diffeomorphic under the similarity transformation $\mathcal{T}(T, \cdot)$ for any invertible matrix with $\det T < 0$.

On a conceptual level, the proof is based on a convex reparameterization of the LQG problem. **Figure 2** shows two examples for $\mathcal{K}$, with one or two connected components.

For PG algorithms and other local search methods, the connectivity of the domain (the set of stabilizing controllers) is important since there are no jumping iterations between different connected components. Nevertheless, in light of Theorem 6 and the fact that the similarity transformation does not change the input–output behavior of a controller, it makes no difference to search over either path-connected component in $\mathcal{K}$ even if $\mathcal{K}$ is not path-connected. In fact, one can further show that any strict sublevel sets of the LQG PO problem have very similar



a System parameters: $A = 3/2$, $B = 1$, $C = 1$ **b** System parameters: $A = 2/3$, $B = 1$, $C = 1$

Figure 2

Two examples of the feasible set $\mathcal{K}$ for LQG control: (*a*) a disconnected feasible set and (*b*) a connected feasible set. Abbreviation: LQG, linear quadratic Gaussian.

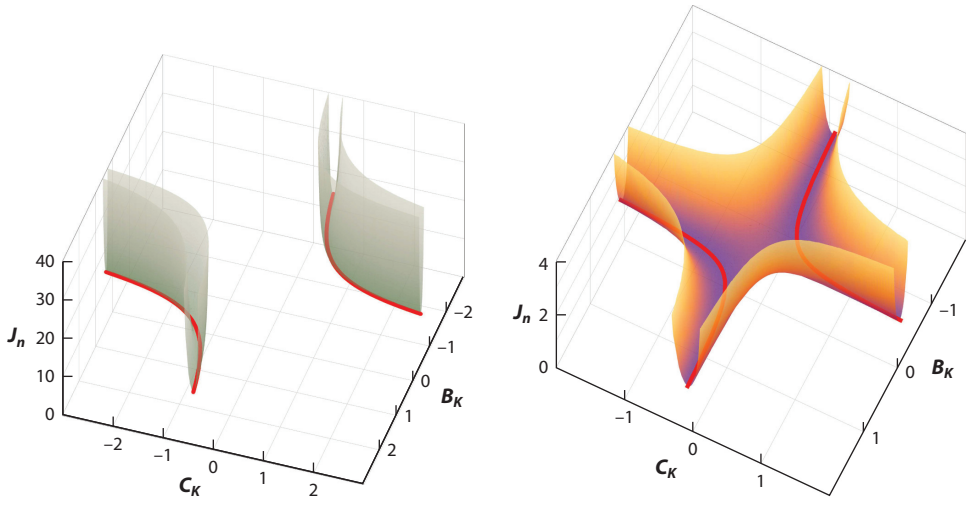
connectivity properties (67). Such observations are encouraging for devising gradient-based local search algorithms for LQG control.

Though the LQG problem has a nice property in terms of the connectivity of $\mathcal{K}$ (given in Equation 14), its optimization landscape is otherwise more complicated than cases we saw earlier. **Figure 3** shows two examples for the LQG cost $J(K)$. First, it is easy to see from the figure that LQG control has nonunique and nonisolated global optima in the state-space form. This feature often adds difficulty in establishing the convergence of PO methods. Second, it is also straightforward to show that the LQG cost is noncoercive. One way to see this is to consider the similarity transformation using λI. By letting $\lambda \to \infty$, we see that $|B_K| \to \infty$, but the cost remains unchanged. Noncoerciveness makes it challenging to establish convergence even to stationary points. Lastly, LQG control could have saddle points that are not optimal. Moreover, if K is a stationary point, then all of its similar controllers $\mathcal{T}(T, K)$ are also stationary points for any nonsingular T. Also, if $K \in \mathcal{K}$ is a nonminimal stationary point, then its minimal reduction could generate an infinite number of additional stationary points. Consequently, it is nontrivial to find an optimal controller, or even certify an optimal controller, through PG methods. Nevertheless, there is one clean case for the certification of the globally optimal points.

Theorem 7. All minimal stationary points[7] $K \in \mathcal{K}$ in the LQG problem in Equation 15 are globally optimal, and they are related to each other by a similarity transform.

The above results indicate that when running the PG methods, if the iterations converge to a minimal stationary point, then a globally optimal controller has been found. However, if the stationary points are nonminimal, then one cannot say much about the optimality of these stationary points. Indeed, Zheng et al. (63) provided examples showing the existence of LQG saddle points. Though there have been recent developments in perturbed gradient methods that can be guaranteed to escape strict saddle points (saddle points with an indefinite Hessian—i.e., an escape direction exists in the second order), there exist LQG instances with saddle points whose

[7] Here, the term minimal refers to being both observable and controllable as a dynamical system.



a System parameters: $A = 3/2$, $B = 1$, $C = 1$ **b** System parameters: $A = 2/3$, $B = 1$, $C = 1$

Figure 3

Nonisolated and disconnected globally optimal LQG controllers. In both cases, we set $Q = 1$, $R = 1$, $V = 1$, and $W = 1$. (*a*) LQG cost for the system in **Figure 2*a*** when fixing $A_K = -0.07318$, for which the set of globally optimal points $\{(B_K, C_K) | B_K C_K = -0.13684\}$ has two connected components. (*b*) LQG cost for the system in **Figure 2*b*** when fixing $A_K = -0.67360$, for which the set of globally optimal points $\{(B_K, C_K) \mid B_K C_K = -1.18113\}$ has two connected components. Abbreviation: LQG, linear quadratic Gaussian.

Hessian is degenerate. These observations pose challenges in analyzing the performance of PG methods applied to LQG control.

A recent work by Zheng et al. (130) introduced a novel perturbed policy gradient method that is capable of escaping various bad stationary points (including high-order saddles). Based on the specific structure of LQG control, this paper uses a reparameterization procedure that converts the iterate from a high-order saddle to a strict saddle, from which the standard randomly perturbed gradient descent method can escape efficiently. It also characterizes the high-order saddles that the proposed algorithm can escape; however, there is still a lack of an end-to-end theorem to characterize the iteration complexity of the algorithm. It remains an open challenge to analyze the performance of PG methods on LQG control by (*a*) establishing conditions under which the algorithms will converge to, at least, stationary points; (*b*) designing effective ways to escape nonoptimal stationary points (at least saddle points); (*c*) characterizing the algorithm complexity of the designed algorithms; and (*d*) developing sample-based methods and analyzing the sample complexity. It is also worth mentioning that Umenberger et al. (68) proved the global convergence of PO for a simpler estimation problem. This topic is discussed further in Section 6.

5.2. Output Feedback and Structured Control

We now shift our attention to another class of synthesis problems with partial observations, namely, output feedback and structured control. First, we would like to point out that policy synthesis on partially available data (not necessarily the underlying state) is of great interest in applications, particularly for large-scale systems. For example, in decentralized control, stabilizing feedback with a particular sparsity pattern is desired; in the output feedback case, one aims to design a stabilizing policy that can be factored with its right multiplicand as the observation map. These problems can be conveniently formalized in the form of Equation 1, where $\mathcal{K}$

becomes a subset of stabilizing (static or dynamic) feedback policies: For both output feedback and structured synthesis, $\mathcal{K}$ is a linearly constrained subset of stabilizing feedback gains. The PO perspective adopted in this survey then immediately offers an algorithm for these problems, namely, a projected first-order update.[8] A natural question is whether such an intuitive generalization has any theoretical guarantees of convergence; the short answer, however, is negative. We now summarize some of our current understanding of why this is the case, intermingled with some more encouraging results.

A major obstacle in guaranteeing convergence to the global optimum is due to the geometry of the corresponding set $\mathcal{K}$—and not only its nonconvexity inherited from the set of stabilizing controllers. Rather, due to the intricate geometry of this set, its intersection with linear subspaces can result in disconnected sets; an analogous phenomenon in the case of LQG control was examined in the previous section. This is a known fact from classical control in the context of output feedback and the root locus method, where the (scalar) feedback gain can undergo intervals of being stabilizing or not; an example is shown in **Figure 4*a***. However, it is surprising that, even for single-input, single-output systems, the number of such connected components was not explicitly characterized until recently. A PO perspective on control synthesis only makes this observation more compelling.

