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

Dhruv Malik[†] **Ashwin Pananjady**[†] **Kush Bhatia**[†]
Koulik Khamaru[⋆] **Peter L. Bartlett**[†,⋆] **Martin J. Wainwright**[†,⋆]
Departments of [†]EECS and [⋆]Statistics, UC Berkeley

Abstract

We study derivative-free methods for policy optimization over the class of linear policies. We focus on characterizing the convergence rate of a canonical stochastic, two-point, derivative-free method for linear-quadratic systems in which the initial state of the system is drawn at random. In particular, we show that for problems with effective dimension D, such a method converges to an ϵ-approximate solution within $\widetilde{\mathcal{O}}(D/\epsilon)$ steps, with multiplicative pre-factors that are explicit lower-order polynomial terms in the curvature parameters of the problem. Along the way, we also derive stochastic zero-order rates for a class of non-convex optimization problems.

1 Introduction

Recent years have witnessed a number of successes in applying modern reinforcement learning (RL) methods to many fields, including robotics [44, 26] and competitive gaming [41, 32]. 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* [13, 21, 27], 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 [32], and TRPO [38], 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 [42] 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 [38], PSNG [35] and evolutionary strategies [37], 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 [1, 2, 3, 9, 11, 12, 16, 17, 45] 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 [24, 48]. 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 and when our goal is to obtain a policy whose cost is at most ϵ-suboptimal. The sample

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

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 ϵ. 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 ϵ, but also the dimension and conditioning of the underlying problem.

Our analysis assumes a distinct form of randomness in the underlying linear system: the initial state is chosen randomly from an unknown distribution, but the linear dynamics at each time step remain deterministic [17]. We refer to this setting as 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 [5, 10, 6], 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 [28] as well as references therein). Non-asymptotic guarantees for model-based control of linear quadratic systems were first obtained by Fiechter [18], 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. [11] proposed a robust alternative to nominal control, showing an improved sample complexity as well as better-behaved 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 [12, 1, 3].

In this paper, we study model-free control of these systems, a problem that has seen some recent work in both the offline [17] and online [2] settings. Most directly relevant to our work is the paper of Fazel et al. [17], who studied the offline setting for the randomly initialized 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. [19] showed that these methods can be analyzed for convex optimization by making an explicit connection to function smoothing, and Agarwal et al. [4] improved some of these convergence rates. Results are also available for strongly convex [23], smooth [20] and convex [33, 14, 47] functions, with Shamir [39, 40] 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 [19], or a *k-point estimate* [4] for some $k \geq 2$. There has also been some recent work on constrained zero-order optimization of high dimensional non-convex functions [7], as well zero-order optimization of non-convex functions satisfying certain smoothness properties that are motivated by statistical estimation [46].

Our contributions In this paper, we study randomly initialized 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 algorithm that we analyze is a standard and practical two-point variant of the stochastic search heuristic. Our main contribution is stated in the following informal theorem (to be stated more precisely in the sequel):

Main Theorem (informal). *In an m-dimensional state space, with high probability one can obtain an ϵ-approximate solution to a linear quadratic system from observing the noisy costs of $\widetilde{\mathcal{O}}(m^2/\epsilon)$ trajectories from the system.*

In our theoretical statements, the multiplicative prefactors are explicit lower-order polynomials of the 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 [17]. We also isolate and sharpen some key properties that are essential to establishing sharp rates of zero-order optimization; as an example, compared to the setting with random-initialization and one-point reward feedback studied by Fazel et al. [17], establishing these properties allows

us to analyze a natural algorithm that improves[1] the dependence of the bound on the error tolerance ϵ from at least $\mathcal{O}\left(1/\epsilon^4\right)$ to $\mathcal{O}\left(1/\epsilon\right)$. 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

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

where ξ 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 the two point oracle model, 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 domain $\mathcal{X}$ is said to have (ϕ, β) locally Lipschitz gradients at $x \in \mathcal{X}$ if for all $y \in \mathcal{X}$ with $\|x - y\|_2 \leq \beta$.*

$$\|\nabla g(y) - \nabla g(x)\|_2 \leq \phi \|y - x\|_2 \tag{2}$$

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 (ϕ, β) 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.

