

# Derivative-Free Methods for Policy Optimization: Guarantees for Linear Quadratic Systems

Dhruv Malik<sup>†</sup>   Ashwin Pananjady<sup>†</sup>   Kush Bhatia<sup>†</sup>  
 Koulik Khamaru<sup>‡</sup>   Peter L. Bartlett<sup>†,‡</sup>   Martin J. Wainwright<sup>†,‡,\*</sup>

Department of Electrical Engineering and Computer Sciences, UC Berkeley<sup>†</sup>  
 Department of Statistics, UC Berkeley<sup>‡</sup>  
 Voleon Group<sup>\*</sup>

December 21, 2018

## Abstract

We study derivative-free methods for policy optimization over the class of linear policies. We focus on characterizing the convergence rate of these methods when applied to linear-quadratic systems, and study various settings of driving noise and reward feedback. We show that these methods provably converge to within any pre-specified tolerance of the optimal policy with a number of zero-order evaluations that is an explicit polynomial of the error tolerance, dimension, and curvature properties of the problem. Our analysis reveals some interesting differences between the settings of additive driving noise and random initialization, as well as the settings of one-point and two-point reward feedback. Our theory is corroborated by extensive simulations of derivative-free methods on these systems. Along the way, we derive convergence rates for stochastic zero-order optimization algorithms when applied to a certain class of non-convex problems.

## 1 Introduction

Recent years have witnessed a number of successes in applying modern reinforcement learning (RL) methods to many fields, including robotics [TFR<sup>+</sup>17, LFDA16] and competitive gaming [S<sup>+</sup>16, M<sup>+</sup>15]. Impressively, most of these successes have been achieved by using general-purpose RL methods that are applicable to a host of problems. Prevalent general-purpose RL approaches can be broadly categorized into: (a) *model-based approaches* [DRF12, GLSL16, LHP<sup>+</sup>15], in which an agent attempts to learn a model for the dynamics by observing the evolution of its state sequence; and (b) *model-free approaches*, including DQN [M<sup>+</sup>15], and TRPO [SLA<sup>+</sup>15]), in which the agent attempts to learn an optimal policy directly, by observing rewards from the environment. While model-free approaches typically require more samples to learn a policy of equivalent accuracy, they are naturally more robust to model mis-specification.

A literature that is closely related to model-free RL is that of *zero-order or derivative-free* methods for stochastic optimization; see the book by Spall [Spa05] for an overview. Here the goal is to optimize an unknown function from noisy observations of its values at judiciously chosen points. While most analytical results in this space apply to convex optimization, many of the procedures themselves rely on moving along randomized approximations to the directional derivatives of the function being optimized, and thus are applicable even to non-convex problems. In the particular context of RL, variants of derivative-free methods, including TRPO [SLA<sup>+</sup>15], PSNG [RLTK17]

and evolutionary strategies [SHC<sup>+</sup>17], have been used to solve highly non-convex optimization problems and have been shown to achieve state-of-the-art performance on various RL tasks.

While many RL algorithms are easy to describe and run in practice, certain theoretical aspects of their behavior remain mysterious, even when they are applied in relatively simple settings. One such setting is the most canonical problem in continuous control, that of controlling a linear dynamical system with quadratic costs via the linear quadratic regulator (LQR). A recent line of work [AYS11, AYLS18, AL18, CHK<sup>+</sup>18, DMM<sup>+</sup>17, DMM<sup>+</sup>18, FTM17, FGKM18, TR18] has sought to delineate the properties and limitations of various RL algorithms in application to LQR problems. An appealing property of LQR systems from an analytical point of view is that the optimal policy is guaranteed to be linear in the states [Kal60, Whi96]. Thus, when the system dynamics are known, as in classical control, the optimal policy can be obtained by solving the discrete-time algebraic Ricatti equation.

In contrast, methods in reinforcement learning target the case of unknown dynamics, and seek to learn an optimal policy on the basis of observations. A basic form of model-free RL for linear quadratic systems involves applying derivative-free methods in the space of linear policies. It can be used even when the only observations possible are the costs from a set of rollouts, each referred to as a sample<sup>1</sup>, and when our goal is to obtain a policy whose cost is at most  $\epsilon$ -suboptimal. The sample complexity of a given method refers to the number of samples, as a function of the problem parameters and tolerance, required to meet a given tolerance  $\epsilon$ . With this context, we are led to the following concrete question: *What is the sample complexity of derivative-free methods for the linear quadratic regulator?* This question underlies the analysis in this paper. In particular, we study a standard derivative-free algorithm in an offline setting and derive explicit bounds on its sample complexity, carefully controlling the dependence on not only the tolerance  $\epsilon$ , but also the dimension and conditioning of the underlying problem.

Our analysis treats two distinct forms of randomness in the underlying linear system. In the first setting—more commonly assumed in practice—the linear updates are driven by an additive noise term [DMM<sup>+</sup>17], whereas in the second setting, the initial state is chosen randomly but the linear dynamics remain deterministic [FGKM18]. We refer to these two settings, respectively, as the *additive noise setting*, and the *randomly initialized setting*. We are now in a position to discuss related work on the problem, and to state our contributions.

**Related work:** Quantitative gaps between model-based and model-free reinforcement learning have been studied extensively in the setting of finite state-action spaces [AJ17, DLB17, AOM17], and several interesting questions here still remain open.

For continuous state-action spaces and in the specific context of the linear quadratic systems, classical system identification has been model-based, with a particular focus on asymptotic results (e.g., see the book [Lju98] as well as references therein). Non-asymptotic guarantees for model-based control of linear quadratic systems were first obtained by Fiechter [Fie97], who studied the offline problem under additive noise and obtained non-asymptotic rates for parameter identification using nominal control procedures. In more recent work, Dean et al. [DMM<sup>+</sup>17] proposed a robust alternative to nominal control, showing an improved sample complexity as well as better-behaved

---

<sup>1</sup>Such an *offline* setting with multiple, restarted rollouts should be contrasted with an online setting, in which the agent interacts continuously with the environment, and no hard resets are allowed. In contrast to the offline setting, the goal in the online setting is to control the system for all time steps while simultaneously learning better policies, and performance is usually measured in terms of *regret*.

policies. The online setting for model-based control of linear quadratic systems has also seen extensive study, with multiple algorithms known to achieve sub-linear regret [DMM<sup>+</sup>18, AYS11, AL18].

In this paper, we study model-free control of these systems, a problem that has seen some recent work in both the offline [FGKM18] and online [AYLS18] settings. Most directly relevant to our work is the paper of Fazel et al. [FGKM18], who studied the offline setting for the randomly initialized variant of the LQR, and showed that a population version of gradient descent, when run on the non-convex LQR cost objective, converges to the global optimum. In order to turn this into a derivative-free algorithm, they constructed near-exact gradient estimates from reward samples and showed that the sample complexity of such a procedure is bounded polynomially in the parameters of the problem; however, the dependence on various parameters is not made explicit in their analysis.

Also of particular relevance to our paper is the extensive literature on zero-order optimization. Flaxman et al. [FKM05] showed that these methods can be analyzed for convex optimization by making an explicit connection to function smoothing, and Agarwal et al. [ADX10] improved some of these convergence rates. Results are also available for strongly convex [JNR12], smooth [GL13] and convex [Nes11, DJWW15, WDBS18] functions, with Shamir [Sha13, Sha17] characterizing the fundamental limits of many problems in this space. Broadly speaking, all of the methods in this literature can be seen as variants of *stochastic search*: they proceed by constructing estimates of directional derivatives of the function from randomly chosen zero order evaluations. In the regime where the function evaluations are stochastic, different convergence rates are obtained based on whether such a procedure uses a *one-point estimate* that is obtained from a single function evaluation [FKM05], or a *k-point estimate* [ADX10] for some  $k \geq 2$ . There has also been some recent work on zero-order optimization of non-convex functions satisfying certain smoothness properties that are motivated by statistical estimation [WBS18].

**Our contributions** In this paper, we study both randomly initialized and additive-noise linear quadratic systems in the offline setting through the lens of derivative-free optimization. Our main contribution is to establish upper bounds on the sample complexity as a function of the dimension, error tolerance, and curvature parameters of the problem instance. In contrast to prior work, the rates that we provide are explicit, and the algorithms that we analyze are standard and practical one-point and two-point variants of the random search heuristic. Our results reveal interesting dichotomies between the settings of one-point and two-point feedback, as well as the models involving random initialization and additive noise. Our main contribution is stated in the following informal theorem (to be stated more precisely in the sequel):

**Main Theorem (informal).** *With high probability, one can obtain an  $\epsilon$ -approximate solution to any linear quadratic system from observing the noisy costs of  $\tilde{O}(1/\epsilon^2)$  trajectories from the system, which can be further reduced to  $\tilde{O}(1/\epsilon)$  trajectories when pairs of costs are observed for each trajectory.*

In our theoretical statements, the multiplicative pre-factors are explicit lower-order polynomials of the dimension of the state space, and curvature properties of the cost function. From a technical standpoint, we build upon some known properties of the LQR cost function established in past work on randomly initialized systems [FGKM18], and establish de novo some analogous properties

for the additive noise setting. We also isolate and sharpen some key properties that are essential to establishing sharp rates of zero-order optimization; as an example, for the setting with random-initialization and one-point reward feedback studied by Fazel et al. [FGKM18], establishing these properties allows us to analyze a natural algorithm that improves<sup>2</sup> the dependence of the bound on the error tolerance  $\epsilon$  from at least  $\mathcal{O}(1/\epsilon^4)$  to  $\mathcal{O}(1/\epsilon^2)$ . Crucially, our analysis is complicated by the fact that we must ensure that the iterates are confined to the region in which the linear system is stable, and such stability considerations introduce additional restrictions on the parameters used in our optimization procedure.

## 2 Background and problem set-up

In this section, we discuss the background related to zero-order optimization and the setup for the linear quadratic control problem.

### 2.1 Optimization background

We first introduce some standard optimization related background and assumptions, and make the zero-order setting precise.

**Stochastic zero-order optimization:** We consider optimization problems of the form

$$\min_{x \in \mathcal{X}} f(x) := \mathbb{E}_{\xi \sim \mathcal{D}} [F(x, \xi)], \quad (1)$$

where  $\xi$  is a zero mean random variable that represents the noise in the problem, and the function  $f$  above can be non-convex in general with a possibly non-convex domain  $\mathcal{X} \subseteq \mathbb{R}^d$ .

In particular, we consider stochastic zero-order optimization methods with oracle access to noisy function evaluations. We operate under two distinct oracle models. The first is the one-point setting, in which the optimizer specifies a point  $x \in \mathcal{X}$ , and an evaluation consists of the random variable  $F(x, \xi)$ . The second is the two-point extension of such a setting, in which the optimizer specifies a pair of points  $(x, y)$ , and obtains the random values  $F(x, \xi)$  and  $F(y, \xi)$ .

**Function properties:** Before defining the optimization problems considered in this paper by instantiating the pair of functions  $(f, F)$ , let us precisely define some standard properties that make repeated appearances in the sequel.

**Definition 1** (Locally Lipschitz Gradients). *A continuously differentiable function  $g$  with bounded domain  $\mathcal{X}$  is said to have  $(\phi, \beta)$  locally Lipschitz gradients at  $x \in \mathcal{X}$  if*

$$\|\nabla g(y) - \nabla g(x)\|_2 \leq \phi \|y - x\|_2 \quad \text{for all } y \in \mathcal{X} \text{ with } \|x - y\|_2 \leq \beta. \quad (2)$$

We often say that  $g$  has locally Lipschitz gradients, by which we mean for each  $x \in \mathcal{X}$  the function  $g$  has locally Lipschitz gradients, albeit with constants  $(\phi, \beta)$  that may depend on  $x$ . This property guarantees that the function  $g$  has at most quadratic growth locally around every point, but the shape of the quadratic and the radius of the ball within which such an approximation holds may depend on the point itself.

---

<sup>2</sup>While the rates established by Fazel et al. [FGKM18] are not explicit, their algorithm is conservative and a bound of order  $1/\epsilon^4$  can be distilled by working through their analysis.

**Definition 2** (Locally Lipschitz Function). *A continuously differentiable function  $g$  with bounded domain  $\mathcal{X}$  is said to be  $(\lambda, \zeta)$  locally Lipschitz at  $x \in \mathcal{X}$  if*

$$|g(y) - g(x)| \leq \lambda \|y - x\|_2 \quad \text{for all } y \in \mathcal{X} \text{ such that } \|x - y\|_2 \leq \zeta. \quad (3)$$

As before, when we say that the function  $g$  is locally Lipschitz, we mean that this condition holds for all  $x \in \mathcal{X}$ , albeit with parameters  $(\lambda, \zeta)$  that may depend on  $x$ . The local Lipschitz property guarantees that the function  $g$  grows no faster than linearly in a local neighborhood around each point.

**Definition 3** (PL Condition). *A continuously differentiable function  $g$  with bounded domain  $\mathcal{X}$  and a finite global minimum  $g^*$  is said to be  $\mu$ -PL if it satisfies the Polyak-Łojasiewicz (PL) inequality with constant  $\mu > 0$ , given by*

$$\|\nabla g(x)\|_2^2 \geq \mu (g(x) - g^*) \quad \text{for all } x \in \mathcal{X}. \quad (4)$$

The PL condition, first introduced by Polyak [Pol64] and Łojasiewicz [Loj63], is a relaxation of the notion of strong convexity. It allows for a certain degree of non-convexity in the function  $g$ . Note that Inequality (4) yields an upper bound on the gap to optimality that is proportional to the squared norm of the gradient. Thus, while the condition admits non-convex functions, it requires that all first-order stationary points also be global minimizers. Karimi et al. [KNS16] recently showed that many standard first-order convex optimization algorithms retain their attractive convergence guarantees over this more general class.

## 2.2 Optimal control background

We now turn to some basic background on optimal control and reinforcement learning. An optimal control problem is specified by a dynamics model and a real-valued cost function. The dynamics model consists of a sequence of functions  $\{h_t(s_t, a_t, z_t)\}_{t \geq 0}$ , which models how the state vector  $s_t$  transitions to the next state  $s_{t+1}$  when a control input  $a_t$  is applied at a timestep  $t$ . The term  $z_t$  captures the noise disturbance in the system. The cost function  $c_t(s_t, a_t)$  specifies the cost incurred by taking an action  $a_t$  in the state  $s_t$ . The goal of the control problem is to find a sequence of control inputs  $\{a_t\}_{t \geq 0}$ , dependent on the history of states  $\mathcal{H}_t := (s_0, s_1, \dots, s_{t-1})$ , so as to solve the optimization problem

$$\min \mathbb{E} \left[ \sum_{t \geq 0} \gamma^t c_t(s_t, a_t) \right] \quad \text{s.t. } s_{t+1} = h_t(s_t, a_t, z_t), \quad (5)$$

where the expectation above is with respect to the noise in the transition dynamics as well as any randomness in the selection of control inputs, and  $0 < \gamma \leq 1$  represents a multiplicative discount factor. A mapping from histories  $\mathcal{H}_t$  to controls  $a_t$  is called a *policy*, and the above minimization is effectively over the space of policies.

There is a distinction to be made here between the classical fully-observed setting in stochastic control in which the dynamics model  $h_t$  is known—in this case, such a problem may be solved (at least in principle) by the Bellman recursion (see, e.g., Bertsekas [Ber05]), and the system identification setting in which the dynamics are completely unknown. We operate in the latter setting, and accommodate the further assumption that even the cost function  $c_t$  is unknown.

In this paper, we assume that the state space is  $m$ -dimensional, and the control space is  $k$ -dimensional, so that  $s_t \in \mathbb{R}^m$  and  $a_t \in \mathbb{R}^k$ . The linear quadratic system specifies particular forms for the dynamics and costs, respectively. In particular, the cost function obeys the quadratic form

$$c_t = s_t^\top Q s_t + a_t^\top R a_t$$

for a pair of positive definite matrices  $(Q, R)$  of the appropriate dimensions. Additionally, the dynamics model is linear in both states and controls, and takes the form

$$s_{t+1} = A s_t + B a_t + z_t,$$

where  $A$  and  $B$  are transition matrices of the appropriate dimension, and the random variable  $z_t$  models additive noise in the problem which is drawn i.i.d. for each  $t$  from a distribution  $\mathcal{D}_{\text{add}}$ . We call this setting the *noisy dynamics* model.

We also consider the *randomly initialized* linear quadratic system without additive noise, in which the state transitions obey

$$s_{t+1} = A s_t + B a_t,$$

and the randomness in the problem comes from choosing the initial state  $s_0$  at random from a distribution  $\mathcal{D}_0$ .

Throughout this paper, we assume<sup>3</sup> that for both distributions  $\mathcal{D} \in \{\mathcal{D}_{\text{add}}, \mathcal{D}_0\}$  and for a random variable  $v \sim \mathcal{D}$ , we have

$$\mathbb{E}[v] = 0, \quad \mathbb{E}[v v^\top] = I, \quad \text{and} \quad \|v\|_2^2 \leq C_m \quad \text{a.s.} \quad (6)$$

While we assume boundedness of the distribution for convenience, our results extend straightforwardly to sub-Gaussian distributions by appealing to high-probability bounds for quadratic forms of sub-Gaussian random vectors [HKZ12] and standard truncation arguments. The final iteration complexity also changes by at most poly-logarithmic factors in the problem parameters; for brevity, we operate under the assumptions (6) throughout the paper and omit standard calculations for sub-Gaussian distributions.

By classical results in optimal control theory [Kal60, Whi96], the optimal controller for the LQR problem under both of these noise models takes the linear form  $a_t = -K^* s_t$ , for some matrix  $K^* \in \mathbb{R}^{k \times m}$ . When the system matrices are known, the controller matrix  $K^*$  can be obtained by solving the discrete-time algebraic Riccati equation [Ric24].

With the knowledge that the optimal policy is an invariant linear transformation of the state, one can reparametrize the LQR objective in terms of the linear class of policies, and focus on optimization procedures that only search over the class of linear policies. Below, we define such a parametrization under the noise models introduced above, and make explicit the connections to the stochastic optimization model (1).

