Model-free linear quadratic regulator

Hesameddin Mohammadi, Mahdi Soltanolkotabi, and Mihailo R. Jovanović

Abstract We review recent results on the convergence and sample complexity of the random search method for the infinite-horizon linear quadratic regulator (LQR) problem with unknown model parameters. This method directly searches over the space of stabilizing feedback gain matrices and, in spite of the lack of convexity, it converges to the globally optimal LQR solution at a linear rate. These results demonstrate that for a model-free method that utilizes two-point gradient estimates, the required simulation time and the total number of function evaluations required for achieving ϵ-accuracy are both $O(\log(1/\epsilon))$.

1 Introduction to a model-free LQR problem

The infinite-horizon LQR problem for continuous-time systems is given by

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

$$\text{subject to} \quad \dot{x} = Ax + Bu, \quad x(0) \sim \mathcal{D}. \tag{1b}$$

Here, $x(t) \in \mathbb{R}^n$ is the state, $u(t) \in \mathbb{R}^m$ is the control input, A and B are constant matrices of appropriate dimensions that determine parameters of the model, Q and R are positive definite matrices, and the expectation is taken over the zero-mean random initial condition $x(0)$ with distribution $\mathcal{D}$. For any controllable pair (A, B), the globally optimal solution to (1) takes the state-feedback form $u(t) = -K^\star x(t)$ and the optimal feedback gain $K^\star \in \mathbb{R}^{m\times n}$ can be obtained by solving the algebraic Riccati equation. However, this approach is not viable for large-scale systems, when prior knowledge of system matrices A and B is not available. In this scenario, an alternative approach is to exploit the linearity of the optimal controller and formulate the LQR problem as a direct search over the set of feedback gain matrices K, namely

Hesameddin Mohammadi · Mahdi Soltanolkotabi · Mihailo R. Jovanović
Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089, USA. e-mail: {hesamedm,soltanol,mihailo}@usc.edu

$$\underset{K}{\text{minimize}}\ f(K) \tag{2}$$

where $f(K)$ is the objective function in (1) associated with the feedback law $u = -Kx$. However, since f is a nonconvex function of K, the analysis of local search optimization algorithms is non-trivial. Furthermore, when the model parameters A and B are not known, the gradient of the objective function f is not accessible and only zeroth-order methods that estimate the gradient can be used.

In this chapter, we review recent results on the convergence and sample complexity of the random search method for optimization problem (2). This problem was recently examined for both discrete-time [1–3] and continuous-time [4] systems. A common theme is that approximations of the gradient $\nabla f(K)$ can be obtained via stochastic simulations of system (1b) [5, 6]. This naturally leads to a zeroth-order optimization approach that attempts to emulate the behavior of gradient descent. To implement this approach, it is essential to have access to approximate function values of the form

$$f_{x_0,\tau}(K) \ := \ \int_0^\tau \Big(x^T(t) Q x(t) \ + \ u^T(t) R u(t) \Big) \, \mathrm{d}t \tag{3}$$

where $x(0) = x_0$ is a random initial condition and $[0, \tau]$ is the finite time horizon. Empirical evidence suggests that the random search method can solve benchmark control problems with state-of-the-art sample efficiency [5]. The fundamental theoretical question is how many function evaluations and what simulation time this method requires to solve problem (2) up to the desired level of accuracy ϵ.

In [4], the above question was answered for the two-point setting in which, for any pair of points K and K', the simulation engine returns the random values $f_{x_0,\tau}(K)$ and $f_{x_0,\tau}(K')$ for some random initial condition x_0. This is in contrast to the one-pint setting in which, at each query, the simulation engine can receive only one specified point K and return the random value $f_{x_0,\tau}(K)$. For convex problems, the gradient estimates obtained in the two-point setting are known to yield faster convergence rates than the one-point setting [7]. However, the two-point setting requires simulations of the system for two different feedback gain matrices under the same initial condition.

For the random search method with one-point gradient estimates to achieve an accuracy level ϵ in solving the LQR problem, reference [8] derived an upper bound on the required number of function evaluations that is proportional to $1/\epsilon^2$. In this chapter, we review the results in [4] which demonstrated that the random search method with two-point gradient estimates converges at a linear rate with high probability. More specifically, the simulation time and the total number of function evaluations that the random search method requires to achieve an accuracy level ϵ are proportional to $\log(1/\epsilon)$. These findings suggest that the use of two-point gradient estimates significantly improves the sample complexity relative to the one-point setting. Finally, while we only focus on continuous-time systems, we note that the proof strategy and results presented here readily extend to discrete-time systems as well [3].

