2020 American Control Conference
Denver, CO, USA, July 1-3, 2020

Random search for learning the linear quadratic regulator

Hesameddin Mohammadi, Mahdi Soltanolkotabi, and Mihailo R. Jovanović

***Abstract*— Many emerging applications involve control of systems with unknown dynamics. As a result, model-free random search techniques that directly search over the space of parameters have become popular. These algorithms often exhibit a competitive sample complexity compared to state-of-the-art techniques. However, due to the nonconvex nature of the underlying optimization problems, the convergence behavior and statistical properties of these approaches are poorly understood. In this paper, we examine the standard linear quadratic regulator problem for continuous-time systems with unknown state-space parameters. We establish theoretical bounds on the sample complexity and prove the linear convergence rate of the random search method.**

***Index Terms*— Linear quadratic regulator, model-free control, nonconvex optimization, Polyak-Lojasiewicz inequality, random search method, reinforcement learning, sample complexity.**

I. Introduction

Reinforcement Learning (RL) approaches often perform well in applications with no control-oriented models [1], [2]. Without even requiring system identification, the class of model-free RL methods prescribe control action only based on estimated values of a cost function [3]–[5]. In spite of empirical success of these techniques, many fundamental questions surrounding convergence and sample complexity remain unanswered even for classical control problems, including the linear quadratic regulator (LQR). In this paper, we make progress in addressing such challenges with a focus on the infinite-horizon LQR problem for *continuous-time* LTI systems.

The globally optimal solution to the LQR problem can be obtained by solving the Riccati equation and efficient numerical schemes with provable convergence guarantees have been developed [6]. However, computing the optimal solution becomes challenging when model is not available. This motivates the use of direct search methods for controller synthesis. In addition to nonconvexity [7], a major challenge in model-free settings is that the gradient of the objective function is unknown so that only zero-order methods can be used to estimate the gradient.

Despite nonconvexity, for *discrete-time* LQR, global convergence guarantees for both gradient descent and random search on the state-feedback gains were provided in [5]. This result exploited observation that the cost function satisfies the Polyak-Lojasiewicz (PL) condition. Recent reference [8] extended this observation to the *continuous-time* LQR problem and established linear convergence for gradient descent. Extensions to the $\mathcal{H}_\infty$-regularized LQR [9] and Markovian jump systems [10] have also been studied.

In this paper, we show that the random search method can solve the *continuous-time* LQR problem with unknown dynamics up to any desired accuracy with high probability (w.h.p.) in polynomial time. Our results provide upper bounds on the sample complexity and quantify how the final accuracy depends on the number of samples and simulation time.

While Reference [5] motivates our work, we study the *continuous-time* LQR problem and, compared to [5], we provide a significant improvement in computational efficiency by reducing the required simulation time for achieving ϵ-accuracy from $O(\mathrm{poly}(1/\epsilon))$ to $O(\log(1/\epsilon))$. We also refer to our more recent works where we established an overall sample complexity of $O(\log(1/\epsilon))$ in the case of two-point gradient estimates for both continuous-time [11] and discrete-time [12] systems.

The paper is structured as follows. In Section II, we revisit the LQR problem and present the random search method. In Section III, we highlight the main result of the paper. In Section IV, we discuss the convergence of gradient descent. In Section V, we quantify the accuracy of the gradient estimate used in random search method. In Section VI, we prove the main convergence result and, in Section VII, we offer concluding remarks and discuss future directions.

II. Problem formulation

Consider the LTI system

$$\dot{x} \,=\, Ax \,+\, Bu, \quad x(0) \,\sim\, \mathcal{D} \tag{1a}$$

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 of appropriate dimensions, and $x(0)$ is a random initial condition with distribution $\mathcal{D}$. The LQR problem associated with system (1a) is given by

$$\underset{x,\,u}{\text{minimize}} \;\; \mathbb{E}\left[\int_0^\infty (x^T(t)Qx(t) + u^T(t)Ru(t))\,\mathrm{d}t\right] \tag{1b}$$

where Q and R are positive definite matrices. For a controllable pair (A, B), the solution to (1) is the linear feedback

$$u \;=\; -K^\star x \;=\; -R^{-1}B^T P^\star x$$

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

$$A^T P^\star \,+\, P^\star A \,+\, Q \,-\, P^\star B R^{-1} B^T P^\star \;=\; 0. \tag{2}$$

When the model parameters A and B are known, the ARE can be solved efficiently via a variety of techniques [13]–[16]. However, these techniques are not directly applicable when the parameters are unknown. One approach to dealing with

Financial support from the National Science Foundation under Awards ECCS-1708906 and ECCS-1809833 is gratefully acknowledged.