Theorem 8. The set of stabilizing output feedback gains for an n-dimensional single-input, single-output system, when nonempty, has at most $\lceil n/2 \rceil$ connected components.

Bu et al. (131) provided the proof of this result (as well as its analogue for the continuous-time case) and the precise characterization of these intervals. Less is known about the number of connected components for multi-input, multi-output feedback, knowledge of which could be useful for initializing PO algorithms. Nevertheless, some topological characterizations of these sets, including sufficient conditions for the connectedness of structured stabilizing gains, have been obtained in the literature (61, 131, 133).

The above topological insights are important in the context of policy search updates when they are required to stay stabilizing. As such, for general structured (including output feedback) synthesis, it is judicious to instead examine convergence to local optima or stationary points. Bu et al. (36) made the first observations in this direction, showing that projected gradient descent has a sublinear convergence to the (first-order) stationary point of structured LQR synthesis; Fatkhullin & Polyak (62) have reported analogous results for output feedback synthesis. Li et al. (40) examined sample-based learning methods for reaching a first-order stationary point using zeroth-order methods for structured synthesis. The sublinear convergence to a stationary point—the moment we impose a linear constraint on the set of stabilizing feedback gains—only hints at the fact that we are not fully utilizing the underlying geometry of the feedback synthesis problem.

These observations have motivated a new line of work on structured synthesis that is also in line with the natural gradient iteration and quasi-Newton method for LQR presented earlier. The key missing insight pertains, of course, to the Hessian and the Riemannian geometry of the set of stabilizing feedback gains as well as their linearly constrained subsets. As it turns out, LQR PO is closely related to iterative approaches for solving the Riccati equation that have subsequently been adopted for data-driven setups. In particular, it can be shown that what is known as the Hewer algorithm for LQR (98) is really a realization of a quasi-Newton update for a particular choice of a step size. In our desire to understand the fundamental limitation of PO for control synthesis,

[8]The projection is used to enforce sparsity/structure patterns. The projection does not involve stability/robustness concerns.

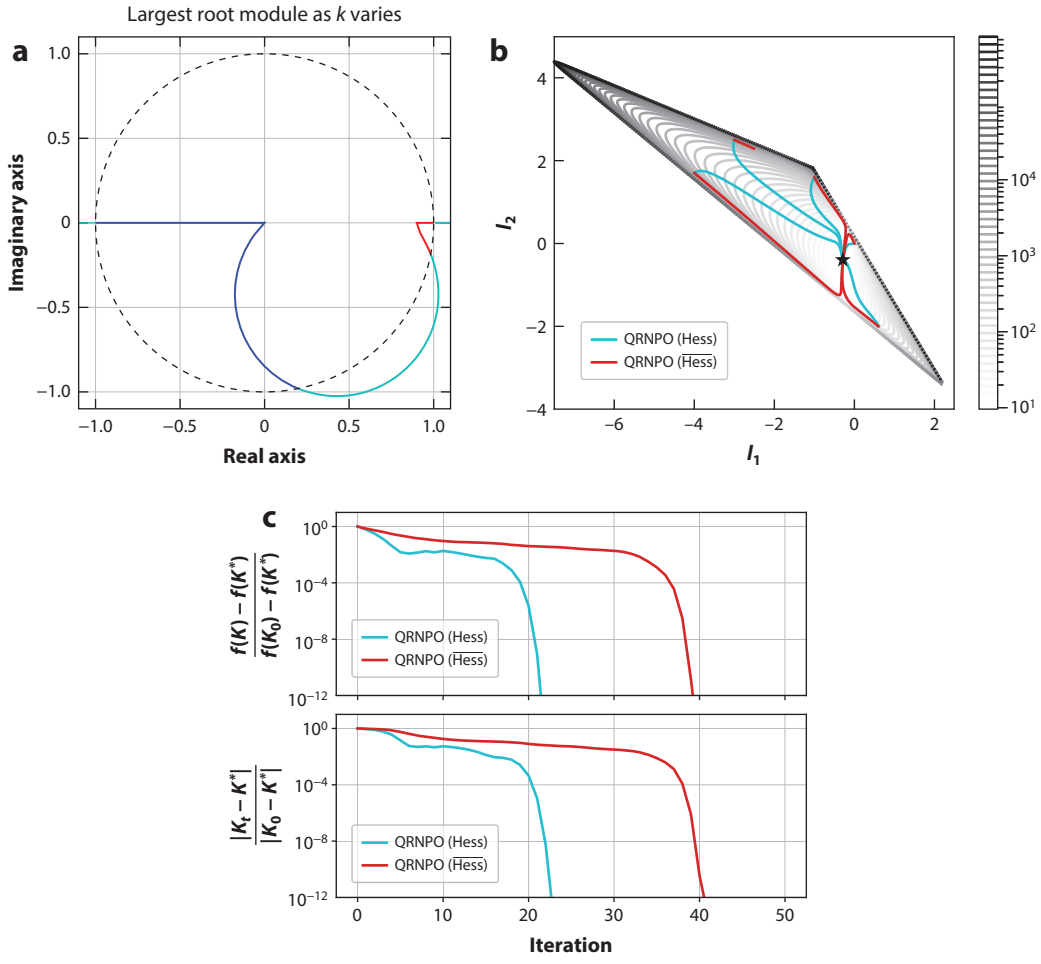


Figure 4

(*a*) The largest modulus root of a (discrete-time) feedback system, with the blue and red segments corresponding to the stabilizing intervals. (*b*) The geometry of output feedback synthesis (feedback gains shown along each axis) with respect to two distinct classes of algorithms (see 132). Hess denotes a scaled Newton update with a Hessian using the Riemannian geometry of $\mathcal{K}$, and $\overline{\text{Hess}}$ denotes a scaled Newton update with a Hessian using the Euclidean geometry of $\mathcal{K}$. (*c*) The average performance of the two algorithms on 100 random instances of the static output feedback problem. Talebi & Mesbahi (132) reported similar results for structured synthesis. Abbreviation: QRNPO, quasi-Riemannian Newton policy optimization. Panel *a* adapted from Reference 131 with permission from IEEE.

it is thus relevant to characterize the Newton update on the set of stabilizing feedback gains as well as its linearly constrained subset. This more geometric question is still relevant in the context of first-order methods, as it provides fundamental insights into how to recover a linear (or even quadratic) rate of guaranteed convergence to stationary or even locally optimal points for PO methods for structured synthesis.

Talebi & Mesbahi (132) have thoroughly analyzed these topics, including proper construction of the Hessian for the LQR problem through the intrinsic Riemannian connection, how to extend this Hessian to linearly constrained subsets $\mathcal{K}$, and how to choose the step size in the corresponding iterations to remain stabilizing. The results of this analysis include the following.

Theorem 9. Suppose that K^* is a nondegenerate local minimum of LQR on the linear constrained subset $\mathcal{K}$. Then there is a neighborhood around K^* and a positive scaling for which the scaled Riemannian Newton policy update remains stabilizing and converges to K^* at a linear and, eventually, quadratic rate.

Figure 4*c* shows a representative scenario for the output feedback problem. Ensuring stability during the course of these iterations, which is particular to control synthesis, often makes the analysis of these algorithms more intricate.

6. THE ROLE OF CONVEX PARAMETERIZATION

There is a large body of literature on the reparameterization of various control problems to represent them as convex problems (22, 24). In this section, we discuss the connections between such convex approaches and PO, showing that linear matrix inequality formulations for control design lead to desired landscape properties for PO.

We have seen successful applications of PO to a range of control problems in the previous sections, in many cases achieving the globally optimal policy. A natural question is whether there is a unified approach to determining when stationary points for PO are global minima. In this section, we revisit the gradient dominance property given in Definition 1 and provide a unified framework to show that some related inequality holds for a large family of PO problems, despite the nonconvexity of the cost $J(K)$ as a function of K. This viewpoint gives insights into the mysterious emergence of gradient dominance (or the Polyak–Łojasiewicz property) in various control problems that are nonconvex in K, providing a general tool to determine when stationary points for nonconvex PO problems are actually global minima.