2 A gradient-based random search method

Herein, we present the gradient descent and the random search method for problem (2). The LQR objective function in (1) associated with the feedback law $u = -Kx$ is determined by

$$f(K) \;:=\; \begin{cases} \text{trace}\left((Q + K^T R\, K)X(K)\right), & K \in \mathcal{S} \\ \infty, & \text{otherwise} \end{cases} \tag{4a}$$

where $\mathcal{S} := \{K \in \mathbb{R}^{m\times n} \mid A - BK \text{ is Hurwitz}\}$ is the set of stabilizing feedback gains K,

$$X(K) \;:=\; \int_0^\infty \mathrm{e}^{(A-BK)t}\, \Omega\, \mathrm{e}^{(A-BK)^T t}\, \mathrm{d}t \;\succ\; 0 \tag{4b}$$

and $\Omega := \mathbb{E}[x(0)x^T(0)] \succ 0$ is the covariance matrix of the initial condition $x(0)$. Controllability of the pair (A, B) guarantees $\mathcal{S} \neq \emptyset$ and, for any $K \in \mathcal{S}$ [9],

$$\nabla f(K) \;=\; 2\left(R\, K - B^T P(K)\right) X(K) \tag{4c}$$

where

$$P(K) \;:=\; \int_0^\infty \mathrm{e}^{(A-BK)^T t}\, (Q \;+\; K^T R\, K)\, \mathrm{e}^{(A-BK)t}\, \mathrm{d}t \;\succ\; 0. \tag{4d}$$

With the explicit expression for $\nabla f(K)$ in (4c), the gradient descent method for problem (2) with the stepsize $\alpha > 0$ is given by

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

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 $\text{vec}(U_i)$ is a random vector uniformly distributed on the sphere $\sqrt{mn}\, S^{mn-1}$.

– Sample an initial condition x_i from the distribution $\mathcal{D}$.

– For $j \in \{1, 2\}$, simulate system (1b) up to time τ with the feedback gain $K_{i,j}$ and initial condition x_i to form $\hat{f}_{i,j} = \displaystyle\int_0^\tau (x^T(t)Qx(t) + u^T(t)Ru(t))\, \mathrm{d}t$.

end for

Output: The gradient estimate $\overline{\nabla} f(K) := \dfrac{1}{2rN} \displaystyle\sum_{i=1}^{N} \left(\hat{f}_{i,1} \;-\; \hat{f}_{i,2}\right) U_i$.

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. To address this challenge, we consider the random search method and study its convergence properties. At each iteration, this method calls Algorithm 1 to form an empirical approximation $\overline{\nabla} f(K^k)$ to the gradient $\nabla f(K^k)$ via simulation of system (1b) for randomly perturbed feedback gains $K^k \pm U_i$, $i = 1, \ldots, N$, that belong to the sphere of dimension $mn - 1$ with radius $\sqrt{mn}$ and updates K^k via

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

Note that the gradient estimation scheme in Algorithm 1 does not require knowledge of system matrices A and B in (1b) but only access to a simulation engine.

3 Main results

We recently analyzed the sample complexity and convergence of the random search method (RS) for the model-free setting in [4]. 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 $\mathcal{S}(a) := \{K \in \mathcal{S} \mid f(K) \leq a\}$, where a is a positive scalar. In particular, it can be shown that, restricted to any sublevel set $\mathcal{S}(a)$, the function f satisfies

$$\textbf{Smoothness:} \quad f(K') - f(K) \leq \langle \nabla f(K), K' - K \rangle + \frac{L_f(a)}{2} \|K - K'\|_F^2$$

$$\textbf{PL condition:} \quad f(K) - f(K^\star) \leq \frac{1}{2\mu_f(a)} \|\nabla f(K)\|_F^2$$

for all K and K' whose line segment is in $\mathcal{S}(a)$, where the smoothness and PL parameters $L_f(a)$ and $\mu_f(a)$ are positive rational functions of a [10]. 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 of system* (1b) *have i.i.d. zero-mean unit-variance entries with bounded sub-Gaussian norm. This implies that any random vector $v \sim \mathcal{D}$ satisfies $\|v_i\|_{\psi_2} \leq \kappa$, for some constant κ and $i = 1, \ldots, n$, where $\|\cdot\|_{\psi_2}$ is the sub-Gaussian norm.*

We note that the zero-mean Gaussian distribution with identity covariance matrix obeys Assumption 1. For Gaussian distributions, the above covariance condition is without loss of generality as we can use a change of variables to make the covariance matrix identity. With these definitions and assumptions in place we are now ready to state the main theoretical result.

