On the Linear Convergence of Random Search for Discrete-Time LQR

Hesameddin Mohammadi, *Graduate Student Member, IEEE*, Mahdi Soltanolkotabi, *Member, IEEE*, and Mihailo R. Jovanović, *Fellow, IEEE*

***Abstract*—Model-free reinforcement learning techniques directly search over the parameter space of controllers. Although this often amounts to solving a nonconvex optimization problem, for benchmark control problems simple local search methods exhibit competitive performance. To understand this phenomenon, we study the discrete-time Linear Quadratic Regulator (LQR) problem with unknown state-space parameters. In spite of the lack of convexity, we establish that the random search method with two-point gradient estimates and a fixed number of roll-outs achieves ϵ-accuracy in $O(\log(1/\epsilon))$ iterations. This significantly improves existing results on the model-free LQR problem which require $O(1/\epsilon)$ total roll-outs.**

***Index Terms*—Data-driven control, linear quadratic regulator, model-free control, nonconvex optimization, random search method, reinforcement learning, sample complexity.**

I. Introduction

WE STUDY the sample complexity and convergence of the random search method for the infinite-horizon discrete-time LQR problem. Random search is a derivative-free optimization algorithm that directly searches over the parameter space of controllers using approximations of the gradient obtained through simulation data. Despite its simplicity, this approach has been used to solve benchmark control problems with state-of-the-art sample efficiency [1], [2]. However, even for the standard LQR problem, many open theoretical questions surround convergence properties and sample complexity of this method mainly because of the lack of convexity.

For *discrete-time* LQR problem, global convergence guarantees were recently provided for gradient descent and the random search method with one-point gradient estimates [3]. The key observation was that the LQR cost satisfies the Polyak-Łojasiewicz (PL) condition which can ensure convergence of gradient descent at a linear rate even for nonconvex problems. This reference also established a bound on the sample complexity of random search for reaching the error tolerance ϵ that requires a number of function evaluations that is proportional to $(1/\epsilon^4)\log(1/\epsilon)$. Extensions to the *continuous-time* LQR [4], [5], the $\mathcal{H}_\infty$ regularized LQR [6], and Markovian jump linear systems [7] have also been made.

Assuming access to the infinite horizon cost, the number of function evaluations for the random search method with one-point estimates was improved to $1/\epsilon^2$ in [8]. Moreover, this reference showed that the use of two-point estimates reduces the number of function evaluations to $1/\epsilon$. Apart from the PL property, these results do not exploit structure of the LQR problem. Our recent work [9] focused on the *continuous-time* LQR problem, and established that the random search method with two-point gradient estimates converges to the optimal solution at a linear rate with high probability. In this letter, we extend the results of [9] to the *discrete-time* case. Relative to the existing literature, our results offer a significant improvement both in terms of the required number of function evaluations and simulation time. Specifically, the total number of function evaluations to achieve an ϵ-accuracy is proportional to $\log(1/\epsilon)$ compared to at least $(1/\epsilon^4)\log(1/\epsilon)$ in [3] and $1/\epsilon$ in [8]. Similarly, the required simulation time is proportional to $\log(1/\epsilon)$; this is in contrast to [3] which requires $\text{poly}(1/\epsilon)$ simulation time.

II. State-Feedback Characterization

Consider the LTI system

$$x^{t+1} = Ax^t + Bu^t, \quad x^0 = \zeta \tag{1a}$$

where $x^t \in \mathbb{R}^n$ is the state, $u^t \in \mathbb{R}^m$ is the control input, A and B are constant matrices, and $x^0 = \zeta$ is a zero-mean random initial condition with distribution $\mathcal{D}$. The LQR problem associated with system (1a) is given by

$$\underset{x,u}{\text{minimize}} \ \mathbb{E}\left[\sum_{t=0}^{\infty}(x^t)^T Q x^t + (u^t)^T R u^t\right] \tag{1b}$$

where Q and R are positive definite matrices and the expectation is taken over $\zeta \sim \mathcal{D}$. For a controllable pair (A, B), the

Manuscript received March 17, 2020; revised May 26, 2020; accepted June 15, 2020. Date of publication July 1, 2020; date of current version July 23, 2020. The work of Hesameddin Mohammadi and Mihailo R. Jovanović was supported in part by the National Science Foundation (NSF) under Award ECCS-1708906 and Award ECCS-1809833. The work of Mahdi Soltanolkotabi was supported in part by the Packard Fellowship in Science and Engineering, in part by the Sloan Research Fellowship in Mathematics, in part by the Google Faculty Research Award, as well as Awards from NSF, and in part by AFOSR Young Investigator Program. Recommended by Senior Editor G. Cherubini. *(Corresponding author: Hesameddin Mohammadi.)*

The authors are with the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: hesamedm@usc.edu; soltanol@usc.edu; mihailo@usc.edu).