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

this situation is to use the linearity of the optimal controller and reformulate the LQR problem as an optimization over feedback gains,

$$\underset{K}{\text{minimize}}\ f(K) \tag{3}$$

$$f(K)\ :=\ \begin{cases} \text{trace}\left((Q+K^TRK)X(K)\right), & K\in\mathcal{S} \\ \infty, & \text{otherwise.}\end{cases}$$

Here, the function $f(K)$ determines the LQR cost in (1b) associated with the linear feedback law $u=-Kx$,

$$\mathcal{S}\ :=\ \{K\in\mathbb{R}^{m\times n}\mid A-BK \text{ is Hurwitz}\} \tag{4}$$

is the set of stabilizing feedback gains, and for any $K\in\mathcal{S}$

$$X(K)\ :=\ \int_0^\infty \mathrm{e}^{(A-BK)t}\,\Omega\,\mathrm{e}^{(A-BK)^Tt}\,\mathrm{d}t \tag{5}$$

where $\Omega := \mathbb{E}[x(0)x^T(0)] \succ 0$ is the correlation matrix associated with the initial condition $x(0)\sim\mathcal{D}$, which we assume to be positive definite. Moreover, since the optimal feedback gain $K^\star = R^{-1}B^TP^\star$ does not depend on the initial condition, without loss of generality, we assume that the random initial condition $x(0)$ is uniformly bounded, i.e., $\|x(0)\|_2\le M$ with probability one. In problem (3), K is the optimization variable, and (A, B, $Q\succ 0$, $R\succ 0$, $\Omega\succ 0$, $M>0$) are the problem parameters.

The formulation of the LQR problem given by (3) has been studied for both continuous-time [6], [8] and discrete-time systems [5], [17]. It also represents a building block for several important control problems including imposing structural constraints (e.g., sparsity) on the feedback gain matrix [18]–[20] and optimal sensor/actuator selection [21]–[23].

In this paper, we analyze the convergence properties of the random search method for solving problem (3) with unknown model parameters [24], [25]. At each iteration, this method forms an empirical approximation $\overline{\nabla} f(K)$ to the gradient of the objective function $f(K)$ via simulations of system (1a) for several randomly perturbed feedback gains $K+U_i$, $i=1,\dots,N$; see Algorithm 1. The random search method then follows the update rule

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

for some stepsize $\alpha>0$.

III. Main result

Even though the optimization problem (3) is nonconvex [7], we demonstrate that, for any desired accuracy, the iterates of (RS) with a suitably selected set of parameters in Algorithm 1 and a constant stepsize α converge to the optimal solution w.h.p. in polynomial time.

Theorem 1: There are positive rational functions $r_0(a)$, …, $r_5(a)$, and $\gamma(a)<1$ such that for any

$$\epsilon\ \le\ \min\{(a\ -\ f^\star)\,(1-\gamma(a)),\, r_0(a)\}$$

if we choose the simulation time τ and smoothing parameter r in Algorithm 1 to satisfy

$$\tau\ \ge\ \frac{1}{r_1(a)}\log\left(\frac{mnr_2(a)}{r\epsilon}\right),\quad r\ \le\ \min\{\frac{\epsilon}{r_3(a)},\, r_4(a)\}$$

then the iterates of (RS) starting from the initial condition $K^0\in\mathcal{S}\subset\mathbb{R}^{m\times n}$ with $f(K^0)=a$ satisfy the error bound

$$f(K^k)\ -\ f^\star\ \le\ (a\ -\ f^\star)\,(\gamma(a))^k\ +\ (1-\gamma(a))^{-1}\epsilon$$

with probability at least $1-k(mn+1)\exp\left(\dfrac{-N\epsilon^2}{(3d/2)^2\ +\ d\epsilon}\right)$. Here, $d := mnr_5(a)/r$, $f^\star=f(K^\star)$, and N is the number of simulations in Algorithm 1 that is performed at each iteration.

Algorithm 1 Gradient estimation

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

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

– Sample a perturbed feedback gain $K_i := K+U_i$, where U_i is uniformly distributed on the sphere $\mathbb{S}_r(0)$ of radius r centered at 0.

– Sample an initial condition $x_i(0)$ with distribution $\mathcal{D}$.

– Simulate system (1a) with the feedback gain K_i and the initial condition $x_i(0)$ up to time τ and construct

$$\hat{f}_i\ :=\ \int_0^\tau (x^T(t)Qx(t)+u^T(t)Ru(t))\,\mathrm{d}t.$$

end for

Output: The gradient estimate $\overline{\nabla} f(K) := \dfrac{mn}{r^2N}\displaystyle\sum_{i=1}^N \hat{f}_i\,U_i$.

The proof of Theorem 1 along with a discussion on the values of parameters r_0, …, r_5, the rate of convergence γ, and the stepsize α are presented in Section VI.

IV. Synthesis with a known model

The random search method in (RS) at each iteration calls Algorithm 1 to estimate the gradient of the objective function

$$\nabla f(K)\ =\ 2\left(RK-B^TP(K)\right)X(K). \tag{6}$$

Here, $P(K)$ is the unique positive definite solution to

$$(A\ -\ BK)^TP\ +\ P(A\ -\ BK)\ =\ -Q\ -\ K^TRK \tag{7}$$

and X is given by (5) [26]. Note that the existence and uniqueness of $P(K)\succ 0$ is equivalent to the closed-loop stability, i.e., $K\in\mathcal{S}$. Replacing the estimate $\overline{\nabla} f(K^k)$ in (RS) with $\nabla f(K^k)$ yields the gradient descent method

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

Although nonconvex, the function f has two main properties that can be used to prove linear convergence of (GD).