Theorem 1 ([4]) *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 obey Assumption 1 and let the simulation time τ, the smoothing constant r, and the number of samples N satisfy*

$$\tau \geq \theta'(a) \log \frac{1}{r\epsilon}, \; r < \min\{r(a), \theta''(a)\sqrt{\epsilon}\}, \; N \geq c(1 + \beta^4 \kappa^4 \, \theta(a) \log^6 n) \, n \tag{5}$$

for some $\beta > 0$ and a desired accuracy $\epsilon > 0$. Then, starting from any $K^0 \in \mathcal{S}(a)$, the method in (RS) *with the constant stepsize $\alpha \leq 1/(32\omega(a)L_f(a))$ achieves $f(K^k) - f(K^\star) \leq \epsilon$ with probability not smaller than $1 - kp - 2kN\mathrm{e}^{-n}$ in at most*

$$k \;\leq\; \left(\log \frac{f(K^0) \,-\, f(K^\star)}{\epsilon}\right) \Bigg/ \left(\log \frac{1}{1 \,-\, \mu_f(a)\alpha/8}\right)$$

iterations. Here, $\mu_f(a)$ and $L_f(a)$ are the PL and smoothness parameters of the function f over the sublevel set $\mathcal{S}(a)$, $p := c'(n^{-\beta} + N^{-\beta} + N\mathrm{e}^{-\frac{n}{8}} + \mathrm{e}^{-c'N})$, $\omega(a) := c''\left(\sqrt{m} + \beta\kappa^2\theta(a)\sqrt{mn}\log n\right)^2$, c, c', and c'' are positive absolute constants, and $\theta(a)$, $\theta'(a)$, $\theta''(a)$ and $r(a)$ are problem-dependent positive parameters.

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 $\log(1/\epsilon)$. This significantly improves existing results for model-free LQR that require $O(1/\epsilon)$ function evaluations [8].

4 Proof sketch

The smoothness of the objective function along with the PL condition are sufficient for the gradient descent method to achieve linear convergence even for nonconvex problems [11]. These properties were recently used to show convergence of gradient descent for both *discrete-time* [1] and *continuous-time* [10] LQR problems.

Theorem 2 ([10]) *Consider the gradient descent method* (GD) *for finding the optimal solution $K^\star$ of problem* (2)*. For any initialization $K^0 \in \mathcal{S}(a)$, with the constant stepsize $\alpha = 1/L_f(a)$, we have*

$$\begin{aligned} f(K^k) \,-\, f(K^\star) \;&\leq\; \gamma^k\big(f(K^0) \,-\, f(K^\star)\big) \\ \|K^k \,-\, K^\star\|_F^2 \;&\leq\; b\,\gamma^k\,\|K^0 \,-\, K^\star\|_F^2 \end{aligned}$$

where $\gamma = 1 - \mu_f(a)/L_f(a)$ is the linear rate of convergence and $b > 0$ depends only on the scalar a and the parameters of the LQR objective function f.

The random search method (RS), however, does not have access to the true value of the gradient $\nabla f(K)$ as Algorithm 1 produces only a biased estimate $\overline{\nabla} f(K)$ of $\nabla f(K)$. Unlike existing techniques that directly work with the gradient estimation error [1, 8], the proof of Theorem 1 is based on showing a high correlation between the gradient estimate and the true gradient. In particular, the proof exploits Proposition 1 that establishes a geometric decrease in the objective value by using an update direction G which is not necessarily close to $\nabla f(K)$ but is well correlated with it.

Proposition 1 (Approximate GD [4]) *If the matrix $G \in \mathbb{R}^{m\times n}$ and the feedback gain $K \in \mathcal{S}(a)$ are such that*

$$\langle G, \nabla f(K)\rangle \;\geq\; \mu_1\|\nabla f(K)\|_F^2, \quad \|G\|_F^2 \;\leq\; \mu_2\|\nabla f(K)\|_F^2 \tag{6}$$

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

$$f(K - \alpha G) - f(K^\star) \leq \big(1 - \mu_f(a)\mu_1\alpha\big)\big(f(K) - f(K^\star)\big)$$

where $L_f(a)$ and $\mu_f(a)$ are the smoothness and PL parameters of the LQR objective function f over $\mathcal{S}(a)$.

To better understand the implications of Proposition 1, let us consider the trivial example $G = \nabla f(K)$. In this case, the inequalities in (6) hold with $\mu_1 = \mu_2 = 1$. Thus, for the stepsize $\alpha = 1/L_f(a)$, we recover the linear convergence rate of $1 - \mu_f(a)/L_f(a)$ that was established for the gradient descent method in Theorem 2.