---

<sup>3</sup>It is important to note that our assumption of identity covariance of the noise distributions can be made without loss of generality: for a problem with non-identity (but full-dimensional) covariance  $\Sigma$ , we may reparametrize the problem with the modifications

$$A' = \Sigma^{-1/2} A \Sigma^{1/2}, \quad B' = \Sigma^{-1/2} B, \quad \text{and} \quad s'_t = \Sigma^{-1/2} s_t \quad \text{for all } t \geq 0,$$

in which case the new problem with states  $s'_t$  and the pair of transition matrices  $(A', B')$  is driven by noise satisfying the assumptions (6).

**Random initialization** For each choice of the (random) initial state  $s_0$ , let  $\mathcal{C}_{\text{init},\gamma}(K; s_0)$  denote the cost of executing a linear policy  $K$  from initial state  $s_0$ , so that

$$\mathcal{C}_{\text{init},\gamma}(K; s_0) := \sum_{t=0}^{\infty} \gamma^t \left( s_t^\top Q s_t + a_t^\top R a_t \right), \quad (7)$$

where we have the noiseless dynamics  $s_{t+1} = A s_t + B a_t$  and  $a_t = -K s_t$  for each  $t \geq 0$ , and  $0 < \gamma \leq 1$ . While  $\mathcal{C}_{\text{init},\gamma}(K; s_0)$  is a random variable that denotes some notion of sample cost, our goal is to minimize the population cost

$$\mathcal{C}_{\text{init},\gamma}(K) := \mathbb{E}_{s_0 \sim \mathcal{D}_0} [\mathcal{C}_{\text{init},\gamma}(K; s_0)] \quad (8)$$

over choices of the policy  $K$ .

**Noisy dynamics** In this case, the noise in the problem is given by the sequence of random variables  $\mathcal{Z} = \{z_t\}_{t \geq 0}$ , and for every instantiation of  $\mathcal{Z} \sim \mathcal{D}_{\text{add}}^{\mathbb{N}} := (\mathcal{D}_{\text{add}} \otimes \mathcal{D}_{\text{add}} \otimes \dots)$ , our sample cost is given by the function

$$\mathcal{C}_{\text{dyn},\gamma}(K; \mathcal{Z}) := \sum_{t=0}^{\infty} \gamma^t \left( s_t^\top Q s_t + a_t^\top R a_t \right),$$

where we have  $s_0 = 0$ , random state evolution  $s_{t+1} = A s_t + B a_t + z_t$  and action  $a_t = -K s_t$  for each  $t \geq 0$ , and  $0 < \gamma < 1$ . In contrast to the random initialization setting, this setting involves a discount factor  $\gamma < 1$ , since this is required to keep the costs finite.

Once again, we are interested in optimizing the population cost function

$$\mathcal{C}_{\text{dyn},\gamma}(K) := \mathbb{E}_{\mathcal{Z} \sim \mathcal{D}_{\text{add}}^{\mathbb{N}}} [\mathcal{C}_{\text{dyn},\gamma}(K; \mathcal{Z})]. \quad (9)$$

From here on, the word policy will always refer to a linear policy, and since we work with this natural parametrization of the cost function, our problem has effective dimension  $D = m \cdot k$ , given by the product of state and control dimensions.

A policy  $K$  is said to stabilize the system  $(A, B)$  if we have  $\rho_{\text{spec}}(A - BK) < 1$ , where  $\rho_{\text{spec}}(\cdot)$  denotes the spectral radius of a matrix. We assume throughout that the LQR system to be optimized is controllable, meaning that there exists some policy  $K$  satisfying the condition  $\rho_{\text{spec}}(A - BK) < 1$ . Furthermore, we assume access to *some* policy  $K_0$  with finite cost (see the related literature [FGKM18, DMM<sup>+</sup>18]); we use such a policy  $K_0$  as an initialization for our algorithms.

### 2.2.1 Some properties of the LQR cost function

Let us turn to establishing properties of the pair of population cost functions  $(\mathcal{C}_{\text{init},\gamma}(K), \mathcal{C}_{\text{dyn},\gamma}(K))$  and their respective sample variants  $(\mathcal{C}_{\text{init},\gamma}(K, s_0), \mathcal{C}_{\text{dyn},\gamma}(K; \mathcal{Z}))$ , in order to place the problem within the context of optimization.

First, it is important to note that both the population cost functions  $(\mathcal{C}_{\text{init},\gamma}(K), \mathcal{C}_{\text{dyn},\gamma}(K))$  are non-convex. In particular, for any unstable policy, the state sequence blows up and the costs become infinite, but as noted by Fazel et al. [FGKM18], the stabilizing region  $\{K : \rho_{\text{spec}}(A - BK) < 1\}$  is non-convex, thereby rendering our optimization problems non-convex.

In spite of this non-convexity, the cost functions exhibit many properties that make them amenable to fast stochastic optimization methods. Variants of the following properties were first established by Fazel et al. [FGKM18] for the random initialization cost function  $\mathcal{C}_{\text{init},\gamma}$ . The following Lemma 1 and Lemma 2 require certain refinements of their claims, which we prove in Appendix A. Lemma 3 follows directly from Lemma 3 in Fazel et al. [FGKM18]. Lemma 4 relates the population cost of the noisy dynamics model to that of the random initialization model in a pointwise sense.

**Lemma 1** (LQR Cost is locally Lipschitz). *Given any linear policy  $K$ , there exist positive scalars  $(\lambda_K, \widetilde{\lambda}_K, \zeta_K)$ , depending on the function value  $\mathcal{C}_{\text{init},\gamma}(K)$ , such that for all policies  $K'$  satisfying  $\|K' - K\|_F \leq \zeta_K$ , and for all initial states  $s_0$ , we have*

$$|\mathcal{C}_{\text{init},\gamma}(K') - \mathcal{C}_{\text{init},\gamma}(K)| \leq \lambda_K \|K' - K\|_F, \text{ and} \quad (10a)$$

$$|\mathcal{C}_{\text{init},\gamma}(K'; s_0) - \mathcal{C}_{\text{init},\gamma}(K; s_0)| \leq \widetilde{\lambda}_K \|K' - K\|_F. \quad (10b)$$

**Lemma 2** (LQR Cost has locally Lipschitz Gradients). *Given any linear policy  $K$ , there exist positive scalars  $(\beta_K, \phi_K)$ , depending on the function value  $\mathcal{C}_{\text{init},\gamma}(K)$ , such that for all policies  $K'$  satisfying  $\|K' - K\|_F \leq \beta_K$ , we have*

$$\|\nabla \mathcal{C}_{\text{init},\gamma}(K') - \nabla \mathcal{C}_{\text{init},\gamma}(K)\|_F \leq \phi_K \|K' - K\|_F. \quad (11)$$

**Lemma 3** (LQR satisfies PL). *There exists a universal constant  $\mu_{\text{lqr}} > 0$  such that for all stable policies  $K$ , we have*

$$\|\nabla \mathcal{C}_{\text{init},\gamma}(K)\|_F^2 \geq \mu_{\text{lqr}} (\mathcal{C}_{\text{init},\gamma}(K) - \mathcal{C}_{\text{init},\gamma}(K^*)),$$

where  $K^*$  is the global minimum of the cost function  $\mathcal{C}_{\text{init},\gamma}$ .

For the sake of exposition, we have stated these properties without specifying the various smoothness and PL constants. Please see Appendix A for explicit expressions for the tuple  $(\lambda_K, \widetilde{\lambda}_K, \phi_K, \beta_K, \zeta_K, \mu_{\text{lqr}})$  as functions of the parameters of the LQR problem.

**Lemma 4** (Equivalence of population costs up to scaling). *For all policies  $K$ , we have*

$$\mathcal{C}_{\text{dyn},\gamma}(K) = \frac{\gamma}{1-\gamma} \mathcal{C}_{\text{init},\gamma}(K).$$

Lemma 4 thus shows that, at least in a population sense, both the noisy dynamics and random initialization models behave identically when driven by noise with the same first two moments. Hence, the properties posited by Lemmas 1, 2, and 3 for the cost function  $\mathcal{C}_{\text{init},\gamma}(K)$  also carry over to the function  $\mathcal{C}_{\text{dyn},\gamma}(K)$ . In particular, the noisy cost function  $\mathcal{C}_{\text{dyn},\gamma}(K)$  is also  $\left(\frac{\gamma}{1-\gamma} \phi_K, \beta_K\right)$  locally smooth and  $\left(\frac{\gamma}{1-\gamma} \lambda_K, \zeta_K\right)$  locally Lipschitz, and also globally  $\frac{\gamma}{1-\gamma} \mu_{\text{lqr}}$ -PL.

## 2.2.2 Stochastic zero-order oracle in LQR

Let us now describe the form of observations that we make in the LQR system. Recall that we are operating in the derivative-free setting, where we have access to only (noisy) function evaluations and not the problem parameters; in particular, the tuple  $(A, B, Q, R)$  that parametrizes the LQR problem is unknown.



Our observations consist of the noisy function evaluations  $\mathcal{C}_{\text{init},\gamma}(K; s_0)$  and  $\mathcal{C}_{\text{dyn},\gamma}(K; \mathcal{Z})$ . We consider both the one-point and two-point settings in the former case. In the one-point setting for the randomly initialized model, a *query* of the function at the point  $K$  obtains the noisy function value  $\mathcal{C}_{\text{init},\gamma}(K; s_0)$  for an initial state  $s_0$  drawn at random from the distribution  $\mathcal{D}_0$ . In the two-point setting, a query of the function at the points  $(K, K')$  obtains the pair of noisy function values  $\mathcal{C}_{\text{init},\gamma}(K; s_0)$  and  $\mathcal{C}_{\text{init},\gamma}(K'; s_0)$  for an initial state  $s_0$  drawn at random; this setting has an immediate operational interpretation as running two policies with the same random initialization. The one-point query model is defined analogously for the noisy dynamics cost  $\mathcal{C}_{\text{dyn},\gamma}$ .

A few points regarding our query model merit discussion. First, note that in the context of the control objective, each query produces a noisy sample of the long term trajectory cost, and so our sample complexity is measured in terms of the number of *rollouts*, or trajectories. Such an assumption is reasonable since the “true” sample complexity that takes into account the length of the trajectories is only larger by a small factor—the truncated, finite cost converges exponentially quickly to the infinite sum for stable policies. Second, we note that while the one-point query model was studied by Fazel et al. [FGKM18] for the random initialization model—albeit with sub-optimal guarantees—we also study a two-point query model, which is known to lead to better dimension-dependence in zero-order stochastic optimization [DJWW15].

### 3 Main results

We now turn to a statement of our main result, which characterizes the convergence rate of a natural derivative-free algorithm for any (population) function that satisfies certain PL and smoothness properties. We thus obtain, as corollaries, rates of zero-order optimization algorithms when applied to the functions  $\mathcal{C}_{\text{init},\gamma}$  and  $\mathcal{C}_{\text{dyn},\gamma}$ .

#### 3.1 Stochastic zero-order algorithm

We analyze a standard zero-order algorithm for stochastic optimization [ADX10, Sha17] in application to the LQR problem. We begin by introducing some notation required to describe this algorithm, operating in the general setting where we want to optimize a function  $f : \mathcal{X} \mapsto \mathbb{R}$  of the form  $f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[F(x; \xi)]$ . Here we assume the inclusion  $\mathcal{X} \subseteq \mathbb{R}^d$ , and let  $\mathcal{D}$  denote a generic source of randomness in the zero-order function evaluation.

The zero-order algorithms that we study here use noisy function evaluations in order to construct near-unbiased estimates of the gradient. Let us now describe how such an estimate is constructed in the one-point and two-point settings. Let  $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$  denote the  $d$ -dimensional unit shell. Let  $\text{Unif}(\mathbb{S}^{d-1})$  denote the uniform distribution over the set  $\mathbb{S}^{d-1}$ .

For a given scalar  $r > 0$  and a random direction  $u \sim \text{Unif}(\mathbb{S}^{d-1})$  chosen independently of the random variable  $\xi$ , consider the one point gradient estimate

$$g_r^1(x, u, \xi) := F(x + ru, \xi) \frac{d}{r} u, \quad (12a)$$

and its two-point analogue

$$g_r^2(x, u, \xi) := [F(x + ru, \xi) - F(x - ru, \xi)] \frac{d}{2r} u. \quad (12b)$$

In both of these cases, the resulting ratios are almost unbiased approximations of the secant ratio that defines the derivative at  $x$ , and these approximations get better and better as the *smoothing radius*  $r$  gets smaller. On the other hand, small values of the radius  $r$  may result in estimates with large variance. Our algorithms make use of such randomized approximations in a sequence of rounds by choosing appropriate values of the radius  $r$ ; the general form of such an algorithm is stated below.

---

**Algorithm 1** Stochastic Zero-Order Method

---

```

1: Given iteration number  $T \geq 1$ , initial point  $x_0 \in \mathcal{X}$ , step size  $\eta > 0$  and smoothing radius  $r > 0$ 
2: for  $t \in \{0, 1, \dots, T-1\}$  do
3:   Sample  $\xi_t \sim \mathcal{D}$  and  $u_t \sim \text{Unif}(\mathbb{S}^{d-1})$ 
4:    $g(x_t) \leftarrow \begin{cases} g_r^1(x_t, u_t, \xi_t) & \text{if operating in one-point setting} \\ g_r^2(x_t, u_t, \xi_t) & \text{if operating in two-point setting.} \end{cases}$ 
5:    $x_{t+1} \leftarrow x_t - \eta g(x_t)$ 
return  $x_T$ 

```

---

### 3.2 Convergence guarantees

We now turn to analyzing Algorithm 1 in the settings of interest. As mentioned before, the difficulty of optimizing the LQR cost functions is governed by multiple factors such as stability, non-convexity of the feasible set, and non-convexity of the objective. Furthermore, the Lipschitz gradient and Lipschitz properties for this cost function only hold locally with the radius of locality depending on the current iterate. Most crucially, the function is infinite outside of the region of stability, and so large steps can have disastrous consequences since we do not have access to a projection oracle that brings us back into the region of stability. It is thus essential to control the behavior of our stochastic, high variance algorithm over the entire course of optimization.

Our strategy to overcome these challenges is to perform a careful martingale analysis, showing that the iterates remain bounded throughout the course of the algorithm; the rate depends, among other things, on the variance of the gradient estimates obtained over the course of the algorithm. By showing that the algorithm remains within the region of finite cost, we can also obtain good bounds on the local Lipschitz constants and gradient smoothness parameters, so that our step-size can be set accordingly.

Let us now introduce some notation in order to make this intuition precise. We operate once again in the setting of general function optimization, i.e., we are interested in optimizing a function  $f(x) = \mathbb{E}_\xi[F(x; \xi)]$  obeying the PL inequality as well as certain local curvature conditions.

Recall that we are given an initial point  $x_0$  with finite cost  $f(x_0)$ ; the global upper bound on the cost that we target in the analysis is set according to the cost  $f(x_0)$  of this initialization. Given the initial gap to optimality  $\Delta_0 := f(x_0) - f(x^*)$ , we define the set

$$\mathcal{G}^0 := \{x \mid f(x) - f(x^*) \leq 10\Delta_0\}, \quad (13)$$

corresponding to points  $x$  whose cost gap is at most ten times the initial cost gap  $\Delta_0$ .

Assume that the function  $f$  is  $(\phi_x, \beta_x)$  locally smooth and  $(\lambda_x, \zeta_x)$  locally Lipschitz at the point  $x$ . Thus, both of these properties hold simultaneously within a neighbourhood of radius

$\rho_x = \min\{\beta_x, \zeta_x\}$  of the point  $x$ . Now define the quantities

$$\phi_0 := \sup_{x \in \mathcal{G}^0} \phi_x, \quad \lambda_0 := \sup_{x \in \mathcal{G}^0} \lambda_x, \quad \text{and} \quad \rho_0 := \inf_{x \in \mathcal{G}^0} \rho_x.$$

By defining these quantities, we have effectively transformed the local properties of the function  $f$  into global properties that hold over the bounded set  $\mathcal{G}^0$ . We also define a convenient functional of these curvature parameters  $\theta_0 := \min\left\{\frac{1}{2\phi_0}, \frac{\rho_0}{\lambda_0}\right\}$ , which simplifies the statements of our results. Importantly, these smoothness properties only hold locally, and so we must also ensure that the steps taken by our algorithm are not too large. This is controlled by both the step-size as well as the norms of our gradient estimate  $g$  computed over the course of the algorithm. Define the uniform bounds

$$G_\infty = \sup_{x \in \mathcal{G}^0} \|g(x)\|_2, \quad \text{and} \quad G_2 = \sup_{x \in \mathcal{G}^0} \mathbb{E} [\|g(x) - \mathbb{E}[g(x) \mid x]\|_2^2]$$

on the point-wise gradient norm and its variance, respectively. Note that these quantities also depend implicitly on the smoothing radius  $r$  and on how the gradient estimate  $g$  is computed.

With this set-up, we are now ready to state the main result regarding the convergence rate of Algorithm 1 on the functions of interest.

**Theorem 1.** *Suppose that the step-size and smoothing radius are chosen so as to satisfy*

$$\eta \leq \min\left\{\frac{\epsilon\mu}{240\phi_0 G_2}, \frac{1}{2\phi_0}, \frac{\rho_0}{G_\infty}\right\}, \quad \text{and} \quad r \leq \min\left\{\frac{\theta_0\mu}{8\phi_0}\sqrt{\frac{\epsilon}{15}}, \frac{1}{2\phi_0}\sqrt{\frac{\epsilon\mu}{30}}, \rho_0\right\}. \quad (14a)$$

*Then for a given error tolerance  $\epsilon$  such that  $\epsilon \log(120\Delta_0/\epsilon) < \frac{10}{3}\Delta_0$ , the iterate  $x_T$  of Algorithm 1 after  $T = \frac{4}{\eta\mu} \log\left(\frac{120\Delta_0}{\epsilon}\right)$  steps satisfies the bound*

$$f(x_T) - f(x^*) \leq \epsilon \quad (14b)$$

*with probability greater than  $3/4$ .*

A few comments on Theorem 1 are in order. First, notice that the algorithm is guaranteed to return an  $\epsilon$ -accurate solution with constant probability. This success probability can be improved to the value  $1 - \delta$ , for any  $\delta \in (0, \frac{1}{4})$ , by running the algorithm  $\tilde{\mathcal{O}}(\log(\frac{1}{\delta}))$  times and choosing the iterate with the smallest final cost. Such procedures to boost constant probability results to high probability ones are standard in the literature on randomized algorithms [MU05, TVar]. Further, the probability bound of  $\frac{3}{4}$  in itself can be sharpened by a slightly more refined analysis with different constants. Additionally, by examining the proof, it can be seen that we establish a result (cf. Proposition 1 in Section 4) that is slightly stronger than Theorem 1, and then obtain the theorem from this more general result. The proof of the theorem itself is relatively short, and makes use of a carefully constructed martingale along with an appropriately defined stopping time. As mentioned before, the main challenge in the proof is to ensure that we have bounded iterates while still preserving the strong convergence properties of zero-order stochastic methods for smooth functions that satisfy the PL property.