Digital Object Identifier 10.1109/LCSYS.2020.3006256

solution to (1) takes a state-feedback form,

$$u^t = -K^\star x^t = -(R + B^T P^\star B)^{-1} B^T P^\star A x^t$$

where $P^\star$ is the unique positive definite solution to the Algebraic Riccati Equation (ARE),

$$A^T P^\star A + Q - A^T P^\star B (R + B^T P^\star B)^{-1} B^T P^\star A = P^\star.$$

When the model parameters A and B are known, the ARE can be solved efficiently via a variety of techniques [10], [11]. However, these techniques are not directly applicable when the matrices A and B are not known. One approach to dealing with the model-free scenario is to use the linearity of the optimal controller and reformulate the LQR problem as an optimization over state-feedback gains,

$$\underset{K}{\text{minimize}}\ f(K) \ := \ \mathbb{E}\big[f_\zeta(K)\big] \tag{2}$$

where $f_\zeta(K) := \big\langle Q + K^T R K, X_\zeta(K)\big\rangle = \zeta^T P(K) \zeta$ and the matrices $P(K)$ and $X_\zeta(K)$ are given by

$$\begin{aligned} P(K) &:= \sum_{t=0}^{\infty} ((A - BK)^T)^t (Q + K^T R K)(A - BK)^t \\ X_\zeta(K) &:= \sum_{t=0}^{\infty} (A - BK)^t \zeta \zeta^T ((A - BK)^T)^t. \end{aligned} \tag{3}$$

Here, $f_\zeta(K)$ determines the LQR cost in (1b) associated with the feedback law $u = -Kx$ and the initial condition $x^0 = \zeta$. A necessary and sufficient condition for the boundedness of $f_\zeta(K)$ for all $\zeta \in \mathbb{R}^n$ is closed-loop stability,

$$K \in \mathcal{S} := \{K \in \mathbb{R}^{m \times n} | \rho(A - BK) < 1\} \tag{4}$$

where $\rho(\cdot)$ is the spectral radius.

For any $K \in \mathcal{S}$, the matrices $P(K)$ and $X_\zeta(K)$ are well-defined and are, respectively, determined by the unique solutions to the Lyapunov equations

$$\mathcal{A}_K^*(P) = -Q - K^T R K, \quad \mathcal{A}_K(X_\zeta) = -\zeta \zeta^T. \tag{5}$$

Here, $\mathcal{A}_K, \mathcal{A}_K^* : \mathbb{S}_n \to \mathbb{S}_n$

$$\mathcal{A}_K(X) = (A - BK) X (A - BK)^T - X \tag{6a}$$

$$\mathcal{A}_K^*(P) = (A - BK)^T P (A - BK) - P \tag{6b}$$

determine the adjoint pair of invertible closed-loop Lyapunov operators acting on the set of symmetric matrices $\mathbb{S}_n \subset \mathbb{R}^{n \times n}$.

The invertibility of $\mathcal{A}_K$ and $\mathcal{A}_K^*$ for $K \in \mathcal{S}$ allows us to express the LQR objective function in (2) as

$$f(K) = \begin{cases} \big\langle Q + K^T R K, X(K)\big\rangle = \langle \Omega, P(K) \rangle, & K \in \mathcal{S} \\ \infty, & \text{otherwise} \end{cases}$$

where

$$X(K) := \mathbb{E}\big[X_\zeta(K)\big] = -\mathcal{A}_K^{-1}(\Omega) \tag{7}$$

and $\Omega := \mathbb{E}[\zeta \zeta^T]$ is the covariance matrix of the initial condition. We assume $\Omega \succ 0$ to ensure that the random vector $\zeta \sim \mathcal{D}$ has energy in all directions. This condition guarantees $f(K) = \infty$ for all $K \notin \mathcal{S}$. Finally, it is well known that for any $K \in \mathcal{S}$, the cone of positive definite matrices is closed under the action of $-\mathcal{A}_K^{-1}$ and $-(\mathcal{A}_K^*)^{-1}$. Thus, from the positive definiteness of the matrices $Q + K^T R K$ and Ω, it follows that $P(K), X(K) \succ 0$ for all $K \in \mathcal{S}$. In (2), K is the optimization variable, and (A, B, $Q \succ 0$, $R \succ 0$, $\Omega \succ 0$) are the problem parameters.

For any feedback gain $K \in \mathcal{S}$, it can be shown that [12]

$$\nabla f_\zeta(K) = E(K) X_\zeta(K), \quad \nabla f(K) = E(K) X(K) \tag{8a}$$

where

$$E(K) := 2\big((R + B^T P(K) B) K - B^T P(K) A\big) \tag{8b}$$

is a fixed matrix that does not depend on the random initial condition ζ. Thus, the randomness of the gradient $\nabla f_\zeta(K)$ arises from the random matrix $X_\zeta(K)$.

Remark 1: The LQR problem for continuous-time systems can be treated in a similar way. In this case, although the Lyapunov operator $\mathcal{A}_K$ has a different definition, the form of the objective function in terms of the matrices $X(K)$ and $P(K)$ and also the form of the gradient in terms of $X(K)$ and $E(K)$ remain unchanged. While this similarity allows for our results to hold for both continuous and discrete-time systems, in this letter we only focus on the latter and refer to [9] for a treatment of continuous-time systems.

III. Random Search

The formulation of the LQR problem given by (2) has been studied for both continuous-time [4], [13] and discrete-time systems [3], [14]. In this letter, we analyze the sample complexity and convergence properties of the random search method for solving problem (2) with unknown model parameters. At each iteration $k \in \mathbb{N}$, the random search method calls Algorithm 1 that forms an empirical approximation $\overline{\nabla} f(K^k)$ to the gradient of the objective function via finite-time simulation of system (1a) for randomly perturbed feedback gains $K^k \pm U_i$, $i = 1, \ldots, N$.