In our convergence analysis, we do not show that $\overline{\nabla} f(K)$ obeys the approximate GD condition in Proposition 2 directly. Instead, we introduce an unbiased estimate $\widehat{\nabla} f(K)$ of the gradient $\nabla f(K)$ in Eq. (8) and establish the approximate GD condition for this estimate. We then show that the approximate gradient $\overline{\nabla} f(K)$ that is utilized in our algorithm remains close to this unbiased estimate. More specifically, we establish the following two key properties: first, for any $\epsilon > 0$, using a simulation time $\tau = O(\log(1/\epsilon))$ and an appropriate smoothing parameter r in Algorithm 1, the estimation bias $\|\widehat{\nabla} f(K) - \overline{\nabla} f(K)\|_F$ can be made smaller than ϵ; and, second, with $N = \tilde{O}(n)$ samples, the unbiased estimate $\widehat{\nabla} f(K)$ becomes well correlated with $\nabla f(K)$ with *high probability*. In particular, the events

$$\begin{aligned} \mathsf{M}_1 &:= \left\{ \left\langle \widehat{\nabla} f(K), \nabla f(K) \right\rangle \geq \mu_1 \|\nabla f(K)\|_F^2 \right\}, \\ \mathsf{M}_2 &:= \left\{ \|\widehat{\nabla} f(K)\|_F^2 \leq \mu_2 \|\nabla f(K)\|_F^2 \right\} \end{aligned} \tag{7}$$

occur with high probability for some positive scalars μ_1 and μ_2.

These two properties combined with Proposition 1 are the key ingredients that were used to analyze convergence of the random search method (RS) and prove Theorem 1. We next present main ideas that were used to establish these properties.

4.1 Controlling the bias

Herein, we define the unbiased estimate $\widehat{\nabla} f(K)$ of the gradient and quantify an upper bound on its distance to the output $\overline{\nabla} f(K)$ of Algorithm 1.To simplify our presentation, for any $K \in \mathbb{R}^{m \times n}$, we define the closed-loop Lyapunov operator $\mathcal{A}_K \colon \mathbb{S}^n \to \mathbb{S}^n$,

$$\mathcal{A}_K(X) := (A - BK)X + X(A - BK)^T$$

where $\mathbb{S}^n$ is the set of symmetric matrices. For $K \in \mathcal{S}$, both $\mathcal{A}_K$ and its adjoint

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

are invertible. Moreover, we can represent the matrices $X(K)$ and $P(K)$ in (4) as

$$X(K) = \mathcal{A}_K^{-1}(-\Omega), \quad P(K) = (\mathcal{A}_K^*)^{-1}(-K^T R K - Q).$$

For any $\tau \geq 0$ and $x_0 \in \mathbb{R}^n$, let $f_{x_0,\tau}(K)$ denote the τ-truncated version of the LQR objective function associated with system (1b) with the initial condition $x(0) = x_0$ and the feedback law $u = -Kx$ as defined in (3). For any $K \in \mathcal{S}$ and $x_0 \in \mathbb{R}^n$, the infinite horizon cost $f_{x_0}(K) := f_{x_0,\infty}(K)$ exists and it satisfies $f(K) = \mathbb{E}_{x_0}[f_{x_0}(K)]$. Furthermore, the gradient of $f_{x_0}(K)$ is given by (cf. (4c))

$$\nabla f_{x_0}(K) \;=\; 2(RK \;-\; B^T P(K))X_{x_0}(K), \quad X_{x_0}(K) := \mathcal{A}_K^{-1}(-x_0 x_0^T).$$

Since the gradients $\nabla f(K)$ and $\nabla f_{x_0}(K)$ are linear in $X(K)$ and $X_{x_0}(K)$, respectively, for the random initial condition $x(0) = x_0$ with $\mathbb{E}[x_0 x_0^T] = \Omega$, if follows that

$$\mathbb{E}_{x_0}[X_{x_0}(K)] \;=\; X(K), \quad \mathbb{E}_{x_0}[\nabla f_{x_0}(K)] \;=\; \nabla f(K).$$

Next, we define the following three estimates of the gradient

$$\begin{aligned}
\overline{\nabla} f(K) &:= \frac{1}{2rN} \sum_{i=1}^{N} \left(f_{x_i,\tau}(K + rU_i) - f_{x_i,\tau}(K - rU_i) \right) U_i \\
\widetilde{\nabla} f(K) &:= \frac{1}{2rN} \sum_{i=1}^{N} \left(f_{x_i}(K + rU_i) - f_{x_i}(K - rU_i) \right) U_i \\
\widehat{\nabla} f(K) &:= \frac{1}{N} \sum_{i=1}^{N} \left\langle \nabla f_{x_i}(K), U_i \right\rangle U_i
\end{aligned} \tag{8}$$