It should be noted that Theorem 1 is a general guarantee: it characterizes the zero-order complexity of optimizing locally smooth functions that satisfy a PL inequality in terms of properties of the gradient estimates obtained over the course of algorithm. In particular, two properties of

Parameter settings Query Model	Smoothing radius $r$	Variance $G_2$	Step-size $\eta$	#queries $T$
One-point LQR (Random initialization/ Noisy dynamics)	$\mathcal{O}(\sqrt{\epsilon})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^2)$	$\tilde{\mathcal{O}}(\epsilon^{-2})$
Two-point LQR (Random initialization)	$\mathcal{O}(\sqrt{\epsilon})$	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon)$	$\tilde{\mathcal{O}}(\epsilon^{-1})$

**Table 1.** Derivative-free complexity of LQR optimization under the two query models, as a function of the final error tolerance  $\epsilon$ . The multiplicative pre-factors are functions of the effective dimension  $D$  and curvature parameters, and differ in the three cases; see the statements of the corollaries below.

these estimates appear: the variance of the estimate, as well as a uniform bound on its size. These quantities, in turn, depend on both the noise in the zero-order evaluations as well as our choice of query model. In the next section, we specialize Theorem 1 so as to derive particular consequences for the LQR models introduced above.

### 3.3 Consequences for LQR optimization

Theorem 1 yields immediate consequences for LQR optimization in various settings, and the dependence of the optimization rates on the tolerance  $\epsilon$  is summarized by Table 1. We state and discuss precise versions of these results below.

First, let us consider the random initialization model. From the various lemmas in Section 2.2.1, we know that the population objective  $\mathcal{C}_{\text{init},\gamma}(K)$  is locally  $(\phi_K, \beta_K)$  smooth and  $(\lambda_K, \zeta_K)$  Lipschitz, and also globally  $\mu_{\text{lqr}}$ -PL. By assumption, we are given a starting point  $K_0$  having finite population cost  $\mathcal{C}_{\text{init},\gamma}(K_0)$ . Proceeding as in the previous section, we may thus define the set

$$\mathcal{G}^{\text{lqr}} := \{K \mid \mathcal{C}_{\text{init},\gamma}(K) - \mathcal{C}_{\text{init},\gamma}(K^*) \leq 10\Delta_0\}, \quad (15)$$

corresponding to point  $x$  whose cost gap is at most ten times the initial cost gap to optimality  $\Delta_0 = \mathcal{C}_{\text{init},\gamma}(K_0) - \mathcal{C}_{\text{init},\gamma}(K^*)$ .

Now define the quantities

$$\phi_{\text{lqr}} := \sup_{K \in \mathcal{G}^{\text{lqr}}} \phi_K, \quad \lambda_{\text{lqr}} := \sup_{K \in \mathcal{G}^{\text{lqr}}} \lambda_K, \quad \text{and} \quad \rho_{\text{lqr}} := \inf_{K \in \mathcal{G}^{\text{lqr}}} \rho_K,$$

thereby transforming the local smoothness properties of the function  $\mathcal{C}_{\text{init},\gamma}$  into global properties that hold over the bounded set  $\mathcal{G}^0$ . Once again, let  $\theta_{\text{lqr}} := \min \left\{ \frac{1}{2\phi_{\text{lqr}}}, \frac{\rho_{\text{lqr}}}{\lambda_{\text{lqr}}} \right\}$  be a functional of these curvature parameters that simplifies the statements of our results.

With this setup, we now establish the following corollaries for derivative-free policy optimization for linear quadratic systems.

**Corollary 1** (One-point, Random initialization). *Suppose that the step-size and smoothing radius are chosen such that*

$$\eta \leq C \min \left\{ \frac{\epsilon \mu_{\text{lqr}} r^2}{\phi_{\text{lqr}} C_m^2 D^2 [\mathcal{C}_{\text{init},\gamma}(K_0)]^2}, \frac{1}{\phi_{\text{lqr}}}, \frac{\rho_{\text{lqr}} r}{C_m D [\mathcal{C}_{\text{init},\gamma}(K_0)]} \right\}, \text{ and}$$

$$r \leq \min \left\{ \frac{\theta_{\text{lqr}} \mu_{\text{lqr}}}{8\phi_{\text{lqr}}} \sqrt{\frac{\epsilon}{15}}, \frac{1}{2\phi_{\text{lqr}}} \sqrt{\frac{\epsilon \mu_{\text{lqr}}}{30}}, \rho_{\text{lqr}}, \frac{10\mathcal{C}_{\text{init},\gamma}(K_0)}{\lambda_{\text{lqr}}} \right\}.$$

Then for any error tolerance  $\epsilon$  such that  $\epsilon \log(120\Delta_0/\epsilon) < \frac{10}{3}\Delta_0$ , running Algorithm 1 for  $T = \frac{4}{\eta\mu} \log\left(\frac{120\Delta_0}{\epsilon}\right)$  iterations yields an iterate  $K_T$  such that

$$\mathcal{C}_{\text{init},\gamma}(K_T) - \mathcal{C}_{\text{init},\gamma}(K^*) \leq \epsilon$$

with probability greater than  $3/4$ .

Leaving a discussion of the dependence on various parameters for later, let us briefly contrast this result with that of Fazel et al. [FGKM18], who also operate in the one-point setting. As mentioned before, a crucial condition that we must ensure is that our iterates never wander into the region of instability, since this, coupled with the non-convexity of the objective, substantially hinders the convergence of the algorithm. Fazel et al. overcome these issues by using a zero-order algorithm that, for each step, follows an extremely accurate approximation to the gradient descent step via a large number of zero-order queries<sup>4</sup>. By estimating the gradient to high accuracy, they are able to guarantee that the cost function on the iterates is monotonically decreasing, thereby ensuring that it never wanders into the unstable region. This close tracking of the population gradient step, combined with their result for population gradient descent on this problem, then yields a polynomial bound on the zero-order complexity.

On the other hand, we operate in the setting in which our estimates of the gradient are obtained through a single one-point evaluation. Such a single evaluation yields a noisy estimate of the gradient, so that we cannot—at least in general—guarantee the descent property above due to randomness in these evaluations; we thus require the uniform control afforded by Theorem 1 in order to ensure convergence.

On a related note, it is important to point out that both of these algorithms fall under the broad umbrella of *minibatch* derivative-free algorithms, where  $k$  zero-order samples are used to estimate the gradient at any point: our algorithm corresponds to the case  $k = 1$ , while that of Fazel et al. corresponds to the case of some large  $k$ . In principle, Theorem 1 allows us to provide a family of results for each value of  $k$ , in which the variance of the gradient is reduced by a factor  $k$ , allowing us to increase our step-size proportionally and converge in  $1/k$ -fraction of the number of iterations (but with the same number of zero-order evaluations in total). However, the convergence guarantee only holds provided the step-size is bounded by the quantity  $\frac{r\rho_{\text{lqr}}}{10\mathcal{C}_{\text{init},\gamma}(K_0)}$ , since we only have local control on the curvature properties of the function. Operationally speaking, this means that for larger step-sizes, we are unable to guarantee stability of the policies obtained over the course of the algorithm. Such a bottleneck is in fact also observed in practice, as shown in Figure 1 for both the one-point and two-point settings.

Let us now turn to the two-point setting, in which we obtain two noisy evaluations per query.

**Corollary 2** (Two-point, Random initialization). *Suppose that the step-size and smoothing radius are chosen so as to satisfy*

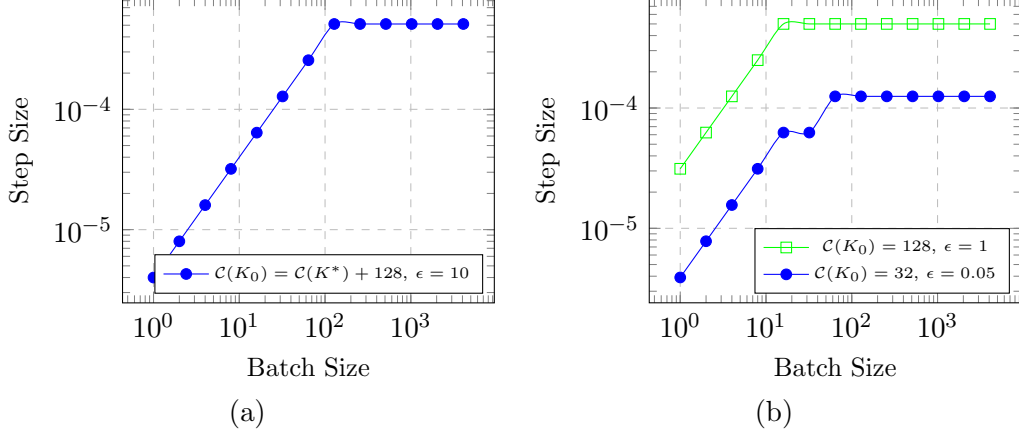
$$\eta \leq \min \left\{ \frac{\epsilon\mu_{\text{lqr}}}{240\phi_{\text{lqr}}D\lambda_{\text{lqr}}^2}, \frac{1}{2\phi_{\text{lqr}}}, \frac{\rho_{\text{lqr}}}{D\lambda_{\text{lqr}}} \right\}, \quad \text{and} \quad r \leq \min \left\{ \frac{\theta_{\text{lqr}}\mu_{\text{lqr}}}{8\phi_{\text{lqr}}} \sqrt{\frac{\epsilon}{15}}, \frac{1}{2\phi_{\text{lqr}}} \sqrt{\frac{\epsilon\mu_{\text{lqr}}}{30}}, \rho_{\text{lqr}} \right\}.$$

Then for any error tolerance  $\epsilon$  such that  $\epsilon \log(120\Delta_0/\epsilon) < \frac{10}{3}\Delta_0$ , running Algorithm 1 for  $T = \frac{4}{\eta\mu} \log\left(\frac{120\Delta_0}{\epsilon}\right)$  iterations yields an iterate  $K_T$  such that

$$\mathcal{C}_{\text{init},\gamma}(K_T) - \mathcal{C}_{\text{init},\gamma}(K^*) \leq \epsilon$$

---

<sup>4</sup>In their analysis, Fazel et al. [FGKM18] require gradient estimates that are  $\mathcal{O}(\epsilon)$  accurate in operator norm, using a value of  $r$  that is  $\mathcal{O}(\epsilon)$ . This leads to a final bound on zero-order complexity that is at least of order  $\epsilon^{-4}$ .



**Figure 1.** Plot of the maximum step-size that allows for convergence, plotted against the size of the mini-batch used to estimate the gradient in randomly initialized LQR with (a) one-point evaluations and (b) two-point evaluations. The step-size plateaus due to stability considerations, leading to a higher zero-order complexity in spite of the lower variance estimates afforded by large batch-sizes. Plots were obtained by averaging 20 runs of Algorithm 1. For more problem details, see Appendix D.

with probability greater than  $3/4$ .

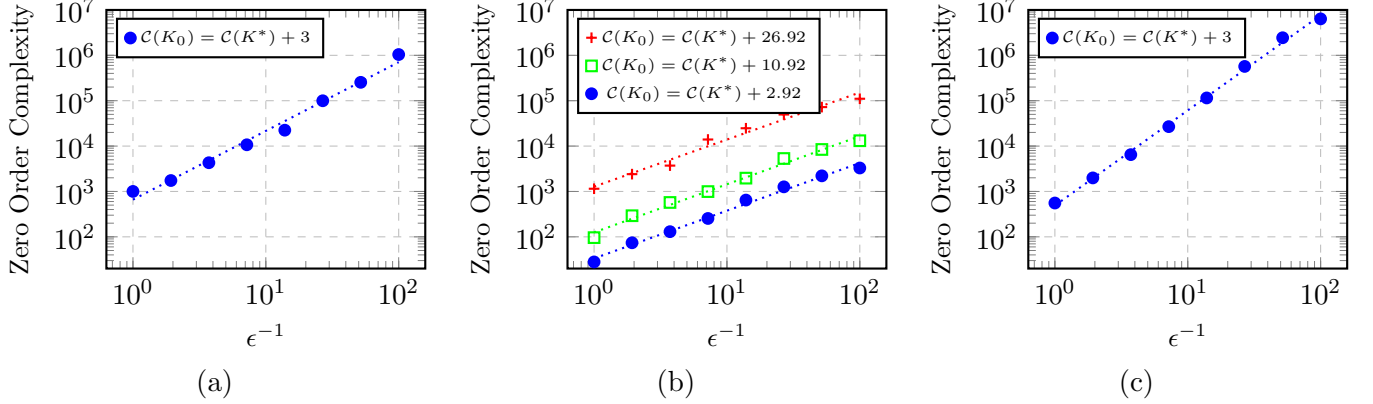
As known from the literature on zero-order optimization in convex settings [DJWW15, Sha17], the two-point query model allows us to substantially reduce the variance of our gradient estimate, thus ensuring much faster convergence than with one-point evaluations. The most salient difference is the fact that we now converge with  $\tilde{\mathcal{O}}(1/\epsilon)$  iterations as opposed to the  $\tilde{\mathcal{O}}(1/\epsilon^2)$  iterations required in Corollary 1. This gap between the two settings is substantial and merits further investigation, but in general, it is clear that two-point evaluations should certainly be used if available. This gap, and other differences, are discussed in the sequel.

Let us now turn to establishing convergence results for the noisy dynamics model in the one-point setting. Note that Lemma 4 provides a way to directly relate the population costs of the random initialization and noisy dynamics models; furthermore, the set  $\mathcal{G}^{\text{lqr}}$  is exactly the same. In addition, since we look at a discounted cost  $\mathcal{C}_{\text{dyn},\gamma}$  in this setting, the corresponding curvature parameters have an inherent dependence on  $\gamma$  which we denote using corresponding subscripts. With an additional computation of the variance and norm of the gradient estimates, we then obtain the following corollary for one-point optimization of the noisy dynamics model. Our statement involves the constants

$$G_{2,\text{lqr}} := \left( \frac{D}{r} \cdot \frac{2C_m}{1 - \sqrt{\gamma}} \right)^2 \cdot \left( \frac{20\mathcal{C}_{\text{dyn},\gamma}(K_0)}{\sigma_{\min}(Q)} \left( \frac{1 - \gamma}{\gamma} \right) \right)^3 \quad \text{and}$$

$$G_{\infty,\text{lqr}} := \frac{D}{r} \cdot \frac{2C_m}{1 - \sqrt{\gamma}} \cdot \left( \frac{20\mathcal{C}_{\text{dyn},\gamma}(K_0)}{\sigma_{\min}(Q)} \left( \frac{1 - \gamma}{\gamma} \right) \right)^{3/2}.$$

**Corollary 3** (One-point, Noisy dynamics). *Suppose that the step-size and smoothing radius are*



**Figure 2.** Number of samples required to reach an error tolerance of  $\epsilon$ , plotted against  $1/\epsilon$ , for (a) Randomly initialized LQR with one-point evaluations (b) Randomly initialized LQR with two-point evaluations for differing values of the initial cost, and (c) Noisy dynamics LQR model with one-point evaluations. We use  $\mathcal{C}$  to denote the population cost in the various cases, and the plots were obtained by averaging 20 runs of Algorithm 1. Each dotted line represents the line of best fit for the corresponding data points. For more problem details, see Appendix D.

chosen so as to satisfy

$$\eta \leq \min \left\{ \frac{\epsilon \mu_{\text{lqr}, \gamma}}{240 \phi_{\text{lqr}, \gamma} G_{2, \text{lqr}}}, \frac{1}{2 \phi_{\text{lqr}, \gamma}}, \frac{\rho_{\text{lqr}, \gamma}}{G_{\infty, \text{lqr}}} \right\}, \text{ and}$$

$$r \leq \min \left\{ \frac{\theta_{\text{lqr}, \gamma} \cdot \mu_{\text{lqr}, \gamma}}{8 \phi_{\text{lqr}, \gamma}} \sqrt{\frac{\epsilon}{15}}, \frac{1}{2 \phi_{\text{lqr}, \gamma}} \sqrt{\frac{\epsilon \cdot \mu_{\text{lqr}, \gamma}}{30}}, \rho_{\text{lqr}, \gamma} \right\}.$$

Then for any error tolerance  $\epsilon$  such that  $\epsilon \log(120 \Delta_0 / \epsilon) < \frac{10}{3} \Delta_0$ , Algorithm 1 with  $T = \frac{4}{\eta \mu_{\text{lqr}, \gamma}} \log \left( \frac{120 \Delta_0}{\epsilon} \right)$  iterations yields an iterate  $K_T$  such that

$$\mathcal{C}_{\text{dyn}, \gamma}(K_T) - \mathcal{C}_{\text{dyn}, \gamma}(K^*) \leq \epsilon$$

with probability greater than  $3/4$ .

Thus, we have shown that the one-point settings for both the random initialization and noisy dynamics models exhibit similar behaviors in the different parameters. Reasoning heuristically, such a behavior is due to the fact that the additional additive noise in the dynamics is quickly damped away by the discount factor, so that the cost is dominated by the noise in the initial iterates. The variance bound, however, is substantially different, and this leads to the differing dependence on the smoothness parameters and dimension of the problem.

Another interesting problem studied in the noisy dynamics model is one of bounding the regret of online procedures. Equipped with a high probability bound on convergence—as opposed to the constant probability bound currently posited by Corollary 3, the offline guarantee and associated algorithm can in principle be turned into a no-regret learner in the online setting. We leave this to future work.

Let us now briefly discuss the dependence of the various bounds on the different parameters of the LQR objective, in the various cases above.