Definition 2 (Locally Lipschitz Function). *A continuously differentiable function g with domain $\mathcal{X}$ is said to be (λ, ζ) locally Lipschitz at $x \in \mathcal{X}$ if for all $y \in \mathcal{X}$ such that $\|x - y\|_2 \leq \zeta$*

$$|g(y) - g(x)| \leq \lambda \|y - x\|_2 \tag{3}$$

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 (λ, ζ) 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 domain $\mathcal{X}$ and a finite global minimum g^* is said to be μ-PL if it satisfies the Polyak-Łojasiewicz (PL) inequality with constant $\mu > 0$, given by*

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

The PL condition, first introduced by Polyak [34] and Łojasiewicz [29], 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. [25] 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 time-step 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 $\mathcal{H}_t := (s_0, s_1, \ldots, s_{t-1})$, so as to solve

[1] While the rates established by Fazel et al. [17] are not explicit, their algorithm is conservative and a bound of order $1/\epsilon^4$ can be distilled by working through their analysis.

the optimization problem

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

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 [8]), 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

$$s_{t+1}=As_t+Ba_t,$$

where A and B are transition matrices of the appropriate dimension. The randomness in the problem comes from choosing the initial state s_0 at random from a distribution $\mathcal{D}$.

Throughout this paper, we assume[2] that for a random variable $v\sim\mathcal{D}$, we have

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

[2]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 Σ, we may re-parametrize the problem with the modifications

$$A'=\Sigma^{-1/2}A\Sigma^{1/2}, \quad B'=\Sigma^{-1/2}B, \text{ and } s_t'=\Sigma^{-1/2}s_t,$$

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).

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 [22] 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 [24, 48], 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 [36].

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).

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

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

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

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

over choices of the policy[3] K.

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

[3]Such a setting should be contrasted with the setting with additive noise, for which we also obtain guarantees; these can be found in the full version of the present paper [30].

condition $\rho_{\mathrm{spec}}(A-BK) < 1$. Furthermore, we assume access to *some* policy K_0 with finite cost (see the related literature [17, 12]); 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 function $\mathcal{C}_{\mathsf{init},\gamma}(K)$ and its sample variant $\mathcal{C}_{\mathsf{init},\gamma}(K, s_0)$, in order to place the problem within the context of optimization.

First, it is important to note that the population cost function $\mathcal{C}_{\mathsf{init},\gamma}(K)$ is non-convex. In particular, for any unstable policy, the state sequence blows up and the costs becomes infinite, but as noted by Fazel et al. [17], the stabilizing region $\{K : \rho_{\mathrm{spec}}(A - BK) < 1\}$ is non-convex, thereby rendering our optimization problems non-convex.

In spite of this non-convexity, the cost function exhibit many properties that make it amenable to fast stochastic optimization methods. Variants of the following properties were first established by Fazel et al. [17] for the random initialization cost function $\mathcal{C}_{\mathsf{init},\gamma}$. The following Lemma 1 and Lemma 2 require certain refinements of their claims, which we prove in Appendix C.

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

$$|\mathcal{C}_{\mathsf{init},\gamma}(K'; s_0) - \mathcal{C}_{\mathsf{init},\gamma}(K; s_0)| \leq \lambda_K \|K' - K\|_F.$$

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

$$\|\nabla \mathcal{C}_{\mathsf{init},\gamma}(K') - \nabla \mathcal{C}_{\mathsf{init},\gamma}(K)\|_F \leq \phi_K \|K' - K\|_F. \quad (10)$$

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

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

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

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

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 $\mathcal{C}_{\mathsf{init},\gamma}(K; s_0)$ and we consider two-point setting. In the two-point setting, a query of the function at the points (K, K') obtains the pair of noisy function values $\mathcal{C}_{\mathsf{init},\gamma}(K; s_0)$ and $\mathcal{C}_{\mathsf{init},\gamma}(K'; s_0)$ 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.

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. [17] for the random initialization model—albeit with sub-optimal guarantees—we study a two-point query model, which is known to lead to better dimension-dependence in zero-order stochastic optimization [14].