Intuitively, the gradient dominance property implies that $J(K)$ is close to the optimal value for any K with small gradient norm, from which one can directly conclude nice optimization landscape/convergence properties.[9] Our goal in this section is to show how to use the existence of convex parameterizations, together with important additional assumptions about the map between the variables in the nonconvex and convex problems, to establish such a desired property or some closely related variant for the nonconvex $J(K)$.

We begin by considering an abstract description of the following pair of problems:

$$\min_{K} \quad J(K) \quad \text{subject to } K \in \mathcal{K}, \tag{33.}$$

$$\min_{L,P,Z} \quad f(L,P,Z) \quad \text{subject to} \quad (L,P,Z) \in \mathcal{S}, \tag{34.}$$

where $\mathcal{K}$ describes the set of desired controllers (typically the set of stabilizing controllers), and $\mathcal{S}$ captures the appropriate constraint sets (typically characterized by some linear matrix inequalities), which are determined for each problem case (see examples below). The following key assumption about the pair of problems in Equations 33 and 34 is critical for Theorem 10.

Assumption 1. The feasible set $\mathcal{S}$ is a convex set, and the function $f(L, P, Z)$ is convex, bounded, and differentiable over $\mathcal{S}$. We assume that any feasible point $(L, P, Z) \in \mathcal{S}$ satisfies $P \succ 0$. In addition, we assume that for all $K \in \mathcal{K}$, we can express $J(K)$ as follows:[10]

$$J(K) = \min_{L,P,Z} \; f(L,P,Z)$$
$$\text{subject to } (L,P,Z) \in \mathcal{S}, \; LP^{-1} = K.$$

[9] Convergence rates will depend on the values of the degree p (for more properties, see 79). In particular, $p = 1$ gives a sublinear convergence rate, and $p = 2$ gives a linear rate.

[10] Note that this assumption needs to hold for all feasible points in the two domains, not only at the optima.

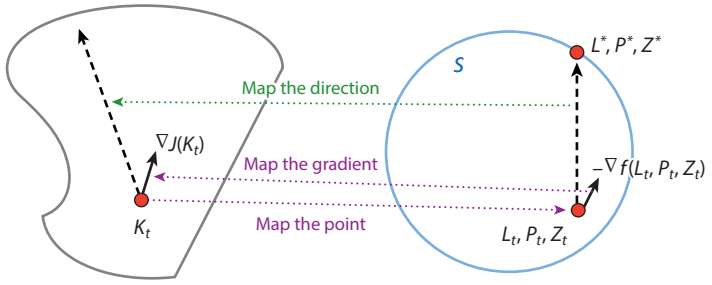


Figure 5

Proof illustration. We can map any policy K to (L, P, Z), and then map the direction $(L^*, P^*, Z^*) - (L, P, Z)$ and the gradient $\nabla f(L, P, Z)$ back to the original K space. Since the problem in (L, P, Z) space is convex, $\langle \nabla f(L, P, Z), (L^*, P^*, Z^*) - (L, P, Z) \rangle$ is less than or equal to 0. We then show that a similar relation holds for the nonconvex problem. Figure adapted from Reference 134 with permission from IEEE.

Recall that $J'(K, V)$ denotes the directional derivative of $J(K)$ along the direction V. When J is differentiable, it holds that $J'(K, V) = \text{trace}(V^\top \nabla J(K))$. We have the following result (modified from Reference 134 to also allow nondifferentiable points).

Theorem 10. Consider the problem pair in Equations 33 and 34. Suppose Assumption 1 holds, and $J(K)$ is either differentiable or subdifferentially regular in $\mathcal{K}$. Let K^* denote a global minimizer of $J(K)$ in $\mathcal{K}$. For any K satisfying $J(K) > J(K^*)$, there exists nonzero V in the descent cone of $\mathcal{K}$ at K, such that the following inequality holds:

$$0 < J(K) - J(K^*) \leq -J'(K, V). \tag{35.}$$

Consequently, any stationary points of J will be global minima.

The above theorem gives a unified sufficient condition ensuring that stationary points of nonconvex PO problems are global minima. The main idea of the proof is illustrated in **Figure 5**. For special problems such as LQR, it is possible to further bound $\|V\|_F$ and apply $\|\nabla J(K)\|_F \geq \left|\nabla J(K)[\frac{V}{\|V\|_F}]\right|$ to show the gradient dominance property with degree $p = 1$ for Equation 17. Now suppose the convex parameterization requires a set of parameters $P_1, \ldots, P_m$; it is possible to develop a more general version of Theorem 10, following the ideas described by Sun & Fazel (134), that allows a map $K = \Phi(P_1, \ldots, P_m)$ as long as Φ has nicely behaved first-order derivatives. Umenberger et al. (68) recently proposed a more involved version of similar constructions as a general framework of differentiable convex lifting.

6.1. Examples

We now provide three examples of the concepts discussed above.

6.1.1. Example 1: discrete-time, infinite-horizon linear quadratic regulator. Consider minimizing the LQR cost $J(K)$ in Equation 5 subject to $K \in \mathcal{K}$, the set of all stabilizing controllers. This problem has the following well-known convex parameterization:

$$\begin{aligned} &\min_{L,P,Z} f(L, P, Z) := \text{trace}(QP) + \text{trace}(ZR) \\ &\text{subject to } P \succ 0, \quad \begin{bmatrix} Z & L \\ L^\top & P \end{bmatrix} \succeq 0, \quad \begin{bmatrix} P - \Sigma & AP + BL \\ (AP + BL)^T & P \end{bmatrix} \succeq 0. \end{aligned} \tag{36.}$$

To check Assumption 1, we add the constraint $K = LP^{-1}$ or $KP = L$ to the problem shown in Equation 36 and simplify the linear matrix inequalities to conclude that the minimum value of

this problem equals $J(K)$ for all feasible K (134). Thus, Theorem 10 applies and can be tailored to show the gradient dominance property for the corresponding nonconvex cost $J(K)$, which allows us to recover the results seen earlier in Section 3.2 (on convergence to a global minimizer) for LQR PO.

The constant in the gradient dominance inequality, Equation 17, can also be bounded in terms of problem matrices, the initial policy cost $J(K^0)$, and a notion of the problem condition number (see 35, 134; for the continuous-time counterpart, see 81, 101). In fact, Mohammadi et al. (81) were the first to leverage existing linear matrix inequality conditions to study the global convergence properties of PO methods, specifically for the continuous-time LQR.

6.1.2. Example 2: $\mathcal{H}_\infty$ state-feedback synthesis. As described in Section 2, this problem can be formulated as PO with $J(K)$ given by Equation 10 and $\mathcal{K}$ given by Equation 3. Under the positive definiteness assumption on (Q, R), the nonstrict version of the bounded real lemma can be used to show that this problem has the following convex parameterization:

$$\min_{L,P,\gamma} f(L,P,\gamma) := \gamma$$

$$\text{subject to } P \succ 0, \quad \begin{bmatrix} -P & 0 & P & (AP-BL)^\top & L^\top \\ 0 & -\gamma I & 0 & I & 0 \\ P & 0 & -\gamma Q^{-1} & 0 & 0 \\ AP-BL & I & 0 & -P & 0 \\ L & 0 & 0 & 0 & -\gamma R^{-1} \end{bmatrix} \preceq 0,$$

where γ is the closed-loop $\mathcal{H}_\infty$ norm. Then, based on Theorem 10 and the subdifferentially regular property of the $\mathcal{H}_\infty$ cost function, we can immediately conclude that any Clarke stationary points for this problem are global minima (54). Then, with the help of coerciveness, Goldstein's subgradient method can be guaranteed to find the global minimum of this problem.

6.1.3. Example 3: output estimation problem with regularization. For lack of coerciveness of the objective in many other control problems beyond LQR, some explicit regularization techniques can help ensure the global convergence of PO methods (which is different from the implicit regularization discussed in Section 4.1). Specifically, the recent work by Umenberger et al. (68) studied continuous-time output estimation (i.e., the filtering problem, which is a fundamental subroutine for LQG control and partially observable control) and established the global convergence of the PG method by regularizing the output estimation objective.