**Dependence on  $\epsilon$ :** Our bounds illustrate two distinct dependences on the tolerance parameter  $\epsilon$ . In particular, the zero-order complexity scales proportional to  $\epsilon^{-2}$  for both one-point settings (Corollaries 1 and 3), but proportional to  $\epsilon^{-1}$  in the two-point setting (Corollary 2). As alluded to before, this distinction arises due to the lower variance of the gradient estimator in the two-point setting. Lemma 1 establishes the Lipschitz property of the LQR cost function for each instantiation of the noise variable  $s_0$ , which ensures that the Lipschitz constant of our *sample* cost function is also bounded; therefore, the noise of the problem reduces as we approach the optimum solution. In contrast, the optimization problem with one-point evaluations becomes more difficult the closer we are to the optimum solution, since the noise remains constant, while the “signal” in the problem (measured by the rate of decrease of the population cost function) reduces as we approach the optimum. The  $O(1/\epsilon^2)$  dependence in the one-point settings is reminiscent of the complexity required to optimize strongly convex and smooth functions [ADX10, Sha13], and it would be interesting if a matching lower bound could also be proved in this LQR setting<sup>5</sup>. Even in the absence of such a lower bound, the one-point setting is strictly worse than the two-point setting even with respect to the other parameters of the problem, which we discuss next. Figure 2 shows the convergence rate of the algorithm in all three settings as a function of  $\epsilon$ , where we confirm that scalings in practice corroborate our theory quite accurately.

**Dependence on dimension:** The dependence on dimension enters once again via our bound on the variance of the gradient estimate, as is typical of many derivative-free procedures [DJWW15, Sha17]. The two-point setting gives rise to the best dimension dependence (linear in  $D$ ), and the reason is similar to why this occurs for convex optimization [Sha17]. It is particularly interesting to compare the dimension dependence to results in model-based control. There, in the noisy dynamics model, the sample complexity scales with the sum of state and control dimensions  $m + k$ , whereas the dependence in the two-point setting is on their product  $D = m \cdot k$ . However, each observation in that setting consists of a state vector of length  $m$ , while here we only get access to scalar cost values, and so in that loose sense, the complexities of the two settings are comparable.

In the one-point setting, the dependence on dimension is significantly poorer, and at least quadratic. This of course ignores other dimension-dependent factors such as  $C_m$ , as well as the curvature parameters  $(\phi_{\text{lqr}}, \lambda_{\text{lqr}}, \mu)$  (see the discussion below).

**Dependence on curvature parameters:** The iteration complexity scales linearly in the smoothness parameter of the problem  $\phi_{\text{lqr}}$ , and quadratically in the other curvature parameters. See Appendix A.3 for precise definitions of these parameters for the LQR problem. In particular, it is worth noting that our tightest bounds for these quantities depend on the dimension of the problem implicitly for some LQR instances, and are actually lower-order polynomials of the initial cost. In practice, however, it is likely that much sharper bounds can be proved on these parameters, e.g., in simulation (see Figure 3), the dependence of the sample complexity on the initial cost is in fact relatively weak—of the order  $\mathcal{C}(K_0)^2$ —and our bounds are clearly not sharp in that sense.

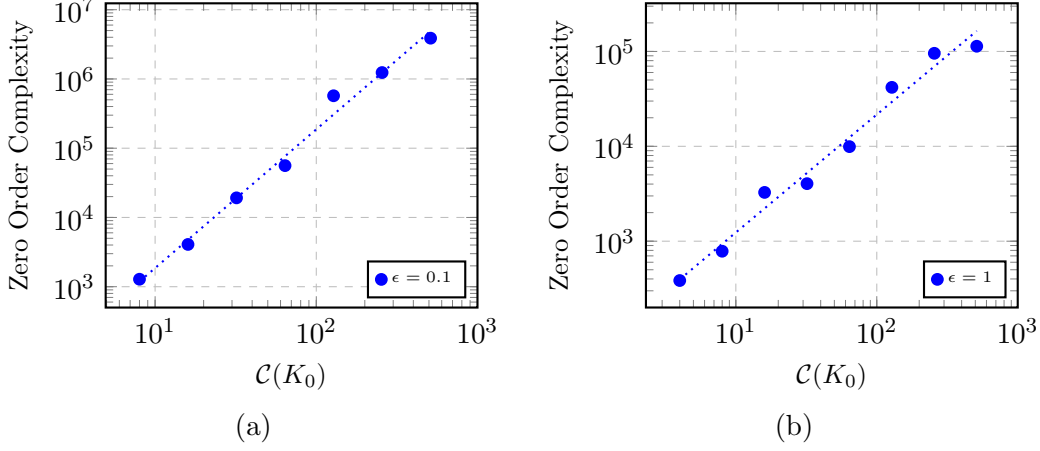
## 4 Proofs of main results

In this section, we provide proofs of Theorem 1, and Corollaries 1, 2, and 3. The proofs of the corollaries require many technical lemmas, whose proofs we postpone to the appendix.

---

<sup>5</sup>Note that this lower bound follows immediately for the class of PL and smooth functions.





**Figure 3.** Number of samples required to reach a fixed error tolerance of  $\epsilon$ , plotted against the cost of the initialization  $K_0$ , for (a) Randomly initialized LQR with two-point evaluations (b) Noisy dynamics LQR with one-point evaluations. The plots were obtained by averaging 20 runs of Algorithm 1. Each dotted line represents the line of best fit for the corresponding data points. For more problem details, see Appendix D.

#### 4.1 Proof of Theorem 1

Recall that by assumption, the population function  $f$  has domain  $\mathcal{X} \subseteq \mathbb{R}^d$  and satisfies the following properties over the restricted domain  $\mathcal{G}^0 \subseteq \mathcal{X}$ , previously defined in Equation (15):

- (a) It has  $(\phi_0, \rho_0)$ -locally Lipschitz gradients,
- (b) It is  $(\lambda_0, \rho_0)$ -locally Lipschitz, and
- (c) It is globally  $\mu$ -PL.

Recall the values of the step-size  $\eta$ , smoothing radius  $r$ , and iteration complexity  $T$  posited by Theorem 1. For ease of exposition, it is helpful to run our stochastic zero-order method on this problem for  $2T$  iterations; we thus obtain a (random) sequence of iterates  $\{x_t\}_{t=0}^{2T}$ . For each  $t = 0, 1, 2, \dots$ , we define the cost error  $\Delta_t = f(x_t) - f(x^*)$ , as well as the stopping time

$$\tau := \min \left\{ t \mid \Delta_t > 10\Delta_0 \right\}. \quad (16)$$

In words, the time  $\tau$  is the index of the first iterate that exits the bounded region  $\mathcal{G}^0$ . The gradient estimate  $g$  at the point  $x \in \mathcal{G}^0$  is assumed to satisfy the bounds

$$\text{var}(g(x)) \leq G_2 \quad \text{and} \quad \|g(x)\|_2 \leq G_\infty \quad \text{almost surely.}$$

With this set up in place, we now state and prove a proposition that is stronger than the assertion of Theorem 1.

**Proposition 1.** *With the parameter settings of Theorem 1, we have*

$$\mathbb{E}[\Delta_T 1_{\tau > T}] \leq \epsilon/20,$$

*and furthermore, the event  $\{\tau > T\}$  occurs with probability greater than  $4/5$ .*

Let us verify that Proposition 1 implies the claim of Theorem 1. We have

$$\begin{aligned}
\mathbb{P}\{\Delta_T \geq \epsilon\} &\leq \mathbb{P}\{\Delta_T 1_{\tau > T} \geq \epsilon\} + \mathbb{P}\{1_{\tau \leq T}\} \\
&\stackrel{(i)}{\leq} \frac{1}{\epsilon} \mathbb{E}[\Delta_T 1_{\tau > T}] + \mathbb{P}\{1_{\tau \leq T}\} \\
&\stackrel{(ii)}{\leq} 1/20 + 1/5 \\
&\leq 1/4,
\end{aligned}$$

where step (i) follows from Markov's inequality, and step (ii) from Proposition 1. Thus, Theorem 1 follows as a direct consequence of Proposition 1, and we dedicate the rest of the proof to establishing Proposition 1.

Let  $\mathbb{E}^t$  to represent the expectation conditioned on the randomness up to time  $t$ . The following lemma bounds the progress of one step of the algorithm:

**Lemma 5.** *Given any function satisfying the previously stated properties, suppose that we run Algorithm 1 with smoothing radius  $r \leq \rho_0$ , and with a step-size  $\eta$  such that  $\|\eta g^t\|_2 \leq \rho_0$  almost surely. Then for any  $t = 0, 1, \dots$  such that  $x_t \in \mathcal{G}^0$ , we have*

$$\mathbb{E}^t[\Delta_{t+1}] \leq \left(1 - \frac{\eta\mu}{4}\right)\Delta_t + \frac{\phi_0\eta^2}{2}G_2 + \eta\mu\frac{\epsilon}{120}. \quad (17)$$

The proof of the lemma is postponed to Section 4.1.1. Taking it as given, let us now establish Proposition 1.

Proposition 1 has two natural parts; let us focus first on proving the bound on the expectation. Let  $\mathcal{F}_t$  denote the  $\sigma$ -field containing all the randomness in the first  $t$  iterates. Conditioning on this  $\sigma$ -field yields

$$\mathbb{E}[\Delta_{t+1} 1_{\tau > t+1} \mid \mathcal{F}_t] \leq \mathbb{E}[\Delta_{t+1} 1_{\tau > t} \mid \mathcal{F}_t] \stackrel{(i)}{=} \mathbb{E}_{\mathcal{F}_t}[\mathbb{E}[\Delta_{t+1} \mid \mathcal{F}_t] 1_{\tau > t}],$$

where step (i) follows since  $\tau$  is a stopping time, and so the random variable  $1_{\tau > t}$  is determined completely by the sigma-field  $\mathcal{F}_t$ .

We now split the proof into two cases.

**Case 1:** Assume that  $\tau > t$ , so that we have the inclusion  $x_t \in \mathcal{G}^0$ . In addition, note that the iterate  $x_{t+1}$  is obtained after a stochastic zero-order step whose size is bounded as

$$\|\eta g^t\|_2 \leq \eta G_\infty \leq \rho_0,$$

where we have used the fact that  $\eta \leq \frac{\rho_0}{G_\infty}$ .

We may thus apply Lemma 5 to obtain

$$\mathbb{E}[\Delta_{t+1} \mid \mathcal{F}_t] \leq \left(1 - \frac{\eta\mu}{4}\right)\Delta_t + \frac{\phi_0\eta^2}{2}G_2 + \eta\mu\frac{\epsilon}{120}. \quad (18a)$$

**Case 2:** In this case, we have  $\tau \leq t$ , so that

$$\mathbb{E}[\Delta_{t+1} \mid \mathcal{F}_t] 1_{\tau > t} = 0. \quad (18b)$$

Now combining the bounds (18a) and (18b) from the two cases yields the inequality

$$\begin{aligned} \mathbb{E}[\Delta_{t+1} \mid \mathcal{F}_t] 1_{\tau > t} &\leq \left\{ \left(1 - \frac{\eta\mu}{4}\right) \Delta_t + \frac{\phi_0 \eta^2}{2} G_2 + \eta\mu \frac{\epsilon}{120} \right\} 1_{\tau > t} \\ &\leq \left(1 - \frac{\eta\mu}{4}\right) \Delta_t 1_{\tau > t} + \frac{\phi_0 \eta^2}{2} G_2 + \eta\mu \frac{\epsilon}{120}. \end{aligned} \quad (19)$$

Taking expectations over the sigma-field  $\mathcal{F}_t$  and then arguing inductively yields

$$\begin{aligned} \mathbb{E}[\Delta_{t+1} 1_{\tau > t+1}] &\leq \left(1 - \frac{\eta\mu}{4}\right)^{t+1} \Delta_0 + \left(\frac{\phi_0 \eta^2}{2} G_2 + \eta\mu \frac{\epsilon}{120}\right) \sum_{i=0}^t \left(1 - \frac{\eta\mu}{4}\right)^i \\ &\leq \left(1 - \frac{\eta\mu}{4}\right)^{t+1} \Delta_0 + 2 \frac{\eta}{\mu} \phi_0 G_2 + \frac{4\epsilon}{120}. \end{aligned}$$

Setting  $t+1 = T$  then establishes the first part of the proposition with substitutions of the various parameters.

We now turn to establishing that  $\mathbb{P}\{\tau > T\} \geq 4/5$ . We do so by setting up a suitable super-martingale on our iterate sequence and appealing to classical maximal inequalities. Recall that we run the algorithm for  $2T$  steps for convenience, and thereby obtain a set of  $2T$  random variables  $\{\Delta_1, \dots, \Delta_{2T}\}$ . With the stopping time  $\tau$  defined as before (16), define the stopped process

$$Y_t := \Delta_{\tau \wedge t} + (2T - t) \left( \frac{\phi_0 \eta^2}{2} G_2 + \eta\mu \frac{\epsilon}{120} \right) \quad \text{for each } t \in [2T].$$

Note that by construction, each random variable  $Y_t$  is non-negative and almost surely bounded.

We claim that  $\{Y_t\}_{t=0}^{2T}$  is a super-martingale. In order to prove this claim, we first write

$$\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t] = \mathbb{E}[\Delta_{\tau \wedge (t+1)} 1_{\tau \leq t} \mid \mathcal{F}_t] + \mathbb{E}[\Delta_{\tau \wedge (t+1)} 1_{\tau > t} \mid \mathcal{F}_t] + (2T - (t+1)) \left( \frac{\phi_0 \eta^2}{2} G_2 + \eta\mu \frac{\epsilon}{120} \right). \quad (20)$$

Beginning by bounding the first term on the right-hand side, we have

$$\mathbb{E}[\Delta_{\tau \wedge (t+1)} 1_{\tau \leq t} \mid \mathcal{F}_t] = \mathbb{E}[\Delta_{\tau \wedge t} 1_{\tau \leq t} \mid \mathcal{F}_t] = \Delta_{\tau \wedge t} 1_{\tau \leq t}. \quad (21a)$$

As for the second term, we have

$$\begin{aligned} \mathbb{E}[\Delta_{\tau \wedge (t+1)} 1_{\tau > t} \mid \mathcal{F}_t] &= \mathbb{E}[\Delta_{t+1} 1_{\tau > t} \mid \mathcal{F}_t] \\ &= \mathbb{E}[\Delta_{t+1} \mid \mathcal{F}_t] 1_{\tau > t} \\ &\stackrel{(iii)}{\leq} \left(1 - \frac{\eta\mu}{4}\right) \Delta_t 1_{\tau > t} + \left(\frac{\phi_0 \eta^2}{2} G_2 + \eta\mu \frac{\epsilon}{120}\right) 1_{\tau > t} \\ &\leq \left(1 - \frac{\eta\mu}{4}\right) \Delta_{\tau \wedge t} 1_{\tau > t} + \frac{\phi_0 \eta^2}{2} G_2 + \eta\mu \frac{\epsilon}{120}, \end{aligned} \quad (21b)$$

where step (iii) follows from using Inequality (19).

Substituting the bounds (21a) and (21b) into our original inequality (20), we find that

$$\begin{aligned}
\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t] &= \mathbb{E}[\Delta_{\tau \wedge (t+1)} 1_{\tau \leq t} \mid \mathcal{F}_t] + \mathbb{E}[\Delta_{\tau \wedge (t+1)} 1_{\tau > t} \mid \mathcal{F}_t] + (2T - (t+1)) \left( \frac{\phi_0 \eta^2}{2} G_2 + \eta \mu \frac{\epsilon}{120} \right) \\
&\leq \Delta_{\tau \wedge t} 1_{\tau \leq t} + (1 - \eta \mu / 4) \Delta_{\tau \wedge t} 1_{\tau > t} + \left( \frac{\phi_0 \eta^2}{2} G_2 + \eta \mu \frac{\epsilon}{120} \right) + (2T - (t+1)) \left( \frac{\phi_0 \eta^2}{2} G_2 + \eta \mu \frac{\epsilon}{120} \right) \\
&\stackrel{(iv)}{\leq} \Delta_{\tau \wedge t} + (2T - t) \left( \frac{\phi_0 \eta^2}{2} G_2 + \eta \mu \frac{\epsilon}{120} \right) \\
&= Y_t,
\end{aligned}$$

where step (iv) follows from the inequality  $\eta \mu \Delta_{\tau \wedge t} \geq 0$ . We have thus verified the super-martingale property.

Finally, applying Doob's maximal inequality for super-martingales (see, e.g., Durrett [Dur10]) yields

$$\begin{aligned}
\Pr\left\{ \max_{t \in [2T]} Y_t \geq \nu \right\} &\leq \frac{\mathbb{E}[Y_0]}{\nu} \\
&= \frac{1}{\nu} \left( \Delta_0 + 2T \left\{ \frac{\phi_0 \eta^2}{2} G_2 + \eta \mu \frac{\epsilon}{120} \right\} \right) \\
&\stackrel{(v)}{=} \frac{1}{\nu} \left( \Delta_0 + \frac{\epsilon}{5} \log(120 \Delta_0 / \epsilon) \right),
\end{aligned}$$

where step (v) follows from the substitutions  $T = \frac{2}{\eta \mu} \log(120 \Delta_0 / \epsilon)$ , and  $\eta = \frac{\epsilon \mu}{120 D \phi_0 \lambda_0}$ . As long as  $\epsilon$  is sufficiently small so as to ensure that  $\epsilon \log(120 \Delta_0 / \epsilon) < 5 \Delta_0$ , setting  $\nu = 10 \Delta_0$  completes the proof.

#### 4.1.1 Proof of Lemma 5

Recall that the domain of the function  $f$  is  $\mathcal{X} \subseteq \mathbb{R}^d$ . For a scalar  $r > 0$ , the smoothed version  $f_r(x)$  is given by  $f_r(x) := \mathbb{E}[f(x + rv)]$ , where the expectation above is taken with respect to the randomness in  $v$ , and  $v$  has uniform distribution on a  $d$ -dimensional ball  $\mathbb{B}^d$  of unit radius. The estimate  $g$  of the gradient  $\nabla f_r$  at  $x$  is given by

$$g(x) = \begin{cases} F(x + ru, \xi) \frac{d}{r} u & \text{if operating in one-point setting} \\ [F(x + ru, \xi) - F(x - ru, \xi)] \frac{d}{2r} u & \text{if operating in two-point setting,} \end{cases}$$

where  $u$  has a uniform distribution on the shell of the sphere  $\mathbb{S}^{d-1}$  of unit radius, and  $\xi$  is sampled at random from  $\mathcal{D}$ . The following result summarizes some useful properties of the smoothed version of  $f$ , and relates it to the gradient estimate  $g$ .