Here, $U_i \in \mathbb{R}^{m\times n}$ are i.i.d. random matrices with $\mathrm{vec}(U_i)$ uniformly distributed on the sphere $\sqrt{mn}\, S^{mn-1}$ and $x_i \in \mathbb{R}^n$ are i.i.d. random initial conditions sampled from distribution $\mathcal{D}$. Note that $\widetilde{\nabla} f(K)$ is the infinite horizon version of $\overline{\nabla} f(K)$ produced by Algorithm 1 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 $\mathbb{E}_{U_1}[\mathrm{vec}(U_1)\mathrm{vec}(U_1)^T] = I$ and $\mathbb{E}_{x_i,U_i}\left[\mathrm{vec}(\widehat{\nabla} f(K))\right] = \mathbb{E}_{U_1}[\langle \nabla f(K), U_1\rangle\, \mathrm{vec}(U_1)] = \mathrm{vec}(\nabla f(K))$.

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 $\overline{\nabla} f(K)$ and $\widehat{\nabla} f(K)$ 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 $\|U\|_F \leq \sqrt{mn}$. Herein, 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 is provided in Lemma 1.

Lemma 1 ([4]) *For any $K \in \mathcal{S}(a)$ and $U \in \mathbb{R}^{m\times n}$ with $\|U\|_F \leq \sqrt{mn}$, $K + r(a)U \in \mathcal{S}(2a)$, 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 $r \leq r(a)$ in Lemma 1, $\widetilde{\nabla} f(K)$ and $\widehat{\nabla} f(K)$ are well defined since the feedback gains $K + rU_i$ are all stabilizing. We next present 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 $\nabla f(K)$. This can be accomplished by bounding the

difference between these two quantities and $\widetilde{\nabla} f(K)$ through the use of the triangle inequality

$$\|\widehat{\nabla} f(K) - \overline{\nabla} f(K)\|_F \;\leq\; \|\widetilde{\nabla} f(K) - \overline{\nabla} f(K)\|_F \;+\; \|\widehat{\nabla} f(K) - \widetilde{\nabla} f(K)\|_F. \qquad (9)$$

Proposition 2 provides an upper bound on each term on the right-hand side of (9).

Proposition 2 ([4]) *For any $K \in \mathcal{S}(a)$ and $r \leq r(a)$, where $r(a)$ is given by Lemma 1,*

$$\|\widetilde{\nabla} f(K) - \overline{\nabla} f(K)\|_F \;\leq\; \frac{\sqrt{mn}\,\max_i \|x_i\|^2}{r}\,\kappa_1(2a)\,\mathrm{e}^{-\tau/\kappa_2(2a)}$$

$$\|\widehat{\nabla} f(K) - \widetilde{\nabla} f(K)\|_F \;\leq\; \frac{(rmn)^2}{2}\,\ell(2a)\max_i \|x_i\|^2$$

where $\ell(a) > 0$, $\kappa_1(a) > 0$, and $\kappa_2(a) > 0$ are polynomials of degree less than 5.

The first term on the right-hand side of (9) corresponds to a bias arising from the finite-time simulation. Proposition 2 shows that although small values of r may yield 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 bound on the estimation error in terms of r. It is also worth mentioning that the third-derivative of the function $f_{x_0}(K)$ is utilized in obtaining the second inequality.

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

In this section, we show that the events M_i in (7) hold with high probability. The key enabler of the proof is that the random inner product $\left\langle \nabla f(K), \widehat{\nabla} f(K) \right\rangle$ is very well concentrated around its mean $\|\nabla f(K)\|_F^2$. We next describe this phenomena in more detail. The proof exploits the problem structure to confine the dependence of $\widehat{\nabla} f(K)$ on the random initial conditions x_i into a zero-mean random vector. In particular, for any given feedback gain $K \in \mathcal{S}$ and initial condition $x_0 \in \mathbb{R}^n$, we have $\nabla f(K) = EX$, and $\nabla f_{x_0}(K) = EX_{x_0}$, where $E := 2(RK - B^T P(K)) \in \mathbb{R}^{m\times n}$ is a fixed matrix, $X = -\mathcal{A}_K^{-1}(\Omega)$, and $X_{x_0} = -\mathcal{A}_K^{-1}(x_0 x_0^T)$. Thus, we can represent the gradient estimate $\widehat{\nabla} f(K)$ as