Intriguingly, the regularization is inspired by the convex reformulation of the output estimation problem (135) and ensures that the convex reformulation map (i.e., the map from the domain of the convex parameters to that of the policy parameter) is surjective. This way, the convex parameterization provably leads to a gradient dominance property in the policy parameter domain, of degree 1. Hence, PO with rebalancing on the regularized output estimation objective enjoys a sublinear convergence rate to the global optimum. This example not only reinforces the power of convex parameterization and its connection to gradient dominance, but also calls for more attention to ensure the nondegeneracy of the reformulation map and actually unleash such power.

6.2. Discussion

Under the framework of convex parameterization for control, the properties of the map between the convex and nonconvex domains, for both costs and feasible sets, are crucial. Even in the unconstrained case, if the map from the convex function to the nonconvex one introduces new stationary

points, new tools will be needed to analyze whether a first-order algorithm can avoid these spurious saddle points.

This leads to a question: What are the most general conditions on this map that preserve the stationary points of the convex function and introduce only additional strict saddles or other benign stationary points? This is an interesting question for future work, for which the LQG problem would provide an interesting case study (see discussion in Section 5.1). It is also interesting to further explore the power of explicit regularization (as in Reference 68 for output estimation problems). Convex parameterizations are also helpful for exploring other properties; for example, the convex parameterization for LQG control can be used to establish landscape properties of the PO with partial observation, such as the connectivity properties of feasible/sublevel sets (63, 67).

7. CHALLENGES AND OUTLOOK

In this article, we have revisited the theoretical foundation of PO for control and surveyed a number of results that highlight properties of PO algorithms on benchmark control problems. Our survey has been inspired by the recent success and wide range of applications of RL. PO provides a bridge between control and RL and can give new insights into the design trade-offs between assumptions and data, as well as model-based and model-free synthesis. Theoretical developments in PO for control can help create a renewed interest in the control community to examine the synthesis of dynamical systems from this perspective, which, in our view, is more integrated with machine learning.

We close our discussion with an outlook on challenges and open questions related to bridging the gap between PO theory and real-world control applications. As an exhaustive list is impossible, we discuss a few challenges and open questions that, naturally, reflect our perspectives.

First, from our discussion, it is evident that the development of PO theory requires further connections between modern optimization theory (which focuses on the convergence and complexity of iterative algorithms) and control theory (which rigorously addresses the notions of optimality, stability, robustness, and safety for closed-loop dynamical systems). For example, it is natural to ask whether leveraging results in nonconvex optimization on the complexity of escaping saddle points (136–138) can give similar guarantees for PO with control-theoretic constraints.

Next, further work is needed on regularization for stability, robustness, and safety. In this article, we have covered only $\mathcal{H}_\infty$ robustness constraints, but there is an extensive system-theoretic literature on how to enforce other types of robustness and safety guarantees for controller design. For example, more general robustness constraints can be formulated via passivity (139), dissipativity (140), or integral quadratic constraints (141). In addition, safety can be induced by modifying the cost function. It is important to investigate how to provably pose similar robust/safety guarantees for direct policy search via either explicit regularization (on the cost/constraints) or implicit regularization (via algorithm selection).

Further work is also needed on nonlinear systems, deep RL, and perception-based control. We have focused on reviewing the PO theory centered around linear systems, and it is our hope that the insights from such study can be used to guide the algorithmic/theoretical developments of PO methods for nonlinear control. Conceptually, nonlinear control design can still be formulated as $\min_{K\in\mathcal{K}} J(K)$, and convergence to stationary points can still be established given coerciveness. However, how to characterize the feasible set $\mathcal{K}$ in the nonlinear control setting is unclear in the first place. Quite often, the stability/robustness constraints hold only locally for nonlinear systems. It is crucial to investigate how to define and characterize feasible policies for nonlinear control problems. An important class of PO problems arise in deep RL for end-to-end perception-based control. The theoretical properties of PO methods for such problems remain largely unknown. In

this case, the geometry of the feasible set can become even more complicated due to the presence of the perception modality.

An additional challenge relates to decentralized control of multiagent systems, which has a long history in control theory (142, 143) and also connects to the partially observable setting, since each single agent cannot observe the full system state. PO theory for such control tasks requires further investigation. There has been some recent progress (39, 40, 61). For example, Feng & Lavaei (61) showed that there can be an exponential number of connected components for the feasible set; Furieri et al. (39) then established the global convergence and sample complexity under the quadratic invariance condition (144). It would be interesting to explore other conditions as well as algorithm design principles that admit the global convergence of PO methods for decentralized control. It is especially imperative to develop PO methods that scale with a large number of agents. PO has also been studied in multiagent game-theoretic settings, including general-sum linear quadratic dynamic games (145), with negative nonconvergence results, and linear quadratic mean-field games (146–149), where the number of agents is very large and approximated by infinity. It would be interesting to further explore the PO theory in other dynamic game settings with control implications.

The integration of model-based and model-free methods will also be an important future direction. Both approaches are important for control design (92): On the one hand, there has been a recent trend of research examining the LQR problem as a benchmark for learning-based control, starting with the work of Dean et al. (150), and it has been shown that model-based methods can be more sample efficient in this case from an asymptotic viewpoint (151); on the other hand, model-free methods can be more flexible for complex tasks, such as perception-based control. Integrating the two will help achieve the best of both worlds, especially for controlling systems that are only partially understood or parameterized. It is also expected that such an integrated approach will lead to developments in new settings that further connect learning and control theory, such as online control with regret guarantees (152–156).

Finally, we note that many new tasks arising in machine learning for control can also be formulated as PO. For example, imitation learning for control can be formulated as PO with control-theoretic constraints (157–160). Similarly, transfer learning for linear control can be studied as PO if we modify the cost function properly (161). In the context of control, PO conveniently provides a general paradigm for formulating imitation learning and transfer learning tasks. It will be interesting to investigate the convergence theory of gradient-based algorithms for such problems.

SUMMARY POINTS

1. Thanks to the coerciveness and gradient dominance properties, policy optimization (PO) for the linear quadratic regulator (LQR) leads to a nonconvex problem that can still be provably solved using the gradient method.
2. In the state-feedback setting, advanced PO methods can be guaranteed to achieve global convergence on linear risk-sensitive/robust control tasks.
3. In the partial observation setting, the optimization landscape provides important clues for the performance of PO methods.
4. There are fundamental connections between the convex formulations of optimal/robust control tasks and the PO landscape.

FUTURE ISSUES

1. Many new results on the complexity of escaping saddles and finding stationary points for unconstrained optimization may be extended to PO.
2. Advanced regularization techniques are needed for robustness and safety in general.
3. PO theory for nonlinear or perception-based control remains largely open.
4. Scalability is an important issue for PO in multiagent decentralized control.
5. More study is needed to integrate model-based and model-free methods.
6. There are new PO problems in machine learning for control (e.g., imitation/transfer learning).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The research of M.M. is supported by grants from the Air Force Office of Scientific Research (FA9550-20-1-0053) and the National Science Foundation (ECCS-2149470); M.M. acknowledges discussions with and contributions from Jingjng Bu, Shahriar Talebi (who also kindly produced **Figure 4*b*,*c***), Sham Kakade, and Rong Ge. The research of N.L. is supported by grants from the Office of Naval Research Young Investigator Program (N00014-19-1-2217), the Air Force Office of Scientific Research Young Investigator Program (FA9550-18-1-0150), and the National Science Foundation (AI Institute 2112085); N.L. acknowledges discussions with and contributions from Yang Zheng, Yujie Tang, and Yingying Li. The research of B.H. is supported by an award from the National Science Foundation (CAREER-2048168); B.H. acknowledges discussions with Peter Seiler, Geir Dullerud, Xingang Guo, Aaron Havens, Darioush Keivan, Yang Zheng, Javad Lavaei, Mihailo Jovanović, and Michael Overton. The research of K.Z. is supported by a Simons–Berkeley Research Fellowship; K.Z. acknowledges discussions with Max Simchowitz. The research of M.F. is supported by grants from the National Science Foundation (TRIPODS II-DMS 2023166, CCF 2007036, CCF 2212261, AI Institute 2112085, and HDR 1934292) and the Office of Naval Research (MURI N0014-16-1-2710); M.F. acknowledges discussions with Yue Sun, Sham Kakade, and Rong Ge. The research of T.B. is supported in part by grants from the Air Force Office of Scientific Research (FA9550-19-1-0353) and the US Army Research Laboratory (cooperative agreement W911NF-17-2-0196).