**Lemma 6.** *The smoothed version  $f_r$  of  $f$  with smoothing radius  $r$  has the following properties:*

- (a)  $\nabla f_r(x) = \mathbb{E}[g(x)]$ .
- (b)  $\|\nabla f_r(x) - \nabla f(x)\|_2 \leq \phi_0 r$ .

Versions of these properties have appeared in past work [FKM05, ADX10, Sha17], but we provide proofs in Appendix C for completeness.

Taking Lemma 6 as given, we now prove Lemma 5. Let  $\mathcal{F}_t$  denote the sigma field generated by the randomness up to iteration  $t$ , and  $\mathbb{E}$  denote the total expectation operator. We define  $\mathbb{E}^t := \mathbb{E}[\cdot | \mathcal{F}_t]$  as the expectation operator conditioned on the sigma field  $\mathcal{F}_t$ . Recall that the function  $f$  is smooth with smoothness parameter  $\phi_0$ , and we have

$$\begin{aligned} \mathbb{E}^t [f(x_{t+1}) - f(x_t)] &\leq \mathbb{E}^t \left[ \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{\phi_0}{2} \|x_{t+1} - x_t\|_2^2 \right] \\ &\stackrel{(i)}{=} -\langle \eta \nabla f(x_t), \nabla f_r(x_t) \rangle + \frac{\phi_0 \eta^2}{2} \mathbb{E}^t [\|g(x_t)\|_2^2] \\ &\stackrel{(ii)}{=} -\eta \|\nabla f(x_t)\|_2^2 + \eta \phi_0 r \|\nabla f(x_t)\|_2 + \frac{\phi_0 \eta^2}{2} \mathbb{E}^t [\|g(x_t)\|_2^2]. \end{aligned}$$

Steps (i) and (ii) above follow from parts (a) and (b), respectively, of Lemma 6. Now make the observation that

$$\begin{aligned} \mathbb{E}^t [\|g(x_t)\|_2^2] &= \text{var}(g(x_t)) + \|\nabla f_r(x_t)\|_2^2 \\ &\leq \text{var}(g(x_t)) + 2\|\nabla f(x_t)\|_2^2 + 2\|\nabla f_r(x_t) - \nabla f(x_t)\|_2^2 \\ &\leq G_2 + 2\|\nabla f(x_t)\|_2^2 + 2(\phi_0 r)^2. \end{aligned}$$

In addition, since the function is locally smooth at the point  $x_t$ , we have

$$\begin{aligned} (\theta - \theta^2 \phi_0/2) \|\nabla f(x_t)\|_2^2 &\leq f(x_t) - f(x_t - \theta \nabla f(x_t)) \\ &\leq f(x_t) - f(x^*), \end{aligned}$$

for some parameter  $\theta$  chosen small enough such that the relation  $\theta \|\nabla f(x_t)\|_2 \leq \rho_0$  holds. We may thus set  $\theta = \theta_0 = \min \left\{ \frac{1}{2\phi_0}, \frac{\rho_0}{\lambda_0} \right\}$  and recall the notation  $\Delta_t = f(x_t) - f(x^*)$  to obtain

$$\begin{aligned} \mathbb{E}^t [\Delta_{t+1}] &\leq -\eta \|\nabla f(x_t)\|_2^2 + \eta \phi_0 r \frac{2}{\theta_0} \Delta_t^{1/2} + \frac{\phi_0 \eta^2}{2} G_2 + \phi_0 \eta^2 (\|\nabla f(x_t)\|_2^2 + (\phi_0 r)^2) \\ &\stackrel{(iii)}{\leq} -\frac{\eta \mu}{2} \Delta_t + 2 \frac{\eta \phi_0 r}{\theta_0} \Delta_t^{1/2} + \frac{\phi_0 \eta^2}{2} G_2 + \phi_0 \eta^2 (\phi_0 r)^2, \\ &\stackrel{(iv)}{\leq} -\frac{\eta \mu}{2} \Delta_t + \frac{\eta \mu}{4} \Delta_t + 4 \frac{\eta (\phi_0 r)^2}{\mu \theta_0^2} + \frac{\phi_0 \eta^2}{2} G_2 + \phi_0 \eta^2 (\phi_0 r)^2, \end{aligned}$$

where step (iii) follows from applying the PL inequality and using the fact that  $\eta \leq \frac{1}{2\phi_0}$ , and step (iv) from the inequality  $2ab \leq a^2 + b^2$  which holds for any pair of scalars  $(a, b)$ .

Recall the assumed bounds on our parameters, namely

$$\eta \leq \min \left\{ \frac{\epsilon \mu}{240 \phi_0}, \frac{1}{2\phi_0} \right\}, \quad \text{and} \quad r \leq \frac{1}{2\phi_0} \min \left\{ \theta_0 \mu \sqrt{\frac{\epsilon}{240}}, \frac{1}{\phi_0} \sqrt{\frac{\epsilon \mu}{30}} \right\}.$$

Using these bounds, we have

$$\mathbb{E}^t [\Delta_{t+1}] \leq -\frac{\eta \mu}{4} \Delta_t + \frac{\phi_0 \eta^2}{2} G_2 + \eta \mu \frac{\epsilon}{120}.$$

Finally, rearranging yields

$$\mathbb{E}^t [\Delta_{t+1}] \leq \left(1 - \frac{\eta\mu}{4}\right) \Delta_t + \frac{\phi_0 \eta^2}{2} G_2 + \eta\mu \frac{\epsilon}{120}, \quad (22)$$

which completes the proof of Lemma 5.

## 4.2 Proof of Corollary 1

Recall the properties of the LQR cost function  $\mathcal{C}_{\text{init},\gamma}$  that were established in Lemmas 1 through 3. Taking these properties as given (see Appendix A for the proofs of the lemmas), the only remaining detail is to establish the bounds

$$G_2 \leq C \left( \frac{D}{r} C_m \mathcal{C}_{\text{init},\gamma}(K_0) \right)^2 \quad \text{and} \quad G_\infty \leq C \frac{D}{r} C_m \mathcal{C}_{\text{init},\gamma}(K_0). \quad (23)$$

In fact, it suffices to prove the second bound in Equation (23), since we have  $G_2 \leq G_\infty^2$ .

Given a unit vector  $u$ , the norm of the gradient estimate can be bounded as

$$\begin{aligned} \|g_t\|_2 &= \frac{D}{r} \mathcal{C}_{\text{init},\gamma}(K_t + ru; s_0) \\ &\stackrel{(i)}{=} \frac{D}{r} s_0^\top P_K s_0 \\ &\leq \frac{D}{r} \|s_0\|_2^2 \|P_K\|_2 \\ &\stackrel{(ii)}{\leq} C_m \frac{D}{r} \mathcal{C}_{\text{init},\gamma}(K_t + ru), \end{aligned}$$

where step (i) follows from the relation (26), and step (ii) from the relation (27), since  $P_K$  is a PSD matrix. Finally, since  $r \leq \rho_{\text{lqr}}$ , the local Lipschitz property of the function  $\mathcal{C}_{\text{init},\gamma}$  yields

$$\begin{aligned} \mathcal{C}_{\text{init},\gamma}(K_t + ru) &\leq \mathcal{C}_{\text{init},\gamma}(K_t) + r\lambda_K \\ &\leq \mathcal{C}_{\text{init},\gamma}(K_t) + r\lambda_{\text{lqr}} \\ &\stackrel{(iii)}{\leq} 10\mathcal{C}_{\text{init},\gamma}(K_0) + 10\mathcal{C}_{\text{init},\gamma}(K_0), \end{aligned}$$

where step (iii) uses the fact that  $K_t \in \mathcal{G}^{\text{lqr}}$  so that  $\mathcal{C}_{\text{init},\gamma}(K_t) \leq 10\mathcal{C}_{\text{init},\gamma}(K_0)$ , and the upper bound  $r \leq \frac{10\mathcal{C}_{\text{init},\gamma}(K_0)}{\lambda_{\text{lqr}}}$ . Putting together the pieces completes the proof.

## 4.3 Proof of Corollary 2

As before, establishing Corollary 2 requires bounds on the values of the pair  $(G_2, G_\infty)$ , since the remaining properties are established in Lemmas 1 through 3.

In particular, let us establish bounds on these quantities for general optimization of a function with a two-point gradient estimate. The following computations closely follow those of Shamir [Sha17].

**Second moment control:** Using the law of iterated expectations, we have

$$\mathbb{E} \left[ \left\| d \frac{F(x + ru, \xi) - F(x - ru, \xi)}{2r} u \right\|_2^2 \right] = \mathbb{E} \left[ \mathbb{E} \left[ \left\| d \frac{F(x + ru, \xi) - F(x - ru, \xi)}{2r} u \right\|_2^2 \middle| \xi \right] \right].$$

Define the placeholder variable  $q$  and now evaluate:

$$\begin{aligned} \mathbb{E} \left[ \left\| d \frac{F(x + ru, \xi) - F(x - ru, \xi)}{2r} u \right\|_2^2 \middle| \xi \right] &= \frac{d^2}{4r^2} \mathbb{E} \left[ (F(x + ru, \xi) - F(x - ru, \xi))^2 \|u\|_2^2 \middle| \xi \right] \\ &\stackrel{(i)}{=} \frac{d^2}{4r^2} \mathbb{E} \left[ (F(x + ru, \xi) - F(x - ru, \xi))^2 \middle| \xi \right] \\ &= \frac{d^2}{4r^2} \mathbb{E} \left[ (F(x + ru, \xi) - q - F(x - ru, \xi) + q)^2 \middle| \xi \right] \\ &\stackrel{(ii)}{\leq} \frac{d^2}{2r^2} \mathbb{E} \left[ (F(x + ru, \xi) - q)^2 + (F(x - ru, \xi) - q)^2 \middle| \xi \right], \end{aligned}$$

where equality (i) follows from the fact that  $u$  is a unit vector and inequality (ii) follows from the inequality  $(a - b)^2 \leq 2(a^2 + b^2)$ . We further simplify this to obtain:

$$\begin{aligned} \mathbb{E} \left[ \left\| d \frac{F(x + ru, \xi) - F(x - ru, \xi)}{2r} u \right\|_2^2 \middle| \xi \right] &\stackrel{(i)}{\leq} \frac{d^2}{r^2} \mathbb{E} \left[ (F(x + ru, \xi) - q)^2 \middle| \xi \right] \\ &\stackrel{(ii)}{\leq} \frac{d^2}{r^2} \sqrt{\mathbb{E} \left[ (F(x + ru, \xi) - q)^4 \middle| \xi \right]}, \end{aligned}$$

where inequality (i) follows from the symmetry of the uniform distribution on the sphere, and inequality (ii) follows from Jensen's Inequality. For a fixed  $\xi$ , we now define  $q = \mathbb{E}[F(x + ru, \xi) | \xi]$ . Substituting this expression yields

$$\begin{aligned} \mathbb{E} \left[ \left\| d \frac{F(x + ru, \xi) - F(x - ru, \xi)}{2r} u \right\|_2^2 \middle| \xi \right] &\leq \frac{d^2}{r^2} \sqrt{\mathbb{E} \left[ (F(x + ru, \xi) - \mathbb{E}[F(x + ru, \xi) | \xi])^4 \middle| \xi \right]} \\ &\stackrel{(i)}{\leq} \frac{d^2}{r^2} \frac{(\lambda r)^2}{d} \\ &= d\lambda^2, \end{aligned}$$

where inequality (i) follows directly from Lemma 9 in Shamir [Sha17]. The lemma can be applied since we are conditioning on  $\xi$ , and all the randomness lies in the selection of  $u$ . We have thus established the claim in part (c).

**Gradient estimates are bounded:** Note that smoothing radius  $r$  satisfies  $r \leq \rho_0$ , where  $\rho_0$  is the radius within which the function is Lipschitz. Consequently, the local Lipschitz property of  $F$  implies that

$$\begin{aligned} \|g_t\|_2 &:= \left| d \frac{F(x_t + ru_t, \xi_t) - F(x_t - ru_t, \xi_t)}{2r} u_t \right| \\ &\leq \left| d \frac{F(x_t + ru_t, \xi_t) - F(x_t, \xi_t)}{2r} u_t \right| + \left| d \frac{F(x_t, \xi_t) - F(x_t - ru_t, \xi_t)}{2r} u_t \right| \\ &\leq d\lambda_0 \frac{2\|ru_t\|_2}{2r} \leq d\lambda_0. \end{aligned}$$

#### 4.4 Proof of Corollary 3

As in Section 4.2, we establish bounds on the values  $G_2$  and  $G_\infty$  for the noisy LQR dynamics model. In particular, we derive a bound on  $G_\infty$  and use the fact that  $G_2 \leq G_\infty^2$  to establish the bound on  $G_2$ . For deriving these bounds, we use properties of the cost function  $\mathcal{C}_{\text{dyn},\gamma}$  and its connections with  $\mathcal{C}_{\text{init},\gamma}$  which are established in Lemma 4 and Lemma 10; the proofs of these are deferred to Appendix B.

In particular, we establish the bounds

$$G_2 \leq \left( \frac{D}{r} \cdot \frac{2C_m}{1 - \sqrt{\gamma}} \right)^2 \cdot \left( \frac{20\mathcal{C}_{\text{dyn},\gamma}(K_0)}{\sigma_{\min}(Q)} \left( \frac{1 - \gamma}{\gamma} \right) \right)^3, \text{ and}$$

$$G_\infty \leq \frac{D}{r} \cdot \frac{2C_m}{1 - \sqrt{\gamma}} \cdot \left( \frac{20\mathcal{C}_{\text{dyn},\gamma}(K_0)}{\sigma_{\min}(Q)} \left( \frac{1 - \gamma}{\gamma} \right) \right)^{3/2}.$$

For any unit vector  $u$ , we have,

$$\begin{aligned} \|g_t\|_2 &= \frac{D}{r} \mathcal{C}_{\text{dyn},\gamma}(K_t + ru; \mathcal{Z}) \\ &\stackrel{(i)}{\leq} \frac{D}{r} \cdot \frac{2C_m}{1 - \sqrt{\gamma}} \cdot \left( \frac{\mathcal{C}_{\text{dyn},\gamma}(K_t + ru)}{\sigma_{\min}(Q)} \left( \frac{1 - \gamma}{\gamma} \right) \right)^{3/2}, \end{aligned}$$

where (i) follows from using the bound in Lemma 10. Finally, using Lemma 4 and since  $r \leq \rho_{\text{lqr},\gamma}$ , the local Lipschitz property of the function  $\mathcal{C}_{\text{init},\gamma}$  yields

$$\begin{aligned} \mathcal{C}_{\text{dyn},\gamma}(K_t + ru) &\leq \frac{\gamma}{1 - \gamma} \cdot \mathcal{C}_{\text{init},\gamma}(K_t + ru) \\ &\leq \frac{\gamma}{1 - \gamma} \cdot (\mathcal{C}_{\text{init},\gamma}(K_t) + r\lambda_{K,\gamma}) \\ &\leq \frac{\gamma}{1 - \gamma} \cdot (\mathcal{C}_{\text{init},\gamma}(K_t) + r\lambda_{\text{lqr},\gamma}) \\ &\stackrel{(i)}{\leq} \frac{\gamma}{1 - \gamma} \cdot (10\mathcal{C}_{\text{init},\gamma}(K_0) + 10\mathcal{C}_{\text{init},\gamma}(K_0)), \end{aligned} \tag{24}$$

where step (i) uses the fact that  $K_t \in \mathcal{G}^{\text{lqr}}$  so that  $\mathcal{C}_{\text{init},\gamma}(K_t) \leq 10\mathcal{C}_{\text{init},\gamma}(K_0)$ , and the upper bound  $r \leq \frac{10\mathcal{C}_{\text{init},\gamma}(K_0)}{\lambda_{\text{lqr},\gamma}}$ . Putting together the pieces completes the proof.

## 5 Discussion

In this paper, we studied the model-free control problem over linear policies through the lens of derivative-free optimization. We derived quantitative convergence rates for various zero-order methods when applied to learn optimal policies based on data from noisy linear systems with quadratic costs. In particular, we showed that one-point and two-point variants of a canonical derivative-free optimization method achieve fast rates of convergence for the non-convex LQR problem. Notably, our proof deals directly with some additional difficulties that are specific to this problem and do not arise in the analysis of typical optimization algorithms. More precisely, our proof involves careful control of both the (potentially) unbounded nature of the cost function, and the non-convexity of the underlying domain. Interestingly, our proof only relies on certain



local properties of the function that can be guaranteed over a bounded set; for this reason, the optimization-theoretic result in this paper (stated as Theorem 1) is more broadly applicable beyond the RL setting.

While this paper analyzes a canonical zero-order optimization algorithm for model-free control of linear quadratic systems, many open questions remain. One such question concerns lower bounds for LQR problems in the model-free setting, thereby showing quantitative gaps between such a setting and that of model-based control. While we conjecture that the convergence bounds of Corollaries 1, 2, and 3 are sharp in terms of their dependence on the error tolerance  $\epsilon$ , establishing this rigorously will require ideas from the extensive literature on lower bounds in zero-order optimization [Sha13]. Another important direction is establish the sharpness (or otherwise) of our bounds in terms of the dimension of the problem, as well as to obtain tight characterizations of the local curvature parameters of the problem around a particular policy  $K$  in terms of the cost at  $K$ .

In the broader context of model-free reinforcement learning as well, there are many open questions. First, a derivative-free algorithm over linear policies is reasonable even in other systems; can we establish provable guarantees over larger classes of problems? Second, there is no need to restrict ourselves to linear policies; in practical RL systems, derivative-free algorithms are run for policies that parametrized in a much more complex fashion. How does the sample complexity of the problem change with the class of policies over which we are optimizing?

## Acknowledgements

This work was partially supported by National Science Foundation grant NSF-DMS-1612948 to MJW. AP was additionally supported in part by NSF CCF-1704967. PLB gratefully acknowledges the support of the NSF through grant IIS-1619362. KB was supported in part by AFOSR through grant FA9550-17-1-0308 and NSF through grant IIS-1619362.