Algorithm 1 does not require knowledge of matrices A and B but only access to a *two-point* simulation engine. The two-point setting means that for any pair of points K and K', the simulation engine can return the random values $f_{\zeta,\tau}(K)$ and $f_{\zeta,\tau}(K')$ for some random initial condition $x^0 = \zeta$, where

$$f_{\zeta,\tau}(K) := \sum_{t=0}^{\tau} (x^t)^T Q x^t + (u^t)^T R u^t \tag{9}$$

is a finite-time random function approximation associated with system (1a), starting from a random initial condition $x^0 = \zeta$, with the state feedback $u = -Kx$ running up to time τ. This is in contrast to the *one-point* setting in which, at each query, the simulation engine can receive only one specified point K and return the random value $f_{\zeta,\tau}(K)$.

Starting from an initial feedback gain $K^0 \in \mathcal{S}$, the random search method uses the gradient estimates obtained via Algorithm 1 to update the iterates according to

$$K^{k+1} := K^k - \alpha \overline{\nabla} f(K^k), \quad K^0 \in \mathcal{S} \tag{RS}$$

for some stepsize $\alpha > 0$. The stabilizing assumption on the initial iterate $K^0 \in \mathcal{S}$ is required in our analysis as we select the input parameters of Algorithm 1 and the stepsize so that all iterates satisfy $K^k \in \mathcal{S}$.

For convex problems, the gradient estimates obtained in the two-point setting are known to yield faster convergence rates than the one-point setting [15]. However, the two-point setting

Algorithm 1 Gradient Estimation

Input: Feedback gain $K \in \mathbb{R}^{m\times n}$, state and control weight matrices Q and R, distribution $\mathcal{D}$, smoothing constant r, simulation time τ, number of random samples N.

for $i = 1$ to N **do**

– Define two perturbed feedback gains $K_{i,1} := K + rU_i$ and $K_{i,2} := K - rU_i$, where $\mathrm{vec}(U_i)$ is a random vector uniformly distributed on the sphere $\sqrt{mn}\,S^{mn-1}$.

– Sample an initial condition ζ^i from distribution $\mathcal{D}$.

– For $j \in \{1, 2\}$, simulate system (1a) up to time τ with the feedback gain $K_{i,j}$ and initial condition ζ_i to form $f_{\zeta^i,\tau}(K_{i,j})$ as in Eq. (9).

end for

Output: The two-point gradient estimate

$$\overline{\nabla} f(K) := \frac{1}{2rN}\sum_{i=1}^{N}\big(f_{\zeta^i,\tau}(K_{i,1}) - f_{\zeta^i,\tau}(K_{i,2})\big)U_i.$$

requires simulations of the system for two different feedback gain matrices under the same initial condition.

IV. Main Result

We analyze the sample complexity and convergence of the random search method (RS) for the model-free setting. Our main convergence result exploits two key properties of the LQR objective function f, namely smoothness and the Polyak-Łojasiewicz (PL) condition over its sublevel sets $\mathcal{S}(a) := \{K \in \mathcal{S} \mid f(K) \le a\}$, where a is a positive scalar. In particular, it can be shown that, restricted to any sublevel set $\mathcal{S}(a)$, the function f is $L_f(a)$-smooth and satisfies the PL condition with parameter $\mu_f(a)$, i.e.,

$$f(K') - f(K) \le \langle \nabla f(K), K' - K\rangle + \frac{L_f(a)}{2}\|K - K'\|_F^2$$

$$f(K) - f(K^\star) \le \frac{1}{2\mu_f(a)}\|\nabla f(K)\|_F^2$$

for all K and K' such that the line segment between them belongs to $\mathcal{S}(a)$, where $L_f(a)$ and $\mu_f(a)$ are positive rational functions of a. This result has been established for both continuous-time [4] and discrete-time [3], [14] LQR problems. We also make the following assumption on the statistical properties of the initial condition.

Assumption 1 (Initial Distribution): Let the distribution $\mathcal{D}$ of the initial condition have i.i.d. zero-mean unit-variance entries with bounded sub-Gaussian norm. For a random vector $\zeta \in \mathbb{R}^n$ distributed according to $\mathcal{D}$, this implies $\mathbb{E}[\zeta] = 0$, $\mathbb{E}[\zeta\zeta^T] = I$, and $\|\zeta_i\|_{\psi_2} \le \kappa$, for some constant κ and $i = 1, \ldots, n$, where $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm [16].

We now state our main theoretical result.

Theorem 1: Consider the random search method (RS) that uses the gradient estimates of Algorithm 1 for finding the optimal solution $K^\star$ of problem (2). Let the initial condition $x^0 \sim \mathcal{D}$ obey Assumption 1 and let the simulation time τ and the number of samples N in Algorithm 1 satisfy

$$\tau \ge \theta'(a)\log(1/\epsilon), \quad N \ge c\Big(1 + \beta^4\kappa^4\theta(a)\log^6 n\Big)n,$$

for some $\beta > 0$ and a desired accuracy $\epsilon > 0$. Then, we can choose a smoothing parameter $r < \theta''(a)\sqrt{\epsilon}$ in Algorithm 1 such that, for any initial condition $K^0 \in \mathcal{S}(a)$, method (RS) with the constant stepsize $\alpha = 1/(\omega(a)L_f(a))$ achieves $f(K^k) - f(K^\star) \le \epsilon$ in at most

$$k \le -\log\Big(\epsilon^{-1}\,(f(K^0) - f(K^\star))\Big)/\log\big(1 - \mu_f(a)\alpha/8\big)$$

iterations. This holds with probability not smaller than

$$1 - c'k(n^{-\beta} + N^{-\beta} + N\mathrm{e}^{-\frac{n}{8}} + \mathrm{e}^{-c'N}).$$