LITERATURE CITED

1. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518:529–33
2. Vinyals O, Babuschkin I, Chung J, Mathieu M, Jaderberg M, et al. 2019. AlphaStar: mastering the real-time strategy game StarCraft II. *DeepMind*, Jan. 24. **https://www.deepmind.com/blog/alphastar-mastering-the-real-time-strategy-game-starcraft-ii**
3. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529:484–89

4. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550:354–59
5. Rajeswaran A, Kumar V, Gupta A, Vezzani G, Schulman J, et al. 2017. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. arXiv:1709.10087 [cs.LG]
6. Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, et al. 2015. Continuous control with deep reinforcement learning. arXiv:1509.02971 [cs.LG]
7. Schulman J, Moritz P, Levine S, Jordan M, Abbeel P. 2015. High-dimensional continuous control using generalized advantage estimation. arXiv:1506.02438 [cs.LG]
8. Sutton RS, McAllester DA, Singh SP, Mansour Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, ed. S Solla, T Leen, K Müller, pp. 1057–63. Cambridge, MA: MIT Press
9. Konda VR, Tsitsiklis JN. 2000. Actor-critic algorithms. In *Advances in Neural Information Processing Systems 12*, ed. S Solla, T Leen, K Müller, pp. 1008–14. Cambridge, MA: MIT Press
10. Schulman J, Levine S, Abbeel P, Jordan M, Moritz P. 2015. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, ed. F Bach, D Blei, pp. 1889–97. Proc. Mach. Learn. Res. 37. N.p.: PMLR
11. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. 2017. Proximal policy optimization algorithms. arXiv:1707.06347 [cs.LG]
12. Lee AX, Nagabandi A, Abbeel P, Levine S. 2019. Stochastic latent actor-critic: deep reinforcement learning with a latent variable model. arXiv:1907.00953 [cs.LG]
13. Yarats D, Zhang A, Kostrikov I, Amos B, Pineau J, Fergus R. 2021. Improving sample efficiency in model-free reinforcement learning from images. *Proc. AAAI Conf. Artif. Intell.* 35:10674–81
14. Yarats D, Fergus R, Lazaric A, Pinto L. 2022. Mastering visual continuous control: improved data-augmented reinforcement learning. In *The Tenth International Conference on Learning Representations*. La Jolla, CA: Int. Conf. Learn. Represent. **https://openreview.net/forum?id=_SJ-_yyes8**
15. Draper CS, Li YT. 1951. *Principles of Optimalizing Control Systems and an Application to the Internal Combustion Engine*. New York: Am. Soc. Mech. Eng.
16. Whitaker HP, Yamron J, Kezer A. 1958. *Design of model-reference adaptive control systems for aircraft*. Rep., Instrum. Lab., Mass. Inst. Technol., Cambridge
17. Kalman RE. 1960. Contributions to the theory of optimal control. *Bol. Soc. Mat. Mex.* 5:102–19
18. Talkin A. 1961. Adaptive servo tracking. *IRE Trans. Autom. Control* 6:167–72
19. Levine W, Athans M. 1970. On the determination of the optimal constant output feedback gains for linear multivariable systems. *IEEE Trans. Autom. Control* 15:44–48
20. Makila P, Toivonen H. 1987. Computational methods for parametric LQ problems—a survey. *IEEE Trans. Autom. Control* 32:658–71
21. Boyd S, Vandenberghe L. 2004. *Convex Optimization*. Cambridge, UK: Cambridge Univ. Press
22. Boyd S, El Ghaoui L, Feron E, Balakrishnan V. 1994. *Linear Matrix Inequalities in System and Control Theory*. Philadelphia: Soc. Ind. Appl. Math.
23. Gahinet P, Apkarian P. 1994. A linear matrix inequality approach to H_∞ control. *Int. J. Robust Nonlinear Control* 4:421–48
24. Scherer C, Wieland S. 2004. *Linear matrix inequalities in control*. Lect. Notes, Dutch Inst. Syst. Control, Delft Univ. Technol., Delft, Neth.
25. Papachristodoulou A, Anderson J, Valmorbida G, Prajna S, Seiler P, et al. 2022. SOSTOOLS: sum of squares optimization toolbox for MATLAB. *University of Oxford*. **http://sysos.eng.ox.ac.uk/sostools**
26. Anderson J, Papachristodoulou A. 2015. Advances in computational Lyapunov analysis using sum-of-squares programming. *Discrete Contin. Dyn. Syst. B* 20:2361–81
27. Rautert T, Sachs EW. 1997. Computational design of optimal output feedback controllers. *SIAM J. Optim.* 7:837–52
28. Apkarian P, Noll D. 2006. Nonsmooth H_∞ synthesis. *IEEE Trans. Autom. Control* 51:71–86
29. Apkarian P, Noll D, Rondepierre A. 2008. Mixed H_2/H_∞ control via nonsmooth optimization. *SIAM J. Control Optim.* 47:1516–46
30. Noll D, Apkarian P. 2005. Spectral bundle methods for non-convex maximum eigenvalue functions: second-order methods. *Math. Program.* 104:729–47

31. Saeki M. 2006. Static output feedback design for H_∞ control by descent method. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 5156–61. Piscataway, NJ: IEEE
32. Gumussoy S, Henrion D, Millstone M, Overton ML. 2009. Multiobjective robust control with HIFOO 2.0. *IFAC Proc. Vol.* 42(6):144–49
33. Arzelier D, Deaconu G, Gumussoy S, Henrion D. 2011. *H2 for HIFOO*. Paper presented at the 3rd International Conference on Control and Optimization with Industrial Applications, Ankara, Turkey, Aug. 22–24
34. Mårtensson K, Rantzer A. 2009. Gradient methods for iterative distributed control synthesis. In *Proceedings of the 48th IEEE Conference on Decision and Control Held Jointly with 2009 28th Chinese Control Conference*, pp. 549–54. Piscataway, NJ: IEEE
35. Fazel M, Ge R, Kakade S, Mesbahi M. 2018. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, ed. J Dy, A Krase, pp. 1467–76. Proc. Mach. Learn. Res. 80. N.p.: PMLR
36. Bu J, Mesbahi A, Fazel M, Mesbahi M. 2019. LQR through the lens of first order methods: discrete-time case. arXiv:1907.08921 [eess.SY]
37. Malik D, Pananjady A, Bhatia K, Khamaru K, Bartlett P, Wainwright M. 2019. Derivative-free methods for policy optimization: guarantees for linear quadratic systems. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, ed. K Chaudhuri, M Sugiyama, pp. 2916–25. Proc. Mach. Learn. Res. 89. N.p.: PMLR
38. Mohammadi H, Zare A, Soltanolkotabi M, Jovanović MR. 2021. Convergence and sample complexity of gradient methods for the model-free linear–quadratic regulator problem. *IEEE Trans. Autom. Control* 67:2435–50
39. Furieri L, Zheng Y, Kamgarpour M. 2020. Learning the globally optimal distributed LQ regulator. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, ed. AM Bayen, A Jadbabaie, G Pappas, PA Parrilo, B Recht, et al., pp. 287–97. Proc. Mach. Learn. Res. 120. N.p.: PMLR
40. Li Y, Tang Y, Zhang R, Li N. 2022. Distributed reinforcement learning for decentralized linear quadratic control: a derivative-free policy optimization approach. *IEEE Trans. Autom. Control* 67:6429–44
41. Hambly B, Xu R, Yang H. 2021. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM J. Control Optim.* 59:3359–91
42. Yang Z, Chen Y, Hong M, Wang Z. 2019. Provably global convergence of actor-critic: a case for linear quadratic regulator with ergodic cost. In *Advances in Neural Information Processing Systems 32*, ed. H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, R Garnett, pp. 8231–33. Red Hook, NY: Curran
43. Jin Z, Schmitt JM, Wen Z. 2020. On the analysis of model-free methods for the linear quadratic regulator. arXiv:2007.03861 [math.OC]
44. Mohammadi H, Soltanolkotabi M, Jovanović MR. 2020. On the linear convergence of random search for discrete-time LQR. *IEEE Control Syst. Lett.* 5:989–94
45. Perdomo J, Umenberger J, Simchowitz M. 2021. Stabilizing dynamical systems via policy gradient methods. In *Advances in Neural Information Processing Systems 34*, ed. M Ranzato, A Beygelzimer, Y Dauphin, PS Liang, J Wortman Vaughan, pp. 29274–86. Red Hook, NY: Curran
46. Ozaslan IK, Mohammadi H, Jovanović MR. 2022. Computing stabilizing feedback gains via a model-free policy gradient method. *IEEE Control Syst. Lett.* 7:407–12
47. Zhao F, Fu X, You K. 2022. On the sample complexity of stabilizing linear systems via policy gradient methods. arXiv:2205.14335 [math.OC]
48. Zhang K, Hu B, Başar T. 2021. Policy optimization for H_2 linear control with H_∞ robustness guarantee: implicit regularization and global convergence. *SIAM J. Control Optim.* 59:4081–109
49. Zhang K, Hu B, Başar T. 2020. On the stability and convergence of robust adversarial reinforcement learning: a case study on linear quadratic systems. In *Advances in Neural Information Processing Systems 33*, ed. H Larochelle, M Ranzato, R Hadsell, MF Balcan, H Lin, pp. 22056–68. Red Hook, NY: Curran
50. Gravell B, Esfahani PM, Summers T. 2020. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Trans. Autom. Control* 66:5283–98
51. Zhang K, Zhang X, Hu B, Başar T. 2021. Derivative-free policy optimization for linear risk-sensitive and robust control design: implicit regularization and sample complexity. In *Advances in Neural*