A. Smoothness and gradient dominance over sublevel sets

The gradient descent method converges linearly $O(\gamma^k)$ for some positive $\gamma<1$ if the objective function is smooth and satisfies the Polyak-Lojasiewicz (PL) condition [27]. These

properties do not hold for the LQR objective function f uniformly over its domain $\mathcal{S}$. However, restricted to any nonempty sublevel set $\mathcal{S}(a) := \{K \in \mathcal{S} | f(K) \leq a\}$, the function f is indeed L-smooth, i.e.,

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

for all K, $K' \in \mathcal{S}(a)$ and it satisfies the PL condition, i.e.,

$$2\mu \left(f(K) - f(K^\star)\right) \leq \|\nabla f(K)\|_F^2$$

for all $K \in \mathcal{S}(a)$ [8]. The explicit dependence of the scalars $L > \mu > 0$ on a was established in [8] where it was shown that L and μ are positive rational functions of a. The PL condition was recently used to show convergence of gradient descent for LQR problem for *discrete-time* systems [5].

B. Linear convergence

Our convergence analysis for the random search method (RS) relies on the convergence of gradient descent. Although nonuniform, the PL condition along with smoothness of the objective function were used in [8, Theorem 2] to show linear convergence of the gradient descent method (GD).

Theorem 2: Let $L > \mu > 0$ be the smoothness and PL parameters of the function f over the sublevel set $\mathcal{S}(a)$. For any initial feedback gain $K^0 \in \mathcal{S}(a)$, the iterates of gradient descent (GD) with stepsize $\alpha \in (0, 1/L]$ satisfy

$$f(K^{k+1}) - f(K^\star) \leq (1 - \alpha\mu)\left(f(K^k) - f(K^\star)\right). \quad (8)$$

V. Gradient estimation

In this section, we analyze the accuracy of the gradient estimate $\overline{\nabla} f(K)$ resulting from Algorithm 1. The problem of estimating the gradient using function values obtained via random sampling has received significant attention for gradient-free optimization [28]. Let U_b and U_s be random variables centered at 0 that are uniformly distributed on the ball $\mathbb{B}_r(0)$ of radius $r > 0$ and its boundary $\mathbb{S}_r(0)$, respectively. For the bounded continuous function f: $\mathbb{R}^{m\times n} \to \mathbb{R}$,

$$\nabla \bar{f}(K) = \frac{mn}{r^2} \mathbb{E}_{U_s}[f(K + U_s)U_s] \quad (9)$$

where

$$\bar{f}(K) := \mathbb{E}_{U_b}[f(K + U_b)] \quad (10)$$

is the r-averaged version of the function $f(K)$ [29, Lemma 2.1]. We use the gradient $\nabla \bar{f}(K)$ as a tool to upper bound the gap between the estimate $\overline{\nabla} f(K)$ produced by Algorithm 1 and the gradient $\nabla f(K)$ via the triangle inequality

$$\|\nabla f(K) - \overline{\nabla} f(K)\|_F \leq \|\nabla f(K) - \nabla \bar{f}(K)\|_F + \|\nabla \bar{f}(K) - \overline{\nabla} f(K)\|_F. \quad (11)$$

The function $f(K)$, however, is not uniformly bounded over the domain $\mathcal{S}$. In what follows, we first establish a sufficient condition for the boundedness of the function $f(K + U)$ for all $U \in \mathbb{B}_r(0)$ to ensure that $\bar{f}(K)$ is well defined and satisfies (9). Then, we derive upper bounds on the terms that appear on the right-hand side of (11) and analyze the accuracy of the gradient estimate.

A. 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., $K + U_i \in \mathcal{S}$. Violating such condition leads to an exponential growth of the state and control signals and invalidates our proof technique which relies on the premise of dealing with bounded values of the objective function $f(K)$ and stable closed-loop systems. In Proposition 1, we establish a radius within which any perturbation of the feedback gain $K \in \mathcal{S}$ remains stabilizing.

Proposition 1: For any feedback gain $K \in \mathcal{S}$, we have

$$\{\hat{K} \in \mathbb{R}^{m\times n} \,|\, \|\hat{K} - K\|_2 < \zeta\} \subset \mathcal{S}$$

where $\zeta := \frac{1}{2}\lambda_{\min}(\Omega)\left(\|B\|_2 \|X(K)\|_2\right)^{-1}$.

The proof of Proposition 1 relies on KYP lemma [30, Lemma 7.4] and the small-gain test [30, Theorem 8.2]. These are standard control-theoretic tools that facilitate stability analysis of linear systems with uncertain parameters. We omit the proof due to page limitations.

The sample feedback gains $K + U_i$ are stabilizing as long as the parameter r in Algorithm 1 is smaller that ζ given by Proposition 1. Moreover, the r-averaged function $\bar{f}(K)$ is well defined and it satisfies (9).

B. Bounding the distance between $\nabla f(K)$ and $\nabla \bar{f}(K)$

From the definition of the function $\bar{f}(K)$ in (10) we have

$$\nabla f(K) - \nabla \bar{f}(K) = \mathbb{E}_{U_b}[\nabla f(K) - \nabla f(K + U_b)] \quad (12)$$

where the random variable U_b is uniformly distributed over the ball $\mathbb{B}_r(0)$. Lemma 1 quantifies a Lipschitz continuity parameter for the gradient $\nabla f(K)$ that allows us to bound the distance $\|\nabla f(K) - \nabla \bar{f}(K)\|_F$. We also provide Lipschitz continuity parameters for the objective function and the matrices $X(K)$ and $P(K)$ that are used in the next subsections.