## A Properties of the randomly initialized LQR problem

In this section, we establish some fundamental properties of the cost function  $\mathcal{C}_{\text{init},\gamma}$ , and provide proofs of Lemmas 1 and 2. As part of these proofs, we provide explicit bounds for the local curvature parameters  $(\tilde{\lambda}_{\text{lqr}}, \lambda_{\text{lqr}}, \rho_{\text{lqr}}, \phi_{\text{lqr}}, \mu_{\text{lqr}})$ . We make frequent use of results established by Fazel et al. [FGKM18], and as mentioned before, Lemmas 1 and 2 are refinements of their results.

**Notation:** In this section, we introduce some shorthand to reduce notational overhead. Throughout, we assume that  $\gamma = 1$ ; the general case is straightforward to obtain with the substitutions

$$A \mapsto \sqrt{\gamma}A, \text{ and } B \mapsto \sqrt{\gamma}B.$$

We also use the shorthand  $\mathcal{C}(K) := \mathcal{C}_{\text{init},\gamma}(K)$  for this section. Much (but not all) of the notation we use overlaps with the notation used in Fazel et al. [FGKM18].

We define the matrix  $P_K$  as the solution to the following fixed point equation:

$$P_K = Q + K^\top R K + (A - BK)^\top P_K (A - BK),$$

and we define the state correlation matrix  $\Sigma_K$  as:

$$\Sigma_K = \mathbb{E} \left[ \sum_{t=0}^{\infty} s_t s_t^\top \right] \quad \text{such that} \quad s_t = (A - BK)s_{t-1}. \quad (25)$$

It is straightforward to see that we have

$$\mathcal{C}(K) = \mathbb{E}[s_0^\top P_K s_0], \quad (26)$$

and we make frequent use of this representation in the sequel.

Recall that we have  $\mathbb{E}[s_0 s_0^\top] = I$ , so that

$$\mathcal{C}(K) = \text{tr}(P_K). \quad (27)$$

Moreover, under this assumption, the cost function  $\mathcal{C}$  satisfies the PL Inequality with PL constant  $\frac{\|\Sigma_{K^*}\|_2}{\sigma_{\min}(R)}$  (see Lemma 3 in the paper [FGKM18]).

Also define the *natural gradient* of the cost function as

$$E_K := 2(R + B^\top P_K B)K - B^\top P_K A,$$

so that we have  $\nabla \mathcal{C}(K) = E_K \Sigma_K$ . For any symmetric matrix  $X$ , the perturbation operators  $\mathcal{T}_K(\cdot)$  and  $\mathcal{F}_K(\cdot)$  are defined as

$$\mathcal{T}_K(X) = \sum_{t=0}^{\infty} (A - BK)^t X [(A - BK)^\top]^t, \quad \text{and} \quad \mathcal{F}_K(X) = (A - BK)X(A - BK)^\top.$$

Finally, the operator norms of the operators  $\mathcal{T}_K(\cdot)$  and  $\mathcal{F}_K(\cdot)$  are defined as

$$\begin{aligned} \|\mathcal{T}_K\|_2 &= \sup_X \frac{\|\mathcal{T}_K(X)\|_2}{\|X\|_2} \quad \text{and} \\ \|\mathcal{F}_K\|_2 &= \sup_X \frac{\|\mathcal{F}_K(X)\|_2}{\|X\|_2}. \end{aligned}$$

### Useful constants:

We now define several polynomials of  $\mathcal{C}(K)$ , which are useful in various proofs in this section.

- $c_{K_1} = \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} \sqrt{(\|R\|_2 + \|B\|_2^2 \mathcal{C}(K))(\mathcal{C}(K) - \mathcal{C}(K^*))}$
- $c_{K_2} = 4 \left( \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} \right)^2 \|Q\|_2 \|B\|_2 (\|A\|_2 + \|B\|_2 c_{K_1} + 1)$
- $c_{K_3} = 8 \left( \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} \right)^2 (c_{K_1})^2 \|R\|_2 \|B\|_2 (\|A\|_2 + \|B\|_2 c_{K_1} + 1)$
- $c_{K_4} = 2 \left( \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} \right)^2 (c_{K_1} + 1) \|R\|_2$
- $c_{K_5} = \sqrt{(\|R\|_2 + \|B\|_2^2 \mathcal{C}(K))(\mathcal{C}(K) - \mathcal{C}(K^*))}$
- $c_{K_6} = \|R\|_F + \|B\|_F^2 (c_{K_1} + 1)(c_{K_2} + c_{K_3} + c_{K_4}) + \|B\|_F^2 \mathcal{C}(K) + \|B\|_F \|A\|_2 (c_{K_2} + c_{K_3} + c_{K_4})$
- $c_{K_7} = 5c_{K_6} \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} + 4c_{K_5} \left( \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} \right)^2 \|B\|_2 (\|A\|_2 + \|B\|_2 c_{K_1}).$

- $c_{K_8} = C_m(c_{K_2} + c_{K_3} + c_{K_4})$ .
- $c_{K_9} = \min \left\{ \frac{\sigma_{\min}(Q)}{4\mathcal{C}(K)\|B\|_2(\|A\|_2 + \|B\|_2 c_{K_1} + 1)}, 1 \right\}$ .

With these definitions at hand, we are now in a position to establish Lemmas 1 and 2.

### A.1 Proof of Lemma 1

Let us restate a precise version of the lemma for convenience.

**Lemma 7.** *For any pair  $(K', K)$  such that  $\|K' - K\|_F \leq c_{K_9}$ , we have*

$$\begin{aligned} |\mathcal{C}(K') - \mathcal{C}(K)| &\leq \left( \frac{m}{C_m} \right) c_{K_8} \|K' - K\|_F, \text{ and} \\ |\mathcal{C}(K', s_0) - \mathcal{C}(K, s_0)| &\leq c_{K_8} \|K' - K\|_F. \end{aligned}$$

Comparing Lemma 7 with the statement of Lemma 1, we have therefore established the relations

$$\begin{aligned} \zeta_K &\geq c_{K_9}, \\ \lambda_K &\leq \left( \frac{m}{C_m} \right) c_{K_8}, \text{ and} \\ \tilde{\lambda}_K &\leq c_{K_8}. \end{aligned}$$

Note that we have  $\lambda_K \leq \tilde{\lambda}_K$ , since  $m \leq C_m$ . Let us now prove Lemma 7.

*Proof.* We have

$$\begin{aligned} |\mathcal{C}(K') - \mathcal{C}(K)| &= \text{tr}(P_{K'}) - \text{tr}(P_K) \\ &\leq m \|P_{K'} - P_K\|_2. \end{aligned}$$

Moreover, the sample cost satisfies the relation

$$\begin{aligned} |\mathcal{C}(K', s_0) - \mathcal{C}(K, s_0)| &= |s_0^\top P_{K'} s_0 - s_0^\top P_K s_0| \\ &= |\text{tr}(s_0^\top (P_{K'} - P_K) s_0)| \\ &\leq \|P_{K'} - P_K\|_2 \|s_0\|_2^2 \\ &\leq \|P_{K'} - P_K\|_2 C_m. \end{aligned} \tag{28}$$

Hence, it remains to bound  $\|P_{K'} - P_K\|_2$ . To this end, substituting the definition of the linear operator  $\mathcal{T}_K$ , we have

$$\begin{aligned} \|P_{K'} - P_K\|_2 &= \|\mathcal{T}_{K'}(Q + (K')^\top R K') - \mathcal{T}_K(Q + K^\top R K)\|_2 \\ &= \|(\mathcal{T}_{K'} - \mathcal{T}_K)(Q + (K')^\top R K') - \mathcal{T}_K(K^\top R K - (K')^\top R K')\|_2 \\ &\leq \|(\mathcal{T}_{K'} - \mathcal{T}_K)Q\|_2 + \|(\mathcal{T}_{K'} - \mathcal{T}_K)(K')^\top R K'\|_2 \\ &\quad + \|\mathcal{T}_K\|_2 \|K^\top R K - (K')^\top R K'\|_2. \end{aligned} \tag{29}$$

We provide upper bounds for the three terms above as follows:

$$\|(\mathcal{T}_{K'} - \mathcal{T}_K)(K')^\top RK'\|_2 \leq c_{K_3} \|K - K'\|_2 \quad (30a)$$

$$\|(\mathcal{T}_{K'} - \mathcal{T}_K)Q\|_2 \leq c_{K_2} \|K - K'\|_2 \quad (30b)$$

$$\|\mathcal{T}_K\|_2 \|K^\top RK - (K')^\top RK'\|_2 \leq c_{K_4} \|K - K'\|_2. \quad (30c)$$

Taking the above bounds as given at the moment, we have from Equation (29) that

$$\|P_{K'} - P_K\|_2 \leq (c_{K_2} + c_{K_3} + c_{K_4}) \|K' - K\|_2, \quad (31)$$

Putting together the pieces completes the proof of Lemma 1.  $\square$

It remains to prove the upper bounds (30a)- (30c).

**Auxiliary bounds:** Proofs of the bounds (30a) through (30c) are based on the following intermediate bounds:

$$\|(K')^\top RK' - K^\top RK\|_2 \leq (c_{K_1} + 1) \|R\|_2 \|K' - K\|_2. \quad (32a)$$

$$\|\mathcal{F}_{K'} - \mathcal{F}_K\|_2 \leq 2\|B\|_2 (\|A\|_2 + \|B\|_2 c_{K_1} + 1) \|K' - K\|_2 \quad (32b)$$

$$\|\mathcal{T}_K\|_2 \leq \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} \quad (32c)$$

$$\|K^\top RK\|_2 \leq c_{K_1}^2 \|R\|_2 \quad (32d)$$

We prove these bounds at the end, but let us complete the rest of the proofs assuming these auxiliary bounds.

**Proof of the bound (30a):** The proof of this upper bound is based on Lemma 20 from the paper [FGKM18]. Accordingly, we start by verifying the following condition for Lemma 20:

$$\|\mathcal{F}_K - \mathcal{F}_{K'}\|_2 \|(K')^\top RK'\|_2 \leq \frac{1}{2}. \quad (33)$$

Observe that our assumption  $\|K' - K\|_F \leq c_{K_9}$ , satisfies the assumption of Lemma 10 in the paper [FGKM18], whence we have

$$\begin{aligned} \|B\|_2 \|K' - K\|_2 &\stackrel{(i)}{\leq} \|B\|_2 \frac{\sigma_{\min}(Q)}{4\mathcal{C}(K)\|B\|_2(\|A\|_2 + \|B\|_2 c_{K_1} + 1)} \\ &\stackrel{(ii)}{\leq} \frac{\sigma_{\min}(Q)}{4\mathcal{C}(K)(\|A - BK\|_2 + 1)} \\ &\stackrel{(iii)}{\leq} \frac{1}{4}. \end{aligned} \quad (34)$$

where step (i) follows by substituting the value of  $c_{K_9}$ , and step (ii) follows since  $\|A - BK\|_2 \leq \|A\|_2 + \|B\|_2 c_{K_1} + 1$ . Step (iii) above follows since  $\mathcal{C}(K) \geq \sigma_{\min}(Q)$ . Combining the last inequality with Lemma 16 in the paper [FGKM18] yields

$$\begin{aligned} \|\mathcal{F}_{K'} - \mathcal{F}_K\|_2 &\leq 2\|A - BK\|_2 \|B\|_2 \|K' - K\|_2 + \|B\|_2^2 \|K' - K\|_2^2 \\ &\leq 2\|B\|_2 (\|A - BK\|_2 + 1) \|K' - K\|_2 \end{aligned}$$

Finally, invoking Lemma 14 from the paper [FGKM18] guarantees that  $\|\mathcal{T}_K\|_2 \leq \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)}$ , and we deduce that

$$\begin{aligned}\|\mathcal{T}_K\|_2 \|\mathcal{F}_{K'} - \mathcal{F}_K\|_2 &\leq \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} 2\|B\|_2(\|A - BK\|_2 + 1) \|K' - K\|_2 \\ &\leq \frac{1}{2},\end{aligned}$$

where the last inequality follows from the assumption  $\|K' - K\|_F \leq c_{K_9}$ .

Now that we have verified that condition 33, invoking Lemma 20 in the paper [FGKM18] yields

$$\begin{aligned}\|(\mathcal{T}_{K'} - \mathcal{T}_K)(K')^\top RK'\|_2 &\leq 2\|\mathcal{T}_K\|_2^2 \|\mathcal{F}_K - \mathcal{F}_{K'}\|_2 \|(K')^\top RK'\|_2 \\ &\leq 2\|\mathcal{T}_K\|_2^2 \|\mathcal{F}_K - \mathcal{F}_{K'}\|_2 \|K^\top RK\|_2 \\ &\quad + 2\|\mathcal{T}_K\|_2^2 \|\mathcal{F}_K - \mathcal{F}_{K'}\|_2 \|(K')^\top RK' - K^\top RK\|_2 \\ &\leq c_{K_3} \|K - K'\|_2,\end{aligned}$$

where the last step above follows by substituting the bounds (32a)- (32d).

**Proof of the bounds (30b) and (30c):** The proof of the bound (30b) is similar to the part (30a) and is based on Lemma 20 from the paper [FGKM18]. More concretely, we have

$$\|(\mathcal{T}_{K'} - \mathcal{T}_K)Q\|_2 \leq 2\|\mathcal{T}_K\|_2^2 \|\mathcal{F}_K - \mathcal{F}_{K'}\|_2 \|Q\|_2 \leq c_{K_2} \|K - K'\|_2$$

where the last step above follows from the bounds (32b) and (32c). The proof of the bound (30c) is a direct consequence of the bounds (32a) and (32c).

### A.1.1 Proofs of the auxiliary bounds

In this section we prove the auxiliary bounds (32a) through to (32d)

Bound (32a): Observe that

$$\begin{aligned}\|K^\top RK - (K')^\top RK'\|_2 &= \|(K' - K)^\top R(K' - K) + (K')^\top RK + K^\top R(K') - 2K^\top RK\|_2 \\ &\leq (2\|R\|_2 \|K\|_2 \|K' - K\|_2 + \|R\|_2 \|K' - K\|_2^2) \\ &\stackrel{(i)}{\leq} (2\|K\|_2 + 1) \|R\|_2 \|K' - K\|_2 \\ &\stackrel{(ii)}{\leq} (2c_{K_1} + 1) \|R\|_2 \|K' - K\|_2.\end{aligned}$$

where step (i) follows since  $\|K - K'\|_2 \leq 1$  by assumption, and step (ii) follows since  $\|K\|_2 \leq c_{K_1}$  (see Lemma 22 in the paper [FGKM18]). This completes the proof of bound (32a).

Bound (32b): In order to prove bound (32b), we invoke Lemma 19 in the paper [FGKM18] to obtain

$$\begin{aligned}\|\mathcal{F}_{K'} - \mathcal{F}_K\|_2 &\leq 2\|A - BK\|_2 \|B\|_2 \|K' - K\|_2 + \|B\|_2^2 \|K' - K\|_2^2 \\ &\stackrel{(iii)}{\leq} 2\|A - BK\|_2 \|B\|_2 \|K' - K\|_2 + \frac{1}{4} \|B\|_2 \|K' - K\|_2 \\ &\leq 2\|B\|_2 (\|A\|_2 + \|B\|_2 c_{K_1} + 1) \|K' - K\|_2\end{aligned}$$

where step (iii) above follows from the upper bound (34). This completes the proof of the bound (32b).

Bound (32c) and (32d): The bound (32c) above follows from Lemma 17 in the paper [FGKM18], whereas the bound (32d) follows from the fact that  $\|K\|_2 \leq c_{K_1}$  (see Lemma 22 in the paper [FGKM18]).

Having established all of our auxiliary bounds, let us now proceed to a proof of Lemma 2.

## A.2 Proof of Lemma 2

Lemma 2 is a consequence of the following result.

**Lemma 8.** *If  $\|K' - K\|_F \leq c_{K_9}$ , then*

$$\|\nabla \mathcal{C}(K') - \nabla \mathcal{C}(K)\|_F \leq c_{K_7} \|K' - K\|_F.$$

Indeed, comparing Lemmas 8 and 2, we have the bounds

$$\beta_K \geq c_{K_9} \quad \text{and} \quad \phi_K \leq c_{K_7}.$$

Let us now prove Lemma 8.

*Proof.* We start by noting that from Lemma 1 we have that the cost function  $\mathcal{C}(K)$  is locally Lipschitz in a ball of  $\zeta_K$  around the point  $K$ . Before moving into the main argument, we mention a few auxiliary results that are helpful in the sequel. We start by invoking Lemma 13 from the paper [FGKM18], whence we have

$$\|P_K\|_2 \leq \mathcal{C}(K) \quad \text{and} \quad \|\Sigma_K\|_2 \leq \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)}.$$

We also have

$$\|A - BK\|_2 \leq \|A\|_2 + \|B\|_2 \|K\|_2 \stackrel{(i)}{\leq} \|A\|_2 + \|B\|_2 c_{K_1} \quad \text{and} \quad (35a)$$

$$\|\Sigma_{K'}\|_2 \leq \|\Sigma_K\|_2 + \|\Sigma_{K'} - \Sigma_K\|_2 \stackrel{(ii)}{\leq} 5 \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)}. \quad (35b)$$

Step (i) above follows since  $\|K\|_2 \leq c_{K_1}$  (see Lemma 22 in the paper [FGKM18]), whereas step (ii) follows since  $\|\Sigma_{K'} - \Sigma_K\|_2 \leq 4 \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)}$  (see Lemma 16 in the paper [FGKM18]).