Information Processing Systems 34, ed. M Ranzato, A Beygelzimer, Y Dauphin, PS Liang, J Wortman Vaughan, pp. 2949–64. Red Hook, NY: Curran
52. Zhao F, You K. 2021. Primal-dual learning for the model-free risk-constrained linear quadratic regulator. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, ed. A Jadbabaie, J Lygeros, GJ Pappas, PA Parrilo, B Recht, et al., pp. 702–14. Proc. Mach. Learn. Res. 144. N.p.: PMLR
53. Zhang Y, Yang Z, Wang Z. 2021. Provably efficient actor-critic for risk-sensitive and robust adversarial RL: a linear-quadratic case. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, ed. A Banerjee, K Fukumizu, pp. 2764–72. Proc. Mach. Learn. Res. 130. N.p.: PMLR
54. Guo X, Hu B. 2022. *Global convergence of direct policy search for state-feedback H_∞ robust control: a revisit of nonsmooth synthesis with Goldstein subdifferential*. Paper presented at the 36th Conference on Neural Information Processing Systems, New Orleans, LA, Nov. 28–Dec. 9
55. Keivan D, Havens A, Seiler P, Dullerud G, Hu B. 2022. Model-free μ synthesis via adversarial reinforcement learning. In *2022 American Control Conference*, pp. 3335–41. Piscataway, NJ: IEEE
56. Jansch-Porto JP, Hu B, Dullerud GE. 2020. Convergence guarantees of policy optimization methods for Markovian jump linear systems. In *2020 American Control Conference*, pp. 2882–87. Piscataway, NJ: IEEE
57. Jansch-Porto JP, Hu B, Dullerud G. 2020. Policy learning of MDPs with mixed continuous/discrete variables: a case study on model-free control of Markovian jump systems. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, ed. AM Bayen, A Jadbabaie, G Pappas, PA Parrilo, B Recht, et al., pp. 947–957. Proc. Mach. Learn. Res. 120. N.p.: PMLR
58. Rathod S, Bhadu M, De A. 2021. Global convergence using policy gradient methods for model-free Markovian jump linear quadratic control. arXiv:2111.15228 [cs.LG]
59. Jansch-Porto JP, Hu B, Dullerud GE. 2022. Policy optimization for Markovian jump linear quadratic control: gradient method and global convergence. *IEEE Trans. Autom. Control*. In press. **https://doi.org/10.1109/TAC.2022.3176439**
60. Qu G, Yu C, Low S, Wierman A. 2021. Exploiting linear models for model-free nonlinear control: a provably convergent policy gradient approach. In *2021 60th IEEE Conference on Decision and Control*, pp. 6539–46. Piscataway, NJ: IEEE
61. Feng H, Lavaei J. 2019. On the exponential number of connected components for the feasible set of optimal decentralized control problems. In *2019 American Control Conference*, pp. 1430–37. Piscataway, NJ: IEEE
62. Fatkhullin I, Polyak B. 2021. Optimizing static linear feedback: gradient method. *SIAM J. Control Optim.* 59:3887–911
63. Zheng Y, Tang Y, Li N. 2021. Analysis of the optimization landscape of linear quadratic Gaussian (LQG) control. arXiv:2102.04393 [math.OC]
64. Duan J, Li J, Zhao L. 2021. Optimization landscape of gradient descent for discrete-time static output feedback. arXiv:2109.13132 [math.OC]
65. Duan J, Cao W, Zheng Y, Zhao L. 2022. On the optimization landscape of dynamical output feedback linear quadratic control. arXiv:2201.09598 [math.OC]
66. Mohammadi H, Soltanolkotabi M, Jovanović MR. 2021. On the lack of gradient domination for linear quadratic Gaussian problems with incomplete state information. In *2021 60th IEEE Conference on Decision and Control*, pp. 1120–24. Piscataway, NJ: IEEE
67. Hu B, Zheng Y. 2022. Connectivity of the feasible and sublevel sets of dynamic output feedback control with robustness constraints. *IEEE Control Syst. Lett.* 7:442–47
68. Umenberger J, Simchowitz M, Perdomo JC, Zhang K, Tedrake R. 2022. Globally convergent policy search over dynamic filters for output estimation. arXiv:2202.11659 [math.OC]
69. Buşoniu L, de Bruin T, Tolić D, Kober J, Palunko I. 2018. Reinforcement learning for control: performance, stability, and deep approximators. *Annu. Rev. Control* 46:8–28
70. Recht B. 2019. A tour of reinforcement learning: the view from continuous control. *Annu. Rev. Control Robot. Auton. Syst.* 2:253–79
71. Matni N, Proutiere A, Rantzer A, Tu S. 2019. From self-tuning regulators to reinforcement learning and back again. In *2019 IEEE 58th Conference on Decision and Control*, pp. 3724–40. Piscataway, NJ: IEEE