3 Main results

Our main result is the analysis of a stochastic zero-order optimization algorithm (Algorithm 1) for the linear quadratic regulator (LQR) problem, for which we provide bounds on the sub-optimality gap that hold with non-trivial probability. We begin by introducing the stochastic zero-order algorithm that we analyze for this setting, previously analyzed by Agarwal et al. [4] and Shamir [40] in the context of convex optimization.

3.1 Stochastic zero-order algorithm

We analyze a standard zero-order algorithm for stochastic optimization [4, 40] in application to the LQR problem. We introduce 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 de-

scribe how such an estimate is constructed in the two-point setting. Let $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ denote the d-dimensional unit shell. Let $\mathrm{Unif}(\mathbb{S}^{d-1})$ denote the uniform distribution over the set $\mathbb{S}^{d-1}$.

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

$$\mathrm{g}(x) := \big[F(x + ru, \xi) - F(x - ru, \xi)\big] \frac{d}{2r} u. \quad (11)$$

We point outside that the gradient estimate $g(x)$ above depends on the random direction u and the smoothing radius r. For notational convenience, in the rest of the paper, we hide the dependence on u and r.

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 in general 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** $t \in \{0, 1, \ldots, T-1\}$ **do**
3: Sample $\xi_t \sim \mathcal{D}$ and $u_t \sim \mathrm{Unif}(\mathbb{S}^{d-1})$
4: $\mathrm{g}(x_t) = \big[F(x + ru_t, \xi_t) - F(x - ru_t, \xi_t)\big] \frac{d}{2r} u_t$
5: $x_{t+1} \leftarrow x_t - \eta \mathrm{g}(x_t)$
return x_T

3.2 Convergence guarantees

We now turn to analyzing Algorithm 1 in the settings of LQR. As mentioned before, the difficulty of optimizing the LQR cost function 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 bounds on the locally Lipschitz and 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 are interested in optimizing a function $f \equiv \mathbb{E}_{s_0}[\mathcal{C}(\cdot; s_0)]$ obeying the PL inequality as well as certain local curvature conditions.

Recall that we are given an initial point K_0 with finite cost $\mathcal{C}(K_0)$; the global upper bound on the cost that we target in the analysis is set according to the cost $\mathcal{C}(K_0)$ of this initialization. Given the initial gap to optimality $\Delta_0 := \mathcal{C}(K_0) - \mathcal{C}(K^*)$, we define the set

$$\mathcal{G}^0 := \big\{K \mid \mathcal{C}(K) - \mathcal{C}(K^*) \leq 10\Delta_0\big\}, \quad (12)$$

corresponding to points K whose cost gap is at most ten times the initial cost gap Δ_0.

Assume that the function $\mathcal{C}$ is (ϕ_K, β_K) locally smooth and (λ_K, ζ_K) locally Lipschitz at the point K. Thus, both of these properties hold simultaneously within a neighborhood of radius $\rho_x = \min\{\beta_K, \zeta_K\}$ of the point K. Now define the quantities

$$\phi_0 := \sup_{K \in \mathcal{G}^0} \phi_K, \quad \lambda_0 := \sup_{K \in \mathcal{G}^0} \lambda_K, \quad \text{and} \quad \rho_0 := \inf_{K \in \mathcal{G}^0} \rho_K.$$

By defining these quantities, we have effectively transformed the local properties of the function $\mathcal{C}$ 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. Additionally, we define

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

where the stochastic gradient $\mathrm{g}(x)$ is defined in step (4) of Algorithm 1. In Appendix A.2, provide the following concrete upper bounds for G_∞ and G_2:

$$G_\infty \leq D\lambda_0 \quad \text{and} \quad G_2 \leq D\lambda_0^2. \quad (13)$$

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

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

$$\eta \leq \min\left\{\frac{\epsilon\mu}{240\phi_0 G_2}, \frac{1}{2\phi_0}, \frac{\rho_0}{G_\infty}\right\}, \qquad \textit{and} \quad (14a)$$

$$r \leq \min\left\{\frac{\theta_0\mu}{8\phi_0}\sqrt{\frac{\epsilon}{15}}, \frac{1}{2\phi_0}\sqrt{\frac{\epsilon\mu}{30}}, \rho_0\right\}. \quad (14b)$$

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

$$\mathcal{C}(K_T) - \mathcal{C}(K^*) \leq \epsilon \tag{14c}$$

with probability greater than 3/4.