Recalling the gradient expression  $\nabla \mathcal{C}(K) = E_K \Sigma_K$ . Let  $K'$  be a policy such that  $\|K' - K\|_F \leq c_{K_9}$ . We have

$$\begin{aligned} \|\nabla \mathcal{C}(K') - \nabla \mathcal{C}(K)\|_F &= \|(E_{K'} - E_K) \Sigma_{K'} + E_K (\Sigma_{K'} - \Sigma_K)\|_F \\ &\leq \|(E_{K'} - E_K)\|_F \|\Sigma_{K'}\|_2 + \|E_K\|_F \|\Sigma_{K'} - \Sigma_K\|_2 \\ &\stackrel{(iii)}{\leq} 5 c_{K_6} \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} \|K' - K\|_F \\ &\quad + 4 c_{K_5} \left( \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} \right)^2 \frac{\|B\|_2 (\|A\|_2 + \|B\|_2 c_{K_1})}{\sigma_{\min}(\Sigma_0)} \|K' - K\|_F. \end{aligned}$$

The upper bound in step (iii) on the term  $\|E_{K'} - E_K\|_F \|\Sigma_{K'}\|_2$  follows from Equation (35b) and from the following upper bound which we prove later:

$$\|E_{K'} - E_K\|_F \leq c_{K_6} \|K' - K\|_F \text{ provided } \|K' - K\|_F \leq c_{K_9}. \quad (36)$$

The upper bound on the term  $\|E_K\|_F \|\Sigma_{K'} - \Sigma_K\|_2$  in step (iii) follows from the fact that  $\|E_K\|_F \leq c_{K_5}$  (see Lemma 11 in the paper [FGKM18]) and from the fact that

$$\begin{aligned} \|\Sigma_{K'} - \Sigma_K\|_2 &\stackrel{(iv)}{\leq} 4 \left( \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} \right)^2 \frac{\|B\|_2 (\|A - BK\|_2 + 1)}{\sigma_{\min}(\Sigma_0)} \|K' - K\|_F \\ &\stackrel{(v)}{\leq} 4 \left( \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} \right)^2 \frac{\|B\|_2 (\|A\|_2 + \|B\|_2 c_{K_1} + 1)}{\sigma_{\min}(\Sigma_0)} \|K' - K\|_F, \end{aligned}$$

where step (iv) follows from Lemma 16 in the paper [FGKM18], and step (v) follows from Inequality (35a).

Putting together the pieces, we conclude that the function  $\nabla \mathcal{C}(K)$  is Lipschitz with constant  $\phi_K$ , where  $\phi_K$  is given by

$$\phi_K = 5c_{K_6} \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} + 4c_{K_5} \left( \frac{\mathcal{C}(K)}{\sigma_{\min}(Q)} \right)^2 \|B\|_2 (\|A\|_2 + \|B\|_2 c_{K_1} + 1) = c_{K_7}.$$

□

It remains to prove Inequality (36).

**Proof of Inequality (36):** From the definition of  $E_K$ , we have

$$\begin{aligned} \|E_{K'} - E_K\|_F &= 2\|(R + B^\top P_{K'} B)K' - B^\top P_{K'} A - (R + B^\top P_K B)K + B^\top P_K A\|_F \\ &= 2\|R(K' - K) + B^\top (P_{K'} - P_K)BK' + B^\top P_K B(K' - K) - B^\top (P_{K'} - P_K)A\|_F \\ &\leq 2\|R\|_F \|K' - K\|_F + 2\|B^\top (P_{K'} - P_K)BK'\|_F \\ &\quad + 2\|B^\top P_K B(K' - K)\|_F + 2\|B^\top (P_{K'} - P_K)A\|_F \end{aligned} \quad (37)$$

We provide upper bounds for the three terms above as follows. First, we have

$$\|B^\top (P_{K'} - P_K)BK'\|_F \leq \|B\|_F^2 (c_{K_1} + 1)(c_{K_2} + c_{K_3} + c_{K_4}) \|K' - K\|_F,$$

which follows from the bound (31), since  $\|K' - K\|_F \leq c_{K_9}$ , and the relation  $\|K'\|_2 \leq \|K\|_2 + \|K' - K\|_2 \leq c_{K_1} + 1$ . The same reasoning also yields the bound

$$\|B^\top (P_{K'} - P_K)A\|_F \leq \|B\|_F \|A\|_2 (c_{K_2} + c_{K_3} + c_{K_4}) \|K' - K\|_F,$$

Finally, since  $\|P_K\|_2 \leq \mathcal{C}(K)$ , we have

$$\|B^\top P_K B(K' - K)\|_F \leq \|B\|_F^2 \mathcal{C}(K) \|K' - K\|_F.$$

Combining the above upper bounds with the upper bound (37) we conclude that

$$\|E_{K'} - E_K\|_F \leq c_{K_6} \|K' - K\|_F,$$

where  $c_{K_6}$  is given by

$$c_{K_6} = 2 \left[ \|R\|_F + \|B\|_F \|A\|_2 (c_{K_2} + c_{K_3} + c_{K_4}) + \|B\|_F^2 ((c_{K_1} + 1)(c_{K_2} + c_{K_3} + c_{K_4}) + \mathcal{C}(K)) \right].$$

### A.3 Explicit bounds on the parameters $(\rho_{\text{lqr}}, \lambda_{\text{lqr}}, \phi_{\text{lqr}})$

In order to ease notation, we define constants  $\widetilde{c}_{K_7}$ ,  $\widetilde{c}_{K_8}$  and  $\widetilde{c}_{K_9}$  by replacing the scalar  $\mathcal{C}(K)$  by  $10\mathcal{C}(K_0) - 9\mathcal{C}(K^*)$  in the definitions of  $c_{K_7}$ ,  $c_{K_8}$  and  $c_{K_9}$  respectively (see Section A).

**Lemma 9.** *The parameters  $\rho_{\text{lqr}}$ ,  $\lambda_{\text{lqr}}$ ,  $\phi_{\text{lqr}}$  satisfy the following bounds*

$$\rho_{\text{lqr}} \geq \widetilde{c}_{K_9}, \quad \phi_{\text{lqr}} \leq \widetilde{c}_{K_7} \quad \text{and} \quad \lambda_{\text{lqr}} \leq \widetilde{c}_{K_8}.$$

*Proof.* Observe that from the definition of the set  $\mathcal{G}^{\text{lqr}}$  we have that for all  $K \in \mathcal{G}^{\text{lqr}}$ , the function value  $\mathcal{C}(K)$  is upper bounded as  $\mathcal{C}(K) \leq 10\mathcal{C}(K_0) - 9\mathcal{C}(K^*)$ . Consequently, for any  $K \in \mathcal{G}^{\text{lqr}}$  and any  $K'$  such that  $\|K' - K\|_F \leq \widetilde{c}_{K_9}$ , we can use Lemma 2 and Lemma 1 respectively to show that the cost function  $\mathcal{C}(K)$  has locally Lipschitz gradients with parameter  $\widetilde{c}_{K_8}$  and the function  $\mathcal{C}(K)$  has locally Lipschitz function values parameter  $\widetilde{c}_{K_7}$ . Combining the last observation with the definitions of  $\rho_{\text{lqr}}$ ,  $\lambda_{\text{lqr}}$  and  $\phi_{\text{lqr}}$  we have that  $\rho_{\text{lqr}} \geq \widetilde{c}_{K_9}$ ,  $\phi_{\text{lqr}} \leq \widetilde{c}_{K_7}$  and  $\lambda_{\text{lqr}} \leq \widetilde{c}_{K_8}$ . This completes the proof.  $\square$

## B Properties of the LQR problem with noisy dynamics

Recall that we consider the infinite horizon discounted LQR problem where the cost function  $\mathcal{C}_{\text{dyn},\gamma}(K; \mathcal{Z})$  and the state transition dynamics are given by

$$\mathcal{C}_{\text{dyn},\gamma}(K; \mathcal{Z}) := \sum_{t \geq 0} \gamma^t \left( s_t^\top Q s_t + a_t^\top R a_t \right) \quad (39)$$

$$s_t = (A - BK)s_{t-1} + z_t, \quad \text{where} \quad s_0 = 0 \quad \text{and} \quad z_t \stackrel{i.i.d.}{\sim} \mathcal{D}_{\text{add}},$$

where  $\gamma \in (0, 1)$  denotes the discount factor. Also recall that the distribution  $\mathcal{D}_{\text{add}}$  has zero mean, identity covariance, and obeys the relation  $\sup \|z_t\|_2^2 \leq C_m$  almost surely.

The goal of this section is two-fold: to prove Lemma 4 that relates the cost functions  $\mathcal{C}_{\text{init},\gamma}$  and  $\mathcal{C}_{\text{dyn},\gamma}$ , and to establish properties of the gradient estimate in the noisy dynamics setting required to prove Corollary 3. In particular, our main results are stated below, with Lemma 4 reproduced for convenience.

**Lemma 4** (Equivalence of costs up to scaling). *For any policy  $K$ , we have*

$$\mathcal{C}_{\text{dyn},\gamma}(K) = \frac{\gamma}{1 - \gamma} \mathcal{C}_{\text{init},\gamma}(K).$$

**Lemma 10.** *For any policy  $K$ , we have the uniform bound*

$$\mathcal{C}_{\text{dyn},\gamma}(K; \mathcal{Z}) \leq \frac{2C_m}{1 - \sqrt{\gamma}} \cdot \left( \frac{\mathcal{C}_{\text{dyn},\gamma}(K)}{\sigma_{\min}(Q)} \left( \frac{1 - \gamma}{\gamma} \right) \right)^{3/2}.$$

Before moving to the proofs of these lemmas, let us now define some additional notation to facilitate the proofs. Let

$$M := Q + K^\top R K, \quad G := (A - BK) \quad \text{and} \quad c_j := \gamma^j \left( \sum_{i=1}^j G^{j-i} z_i \right)^\top M \left( \sum_{i=1}^j G^{j-i} z_i \right).$$



Also define the cumulative cost up to time  $t$  by  $\mathcal{C}^t = \sum_{j=1}^t c_j$ , so that a simple computation yields the relation  $\mathcal{C}_{\text{dyn},\gamma}(K; \mathcal{Z}) = \lim_{t \rightarrow \infty} \mathcal{C}^t$ .

Additionally, define the matrix  $X_{K,t}$  via its partition into  $t^2$  blocks  $X_{K,t}^{i,j} \in \mathbb{R}^{m \times m}$  for each pair  $(i, j) \in [t] \times [t]$ , as

$$X_{K,t} = \begin{bmatrix} X_{K,t}^{1,1} & X_{K,t}^{1,2} & \cdots & X_{K,t}^{1,t} \\ X_{K,t}^{2,1} & X_{K,t}^{2,2} & \cdots & X_{K,t}^{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ X_{K,t}^{t,1} & X_{K,t}^{t,2} & \cdots & X_{K,t}^{t,t} \end{bmatrix}.$$

Each sub-block  $X_{K,t}^{i,j}$  of  $X_{K,t}$  is given by

$$\begin{aligned} X_{K,t}^{i,j} &= \sum_{k=j}^t \gamma^k (G^{k-i})^\top M G^{k-j} \text{ if } j \geq i, \\ X_{K,t}^{i,j} &= \sum_{k=i}^t \gamma^k (G^{k-i})^\top M G^{k-j} \text{ if } j < i. \end{aligned} \tag{40}$$

Using this matrix notation, a simple computation yields

$$\mathcal{C}^t = \sum_{\substack{i \in [t] \\ j \in [t]}} z_i^\top X_{K,t}^{i,j} z_j.$$

Finally, define the discounted state correlation matrix as

$$\Sigma_{K,\gamma} = \sum_{k=0}^{\infty} (\sqrt{\gamma}A - \sqrt{\gamma}BK)^k ((\sqrt{\gamma}A - \sqrt{\gamma}BK)^k)^\top,$$

and note that this matrix is equal to  $\Sigma_K$  from Equation (25) in Appendix A with the pair of matrices  $(A, B)$  replaced by  $(\sqrt{\gamma}A, \sqrt{\gamma}B)$ .

The following technical lemma is required for the argument.

**Lemma 11.** *For any policy  $K$  and discount factor  $\gamma \in (0, 1)$ , we have*

$$\text{tr} [\Sigma_{K,\gamma}] = \text{tr} \left[ \sum_{k=0}^{\infty} G_\gamma^k (G_\gamma^k)^\top \right] \leq \frac{\mathcal{C}_{\text{dyn},\gamma}(K)}{\sigma_{\min}(Q)} \left( \frac{1-\gamma}{\gamma} \right), \tag{41a}$$

$$\sum_{j=0}^{\infty} \|G_\gamma^j\|_2^2 \leq \frac{\mathcal{C}_{\text{dyn},\gamma}(K)}{\sigma_{\min}(Q)} \left( \frac{1-\gamma}{\gamma} \right), \text{ and} \tag{41b}$$

$$\sum_{j=0}^{\infty} \|\gamma^j G^j\|_2 \leq \frac{(\text{tr} [\Sigma_{K,\gamma}])^{1/2}}{1 - \sqrt{\gamma}}. \tag{41c}$$

See Section B.3 for the proof of this auxiliary claim.

With this set-up, we are now equipped to prove Lemmas 4 and 10.

## B.1 Proof of Lemma 4

Working with the cumulative cost, we have

$$\begin{aligned}\mathbb{E}[\mathcal{C}^t] &= \mathbb{E} \left[ \sum_{\substack{i \in [t] \\ j \in [t]}} \text{tr}(X_{K,t}^{i,j} z_j z_i^\top) \right] \\ &= \sum_{i=1}^t \text{tr} \left( X_{K,t}^{i,i} \right),\end{aligned}$$

where we have used the fact that  $\mathbb{E}[z_j z_i^\top] = \mathbb{I}_{i=j} I$ .

Substituting the definition of the matrix  $X_{K,t}^{i,i}$ , we have

$$\begin{aligned}\mathbb{E}[\mathcal{C}^t] &= \sum_{i=1}^t \text{tr} \left[ \sum_{k=i}^t \gamma^k (G^{k-i})^\top M G^{k-i} \right] \\ &= \sum_{i=1}^t \gamma^i \text{tr} \left[ \sum_{k=0}^{t-i} (G_\gamma^k)^\top M G_\gamma^k \right].\end{aligned}$$

Now for each fixed summand above, taking  $t \rightarrow \infty$  yields

$$\text{tr} \left[ \sum_{k=0}^{\infty} (G_\gamma^k)^\top M G_\gamma^k \right] = \text{tr} [M \Sigma_{K,\gamma}],$$

where we have used the cyclic property of the trace.

Putting together the pieces, we have

$$\begin{aligned}\mathcal{C}_{\text{dyn},\gamma}(K) &= \sum_{i=1}^{\infty} \gamma^i \text{tr} [M \Sigma_{K,\gamma}] \\ &= \left( \frac{\gamma}{1-\gamma} \right) \text{tr} [M \Sigma_{K,\gamma}] \\ &= \left( \frac{\gamma}{1-\gamma} \right) \cdot \mathcal{C}_{\text{init},\gamma}(K),\end{aligned}$$

thereby establishing Lemma 4.

## B.2 Proof of Lemma 10

As before, let us begin by analyzing the cumulative cost up to time  $t$ , and write

$$\mathcal{C}^t = \sum_{\substack{i \in [t] \\ j \in [t]}} z_i^\top X_{K,t}^{i,j} z_j \stackrel{(i)}{\leq} C_m \sum_{\substack{i \in [t] \\ j \in [t]}} \|X_{K,t}^{i,j}\|_2 = C_m \left( \sum_{i=1}^t \sum_{j \geq i} \|X_{K,t}^{i,j}\|_2 + \sum_{j=1}^t \sum_{i > j} \|X_{K,t}^{i,j}\|_2 \right), \quad (42)$$

where in step (i), we have used the fact that  $\|z_i\|_2 \|z_j\|_2 \leq C_m$ .

Bounding the first term on the RHS of Equation (42), we have

$$\begin{aligned}
\sum_{i=1}^t \sum_{j \geq i} \|X_{K,t}^{i,j}\|_2 &= \sum_{i=1}^t \sum_{j=i}^t \left\| \sum_{k=j}^t \gamma^k (G^{k-i})^\top M G^{k-j} \right\|_2 \\
&\leq \sum_{i=1}^t \sum_{j=i}^t \|\gamma^j G^{j-i}\|_2 \cdot \left\| \sum_{k=j}^t \gamma^{k-j} (G^{k-j})^\top M G^{k-j} \right\|_2 \\
&= \sum_{i=1}^t \sum_{j=i}^t \|\gamma^j G^{j-i}\|_2 \cdot \left\| \sum_{k=0}^{t-j} (G_\gamma^k)^\top M G_\gamma^k \right\|_2.
\end{aligned}$$

By symmetry, an identical argument bounds the second term of Equation (42) to yield the uniform bound

$$\begin{aligned}
\mathcal{C}^t &\leq 2C_m \sum_{i=1}^t \sum_{j=i}^t \|\gamma^j G^{j-i}\|_2 \cdot \left\| \sum_{k=0}^{t-j} (G_\gamma^k)^\top M G_\gamma^k \right\|_2 \\
&\stackrel{(ii)}{\leq} 2C_m \sum_{i=1}^t \sum_{j=i}^t \|\gamma^j G^{j-i}\|_2 \cdot \text{tr} \left( \sum_{k=0}^{\infty} (G_\gamma^k)^\top M G_\gamma^k \right) \\
&\stackrel{(iii)}{\leq} 2C_m \cdot \left( \frac{(\text{tr} [\Sigma_{K,\gamma}])^{1/2}}{1 - \sqrt{\gamma}} \right) \cdot \left( \frac{\mathcal{C}_{\text{dyn},\gamma}(K)}{\sigma_{\min}(Q)} \left( \frac{1 - \gamma}{\gamma} \right) \right) \\
&\stackrel{(iv)}{\leq} \frac{2C_m}{1 - \sqrt{\gamma}} \cdot \left( \frac{\mathcal{C}_{\text{dyn},\gamma}(K)}{\sigma_{\min}(Q)} \left( \frac{1 - \gamma}{\gamma} \right) \right)^{3/2},
\end{aligned}$$

where in step (ii), we have used the PSD nature of the matrices being summed, and steps (iii) and (iv) follow from inequalities (41a) and (41c) of Lemma 11, respectively. Since the above relation holds for all  $t$ , we can take the limit  $t \rightarrow +\infty$  on the left-hand side so as to obtain the claim of Lemma 10.

### B.3 Proof of Lemma 11

In this section we prove the auxiliary bounds (41a) through (41c).

**Proof of the bound (41a):** Following the proof of Lemma 4, we have

$$\begin{aligned}
\mathcal{C}_{\text{dyn},\gamma}(K) &= \left( \frac{\gamma}{1 - \gamma} \right) \text{tr} \left[ M \Sigma_{K,\gamma} \right] \\
&= \left( \frac{\gamma}{1 - \gamma} \right) \text{tr} \left[ (Q + K^\top R K) \Sigma_{K,\gamma} \right] \\
&\stackrel{(i)}{\geq} \left( \frac{\gamma}{1 - \gamma} \right) \sigma_{\min}(Q) \text{tr}(\Sigma_{K,\gamma}),
\end{aligned}$$

where (i) follows from Von Neumann's trace inequality. Multiplying both sides above by  $\frac{1 - \gamma}{\gamma \cdot \sigma_{\min}(Q)}$  completes the proof.