Lemma 1: For any $K \in \mathcal{S}$ and $\hat{K} \in \mathbb{R}^{m\times n}$ such that $\|\hat{K} - K\|_2 < \delta$, with

$$\delta := \frac{1}{4\|B\|_F} \min\left\{\frac{\lambda_{\min}(\Omega)}{\operatorname{trace}(X(K))}, \frac{\lambda_{\min}(Q)}{\operatorname{trace}(P(K))}\right\}$$

the feedback gain matrix $\hat{K} \in \mathcal{S}$, and

$$\|X(\hat{K}) - X(K)\|_F \leq \epsilon_1 \|\hat{K} - K\|_2 \quad (13a)$$

$$|f(\hat{K}) - f(K)| \leq \epsilon_2 \|\hat{K} - K\|_2 \quad (13b)$$

$$\|P(\hat{K}) - P(K)\|_F \leq \epsilon_3 \|\hat{K} - K\|_2 \quad (13c)$$

$$\|\nabla f(\hat{K}) - \nabla f(K)\|_F \leq \epsilon_4 \|\hat{K} - K\|_2 \quad (13d)$$

where $X(K)$ and $P(K)$ are given by (5) and (7), respectively. Furthermore, the parameters ϵ_i which only depend on K and problem data are given by $\epsilon_1 := \|X(K)\|_2/\delta$, $\epsilon_3 := 2\operatorname{trace}(P)(2\|P\|_2\|B\|_F + (\delta + 2\|K\|_2)\|R\|_F)/\lambda_{\min}(Q)$, $\epsilon_2 := \epsilon_3\|\Omega\|_F$, $\epsilon_4 := 2(\epsilon_1\|K\|_2 + 2\|X(K)\|_2)\|R\|_F + 2\epsilon_1(\|P(K)\|_2 + 2\epsilon_3\|X(K)\|_2)\|B\|_F$.

Lemma 1 combines the stability margin established in Proposition 1 with bounds on the norm of inverse Lyapunov

operator. We omit the proof due to page limitations.

Using Lemma 1, we can bound the right-hand side of equation (12). This leads to the next proposition.

Proposition 2: For any $K \in \mathcal{S}$ and $r < \delta$, we have $\|\nabla f(K) - \nabla \bar{f}(K)\|_F \leq \epsilon_4 r$, where $\bar{f}(K)$ is the r-averaged version of $f(K)$ and the parameters δ and ϵ_4 provided in Lemma 1 only depend on K and problem data.

Proof: Since $\|U_b\|_2 \leq \|U_b\|_F \leq r$, Lemma 1 implies that, for $r < \delta$, inequality (13d) holds with $\hat{K} := K + U_b$. This yields $\|\nabla f(K) - \nabla f(K + U_b)\|_F \leq \epsilon_4 \|U_b\|_2 \leq \epsilon_4 r$. Taking expectation and using the triangle inequality on (12) completes the proof. ∎

C. Bounding the distance between $\nabla \bar{f}(K)$ and $\overline{\nabla} f(K)$

The output $\overline{\nabla} f(K)$ of Algorithm 1 is a biased estimator of $\nabla \bar{f}(K)$. We next address the resulting bias and variance.

1) Bias: The bias arises from finite-time approximation in the simulation step of Algorithm 1. To illustrate this, let us define the τ-truncated versions of the objective function $f(K)$ and the matrix $X(K)$ in (5) as

$$f_\tau(K) \;:=\; \mathrm{trace}\left((Q \,+\, K^T R\, K)\, X_\tau(K)\right) \tag{14a}$$

$$X_\tau(K) \;:=\; \int_0^\tau \mathrm{e}^{(A\,-\,BK)t}\, \Omega\, \mathrm{e}^{(A\,-\,BK)^T t}\, \mathrm{d}t. \tag{14b}$$

Using the solution $x(t) = \mathrm{e}^{(A-BK)t} x(0)$ of the closed-loop system, it is straightforward to verify that

$$f_\tau(K) \;=\; \mathbb{E}_{x(0)} \left[\int_0^\tau (x^T(t) Q x(t) + u^T(t) R u(t))\, \mathrm{d}t \right].$$

Thus, based on sampling distribution of the random gains U_i and the initial conditions $x_i(0)$ in Algorithm 1, it follows that the mean value of the gradient estimate $\overline{\nabla} f(K)$ satisfies

$$\begin{aligned} \mathbb{E}\left[\overline{\nabla} f(K)\right] \;&=\; \frac{mn}{r^2 N} \sum_{i\,=\,1}^{N} \mathbb{E}\left[\hat{f}_i\, U_i\right] \\ &=\; \frac{mn}{r^2}\; \mathbb{E}_{U_s}[f_\tau(K + U_s) U_s]. \end{aligned} \tag{15}$$

Here, $\hat{f}_i$ is the cost associated with the ith simulation in Algorithm 1 with the sample feedback gain $K + U_i$ and U_s is uniformly distributed on $\mathbb{S}_r(0)$. Now, subtracting (15) from (9), we can represent the bias term as

$$\begin{aligned} \nabla \bar{f}(K) \;-\; \mathbb{E}[\overline{\nabla} f(K)] \;=& \\ \frac{mn}{r^2}\; \mathbb{E}_{U_s}[(f_\tau(K + U_s) - f(K + U_s))\, U_s]. \end{aligned} \tag{16}$$

In Proposition 3, we use this equation to establish an exponentially vanishing upper bound on the bias.