Here, $\omega(a) := c''(\sqrt{m} + \beta\kappa^2\theta(a)\sqrt{mn}\log n)^2$, the positive scalars c, c', and c'' are absolute constants, $\mu_f(a)$ and $L_f(a)$ are the PL and smoothness parameters of f over the sublevel set $\mathcal{S}(a)$, and θ, θ', and θ'' are positive polynomials that depend only on the parameters of the LQR problem.

For a desired accuracy level $\epsilon > 0$, Theorem 1 shows that the random search iterates (RS) with constant stepsize (that does not depend on ϵ) reach an accuracy level ϵ at a linear rate (i.e., in at most $O(\log(1/\epsilon))$ iterations) with high probability. Furthermore, the total number of function evaluations and the simulation time required to achieve an accuracy level ϵ are proportional to $\log(1/\epsilon)$. As stated earlier, this significantly improves the existing results for discrete-time LQR [3], [8] that require $O(1/\epsilon)$ function evaluations and poly$(1/\epsilon)$ simulation time.

V. Proof Sketch

In this section, we present our proof strategy for the main result of this letter. The proof of technical results are omitted due to page limitations. The smoothness of the objective function along with the PL condition are sufficient for the gradient descent method with a suitable stepsize α,

$$K^{k+1} := K^k - \alpha\nabla f(K^k), \quad K^0 \in \mathcal{S} \qquad \text{(GD)}$$

to achieve linear convergence even for nonconvex problems [17]. These properties were recently used to show convergence of gradient descent for both discrete-time [3] and continuous-time [4] LQR problems. In the model-free setting, the gradient descent method is not directly implementable because computing the gradient $\nabla f(K)$ requires knowledge of system parameters A and B. The random search method (RS) resolves this issue by using the *gradient estimate* $\overline{\nabla} f(K)$ obtained via Algorithm 1. One approach to the convergence analysis of random search is to first use a large number of samples N in order to make the estimation error small, and then relate the iterates of (RS) to that of gradient descent. It has been shown that achieving $\|\overline{\nabla} f(K) - \nabla f(K)\|_F \le \epsilon$ takes $N = O(1/\epsilon^4)$ samples [3]; see also [5, Th. 3] for the continuous-time LQR. This upper bound unfortunately leads to a sample complexity bound that grows polynomially with $1/\epsilon$. To improve this result, we take an alternative route and give up on the objective of controlling the gradient estimation error. In particular, by exploiting the problem structure, we show that with a fixed number of samples $N = \tilde{O}(n)$, where n denotes the number of states, the estimate $\overline{\nabla} f(K)$ concentrates with *high probability* when projected to the direction of $\nabla f(K)$.

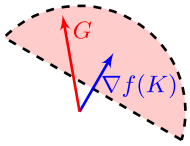

Fig. 1. The intersection of the half-space and the ball parameterized by μ_1 and μ_2, respectively, in Proposition 1. If an update direction G lies within this region, then taking one step along $-G$ with a constant stepsize α yields a geometric decrease in the objective value.

In what follows, we first establish that for any $\epsilon > 0$, using a simulation time $\tau = O(\log(1/\epsilon))$ and an appropriate smoothing parameter r in Algorithm 1, the estimate $\overline{\nabla} f(K)$ can be made ϵ-close to an unbiased estimate $\widehat{\nabla} f(K)$ of the gradient with high probability, $\|\overline{\nabla} f(K) - \widehat{\nabla} f(K)\|_F \le \epsilon$, where the definition of $\widehat{\nabla} f(K)$ is given in Eq. (11). We call this distance the *estimation bias*. We then show that, for a large number of samples N, our unbiased estimate $\widehat{\nabla} f(K)$ becomes highly correlated with the gradient. In particular, we establish that the following two events

$$\mathsf{M}_1 := \left\{ \left\langle \widehat{\nabla} f(K), \nabla f(K) \right\rangle \ge \mu_1 \|\nabla f(K)\|_F^2 \right\} \tag{10a}$$

$$\mathsf{M}_2 := \left\{ \|\widehat{\nabla} f(K)\|_F^2 \le \mu_2 \|\nabla f(K)\|_F^2 \right\} \tag{10b}$$

occur with high probability for some positive scalars μ_1 and μ_2. To justify the definition of these events, let us first demonstrate that the gradient estimate $\widehat{\nabla} f(K)$ can be used to decrease the objective error by a geometric factor if both M_1 and M_2 occur.

Proposition 1: If $G \in \mathbb{R}^{m\times n}$ and $K \in \mathcal{S}(a)$ are such that $\langle G, \nabla f(K)\rangle \ge \mu_1 \|\nabla f(K)\|_F^2$ and $\|G\|_F^2 \le \mu_2 \|\nabla f(K)\|_F^2$ for some scalars $\mu_1, \mu_2 > 0$, then $K - \alpha G \in \mathcal{S}(a)$ for all $\alpha \in [0, \mu_1/(\mu_2 L_f(a))]$, and $f(K - \alpha G) - f(K^\star) \le (1 - \mu_f(a)\mu_1\alpha)(f(K) - f(K^\star))$, where $L_f(a)$ and $\mu_f(a)$ are the smoothness and PL parameters of f over $\mathcal{S}(a)$.