$$\widehat{\nabla} f(K) \;=\; \frac{1}{N}\sum_{i=1}^{N} \left\langle EX_{x_i}, U_i \right\rangle U_i \;=\; \widehat{\nabla}_1 \;+\; \widehat{\nabla}_2$$

where $\widehat{\nabla}_1 := \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \left\langle E(X_{x_i} - X), U_i \right\rangle U_i$ and $\widehat{\nabla}_2 := \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \langle \nabla f(K), U_i \rangle\, U_i$. Note that $\widehat{\nabla}_2$ does not depend on the initial conditions x_i. Moreover, from $\mathbb{E}[X_{x_i}] = X$ and the independence of X_{x_i} and U_i, we have $\mathbb{E}[\widehat{\nabla}_1] = 0$ and $\mathbb{E}[\widehat{\nabla}_2] = \nabla f(K)$.

We next present the key technical results that were used to study the probability of the events M_i in (7) for suitable values of μ_1 and μ_2. These results are obtained

using standard tools from non-asymptotic statistical analysis of the concentration of random variables around their average; see a recent book [12] for a comprehensive discussion. Herein, we use c, c', c'', etc. to denote positive absolute constants.

Proposition 3 can be used to show that with enough samples $N = \tilde{O}(n)$, the inner product of the zero-mean term $\widehat{\nabla}_1$ and the gradient $\nabla f(K)$ can be controlled with high probability. This result is the key for analyzing the probability of the event M_1.

Proposition 3 ([4]) *Let $X_1, \ldots, X_N \in \mathbb{R}^{n\times n}$ be i.i.d. random matrices distributed according to $\mathcal{M}(xx^T)$, where $x \in \mathbb{R}^n$ is a random vector whose distribution obeys Assumption 1 and $\mathcal{M}$ is a linear operator, and let $X := \mathbb{E}[X_1] = \mathcal{M}(I)$. Also, let $U_1, \ldots, U_N \in \mathbb{R}^{m\times n}$ be i.i.d. random matrices with $\mathrm{vec}(U_i)$ uniformly distributed on the sphere $\sqrt{mn}\, S^{mn-1}$. For any $E \in \mathbb{R}^{m\times n}$ and positive scalars δ and β,*

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}\left\langle E\left(X_i - X\right), U_i\right\rangle \left\langle EX, U_i\right\rangle\right| \leq \delta\|EX\|_F\|E\|_F\right\} \geq 1 - C'N^{-\beta} - 4N\mathrm{e}^{-\frac{n}{8}}$$

if $N \geq C(\beta^2\kappa^2/\delta)^2\left(\|\mathcal{M}^*\|_2 + \|\mathcal{M}^*\|_S\right)^2 n\log^6 n$, where $\|\cdot\|_S$ denotes the spectral induced norm.

The proof of Proposition 3 exploits the Hanson-Wright inequality along with a well-known upper bound on the norm of random matrices [13, Theorems 1.1 and 3.2]. In Proposition 4, we present a technical result that can be used to show that $\left\langle \widehat{\nabla}_2, \nabla f(K)\right\rangle$ concentrates with high probability around its average $\|\nabla f(K)\|_F^2$.

Proposition 4 ([4]) *Let $U_1, \ldots, U_N \in \mathbb{R}^{m\times n}$ be i.i.d. random matrices with each $\mathrm{vec}(U_i)$ uniformly distributed on the sphere $\sqrt{mn}\, S^{mn-1}$. Then, for any $W \in \mathbb{R}^{m\times n}$ and scalar $t \in (0, 1]$, we have*

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N}\langle W, U_i\rangle^2 \;<\; (1-t)\|W\|_F^2\right\} \;\leq\; 2\,\mathrm{e}^{-cNt^2}.$$

The proof of Proposition 3 relies on the Bernstein inequality [12, Corollary 2.8.3]. Using Propositions 3 and 4, it is straightforward to show that the event M_1 occurs with high probability.

Next, we turn our attention to quantifying the probability of the event M_2 in (7). Proposition 5 presents a technical result that can be used to quantify a high probability upper bound on $\|\widehat{\nabla}_1\|_F/\|\nabla f(K)\|$. This result is analogous to Proposition 3 and it can be used to study the event M_2.