72. Jacobson D. 1973. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Trans. Autom. Control* 18:124–31
73. Zhou K, Doyle JC, Glover K. 1996. *Robust and Optimal Control*. Upper Saddle River, NJ: Prentice Hall
74. Mustafa D. 1989. Relations between maximum-entropy/H_∞ control and combined H_∞ /LQG control. *Syst. Control Lett.* 12:193–203
75. Mustafa D, Bernstein DS. 1991. LQG cost bounds in discrete-time H_2/H_∞ control. *Trans. Inst. Meas. Control* 13:269–75
76. Peres PL, Geromel JC. 1994. An alternate numerical solution to the linear quadratic problem. *IEEE Trans. Autom. Control* 39:198–202
77. Nesterov Y, Polyak BT. 2006. Cubic regularization of Newton method and its global performance. *Math. Program.* 108:177–205
78. Karimi H, Nutini J, Schmidt M. 2016. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, ed. P Frasconi, N Landwehr, G Manco, J Vreeken, pp. 795–811. Cham, Switz.: Springer
79. Li G, Pong TK. 2018. Calculus of the exponent of Kurdyka–Lojasiewicz inequality and its applications to linear convergence of first-order methods. *Found. Comput. Math.* 18:1199–232
80. Bauschke HH, Combettes PL. 2011. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Cham, Switz.: Springer
81. Mohammadi H, Zare A, Soltanolkotabi M, Jovanović MR. 2019. Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator. In *2019 58th IEEE Conference on Decision and Control*, pp. 7474–79. Piscataway, NJ: IEEE
82. Bu J, Mesbahi M. 2020. Global convergence of policy gradient algorithms for indefinite least squares stationary optimal control. *IEEE Control Syst. Lett.* 4:638–43
83. Lamperski A. 2020. Computing stabilizing linear controllers via policy iteration. In *2020 59th IEEE Conference on Decision and Control*, pp. 1902–07. Piscataway, NJ: IEEE
84. Nesterov Y, Spokoiny V. 2017. Random gradient-free minimization of convex functions. *Found. Comput. Math.* 17:527–66
85. Duchi JC, Jordan MI, Wainwright MJ, Wibisono A. 2015. Optimal rates for zero-order convex optimization: the power of two function evaluations. *IEEE Trans. Inf. Theory* 61:2788–806
86. Shamir O. 2013. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proceedings of the 26th Annual Conference on Learning Theory*, ed. S Shalev-Shwartz, I Steinwart, pp. 3–24. Proc. Mach. Learn. Res. 30. N.p.: PMLR
87. Ghadimi S, Lan G. 2013. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.* 23:2341–68
88. Balasubramanian K, Ghadimi S. 2022. Zeroth-order nonconvex stochastic optimization: handling constraints, high dimensionality, and saddle points. *Found. Comput. Math.* 22:35–76
89. Tang Y, Ren Z, Li N. 2020. Zeroth-order feedback optimization for cooperative multi-agent systems. In *2020 59th IEEE Conference on Decision and Control*, pp. 3649–56. Piscataway, NJ: IEEE
90. Talebi S, Alemzadeh S, Rahimi N, Mesbahi M. 2021. On regularizability and its application to online control of unstable LTI systems. *IEEE Trans. Autom. Control* 67:6413–28
91. Ziemann I, Tsiamis A, Sandberg H, Matni N. 2022. How are policy gradient methods affected by the limits of control? arXiv:2206.06863 [math.OC]
92. Tsiamis A, Ziemann I, Matni N, Pappas GJ. 2022. Statistical learning theory for control: a finite sample perspective. arXiv:2209.05423 [eess.SY]
93. Hjalmarsson H, Gevers M, Gunnarsson S, Lequin O. 1998. Iterative feedback tuning: theory and applications. *IEEE Control Syst. Mag.* 18(4):26–41
94. Hjalmarsson H. 2002. Iterative feedback tuning—an overview. *Int. J. Adapt. Control Signal Process.* 16:373–95
95. Kakade SM. 2002. A natural policy gradient. In *Advances in Neural Information Processing Systems 14*, ed. T Dietterich, S Becker, Z Ghahramani, pp. 1531–38. Cambridge, MA: MIT Press
96. Bradtke SJ, Ydstie BE, Barto AG. 1994. Adaptive linear quadratic control using policy iteration. In *Proceedings of the 1994 American Control Conference*, Vol. 3, pp. 3475–79. Piscataway, NJ: IEEE

97. Kleinman D. 1968. On an iterative technique for Riccati equation computations. *IEEE Trans. Autom. Control* 13:114–15
98. Hewer G. 1971. An iterative technique for the computation of the steady state gains for the discrete optimal regulator. *IEEE Trans. Autom. Control* 16:382–84
99. Lagoudakis MG, Parr R. 2003. Least-squares policy iteration. *J. Mach. Learn. Res.* 4:1107–49
100. Krauth K, Tu S, Recht B. 2019. Finite-time analysis of approximate policy iteration for the linear quadratic regulator. In *Advances in Neural Information Processing Systems 32*, ed. H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, R Garnett, pp. 8514–24. Red Hook, NY: Curran
101. Bu J, Mesbahi A, Mesbahi M. 2020. Policy gradient-based algorithms for continuous-time linear quadratic control. arXiv:2006.09178 [eess.SY]
102. Bertsekas DP. 1997. Nonlinear programming. *J. Oper. Res. Soc.* 48:334
103. Burke JV, Curtis FE, Lewis AS, Overton ML, Simões LE. 2020. Gradient sampling methods for nonsmooth optimization. In *Numerical Nonsmooth Optimization*, ed. A Bagirov, M Gaudioso, N Karmitsa, M Mäkelä, S Taheri, pp. 202–25. Cham, Switz.: Springer
104. Wu HN, Luo B. 2012. Neural network based online simultaneous policy update algorithm for solving the HJI equation in nonlinear H_∞ control. *IEEE Trans. Neural Netw. Learn. Syst.* 23:1884–95
105. Luo B, Wu HN, Huang T. 2014. Off-policy reinforcement learning for H_∞ control design. *IEEE Trans. Cybernet.* 45:65–76
106. Kiumarsi B, Lewis FL, Jiang ZP. 2017. H_∞ control of linear discrete-time systems: off-policy reinforcement learning. *Automatica* 78:144–52
107. Kubo M, Banno R, Manabe H, Minoji M. 2019. Implicit regularization in over-parameterized neural networks. arXiv:1903.01997 [cs.LG]
108. Ma C, Wang K, Chi Y, Chen Y. 2017. Implicit regularization in nonconvex statistical estimation: gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. arXiv:1711.10467 [cs.LG]
109. Chen Y, Wainwright MJ. 2015. Fast low-rank estimation by projected gradient descent: general statistical and algorithmic guarantees. arXiv:1509.03025 [math.ST]
110. Zheng Q, Lafferty J. 2016. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. arXiv:1605.07051 [stat.ML]
111. Başar T, Bernhard P. 1995. *H^∞-Optimal Control and Related Minimax Design Problems*. Boston: Birkhäuser. 2nd ed.
112. Rantzer A. 1996. On the Kalman–Yakubovich–Popov lemma. *Syst. Control Lett.* 28:7–10
113. Al-Tamimi A, Lewis FL, Abu-Khalaf M. 2007. Model-free *Q*-learning designs for linear discrete-time zero-sum games with application to H-infinity control. *Automatica* 43:473–81
114. Zhang K, Yang Z, Başar T. 2019. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems 32*, ed. H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, R Garnett, pp. 11570–82. Red Hook, NY: Curran
115. Gravell B, Ganapathy K, Summers T. 2020. Policy iteration for linear quadratic games with stochastic parameters. *IEEE Control Syst. Lett.* 5:307–12
116. Zhang J, Yang Z, Zhou Z, Wang Z. 2021. Provably sample efficient reinforcement learning in competitive linear quadratic systems. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, ed. A Jadbabaie, J Lygeros, GJ Pappas, PA Parrilo, B Recht, et al., pp. 597–98. Proc. Mach. Learn. Res. 144. N.p.: PMLR
117. Zhang K, Yang Z, Başar T. 2021. Multi-agent reinforcement learning: a selective overview of theories and algorithms. In *Handbook of Reinforcement Learning and Control*, ed. KG Vamvoudakis, Y Wan, FL Lewis, D Cansever, pp. 321–84. Cham, Switz.: Springer
118. Başar T, Olsder G. 1999. *Dynamic Noncooperative Game Theory*. Philadelphia: Soc. Ind. Appl. Math. 2nd ed.
119. Bu J, Ratliff LJ, Mesbahi M. 2019. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. arXiv:1911.04672 [eess.SY]
120. Morimoto J, Doya K. 2005. Robust reinforcement learning. *Neural Comput.* 17:335–59