Proposition 1 demonstrates that, conditioned on the events M_1 and M_2, the unbiased estimate $\widehat{\nabla} f(K)$ yields a simple descent-based algorithm that has linear convergence. Fig. 1 illustrates the region parameterized by μ_1 and μ_2 in Proposition 1. This region has a different geometry than ϵ-neighborhoods of the gradient. A gradient estimate G can have an accuracy of $O(\nabla f(K))$ and still belong to this region. We leverage this fact in our convergence analysis which only requires the gradient estimate $\widehat{\nabla} f(K)$ to be in such a region for certain parameters μ_1 and μ_2 and not necessarily within an ϵ-neighborhood of the gradient.

A. Controlling the Bias

Herein, we define the unbiased estimate $\widehat{\nabla} f(K)$ of the gradient and establish an upper bound on its distance to the output $\overline{\nabla} f(K)$ of Algorithm 1

$$\overline{\nabla} f(K) := \frac{1}{2rN} \sum_{i=1}^{N} \left(f_{\zeta^i,\tau}(K + rU_i) - f_{\zeta^i,\tau}(K - rU_i) \right) U_i$$

$$\widetilde{\nabla} f(K) := \frac{1}{2rN} \sum_{i=1}^{N} \left(f_{\zeta^i}(K + rU_i) - f_{\zeta^i}(K - rU_i) \right) U_i$$

$$\widehat{\nabla} f(K) := \frac{1}{N} \sum_{i=1}^{N} \left\langle \nabla f_{\zeta^i}(K), U_i \right\rangle U_i. \tag{11}$$

Here, $U_i \in \mathbb{R}^{m\times n}$ are i.i.d. random matrices whose vectorized form $\mathrm{vec}(U_i)$ are uniformly distributed on the sphere $\sqrt{mn}\, S^{mn-1}$ and $\zeta^i \in \mathbb{R}^n$ are i.i.d. random initial conditions sampled from distribution $\mathcal{D}$. Note that $\widetilde{\nabla} f(K)$ is the infinite horizon version of $\overline{\nabla} f(K)$ and $\widehat{\nabla} f(K)$ is an unbiased estimate of $\nabla f(K)$. The fact that $\mathbb{E}[\widehat{\nabla} f(K)] = \nabla f(K)$ follows from

$$\begin{aligned} \mathbb{E}_{\zeta^i, U_i}\left[\mathrm{vec}(\widehat{\nabla} f(K))\right] &= \mathbb{E}_{U_1}\left[\langle \nabla f(K), U_1\rangle \mathrm{vec}(U_1)\right] \\ &= \mathbb{E}_{U_1}\left[\mathrm{vec}(U_1)\mathrm{vec}(U_1)^T\right]\mathrm{vec}(\nabla f(K)) = \mathrm{vec}(\nabla f(K)). \end{aligned}$$

Local boundedness of the function $f(K)$: An important requirement for the gradient estimation scheme in Algorithm 1 is the stability of the perturbed closed-loop systems, i.e., $K \pm rU_i \in \mathcal{S}$; violating this condition leads to an exponential growth of the state and control signals. Moreover, this condition is necessary and sufficient for $\widetilde{\nabla} f(K)$ to be well defined. It can be shown that for any sublevel set $\mathcal{S}(a)$, there exists a positive radius r such that $K + rU \in \mathcal{S}$ for all $K \in \mathcal{S}(a)$ and $U \in \mathbb{R}^{m\times n}$ with $\|U\|_F \le \sqrt{mn}$. In this letter, we further require that r is small enough so that $K \pm rU_i \in \mathcal{S}(2a)$ for all $K \in \mathcal{S}(a)$. Such upper bound on r can be provided using the upper bound on the cost difference established in [3, Lemma 24]. A similar result has been established for the continuous-time LQR problem using the small-gain theorem and the KYP lemma [9].

Lemma 1: For any $K \in \mathcal{S}(a)$ and $U \in \mathbb{R}^{m\times n}$ with $\|U\|_F \le \sqrt{mn}$, $K + r(a)U \in \mathcal{S}(2a)$, where $r(a) := \tilde{c}/a$ for some constant $\tilde{c} > 0$ that depends on the problem data.

Note that for any $K \in \mathcal{S}(a)$ and $r \le r(a)$ in Lemma 1, $\widetilde{\nabla} f(K)$ is well defined since the feedback gains $K \pm rU_i$ are all stabilizing. We next establish an upper bound on the difference between the output $\overline{\nabla} f(K)$ of Algorithm 1 and the unbiased estimate $\widehat{\nabla} f(K)$ of the gradient $\nabla f(K)$. We accomplish this by bounding the difference between these two quantities and $\widetilde{\nabla} f(K)$ using the triangle inequality

$$\begin{aligned} \|\widehat{\nabla} f(K) - \overline{\nabla} f(K)\|_F \le \|\widetilde{\nabla} f(K) - \overline{\nabla} f(K)\|_F \\ + \|\widehat{\nabla} f(K) - \widetilde{\nabla} f(K)\|_F. \end{aligned} \tag{12}$$

Proposition 2 provides an upper bound on each term on the right-hand side of the above inequality.