The proof of the theorem is deferred to Appendix A.

A few comments on Theorem 1 are in order. First, notice that the algorithm is guaranteed to return an ϵ-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 $\widetilde{\mathcal{O}}(\log(\frac{1}{\delta}))$ 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 [31, 43]. 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 Appendix A) 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. We now discuss the dependence of the sample complexity on the various parameters in more detail.

Dependence on ϵ: Our bound shows that we have a $\widetilde{\mathcal{O}}(\frac{1}{\epsilon})$ convergence rate. This fast rate arises due to the relatively low variance of the gradient estimator in the two-point setting, which is independent of the smoothing radius r, as has been frequently noted in past literature on zero order optimization [4, 40, 14]. 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, enabling fast convergence. See Figure 1(a) for a numerical confirmation of this scaling.

Dependence on dimension: The dependence on dimension enters our bound via the variance of the gradient estimate, as is typical of many derivative-free procedures [14, 40]. The two-point setting gives rise to a dimension dependence which is linear in D (the dimension of our optimization variable), and the reason is similar to why this occurs for convex optimization [40]. It is particularly interesting to compare the dimension dependence to results in model-based control [11] with noisy dynamics. There, 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.

We observe that other dimension-dependent quantities such as C_m sit implicitly in the bounds we have derived for the curvature parameters (ϕ_0, λ_0, μ).

Dependence on curvature parameters: The iteration complexity scales linearly in the smoothness parameter of the problem ϕ_0, and quadratically in the other curvature parameters λ_0 and μ. See Appendix C.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 dimension-dependent quantity $\mathcal{C}(K_0)$. In practice, however, it is likely that much sharper bounds can be proved on these parameters, e.g., in simulation (see Figure 1(c)), 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. An important direction for future work is to derive tight bounds on the dependence on dimension and $\mathcal{C}(K)$ for the curvature parameters.

4 Experimental Results

In this section, we present experiments to examine the accuracy of our theoretical results, and compare the derivative-free approach to a variant of the algorithm of Fazel et al. [17]. The algorithm given in Fazel et al. [17] uses a one point estimate of the gradient that suffers from a very high variance. Therefore, and in the spirit of keeping our comparison fair, we instead investigate the performance of a batch version of the two-point gradient estimate, which for large batch-sizes, resembles the algorithm of Fazel et al. [17] in spirit, since we obtain high-accuracy estimates of the gradients[4]. Importantly, note that Algorithm 1 corresponds to using a batch size of 1 at each time step.

We conducted three different experiments. The first was to verify that the scaling of the iteration complexity with the parameter ϵ is indeed the $\widetilde{\mathcal{O}}(\frac{1}{\epsilon})$ rate that Theorem 1 predicts. The second was to better understand the scaling of the iteration complexity with the parameter $\mathcal{C}(K_0)$, as well as to see how batching at each time step affects performance. The third experiment was used to compare the scaling of the step size

[4]We remark that Fazel et al. also study a zero order natural gradient algorithm. We do not compare to this since it requires access to a stronger oracle model.