Proposition 3: For any sublevel set $\mathcal{S}(a)$, there are positive parameters $\bar{\kappa}_2$, $\bar{\kappa}_3$, and θ such that the output $\overline{\nabla} f(K)$ of Algorithm 1 with $K \in \mathcal{S}(a)$ and $r < \theta$ satisfies

$$\|\nabla \bar{f}(K) \;-\; \mathbb{E}\left[\overline{\nabla} f(K)\right]\|_F \;\leq\; (mn \bar{\kappa}_3 / r)\; \mathrm{e}^{-\bar{\kappa}_2 \tau} \tag{17}$$

where $\bar{f}(K)$ is the r-averaged version of the function $f(K)$. The parameters $\bar{\kappa}_2$, $\bar{\kappa}_3$, and θ are rational functions of a. These parameters are discussed in the proof; see Appendix A.

2) Variance: We use concentration results to establish a probabilistic bound on the norm of the random matrix

$$G \;:=\; \overline{\nabla} f(K) \;-\; \mathbb{E}\left[\overline{\nabla} f(K)\right]$$

where $\overline{\nabla} f(K) = (mn/(r^2 N)) \sum_{i\,=\,1}^{N} \hat{f}_i\, U_i$ is the output of Algorithm 1. In particular, we can express G as the sum of N zero-mean i.i.d. random matrices, $G = \sum_i V_i$,

$$V_i \;:=\; \frac{1}{N} \left(\frac{mn}{r^2}\, \hat{f}_i U_i \;-\; \mathbb{E}\left[\overline{\nabla} f(K)\right]\right).$$

This allows us to employ the Bernstein inequality [31, Theorem 1.6.2] to show that $\|G\|_F$ can be made arbitrary small by choosing a large number of samples N.

Proposition 4: There exists a positive rational function $\theta(a)$ such that, for any sublevel set $\mathcal{S}(a)$ of the objective function $f(K)$, the output of Algorithm 1 with $r < \theta(a)$, $K \in \mathcal{S}(a) \subset \mathbb{R}^{m \times n}$, and N samples satisfies $\|G\|_F \leq \epsilon$ with probability at least $1 - (mn + 1) \exp\left(\dfrac{-N\epsilon^2}{(\frac{mnl}{2r})^2 + \frac{mnl\epsilon}{3r}}\right)$, where $l := 4aM^2/\lambda_{\min}(\Omega)$ and M upper bounds $\|x_0\|_2$.

Proof: See Appendix B. ∎

D. Total error

Herein, we bound the accuracy of the gradient estimate $\overline{\nabla} f(K)$ as a function of the parameters in Algorithm 1. From inequality (11) and the triangle inequality we have

$$\begin{aligned} \|\nabla f(K) \;-\; \overline{\nabla} f(K)\|_F \;\leq\; \|\nabla f(K) \;-\; \nabla \bar{f}(K)\|_F \;+ \\ \|\nabla \bar{f}(K) \;-\; \mathbb{E}[\overline{\nabla} f(K)]\|_F \;+\; \|\overline{\nabla} f(K) \;-\; \mathbb{E}[\overline{\nabla} f(K)]\|_F. \end{aligned} \tag{18}$$

Theorem 3 combines Propositions 2, 3 and 4 to bound the terms on the right-hand side of the above inequality. We omit the proof due to page limitations.

Theorem 3: There exist positive rational functions $\bar{\kappa}(a)$, $\bar{\kappa}'(a)$, $\bar{\theta}(a)$, and $\bar{\theta}'(a)$ such that for any $K \in \mathcal{S}(a)$, the output of Algorithm 1 with

$$\tau \;\geq\; \frac{1}{\bar{\kappa}(a)} \log\left(\frac{3mn\bar{\kappa}'(a)}{r\epsilon}\right), \quad r \;<\; \min\{\frac{\epsilon}{\bar{\theta}(a)}, \bar{\theta}'(a)\}$$

satisfies $\|\nabla f(K) - \overline{\nabla} f(K)\|_F \;\leq\; \epsilon$ with probability at least

$$1 \;-\; (mn + 1) \exp\left(\frac{-N\epsilon^2}{(\frac{3mnl}{2r})^2 + \frac{mnl\epsilon}{r}}\right)$$

where $l := 4a\, M^2/\lambda_{\min}(\Omega)$, M is an upper bound on $\|x_0\|_2$, and mn is the number of entries in K.

VI. Control synthesis with an unknown model

In this section, we analyze the random search algorithm in (RS). Theorem 3 proves that the parameters in Algorithm 1 can be selected to achieve any desired accuracy for the gradient estimate with high probability. This allows us to relate the iterates of (RS) to those of gradient descent (GD) to deduce convergence of (RS) from the linear convergence

of (GD) established in Theorem 2. We use the notation introduced in Section V to present our main convergence result. Theorem 4 is a more formal restatement of Theorem 1.