Proposition 2: For any $K \in \mathcal{S}(a)$ and $r \le r(a)$, where $r(a)$ is given by Lemma 1,

$$\|\widetilde{\nabla} f(K) - \overline{\nabla} f(K)\|_F \le \frac{\sqrt{mn}\eta}{r} \kappa_1(2a)(1 - \kappa_2(2a))^\tau$$

$$\|\widehat{\nabla} f(K) - \widetilde{\nabla} f(K)\|_F \le \frac{(rmn)^2 \eta}{2} \ell(2a)$$

where $\eta := \max_i \|\zeta^i\|^2$, and $\ell(a) > 0$, $\kappa_1(a) > 0$, and $1 > \kappa_2(a) > 0$ are rational functions that depend on the problem data.

The first term on the right-hand side of (12) corresponds to a bias arising from the finite-time simulation. Proposition 2 shows that although small values of r may result in a large $\|\widetilde{\nabla} f(K) - \overline{\nabla} f(K)\|_F$, because of the exponential dependence of the upper bound on the simulation time τ, this error can be controlled by increasing τ. In addition, since $\widehat{\nabla} f(K)$ is independent of the parameter r, this result provides a quadratic

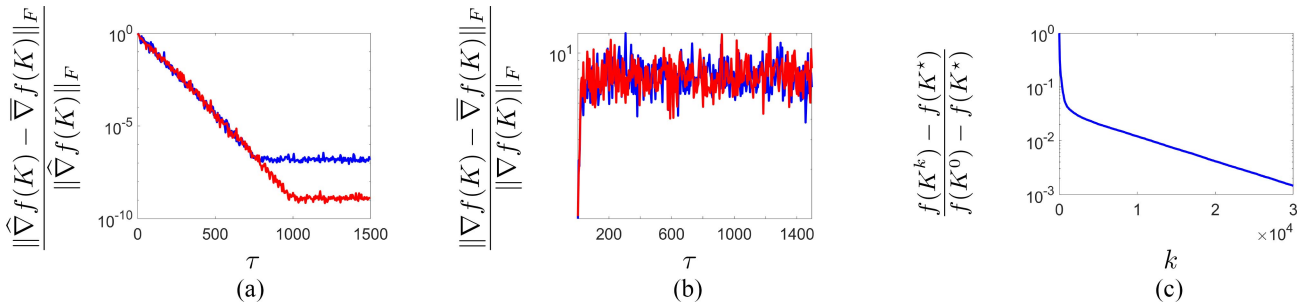

Fig. 2. (a) Bias in gradient estimation; (b) total error in gradient estimation as functions of the simulation time τ. The blue and red curves correspond to two values of the smoothing parameter $r = 10^{-4}$ and $r = 10^{-6}$, respectively. (c) Convergence curve of the random search method (RS).

bound on the estimation error in terms of r. It is also worth mentioning that the third derivative of the function $f_\zeta(K)$ is utilized in obtaining the second inequality.

B. Correlation of $\widehat{\nabla} f(K)$ and $\nabla f(K)$

We establish that under Assumption 1 on the initial distribution, with large enough number of samples $N = \tilde{O}(n)$, the events M_1 and M_2 with $\mu_1 := 1/4$ and

$$\mu_2 := Cm\left(\beta\kappa^2 \frac{\|(\mathcal{A}_K^*)^{-1}\|_2 + \|(\mathcal{A}_K^*)^{-1}\|_S}{\lambda_{\min}(X(K))}\sqrt{n}\log n + 1\right)^2 \quad (13)$$

occur with high probability, where κ is an upper bound on the ψ_2-norm of the entries of ζ^i, $\beta > 0$ is a parameter that determines the failure probability, C is a positive absolute constant, and for an operator $\mathcal{M}$,

$$\|\mathcal{M}\|_2 := \sup_M \frac{\|\mathcal{M}(M)\|_F}{\|M\|_F}, \quad \|\mathcal{M}\|_S := \sup_M \frac{\|\mathcal{M}(M)\|_2}{\|M\|_2}.$$

We note that these parameters do not depend on the desired accuracy-level ϵ. Moreover, since the sub-level sets of the function $f(K)$ are compact [14], $\|(\mathcal{A}_K^*)^{-1}\|$ is a continuous function of K, and $X(K) \succeq \Omega$, we can uniformly upper bound μ_2 over any sublevel set $\mathcal{S}(a)$. Such bound has also been discussed and analytically quantified for the continuous-time LQR problem [9].

Our approach to accomplishing the above task exploits the problem structure, which allows for confining the dependence of $\widehat{\nabla} f(K)$ on the random initial conditions ζ^i into the zero-mean random matrices $X_{\zeta^i} - X$, where $X_{\zeta^i} := X_{\zeta^i}(K)$ and $X := X(K)$ are given by (3) and (7), respectively. In particular, for any given feedback gain $K \in \mathcal{S}$, we can use the form of gradient (8) to write

$$\widehat{\nabla} f(K) = \frac{1}{N}\sum_{i=1}^{N} \langle E X_{\zeta^i}, U_i\rangle U_i = \widehat{\nabla}_1 + \widehat{\nabla}_2$$

where $\widehat{\nabla}_1 := (1/N)\sum_{i=1}^N \langle E(X_{\zeta^i} - X), U_i\rangle U_i$, $\widehat{\nabla}_2 := (1/N)\sum_{i=1}^N \langle \nabla f(K), U_i\rangle U_i$, and the matrix $E := E(K)$ is given by (8b). It is now easy to verify that $\mathbb{E}[\widehat{\nabla}_1] = 0$ and $\mathbb{E}[\widehat{\nabla}_2] = \nabla f(K)$. Furthermore, only the term $\widehat{\nabla}_1$ depends on the initial conditions ζ^i.