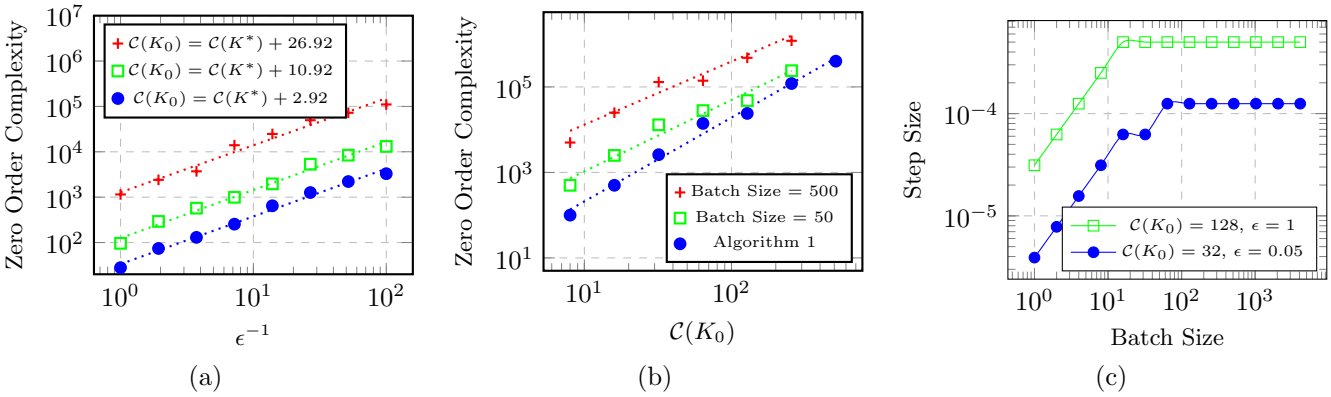

Figure 1. Number of samples required to reach an error tolerance of ϵ, (a) plotted against increasing values of $1/\epsilon$ for different initialization values and (b) plotted against increasing values of $\mathcal{C}(K_0)$ for different batch sizes. In (c), we depict the maximum step size that allows for convergence, plotted against the size of the minibatch used to estimate the gradient. Each dotted line represents the line of best fit for the corresponding data points. For more problem details, see Appendix D.

required to converge to optimum, with the batch size. We used a 3×3 LQR problem for these experiments with $\mathcal{C}(K^*) = 5.08$; the other parameters are specified in Appendix D due to space constraints. In all of our experiments, the step-size was manually tuned to ensure the fastest possible rate of convergence.

We depict the results of the first experiment in Figure 1(a), where we average over 20 runs of the algorithm. The best fit lines for each setting of the initialization accuracy $\mathcal{C}(K_0)$ are plotted as dotted lines. The plot confirms that the empirical scaling is roughly $\mathcal{O}(\frac{1}{\epsilon})$, as predicted by Theorem 1. We additionally verified this on different LQR problem instances; the results are in the Appendix D.

The results from the second experiment are shown in Figure 1(b), where we also test multiple batch-sizes for the problem. Once again, the best fit lines for these data points are shown as dotted lines. The plot reveals that the scaling with initial cost is of the order $\mathcal{O}(\mathcal{C}(K_0)^2)$. This reveals a gap between the dependency on $\mathcal{C}(K_0)$ of the Lipschitz and smoothness constants that our theoretical analysis provides, and the true behavior of the problem. Another interesting aspect of these plots is the behavior of batching for this problem. Typically, we would expect that using a larger batch size corresponds to being able to use a proportionally larger step size for the problem, and so the total zero order complexity while using different batch sizes should remain more or less constant. However, the plots show that in the LQR problem, a larger batch size requires more zero-order evaluations.

We explored this phenomenon further in the third experiment, and our plots in Figure 1(c) show the maximum possible step-size that can be set to ensure convergence. Notably, the step size is limited not just by the variance of the gradient estimate (i.e. the size of the batch), but additionally by stability considerations in the problem. A larger batch does not help if the step size is too large and one eventually takes a step out of the region of stability, and our plots corroborate this intuition. Further details are provided in Appendix D.

Discussion: We showed that under a two-point evaluation oracle, a canonical derivative-free optimization method achieves a fast rate 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 – in particular, we handle both the unboundedness of the cost, and the non-convexity of the domain directly. Interestingly, our proof only relies on certain local properties of the function that can be guaranteed over a bounded set, and so our analysis is more broadly applicable.

While this paper analyzes a canonical zero-order optimization algorithm for model-free control of linear quadratic systems, many open questions remain. First, in order to carry out a fair comparison between model-based and model-free RL in this setting, it is important to analyze this algorithm for linear quadratic systems with additive noise in the dynamics. Lower bounds in the model-free setting are also interesting in this regard, and likely to borrow from the extensive literature on lower bounds in zero-order optimization [39]. 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 we are optimizing over?

References

[1] Y. Abbasi-Yadkori and C. 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.

[2] Y. Abbasi-Yadkori, N. Lazic, and C. Szepesvári. Regret bounds for model-free linear quadratic control. *arXiv preprint arXiv:1804.06021*, 2018.

[3] M. Abeille and A. 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.