Theorem 4: Let $\bar{\kappa}(a)$, $\bar{\kappa}'(a)$, $\bar{\theta}(a)$, and $\bar{\theta}'(a)$ be positive rational functions as in Theorem 3. Let stepsize $\alpha \in (0, 1/L]$ and let $\gamma = 1 - \alpha\mu$, where L and μ are the smoothness and PL parameters of the function f over its sublevel set $\mathcal{S}(a)$. There are positive rational functions $\bar{\delta}(a)$ and $\bar{\epsilon}_2(a)$ such that for any $K^0 \in \mathcal{S}(a)$ and $\epsilon \leq \min\{(a - f^\star)(1-\gamma),\ \bar{\delta}\bar{\epsilon}_2\}$, if we choose the simulation time τ and smoothing parameter r in Algorithm 1 to satisfy

$$\tau \ \geq\ \frac{1}{\bar{\kappa}} \log\left(\frac{3mn\bar{\kappa}'\bar{\epsilon}_2\alpha}{r\,\epsilon}\right), \quad r \ <\ \min\{\frac{\epsilon}{\bar{\theta}\,\bar{\epsilon}_2\alpha},\ \bar{\theta}'\}$$

then the iterates of (RS) satisfy

$$f(K^k) \ -\ f^\star \ \leq\ (a \ -\ f^\star)\,\gamma^k \ +\ (1-\gamma)^{-1}\epsilon$$

with probability at least

$$1 \ -\ k\,(mn+1)\exp\left(\frac{-N\epsilon^2}{(\frac{3mnl\bar{\epsilon}_2\alpha}{2r})^2 + \frac{mnl\bar{\epsilon}_2\alpha\epsilon}{r}}\right).$$

Here, $f^\star := f(K^\star)$, N is the number of samples per iteration, and $l := 4a\,M^2/\lambda_{\min}(\Omega)$, where M upper bounds $\|x_0\|_2$.

Proof: For any $\alpha \in (0, 1/L]$, by Theorem 2

$$f(K') \ -\ f^\star \ \leq\ \big(f(K^0) \ -\ f^\star\big)\,\gamma \tag{19}$$

where $K' = K^0 - \alpha\nabla f(K)$ is the first iteration of gradient descent method (GD). As we discuss in the proof of Proposition 3, the constants δ and ϵ_2 in Lemma 1 can be uniformly lower and upper bounded over the sublevel set $\mathcal{S}(a)$ by positive rational functions $\bar{\delta}(a)$ and $\bar{\epsilon_2}(a)$. By Theorem 3, we can choose parameters in Algorithm 1 such that

$$\|\overline{\nabla} f(K^0) - \nabla f(K^0)\|_F \ \leq\ \epsilon/(\bar{\epsilon}_2\alpha)$$

w.h.p. Thus, noting that $K' \in \mathcal{S}(a)$, for any $\epsilon \leq \bar{\delta}\,\bar{\epsilon}_2$,

$$\|K^1 - K'\|_2 \ =\ \alpha\|\overline{\nabla} f(K^0) - \nabla f(K^0)\|_2 \ \leq\ \epsilon/\bar{\epsilon}_2 \ \leq\ \bar{\delta}$$

w.h.p. By Lemma 1, it follows that $|f(K^1) - f(K')| \leq \epsilon$ w.h.p. and, thus, from (19) and the triangle inequality, for any $\epsilon \leq (a - f^\star)(1-\gamma)$, we obtain

$$f(K^1) - f^\star \ \leq\ \epsilon + \big(f(K^0) \ -\ f^\star\big)\,\gamma \ \leq\ (a \ -\ f^\star)\,. \tag{20}$$

This implies K^1 remains in $\mathcal{S}(a)$ and we can use induction on the first inequality in (20) to show that w.h.p.

$$f(K^k) - f^\star \ \leq\ \big(f(K^0) \ -\ f^\star\big)\,\gamma^k \ +\ \epsilon\sum_{i=0}^{k-1}\gamma^i.$$

Finally, the probability bound is obtained by applying the union bound on Theorem 3 for the first k iterations. ∎

VII. Concluding remarks

We establish a bound on the sample complexity of a random search method for solving the continuous-time LQR problem that directly searches for the controller over the nonconvex set of stabilizing feedback gains. Our results demonstrate that with a simulation time of $O(\log(1/\epsilon))$, the random search method achieves an accuracy level ϵ at a linear rate provided that we have enough samples, $N = \text{poly}\,(1/\epsilon)$. In our more recent work [11], we have improved this result by eliminating the dependence of N on ϵ.

Appendix

A. Bounding the bias: proof of Proposition 3

We first present a technical lemma whose proof is omitted due to page limitations.