1) Quantifying the Probability of M_1: We exploit results from modern high-dimensional statistics on the non-asymptotic analysis of the concentration of random quantities around their mean [16]. Our approach to analyzing the event M_1 consists of two steps. First, we establish that the zero-mean random variable $\langle \widehat{\nabla}_1, \nabla f(K)\rangle$ *highly* concentrates around zero with a large enough number of samples $N = \tilde{O}(n)$. Our proof technique relies on the Hanson-Wright inequality [18, Th. 1.1]. Next, we study the concentration of the random variable $\langle \widehat{\nabla}_2, \nabla f(K)\rangle$ around its mean $\|\nabla f(K)\|_F^2$. The key enabler here is the Bernstein inequality [16, Corollary 2.8.3]. This leads to the next proposition.

Proposition 3: Under Assumption 1, for any stabilizing feedback gain $K \in \mathcal{S}$ and positive scalar β, if

$$N \geq C_1 \frac{\beta^4\kappa^4}{\lambda_{\min}^2(X)}\left(\|(\mathcal{A}_K^*)^{-1}\|_2 + \|(\mathcal{A}_K^*)^{-1}\|_S\right)^2 n\log^6 n$$

then the event M_1 in (10) with $\mu_1 := 1/4$ satisfies

$$\mathbb{P}(\mathsf{M}_1) \geq 1 - C_2 N^{-\beta} - 4N\mathrm{e}^{-\frac{n}{8}} - 2\mathrm{e}^{-C_3 N}.$$

2) Quantifying the Probability of M_2: Similarly, we analyze the event M_2 in two steps. We establish upper bounds on the ratio $\|\widehat{\nabla}_i\|_F/\|\nabla f(K)\|_F$, for $i = \{1, 2\}$, that hold with high probability, and use the triangle inequality

$$\frac{\|\widehat{\nabla}_1\|_F}{\|\nabla f(K)\|_F} + \frac{\|\widehat{\nabla}_2\|_F}{\|\nabla f(K)\|_F} \geq \frac{\|\widehat{\nabla} f(K)\|_F}{\|\nabla f(K)\|_F}.$$

Our results are summarized in the next proposition.

Proposition 4: Under Assumption 1, for any $K \in \mathcal{S}$, scalar $\beta > 0$, and $N \geq C_4 n$, the event M_2 in (10) with μ_2 given by (13) satisfies $\mathbb{P}(\mathsf{M}_2) \geq 1 - C_6(n^{-\beta} + N\mathrm{e}^{-\frac{n}{8}} + \mathrm{e}^{-C_7 N})$.

VI. Computational Experiments

We consider a system with $s = 10$ inverted pendula on force-controlled carts that are connected by springs and dampers; see Fig. 3. We set all masses, pendula lengths, spring and damping constants to unity and let the state vector $x := [\theta^T \omega^T p^T v^T]^T$ contain the angle and angular velocity of pendula as well as position and velocity of masses. Linearizing around the equilibrium point yields the continuous-time system $\dot{x} = A_c x + B_c u$, where

$$A_c = \begin{bmatrix} 0 & I & 0 & 0 \\ 20I & 0 & T & T \\ 0 & 0 & 0 & I \\ -10I & 0 & -T & -T \end{bmatrix}, \quad B_c = \begin{bmatrix} 0 \\ -I \\ 0 \\ I \end{bmatrix}.$$

Here, 0 and I are $s \times s$ zero and identity matrices, and T is a Toeplitz matrix with 2 on the main diagonal, -1 on the

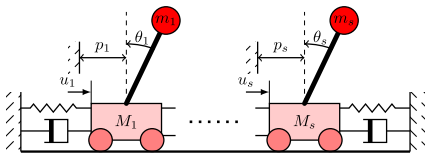

Fig. 3. An interconnected system of inverted pendula on carts.

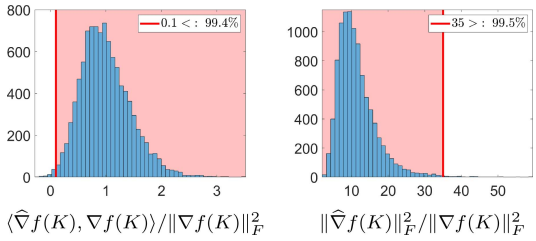

Fig. 4. Histograms of two algorithmic quantities associated with the events M_1 and M_2 given by (10). The red lines demonstrate that M_1 with $\mu_1 = 0.1$ and M_2 with $\mu_2 = 35$ occur in more than 99% of trials.

first upper and lower sub-diagonals, and zero elsewhere. We discretize this system with sampling time $t_s = 0.1$, which yields Eq. (1a) with $A = \mathrm{e}^{A_c t_s}$ and $B = \int_0^{t_s} \mathrm{e}^{A_c t} B_c \mathrm{d}t$. Since the open-loop system is unstable, we use a stabilizing feedback gain $K^0 = [-50I \ -10I \ -5I \ -5I]$ as a starting point for the random search method and choose $Q = \mathrm{blkdiag}(10I, I, I, I)$ and $R = I$ in the LQR cost. We also let the initial conditions ζ^i in Algorithm 1 be standard normal and use $N = n = 2s$ samples.