[4] A. Agarwal, O. Dekel, and L. 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.

[5] S. Agrawal and R. Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.

[6] M. G. Azar, I. Osband, and R. 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.

[7] K. Balasubramanian and S. Ghadimi. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3455–3464. Curran Associates, Inc., 2018.

[8] D. P. Bertsekas. *Dynamic programming and optimal control. Vol. I.* Athena Scientific, Belmont, MA, third edition, 2005. ISBN 1-886529-26-4.

[9] A. Cohen, A. Hasidim, T. Koren, N. Lazic, Y. Mansour, and K. 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.

[10] C. Dann, T. Lattimore, and E. Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

[11] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu. On the sample complexity of the linear quadratic regulator. *arXiv preprint arXiv:1710.01688*, 2017.

[12] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. *arXiv preprint arXiv:1805.09388*, 2018.

[13] M. P. Deisenroth, C. E. Rasmussen, and D. Fox. Learning to control a low-cost manipulator using data-efficient reinforcement learning. In *Robotics: Science and Systems*, volume 7, pages 57–64, 2012.

[14] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Trans. Information Theory*, 61(5): 2788–2806, 2015.

[15] R. Durrett. *Probability: theory and examples*. Cambridge university press, 2010.

[16] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis. Finite time analysis of optimal adaptive policies for linear-quadratic systems. *arXiv preprint arXiv:1711.07230*, 2017.

[17] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1466–1475, 2018.

[18] C.-N. 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. ISBN 0-89791-891-6.

[19] A. Flaxman, A. Kalai, and B. 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. ISBN 0-89871-585-7.

[20] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[21] S. Gu, T. Lillicrap, I. Sutskever, and S. Levine. Continuous deep q-learning with model-based acceleration. In *International Conference on Machine Learning*, pages 2829–2838, 2016.

[22] D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17:no. 52, 6, 2012. ISSN 1083-589X.

[23] K. G. Jamieson, R. Nowak, and B. Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems 25*, pages 2672–2680. Curran Associates, Inc., 2012.

[24] R. E. Kalman. Contributions to the theory of optimal control. *Boletin de la Sociedad Matematica Mexicana*, 5:102–119, 1960.

[25] H. Karimi, J. Nutini, and M. 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. ISBN 978-3-319-46127-4.

[26] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17 (1):1334–1373, 2016.

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

[28] L. Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

[29] S. Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, pages 87–89, 1963.

[30] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. L. Bartlett, and M. J. Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *CoRR*, abs/1812.08305, 2018.

[31] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, Cambridge, 2005. ISBN 0521835402 9780521835404.

[32] V. Mnih et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 00280836.

[33] Y. Nesterov. Random gradient-free minimization of convex functions. CORE Discussion Papers 2011001, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2011.

[34] B. T. Polyak. Gradient methods for solving equations and inequalities. *USSR Computational Mathematics and Mathematical Physics*, 4(6):17 – 32, 1964. ISSN 0041-5553.

[35] A. Rajeswaran, K. Lowrey, E. V. Todorov, and S. M. Kakade. Towards generalization and simplicity in continuous control. In *Advances in Neural Information Processing Systems 30*, pages 6550–6561. 2017.

[36] J. Riccati. Animadversiones in aequationes differentiales secundi gradus. *Acta Eruditorum Lipsiae*, 1724.

[37] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

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

[39] O. Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24, 2013.

[40] O. Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *J. Mach. Learn. Res.*, 18(1):1703–1713, Jan. 2017. ISSN 1532-4435.

[41] D. Silver et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529 (7587):484–489, January 2016.

[42] J. C. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

[43] Y. S. Tan and R. Vershynin. Phase retrieval via randomized kaczmarz: theoretical guarantees. *Information and Inference: A Journal of the IMA*, to appear.

[44] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. 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.

[45] S. Tu and B. 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.

[46] Y. Wang, S. Balakrishnan, and A. Singh. Optimization of smooth functions with noisy observations: Local minimax rates. *arXiv preprint arXiv:1803.08586*, 2018.

[47] Y. Wang, S. S. Du, S. Balakrishnan, and A. 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.

[48] P. Whittle. *Optimal control: Basics and Beyond*. Wiley and Sons, Chichester, England, 1996.