Lemma 2: For any $K \in \mathcal{S}$ and $\tau > 0$, the τ-truncated matrix $X_\tau(K)$ and objective function $f_\tau(K)$ in (14) satisfy

$$\|X(K) - X_\tau(K)\|_F \ \leq\ \kappa_1\,\mathrm{e}^{-\kappa_2\tau} \tag{21a}$$

$$f(K) - f_\tau(K) \ \leq\ \kappa_3\,\mathrm{e}^{-\kappa_2\tau} \tag{21b}$$

where $X(K)$ is given by (5) and the constants are given by

$$\kappa_1 \ :=\ \frac{\|\Omega\|_F\|X(K)\|_2^2}{\lambda_{\min}(\Omega)\lambda_{\min}(X(K))}, \quad \kappa_2 \ :=\ \frac{\lambda_{\min}(\Omega)}{\|X(K)\|_2}$$

$$\kappa_3 \ :=\ \kappa_1(\|Q\|_2 + \|R\|_2\|K\|_2^2).$$

It has been shown that over the sublevel set $\mathcal{S}(a)$, the values of $\|K\|_F$, $\text{trace}(X(K))$ and $1/\lambda_{\min}(X(K))$ are upper bounded by linear functions of a [8]. Using these bounds, it is straightforward to verify that the parameters δ and ϵ_2 in Lemma 1 can be uniformly lower and upper bounded by rational functions of a for all $K \in \mathcal{S}(a)$. Let $\bar{\delta}$ and $\bar{\epsilon}_2$ be two of such bounds, i.e., for any $\delta \leq \bar{\delta}$ and $\epsilon_2 \geq \bar{\epsilon_2}$, Lemma 1 holds for all $K \in \mathcal{S}(a)$. Similarly, the parameters κ_2 and κ_3 in Lemma 2 can also be lower and upper bounded uniformly over the sublevel set $\mathcal{S}(2a)$ by polynomial functions of a. We also let $\bar{\kappa}_2$ and $\bar{\kappa}_3$ be two of such bounds, i.e., for any $\kappa_2 \leq \bar{\kappa}_2$ and $\kappa_3 \geq \bar{\kappa}_3$, Lemma 2 holds for all $K \in \mathcal{S}(2a)$.

If we restrict the smoothing parameter r to satisfy the inequality $r < \min\{\bar{\delta},\ a/\bar{\epsilon}_2\}$, then from (13b) in Lemma 1 with $\delta := \bar{\delta}$ and $\epsilon_2 :=\ \bar{\epsilon}_2$, it follows that

$$f(K + U_s) - f(K) \ \leq\ \bar{\epsilon}_2\,\|U_s\|_2 \ \leq\ \bar{\epsilon}_2\,r \ \leq\ a \tag{22}$$

for all $K \in \mathcal{S}(a)$. Thus, from the triangle inequality we obtain that $K + U_s \in \mathcal{S}(2a)$ for all $U_s \in \mathbb{S}_r(0)$ and $K \in \mathcal{S}(a)$. This allows us to use Lemma 2 for all feedback gains $K + U_s$ and $\kappa_2 := \bar{\kappa}_2$ and $\kappa_3 := \bar{\kappa}_3$ to obtain that

$$\|\nabla\bar{f}(K) \ -\ \mathbb{E}[\overline{\nabla} f(K)]\|_F \ \leq$$

$$\frac{mn}{r^2}\,\mathbb{E}[(f_\tau(K+U_s) - f(K+U_s))\,\|U_s\|_F] \leq \frac{mn\bar{\kappa}_3}{r}\,\mathrm{e}^{-\bar{\kappa}_2\tau}.$$

Here, the first inequality follows from applying the triangle inequality on (16), and the second follows from Lemma 2.

B. Proof of Proposition 4

Since V_i are i.i.d., the Bernstein's inequality [31, Theorem 1.6.2] implies $\|G\|_F \leq \epsilon$ with probability not smaller than

$$1 \ -\ (mn+1)\exp\big(-(\frac{m^2n^2v'}{N\epsilon^2r^4} + \frac{mnl'\epsilon}{3N\epsilon^2r^2})^{-1}\big) \tag{23}$$

where l' and v' are the bounds $\|\hat{f}_i U_i - \mathbb{E}[\hat{f}_i U_i]\|_F \leq l'$, $\mathbb{E}[\|\hat{f}_i U_i - \mathbb{E}[\hat{f}_i U_i]\|_F^2] \leq v'$. To quantify v' and l', we can restrict r to be smaller than a rational function $\theta(a)$ to ensure $K + U_i \in \mathcal{S}(2a)$; cf. (22). This allows us to write

$$\|\hat{f}_i U_i - \mathbb{E}[\hat{f}_i U_i]\|_F \;\leq\; 2\,r\,\max(\hat{f}_i) \;\leq$$
$$2\,r\,\max\left((x_i(0))^T P(K+U_i)x_i(0)\right) \;\leq\; \frac{4\,r\,a\,M^2}{\lambda_{\min}(\Omega)} \;=\; r\,l$$

where $l := 4\,a\,M^2/\lambda_{\min}(\Omega)$. Similarly, for the second term,

$$\mathbb{E}\left[\|\hat{f}_i U_i - \mathbb{E}[\hat{f}_i U_i]\|_F^2\right] \;=\; \mathbb{E}\left[\|\hat{f}_i U_i\|_F^2\right] - \|\,\mathbb{E}[\hat{f}_i U_i]\|_F^2$$
$$\leq\; r^2 \max \hat{f}_i^2 \;\leq\; r^2 \max\left((x_i(0))^T P(K+U_i)x_i(0)\right)^2$$
$$\leq\; 4a^2 r^2 M^4/\lambda_{\min}^2(\Omega) \;=\; (r\,l/2)^2.$$

These bounds in conjunction with (23) complete the proof.