Figure 2(a) illustrates the dependence of the relative error $\|\widehat{\nabla}f(K) - \overline{\nabla}f(K)\|_F/\|\widehat{\nabla}f(K)\|_F$ on the simulation time τ for $K = K^0 = [-50I \ -10I \ -5I \ -5I]$ and two values of smoothing parameter $r = 10^{-4}$ (blue) and $r = 10^{-6}$ (red). We see an exponential decrease in error for small values of τ and note that the error does not pass a saturation level determined by the smoothing parameter $r > 0$. We also observe that as r decreases, this saturation level becomes smaller. These observations are in harmony with the results established in Proposition 2. This should be compared and contrasted with Fig. 2(b), which demonstrates that the relative error with respect to the true gradient does not vanish with increase in the simulation time τ.

In spite of this significant error, the key observation that allows us to establish the linear convergence of the random search method in Theorem 1 is that the gradient estimate has a high correlation with the true gradient. Figure 4 shows histograms of two algorithmic quantities associated with the events M_1 and M_2 given by (10). The red lines demonstrate that M_1 with $\mu_1 = 0.1$ and M_2 with $\mu_2 = 35$ occur in more than 99% of trials; see Propositions 3 and 4.

Figure 2(c) illustrates the convergence curve of the random search method (RS) with stepsize $\alpha = 10^{-5}$, $r = 10^{-5}$, and $\tau = 1000$ in Algorithm 1. This figure confirms linear convergence of (RS) established in Theorem 1.

VII. Concluding Remarks

In this letter, we studied the convergence and sample complexity of the random search method with two-point gradient estimates for the discrete-time LQR problem. Despite nonconvexity, we established that the random search method with a fixed number of roll-outs $N = \tilde{O}(n)$ per iteration achieves ϵ-accuracy in $O(\log(1/\epsilon))$ iterations. This significantly improves existing results on the model-free LQR which require $O(1/\epsilon)$ total roll-outs. Our ongoing research directions include: (i) providing theoretical guarantees for the convergence of gradient-based methods for sparsity-promoting and structured control synthesis [19]; and (ii) extension to nonlinear systems via successive linearization techniques.

References

[1] H. Mania, A. Guy, and B. Recht, "Simple random search provides a competitive approach to reinforcement learning," 2018. [Online]. Available: arXiv:1803.07055.
[2] B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 2, pp. 253–279, May 2019.
[3] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1467–1476.
[4] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, "Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator," in *Proc. 58th IEEE Conf. Decis. Control*, 2019, pp. 7474–7479.
[5] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović, "Random search for learning the linear quadratic regulator," in *Proc. Amer. Control Conf.*, 2020, pp. 4798–4803.
[6] K. Zhang, B. Hu, and T. Başar, "Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence," 2018. [Online]. Available: arXiv:1910.09496.
[7] J. P. Jansch-Porto, B. Hu, and G. E. Dullerud, "Convergence guarantees of policy optimization methods for Markovian jump linear systems," 2020. [Online]. Available: arXiv:2002.04090.
[8] D. Malik, A. Panajady, K. Bhatia, K. Khamaru, P. L. Bartlett, and M. J. Wainwright, "Derivative-free methods for policy optimization: Guarantees for linear-quadratic systems," in *Proc. AISTATS: Conf. AI Stat.*, 2019, pp. 2916–2925.
[9] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, "Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem," 2019. [Online]. Available: arXiv:1912.11899.
[10] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator," *IEEE Trans. Autom. Control*, vol. AC-16, no. 4, pp. 382–384, Aug. 1971.
[11] S. Bittanti, A. J. Laub, and J. C. Willems, *The Riccati Equation*. Berlin, Germany: Springer-Verlag, 2012.
[12] K. Mårtensson, "Gradient methods for large-scale and distributed linear quadratic control," Ph.D. dissertation, Dept. Autom. Control, Lund Inst. Technol., Lund Univ., Lund, Sweden, 2012.
[13] B. Anderson and J. Moore, *Optimal Control; Linear Quadratic Methods*. New York, NY, USA: Prentice-Hall, 1990.
[14] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, "LQR through the lens of first order methods: Discrete-time case," 2019. [Online]. Available: arXiv:1907.08921.
[15] 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. Inf. Theory*, vol. 61, no. 5, pp. 2788–2806, May 2015.
[16] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*. Cambridge, U.K.: Cambridge Univ. Press, 2018.
[17] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition," in *Proc. Eur. Conf. Mach. Learn.*, 2016, pp. 795–811.
[18] M. Rudelson and R. Vershynin, "Hanson-Wright inequality and sub-Gaussian concentration," *Electron. Commun. Probab.*, vol. 18, no. 82, pp. 1–9, 2013.
[19] F. Lin, M. Fardad, and M. R. Jovanović, "Design of optimal sparse feedback gains via the alternating direction method of multipliers," *IEEE Trans. Autom. Control*, vol. 58, no. 9, pp. 2426–2431, Sep. 2013.