121. Pinto L, Davidson J, Sukthankar R, Gupta A. 2017. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, ed. D Precup, YW Teh, pp. 2817–26. Proc. Mach. Learn. Res. 70. N.p.: PMLR
122. Dullerud G, Paganini F. 1999. *A Course in Robust Control Theory: A Convex Approach*. New York: Springer
123. Goldstein A. 1977. Optimization of Lipschitz continuous functions. *Math. Program.* 13:14–22
124. Turchetta M, Krause A, Trimpe S. 2020. Robust model-free reinforcement learning with multi-objective Bayesian optimization. In *2020 IEEE International Conference on Robotics and Automation*, pp. 10702–8. Piscataway, NJ: IEEE
125. Pang B, Jiang ZP. 2021. Robust reinforcement learning: a case study in linear quadratic regulation. *Proc. AAAI Conf. Artif. Intell.* 35:9303–11
126. Pang B, Bian T, Jiang ZP. 2021. Robust policy iteration for continuous-time linear quadratic regulation. *IEEE Trans. Autom. Control* 67:504–11
127. Venkataraman HK, Seiler PJ. 2019. Recovering robustness in model-free reinforcement learning. In *2019 American Control Conference*, pp. 4210–16. Piscataway, NJ: IEEE
128. Zhao F, You K, Başar T. 2021. Infinite-horizon risk-constrained linear quadratic regulator with average cost. In *2021 60th IEEE Conference on Decision and Control*, pp. 390–95. Piscataway, NJ: IEEE
129. Zhao F, You K, Başar T. 2021. Global convergence of policy gradient primal-dual methods for risk-constrained LQRs. arXiv:2104.04901 [math.OC]
130. Zheng Y, Sun Y, Fazel M, Li N. 2022. Escaping high-order saddles in policy optimization for linear quadratic Gaussian (LQG) control. arXiv:2204.00912 [math.OC]
131. Bu J, Mesbahi A, Mesbahi M. 2021. On topological properties of the set of stabilizing feedback gains. *IEEE Trans. Autom. Control* 66:730–44
132. Talebi S, Mesbahi M. 2022. Policy optimization over submanifolds for constrained feedback synthesis. *IEEE Trans. Autom. Control*. In press
133. Ding Y, Feng H, Lavaei J. 2019. Aggressive local search for constrained optimal control problems with many local minima. arXiv:1903.08634 [math.OC]
134. Sun Y, Fazel M. 2021. Learning optimal controllers by policy gradient: global optimality via convex parameterization. In *2021 60th IEEE Conference on Decision and Control*, pp. 4576–81. Piscataway, NJ: IEEE
135. Scherer C, Gahinet P, Chilali M. 1997. Multiobjective output-feedback control via LMI optimization. *IEEE Trans. Autom. Control* 42:896–911
136. Jin C, Ge R, Netrapalli P, Kakade SM, Jordan MI. 2017. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, ed. D Precup, YW Teh, pp. 1724–32. Proc. Mach. Learn. Res. 70. N.p.: PMLR
137. Sun Y, Flammarion N, Fazel M. 2019. Escaping from saddle points on Riemannian manifolds. In *Advances in Neural Information Processing Systems 32*, ed. H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, R Garnett, pp. 7244–54. Red Hook, NY: Curran
138. Ren Z, Tang Y, Li N. 2022. Escaping saddle points in zeroth-order optimization: two function evaluations suffice. arXiv:2209.13555 [math.OC]
139. van der Schaft A. 2000. *L_2-Gain and Passivity Techniques in Nonlinear Control*. London: Springer
140. Willems J. 1972. Dissipative dynamical systems part I: general theory. *Arch. Ration. Mech. Anal.* 45:321–51
141. Megretski A, Rantzer A. 1997. System analysis via integral quadratic constraints. *IEEE Trans. Autom. Control* 42:819–30
142. Sandell N, Varaiya P, Athans M. 1975. A survey of decentralized control methods for large scale systems. In *Systems Engineering for Power: Status and Prospects*, pp. 334–35. Washington, DC: US Energy Res. Dev. Adm.
143. Tsitsiklis JN. 1984. *Problems in decentralized decision making and computation*. PhD Thesis, Mass. Inst. Technol., Cambridge
144. Rotkowitz M, Lall S. 2005. A characterization of convex problems in decentralized control. *IEEE Trans. Autom. Control* 50:1984–96
145. Mazumdar E, Ratliff LJ, Jordan MI, Sastry SS. 2019. Policy-gradient algorithms have no guarantees of convergence in linear quadratic games. arXiv:1907.03712 [cs.LG]

146. Fu Z, Yang Z, Chen Y, Wang Z. 2019. Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games. In *The Eighth International Conference on Learning Representations*. La Jolla, CA: Int. Conf. Learn. Represent. **https://openreview.net/forum?id=H1lhqpEYPr**
147. Carmona R, Laurière M, Tan Z. 2019. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. arXiv:1910.04295 [math.OC]
148. Wang W, Han J, Yang Z, Wang Z. 2021. Global convergence of policy gradient for linear-quadratic mean-field control/game in continuous time. In *Proceedings of the 38th International Conference on Machine Learning*, ed. M Meila, T Zhang, pp. 10772–82. Proc. Mach. Learn. Res. 139. N.p.: PMLR
149. Carmona R, Hamidouche K, Laurière M, Tan Z. 2020. Policy optimization for linear-quadratic zero-sum mean-field type games. In *2020 Conference on Decision and Control*, pp. 1038–43. Piscataway, NJ: IEEE
150. Dean S, Mania H, Matni N, Recht B, Tu S. 2017. On the sample complexity of the linear quadratic regulator. *Found. Comput. Math.* 20:633–79
151. Tu S, Recht B. 2019. The gap between model-based and model-free methods on the linear quadratic regulator: an asymptotic viewpoint. In *Proceedings of the Thirty-Second Conference on Learning Theory*, ed. A Beygelzimer, D Hsu, pp. 3036–83. Proc. Mach. Learn. Res. 99. N.p.: PMLR
152. Lale S, Azizzadenesheli K, Hassibi B, Anandkumar A. 2020. Explore more and improve regret in linear quadratic regulators. arXiv:2007.12291 [cs.LG]
153. Chen X, Hazan E. 2021. Black-box control for linear dynamical systems. In *Proceedings of Thirty Fourth Conference on Learning Theory*, ed. M Belkin, S Kpotufe, pp. 1114–43. Proc. Mach. Learn. Res. 134. N.p.: PMLR
154. Simchowitz M, Foster D. 2020. Naive exploration is optimal for online LQR. In *Proceedings of the 37th International Conference on Machine Learning*, ed. H Daumé III, A Singh, pp. 8937–48. Proc. Mach. Learn. Res. 119. N.p.: PMLR
155. Simchowitz M, Singh K, Hazan E. 2020. Improper learning for non-stochastic control. In *Proceedings of 33rd Conference on Learning Theory*, ed. J Abernethy, S Agarwal, pp. 3320–36. Proc. Mach. Learn. Res. 125. N.p.: PMLR
156. Agarwal N, Bullins B, Hazan E, Kakade S, Singh K. 2019. Online control with adversarial disturbances. In *Proceedings of the 36th International Conference on Machine Learning*, ed. K Chaudhuri, R Salakhutdinov, pp. 111–19. Proc. Mach. Learn. Res. 97. N.p.: PMLR
157. Palan M, Barratt S, McCauley A, Sadigh D, Sindhwani V, Boyd S. 2020. Fitting a linear control policy to demonstrations with a Kalman constraint. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, ed. AM Bayen, A Jadbabaie, G Pappas, PA Parrilo, B Recht, et al., pp. 374–83. Proc. Mach. Learn. Res. 120. N.p.: PMLR
158. Havens A, Hu B. 2021. On imitation learning of linear control policies: enforcing stability and robustness constraints via LMI conditions. In *2021 American Control Conference*, pp. 882–87. Piscataway, NJ: IEEE
159. Yin H, Seiler P, Jin M, Arcak M. 2021. Imitation learning with stability and safety guarantees. *IEEE Control Syst. Lett.* 6:409–14
160. Tu S, Robey A, Zhang T, Matni N. 2022. On the sample complexity of stability constrained imitation learning. In *Proceedings of the 4th Annual Learning for Dynamics and Control Conference*, ed. R Firoozi, N Mehr, E Yel, R Antonova, J Bohg, et al., pp. 180–91. Proc. Mach. Learn. Res. 168. N.p.: PMLR
161. Molybog I, Lavaei J. 2021. When does MAML objective have benign landscape? In *2021 IEEE Conference on Control Technology and Applications*, pp. 220–27. Piscataway, NJ: IEEE