REFERENCES

[1] A. Nagabandi, G. Kahn, R. Fearing, and S. Levine, "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning," in *IEEE Int Conf. Robot. Autom.*, 2018, pp. 7559–7566.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," 2013, arXiv:1312.5602.

[3] D. Bertsekas, "Approximate policy iteration: A survey and some new methods," *J. Control Theory Appl.*, vol. 9, no. 3, pp. 310–335, 2011.

[4] Y. Abbasi-Yadkori, N. Lazic, and C. Szepesvári, "Model-free linear quadratic control via reduction to expert prediction," in *Proc. Mach. Learn. Res.*, vol. 89. PMLR, 2019, pp. 3108–3117.

[5] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proc. Int'l Conf. Machine Learning*, 2018, pp. 1467–1476.

[6] B. Anderson and J. Moore, *Optimal Control; Linear Quadratic Methods*. New York, NY: Prentice Hall, 1990.

[7] J. Ackermann, "Parameter space design of robust control systems," *IEEE Trans. Automat. Control*, vol. 25, no. 6, pp. 1058–1072, 1980.

[8] 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 *Proceedings of the 58th IEEE Conference on Decision and Control*, Nice, France, 2019, pp. 7474–7479.

[9] 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, arXiv:1910.09496.

[10] J. P. Jansch-Porto, B. Hu, and G. E. Dullerud, "Convergence guarantees of policy optimization methods for markovian jump linear systems," 2020, arXiv:2002.04090.

[11] 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," *IEEE Trans. Automat. Control*, 2019, submitted; also arXiv:1912.11899.

[12] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović, "On the linear convergence of random search for discrete-time lqr," *Syst. Contr. Lett.*, 2020, submitted.

[13] D. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Trans. Automat. Control*, vol. 13, no. 1, pp. 114–115, 1968.

[14] S. Bittanti, A. J. Laub, and J. C. Willems, *The Riccati Equation*. Berlin, Germany: Springer-Verlag, 2012.

[15] P. L. D. Peres and J. C. Geromel, "An alternate numerical solution to the linear quadratic problem," *IEEE Trans. Automat. Control*, vol. 39, no. 1, pp. 198–202, 1994.

[16] V. Balakrishnan and L. Vandenberghe, "Semidefinite programming duality and linear time-invariant systems," *IEEE Trans. Automat. Control*, vol. 48, no. 1, pp. 30–41, 2003.

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

[18] F. Lin, M. Fardad, and M. R. Jovanović, "Augmented Lagrangian approach to design of structured optimal state feedback gains," *IEEE Trans. Automat. Control*, vol. 56, no. 12, pp. 2923–2929, December 2011.

[19] F. Lin, M. Fardad, and M. R. Jovanović, "Design of optimal sparse feedback gains via the alternating direction method of multipliers," *IEEE Trans. Automat. Control*, vol. 58, no. 9, pp. 2426–2431, 2013.

[20] M. R. Jovanović and N. K. Dhingra, "Controller architectures: tradeoffs between performance and structure," *Eur. J. Control*, vol. 30, pp. 76–91, July 2016.

[21] B. Polyak, M. Khlebnikov, and P. Shcherbakov, "An LMI approach to structured sparse feedback design in linear control systems," in *Proceedings of the 2013 European Control Conference*, 2013, pp. 833–838.

[22] N. K. Dhingra, M. R. Jovanović, and Z. Q. Luo, "An ADMM algorithm for optimal sensor and actuator selection," in *Proceedings of the 53rd IEEE Conference on Decision and Control*, Los Angeles, CA, 2014, pp. 4039–4044.

[23] A. Zare, H. Mohammadi, N. K. Dhingra, T. T. Georgiou, and M. R. Jovanović, "Proximal algorithms for large-scale statistical modeling and sensor/actuator selection," *IEEE Trans. Automat. Control*, 2019, doi:10.1109/TAC.2019.2948268; also arXiv:1807.01739.

[24] H. Mania, A. Guy, and B. Recht, "Simple random search provides a competitive approach to reinforcement learning," in *Proc. Neural Information Processing (NeurIPS)*, 2018.

[25] B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 2, pp. 253–279, 2019.

[26] H. Horisberger and P. Belanger, "Solution of the optimal constant output feedback problem by conjugate gradients," *IEEE Trans. Automat. Control*, vol. 19, no. 4, pp. 434–435, 1974.

[27] B. Polyak, "Gradient methods for minimizing functionals," *Zh. Vychisl. Mat. Mat. Fiz.*, vol. 3, no. 4, pp. 643–653, 1963.

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

[29] A. Flaxman, A. Kalai, and B. McMahan, "Online convex optimization in the bandit setting: gradient descent without a gradient," in *In 16th ACM-SIAM Symp. on Discerete Algorithms (SODA)*, 2005, pp. 385–394.

[30] G. E. Dullerud and F. Paganini, *A course in robust control theory: a convex approach*. New York: Springer-Verlag, 2000.

[31] J. Tropp, "An introduction to matrix concentration inequalities," *Foundations and Trends in Machine Learning*, vol. 8, no. 1-2, pp. 1–230, 2015.