Proposition 5 ([4]) *Let X_i and U_i with $i = 1, \ldots, N$ be random matrices defined in Lemma 3, $X := \mathbb{E}[X_1]$, and let $N \geq c_0 n$. For any $E \in \mathbb{R}^{m\times n}$ and $\beta > 0$, we have*

$$\frac{1}{N}\left\|\sum_{i=1}^{N}\left\langle E\left(X_i - X\right), U_i\right\rangle U_i\right\|_F \;\leq\; c_1\beta\,\kappa^2(\|\mathcal{M}^*\|_2 + \|\mathcal{M}^*\|_S)\|E\|_F\sqrt{mn}\;\log n$$

with probability not smaller than $1 - c_2(n^{-\beta} + N\mathrm{e}^{-\frac{n}{8}})$.

In Proposition 6, we present a technical result that can be used to study $\|\widehat{\nabla}_2\|_F / \|\nabla f(K)\|$.

Proposition 6 ([4]) *Let $U_1, \ldots, U_N \in \mathbb{R}^{m\times n}$ be i.i.d. random matrices with* $\mathrm{vec}(U_i)$ *uniformly distributed on the sphere $\sqrt{mn}\, S^{mn-1}$ and let $N \geq Cn$. Then, for any $W \in \mathbb{R}^{m\times n}$,*

$$\mathbb{P}\left\{ \frac{1}{N} \| \sum_{j=1}^{N} \left\langle W, U_j \right\rangle U_j \|_F \; > \; C' \sqrt{m} \|W\|_F \right\} \; \leq \; 2N \mathrm{e}^{-\frac{mn}{8}} \; + \; 2\mathrm{e}^{-\hat{c}N}.$$

Using Propositions 5 and 6, it is straightforward to show that the event M_2 occurs with high probability.

5 An example

We consider a mass-spring-damper system with $s = 10$ masses, where we set all mass, spring, and damping constants to unity. In state-space representation (1b), the state $x = [\, p^T \; v^T \,]^T$ contains the position and velocity vectors and the dynamic and input matrices are given by

$$A \; = \; \begin{bmatrix} 0 & I \\ -T & -T \end{bmatrix}, \;\; B \; = \; \begin{bmatrix} 0 \\ I \end{bmatrix}$$

where 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 first super and sub-diagonals, and zero elsewhere. In this example, the A-matrix is Hurwitz and the objective of control is to optimize the LQR cost with Q and R equal to identity. We also let the initial conditions x_i in Algorithm 1 be standard normal and use $N = n = 2s$ samples.

For two values of the smoothing parameter $r = 10^{-4}$ (blue) and $r = 10^{-5}$ (red), and for $K = 0$, we illustrate in Figure 1(a) the dependence of the relative error $\|\widehat{\nabla} f(K) - \overline{\nabla} f(K)\|_F / \|\widehat{\nabla} f(K)\|_F$, and in Figure 1(b), that of the total relative error $\|\nabla f(K) - \overline{\nabla} f(K)\|_F / \|\nabla f(K)\|_F$ on the simulation time τ. In Figure 1(a), we observe an exponential decrease in error for small values of τ. In addition, the error does not pass a saturation level which is determined by r. We also see that, as r decreases, this saturation level becomes smaller. These observations are in harmony with the theoretical developments presented in this chapter; in particular, Proposition 2 coupled with the triangle inequality yield

$$\|\widehat{\nabla} f(K) \; - \; \overline{\nabla} f(K)\|_F \; \leq \; \left(\frac{\sqrt{mn}}{r} \, \kappa_1(2a) \, \mathrm{e}^{-\kappa_2(2a)\tau} \; + \; \frac{r^2 m^2 n^2}{2} \, \ell(2a) \right) \max_i \|x_i\|^2.$$

This upper bound clearly captures the exponential dependence of the bias on the simulation time τ as well as the saturation level that depends quadratically on the smoothing parameter r.

On the other hand, in Figure 1(b), we observe that the distance between the approximate gradient $\overline{\nabla} f(K)$ and the true gradient is rather large. In contrast to the existing results which rely on the use of the estimation error shown in Figure 1(b),

Proposition 2 shows that the simulated gradient $\overline{\nabla} f(K)$ is close to the gradient estimate $\widehat{\nabla} f(K)$, which although is not close to the true gradient $\nabla f(K)$, is highly correlated with it. This is sufficient for establishing convergence guarantees and reducing both sample complexity and simulation time to $O(\log(1/\epsilon))$.