**Proof of the bound (41b):** Observe that for any  $j$ , there exists some unit vector  $v_j$  such that  $\|G_\gamma^j\|_2 = \|G_\gamma^j v_j\|_2$ . Using this fact, we have

$$\begin{aligned} \sum_{j=0}^{\infty} \|G_\gamma^j\|_2^2 &= \sum_{j=0}^{\infty} \|G_\gamma^j v_j\|_2^2 = \sum_{j=0}^{\infty} \text{tr} \left[ (G_\gamma^j)^\top G_\gamma^j v_j v_j^\top \right] \\ &\stackrel{(i)}{\leq} \sum_{j=0}^{\infty} \text{tr} \left[ (G_\gamma^j)^\top G_\gamma^j \right] \cdot \|v_j v_j^\top\|_2 \\ &\stackrel{(ii)}{=} \text{tr} [\Sigma_{K,\gamma}] \end{aligned}$$

where step (i) follows from Von Neumann's trace inequality and (ii) follows from the definition of  $\Sigma_{K,\gamma}$ . Applying the bound from Equation (41a) completes the proof.

**Proof of the bound (41c):** Similar to the proof of (41b), observe that,

$$\begin{aligned} \sum_{j=0}^{\infty} \|\gamma^j G^j\|_2 &= \sum_{j=0}^{\infty} \gamma^{j/2} \left( \text{tr} \left[ (G_\gamma^j)^\top G_\gamma^j v_j v_j^\top \right] \right)^{1/2} \\ &\leq \sum_{j=0}^{\infty} \gamma^{j/2} \left( \text{tr} \left[ (G_\gamma^j)^\top G_\gamma^j \right] \right)^{1/2} \\ &\stackrel{(i)}{\leq} \sum_{j=0}^{\infty} \gamma^{j/2} (\text{tr} [\Sigma_{K,\gamma}])^{1/2} \\ &= \frac{(\text{tr} [\Sigma_{K,\gamma}])^{1/2}}{1 - \sqrt{\gamma}}, \end{aligned}$$

where step (i) follows from using  $\text{tr} \left[ (G_\gamma^j)^\top G_\gamma^j \right] \leq \sum_{j=0}^{\infty} \text{tr} \left[ (G_\gamma^j)^\top G_\gamma^j \right] = \text{tr} [\Sigma_{K,\gamma}]$ .

## C Proof of Lemma 6

We now provide the proof of Lemma 6, splitting our analysis into the two separate claims.

### C.1 Proof of part (a)

Unwrapping the definition of  $\nabla f_r(x)$  yields

$$\begin{aligned} \nabla f_r(x) &\stackrel{(i)}{=} \frac{d}{r} \mathbb{E}[f(x + ru)u] \\ &= \frac{d}{2r} (\mathbb{E}[f(x + ru)u] + \mathbb{E}[f(x - ru)u]) \\ &\stackrel{(ii)}{=} \frac{d}{2r} (\mathbb{E}[f(x + ru)u] - \mathbb{E}[f(x - ru)u]) \\ &= \frac{d}{2r} \mathbb{E}[f(x + ru)u - f(x - ru)u], \end{aligned}$$

where equality (i) follows from Lemma 1 in Flaxman et al. [FKM05], and equality (ii) follows from the symmetry of the uniform distribution on the shell  $\mathbb{S}^{d-1}$ . Now observe that

$$\begin{aligned}\mathbb{E}[F(x + ru, \xi)u - F(x - ru, \xi)u] &= \mathbb{E}\left[\mathbb{E}[F(x + ru, \xi) - F(x - ru, \xi)u|u]\right] \\ &\stackrel{(i)}{=} \mathbb{E}\left[f(x + ru)u - f(x - ru)u\right],\end{aligned}$$

where equality (i) follows from the assumption that  $f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[F(x, \xi)]$ . Putting the equations together establishes the claim in part (a).  $\square$

**Proof of Lemma 6, part (b)** Observe that

$$\begin{aligned}\|\nabla f_r(x) - \nabla f(x)\|_2 &= \|\nabla \mathbb{E}[f(x + rv)] - \nabla f(x)\|_2 \\ &= \|\mathbb{E}[\nabla[f(x + rv) - \nabla f(x)]]\|_2 \\ &\stackrel{(i)}{\leq} \mathbb{E}[\|\nabla[f(x + rv) - \nabla f(x)]\|_2] \\ &\stackrel{(ii)}{\leq} \phi_0 r,\end{aligned}$$

where inequality (i) above follows from Jensen's inequality, whereas step (ii) follows since  $r \leq \rho$  and  $\nabla f$  is locally Lipschitz continuous with parameter  $\phi_0$ .  $\square$

## D Experimental Details & Additional Experiments

For each LQR problem used, the initial  $K_0$  was picked by randomly perturbing the entries of  $K^*$ . The step size was tuned manually and the smoothing radius was always chosen to be the minimum of  $\sqrt{\epsilon}$  and the largest value required to ensure stability. The rollout length was also tuned manually until the cost from a rollout converged arbitrarily close to the true value.

### D.1 Details of Experiments from Section 3

To generate the plot in Figure 1 (a), we used the following one dimensional LQR problem:

$$A = 5, \quad B = 0.33, \quad Q = 1, \quad R = 1,$$

where we operated in the one-point random initialization setting, the initial state was sampled uniformly at random from the set  $\{4, 5, 6\}$ , and the discount factor was set to 1.

To generate the plots in Figure 1 (b), Figure 2 (b) and Figure 3 (a), we used the following LQR problem:

$$A = \begin{bmatrix} 1 & 0 & -10 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & -10 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}, \quad Q = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}, \quad R = \begin{bmatrix} 5 & -3 & 0 \\ -3 & 5 & -2 \\ 0 & -2 & 5 \end{bmatrix},$$

where we operated in the two-point random initialization setting, the initial state was sampled uniformly at random from the canonical basis vectors, and the discount factor was set to 1.

To generate the plots in Figure 2 (a) and 2 (c), we used the following LQR problem:

$$A = 0.1 \times I \quad B = 0.01 \times I \quad Q = 100 \times I \quad R = 100 \times I,$$

where  $I$  represents the  $3 \times 3$  identity matrix. For Figure 2 (a) we operated in the random initialization setting, and used initial states which were sampled uniformly at random from the rows of the matrix  $\frac{\sqrt{3}}{25} \times I$ . For Figure 2 (c), we operated in the one-point additive noise setting. Here the initial state was set to the zero vector, and we used additive noise at each timestep sampled from a zero mean Gaussian with covariance matrix  $\frac{1}{25} \times I$ . In both settings, the discount factor was set to 0.9. For this example, the population level costs in the two settings are equal up to a constant scaling factor.

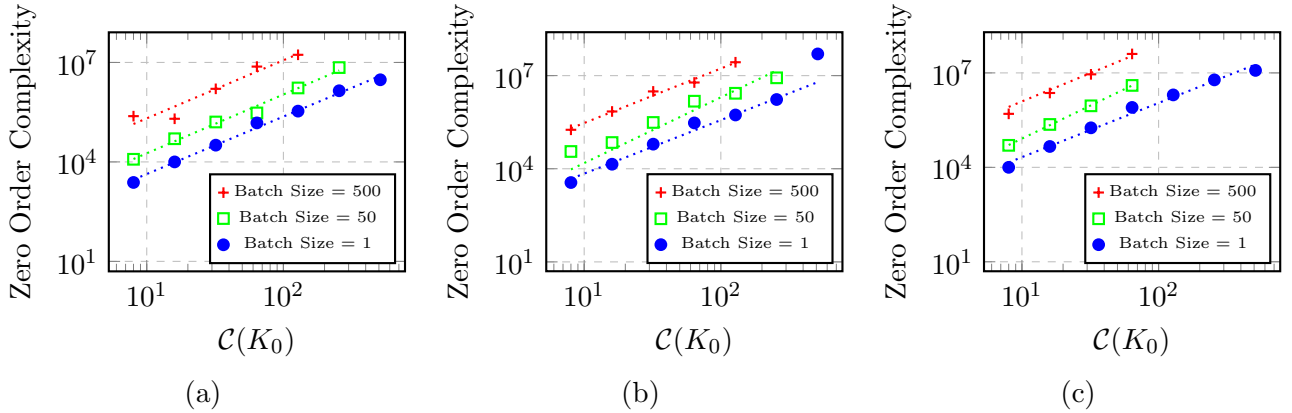
To generate the plot in Figure 3 (b), we used the following LQR problem:

$$A = 0.1 \times I \quad B = 0.01 \times I \quad Q = 25 \times I \quad R = 25 \times I,$$

where  $I$  represents the  $3 \times 3$  identity matrix. We operated in the one-point additive noise setting. The initial state was set to the zero vector, and we used additive noise at each timestep sampled from a zero mean Gaussian with covariance matrix  $\frac{1}{25} \times I$ . The discount factor was set to 0.9.

## D.2 Additional Experiments

In the two point random initialization setting, we performed experiments on several additional LQR instances to test the robustness of the behavior observed in Figures 1 and 3. For ease in notation, we use  $\mathcal{C}$  to denote the population cost for the remainder of this section. Note that for all figures shown in this section, each dotted line represents the line of best fit for its corresponding data points, as in Figures 2 and 3. Using the same example used to generate the plots in Figure 2 (b) and Figure 3 (a), we tested the performance of our two-point algorithm with different values of  $\epsilon$  and  $\mathcal{C}(K_0)$ .



**Figure 4.** Scaling of complexity vs.  $\mathcal{C}(K_0)$  while using minibatches of size 1, 50 and 500, to achieve an error tolerance of (a)  $\epsilon = 0.1$ , (b)  $\epsilon = 0.05$  and (c)  $\epsilon = 0.01$ . Due to the prohibitive complexity when using batches of size 50 and 500, we omit data points for large values of  $\mathcal{C}(K_0)$ .

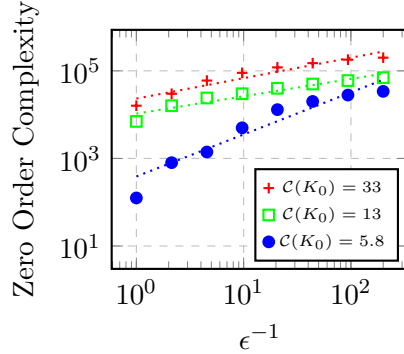
In Figure 4 (a) (b) and (c), we plot the scaling of the zero-order complexity with  $\mathcal{C}(K_0)$  for different values of the tolerance  $\epsilon$ , and each figure additionally contains plots for different values of

the batch-size. We observe that the scaling of our algorithm with respect to  $\mathcal{C}(K_0)$  is approximately on the order of  $\mathcal{O}(\mathcal{C}(K_0)^2)$ , suggesting that our bounds for the Lipschitz and smoothness constants are not sharp in this respect. The same plots also demonstrate that using larger batch sizes is often suboptimal: while the step size can be increased with increasing batch-size, it eventually plateaus due to stability considerations, leading to higher overall zero-order complexity.

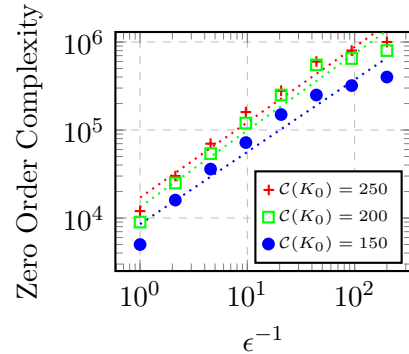
We also ran our algorithm on the following problem introduced by Dean et al. [DMM<sup>+</sup>17], who used this example in their study of model based control methods for the LQR problem. Consider the LQR problem defined by:

$$A = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix}, \quad B = I, \quad Q = 10^{-3} \times I, \quad R = I.$$

For three different values of  $\mathcal{C}(K_0)$ , we picked 8 evenly spaced (logarithmic scale) values of  $\epsilon$  in the interval  $(0.005, 1)$ . The initial state was sampled uniformly at random from  $\{[5, 0, 0], [5, 5, 5], [0, 0, 5]\}$ . The cost of the optimal policy in our example was  $\mathcal{C}(K^*) = 2.36$ . We then measured the total zero order complexity required to attain  $\epsilon$  convergence. These results are plotted in Figure 5.



**Figure 5.** Scaling of complexity vs.  $\epsilon^{-1}$  in LQR instance from Dean et al. [DMM<sup>+</sup>17]



**Figure 6.** Scaling of complexity vs.  $\epsilon^{-1}$  in randomly generated  $8 \times 8$  example.

Finally, we also obtained data for the scaling with respect to  $\epsilon$  on an example in slightly higher dimensions, to empirically verify the fact that our algorithm can be used for LQR problems larger than  $3 \times 3$ . We randomly generated  $A$ ,  $B$ ,  $Q$  and  $R$  as  $8 \times 8$  matrices. Each entry of  $A$  was independently sampled from the Gaussian distribution  $\mathcal{N}(2, 1)$ , and each entry of  $B$  was independently sampled from the Gaussian distribution  $\mathcal{N}(0, 1)$ . To generate each of  $Q$  and  $R$ , we generated a matrix where each entry was independently sampled from the Gaussian distribution  $\mathcal{N}(5, 1)$ , then symmetrized the matrix by adding it to its transpose, finally adding  $10I$  to ensure positive definiteness. The initial states were sampled uniformly at random from the columns of the  $8 \times 8$  identity matrix. For three different values of  $\mathcal{C}(K_0)$ , we picked 8 evenly spaced (logarithmic scale) values of  $\epsilon$  in the interval  $(0.005, 1)$ . We then measured the total zero order complexity required to attain  $\epsilon$  convergence. These results are plotted in Figure 6.

## References

- [ADX10] Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, June 2010.
- [AJ17] Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.
- [AL18] Marc Abeille and Alessandro Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1–9, 2018.
- [AOM17] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.
- [AYLS18] Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Regret bounds for model-free linear quadratic control. *arXiv preprint arXiv:1804.06021*, 2018.
- [AYS11] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.
- [Ber05] Dimitri P Bertsekas. *Dynamic programming and optimal control. Vol. I*. Athena Scientific, Belmont, MA, third edition, 2005.
- [CHK<sup>+</sup>18] Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1029–1038, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.
- [DJWW15] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Trans. Information Theory*, 61(5):2788–2806, 2015.
- [DLB17] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.
- [DMM<sup>+</sup>17] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *arXiv preprint arXiv:1710.01688*, 2017.
- [DMM<sup>+</sup>18] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. *arXiv preprint arXiv:1805.09388*, 2018.



- [DRF12] Mark P Deisenroth, Carl E Rasmussen, and Dieter Fox. Learning to control a low-cost manipulator using data-efficient reinforcement learning. In *Robotics: Science and Systems*, volume 7, pages 57–64, 2012.
- [Dur10] Rick Durrett. *Probability: theory and examples*. Cambridge university press, 2010.
- [FGKM18] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018*, pages 1466–1475, 2018.
- [Fie97] Claude-Nicolas Fiechter. PAC adaptive control of linear systems. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory, COLT '97*, pages 72–80, New York, NY, USA, 1997. ACM.
- [FKM05] Abraham Flaxman, Adam Kalai, and Brendan McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, January 2005.
- [FTM17] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time analysis of optimal adaptive policies for linear-quadratic systems. *arXiv preprint arXiv:1711.07230*, 2017.
- [GL13] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- [GLSL16] Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous deep q-learning with model-based acceleration. In *International Conference on Machine Learning*, pages 2829–2838, 2016.
- [HKZ12] Daniel Hsu, Sham M Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17:no. 52, 6, 2012.
- [JNR12] Kevin G Jamieson, Robert Nowak, and Ben Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems 25*, pages 2672–2680. Curran Associates, Inc., 2012.
- [Kal60] Rudolf E Kalman. Contributions to the theory of optimal control. *Boletín de la Sociedad Matemática Mexicana*, 5:102–119, 1960.
- [KNS16] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition. In *European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9851, ECML PKDD 2016*, pages 795–811, Berlin, Heidelberg, 2016. Springer-Verlag.
- [LFDA16] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

- [LHP<sup>+</sup>15] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.
- [Lju98] Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.
- [Loj63] Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, pages 87–89, 1963.
- [M<sup>+</sup>15] Volodymyr Mnih et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015.
- [MU05] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, Cambridge, 2005.
- [Nes11] Yurii Nesterov. Random gradient-free minimization of convex functions. CORE Discussion Papers 2011001, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2011.
- [Pol64] Boris T Polyak. Gradient methods for solving equations and inequalities. *USSR Computational Mathematics and Mathematical Physics*, 4(6):17 – 32, 1964.
- [Ric24] Jacopo Riccati. Animadversiones in aequationes differentiales secundi gradus. *Acta Eruditorum Lipsiae*, 1724.
- [RLTK17] Aravind Rajeswaran, Kendall Lowrey, Emanuel V Todorov, and Sham M Kakade. Towards generalization and simplicity in continuous control. In *Advances in Neural Information Processing Systems 30*, pages 6550–6561. 2017.
- [S<sup>+</sup>16] David Silver et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016.
- [Sha13] Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24, 2013.
- [Sha17] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *J. Mach. Learn. Res.*, 18(1):1703–1713, January 2017.
- [SHC<sup>+</sup>17] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.
- [SLA<sup>+</sup>15] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.
- [Spa05] James C Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

- [TFR<sup>+</sup>17] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*, pages 23–30. IEEE, 2017.
- [TR18] Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *ICML*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 5012–5021. JMLR.org, 2018.
- [TVar] Yan Shuo Tan and Roman Vershynin. Phase retrieval via randomized kaczmarz: theoretical guarantees. *Information and Inference: A Journal of the IMA*, to appear.
- [WBS18] Yining Wang, Sivaraman Balakrishnan, and Aarti Singh. Optimization of smooth functions with noisy observations: Local minimax rates. *arXiv preprint arXiv:1803.08586*, 2018.
- [WDBS18] Yining Wang, Simon S Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1356–1365, 2018.
- [Whi96] Peter Whittle. *Optimal control: Basics and Beyond*. Wiley and Sons, Chichester, England, 1996.