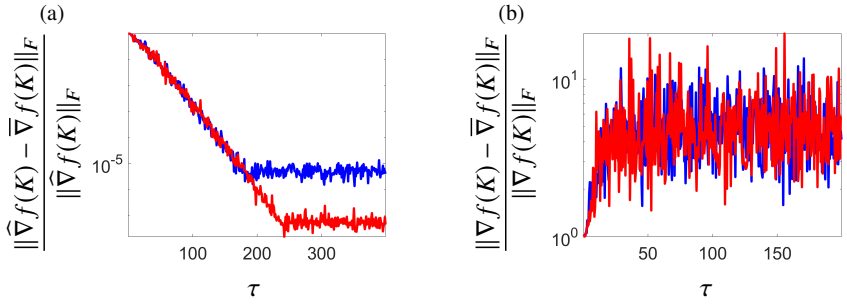

Fig. 1 (a) Bias in gradient estimation and (b) total error in gradient estimation as functions of the simulation time τ. The blue and red curves correspond to two values of the smoothing parameter $r = 10^{-4}$ and $r = 10^{-5}$, respectively.

Finally, Figure 2 demonstrates linear convergence of the random search method (RS) with stepsize $\alpha = 10^{-4}$, and ($r = 10^{-5}$, $\tau = 200$) in Algorithm 1, as established in Theorem 1.

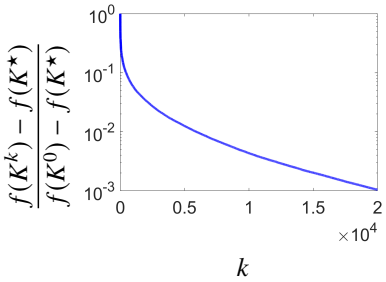

Fig. 2 Convergence curve of the random search method (RS).

6 Thoughts and outlook

For the *discrete-time* LQR problem, global convergence guarantees were provided in [1] for gradient decent and the random search methods with one-point gradient estimates. Furthermore, a bound on the sample complexity for reaching the error tolerance ϵ that requires a number of function evaluations that is at-least proportional to $(1/\epsilon^4)\log(1/\epsilon)$ was established. If one has access to the infinite horizon cost values, i.e., if $\tau = \infty$, the number of function evaluations for the random search method with one-point estimates was improved to $1/\epsilon^2$ in [8]. Moreover, this work showed that the use of two-point estimates reduces the number of function evaluations to $1/\epsilon$.

In this chapter, we focus on the *continuous-time* LQR problem and summarize the results presented in [4, 10, 14, 15]. These recent references demonstrate that the random search method with two-point gradient estimates converges to the optimal solution at a linear rate with high probability. Relative to the existing literature, a significant improvement is offered both in terms of the required function evaluations and simulation time. Specifically, the total number of function evaluations required to achieve an accuracy level ϵ is proportional to $\log(1/\epsilon)$ compared to at least $(1/\epsilon^4)\log(1/\epsilon)$ in [1] and $1/\epsilon$ in [8]. Similarly, the simulation time required to achieve an accuracy level ϵ is proportional to $\log(1/\epsilon)$; this is in contrast to [1] which requires poly$(1/\epsilon)$ simulation time and [8] which assumes an infinite simulation time. We refer the reader to [4] for a comprehensive discussion along with all technical details that are omitted here for brevity.

References

1. 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.
2. J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, "LQR through the lens of first order methods: Discrete-time case," 2019, arXiv:1907.08921.
3. H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović, "On the linear convergence of random search for discrete-time LQR," *IEEE Control Syst. Lett.*, 2020, in press; doi:10.1109/LCSYS.2020.3006256.
4. 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, conditionally accepted; also arXiv:1912.11899.
5. H. Mania, A. Guy, and B. Recht, "Simple random search provides a competitive approach to reinforcement learning," in *Proc. Neural Information Processing (NeurIPS)*, 2018.
6. B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 2, pp. 253–279, 2019.
7. 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, 2015.
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," *J. Mach. Learn. Res.*, vol. 51, pp. 1–51, 2020.
9. W. S. Levine and M. Athans, "On the determination of the optimal constant output feedback gains for linear multivariable systems," *IEEE Trans. Automat. Control*, vol. 15, no. 1, pp. 44–48, 1970.
10. 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.
11. H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition," in *In European Conference on Machine Learning*, 2016, pp. 795–811.
12. R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018, vol. 47.
13. M. Rudelson and R. Vershynin, "Hanson-Wright inequality and sub-Gaussian concentration," *Electron. Commun. Probab.*, vol. 18, 2013.
14. H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović, "Random search for learning the linear quadratic regulator," in *Proceedings of the 2020 American Control Conference*, Denver, CO, 2020, pp. 4798–4803.
15. H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović, "Learning the model-free linear quadratic regulator via random search," in *Proceedings of Machine Learning Research, 2nd Annual Conference on Learning for Dynamics and Control*, vol. 120, Berkeley, CA, 2020, pp. 1–9.