Learning the model-free linear quadratic regulator via random search

Hesameddin Mohammadi HESAMEDM@USC.COM

Mahdi Soltanolkotabi SOLTANOL@USC.COM

Mihailo R. Jovanović MIHAILO@USC.COM

Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089.

Abstract

Model-free reinforcement learning techniques attempt to find an optimal control action for an unknown dynamical system by directly searching over the parameter space of controllers. The convergence behavior and statistical properties of these approaches are often poorly understood because of the nonconvex nature of the underlying optimization problems as well as the lack of exact gradient computation. In this paper, we examine the standard infinite-horizon linear quadratic regulator problem for continuous-time systems with unknown state-space parameters. We provide theoretical bounds on the convergence rate and sample complexity of a random search method. Our results demonstrate that the required simulation time for achieving ϵ-accuracy in a model-free setup and the total number of function evaluations are both of $O(\log{(1/\epsilon)})$.

Keywords: Data-driven control, linear quadratic regulator, model-free control, nonconvex optimization, random search method, reinforcement learning, sample complexity.

1. Introduction

In many emerging applications, control-oriented models are not readily available and classical approaches from optimal control may not be directly applicable. This challenge has led to the emergence of Reinforcement Learning (RL) approaches that often perform well in practice. Examples include learning complex locomotion tasks via neural network dynamics (Nagabandi et al., 2018) and playing Atari games based on images using deep-RL (Mnih et al., 2013).

RL approaches can be broadly divided into model based (Dean et al., 2017; Simchowitz et al., 2018) and model free (Bertsekas, 2011; Abbasi-Yadkori et al., 2019). While model-based RL uses data to obtain approximations of the underlying dynamics, its model-free counterpart prescribes control actions based on estimated values of a cost function without attempting to identify a model. In spite of the empirical success of RL in a variety of domains, the mathematical understanding of these techniques is still in its infancy. Because of the interactive and nonconvex nature of these algorithms, fundamental questions surrounding convergence and sample complexity remain unanswered even for classical control problems, including the linear quadratic regulator (LQR). In this paper, we take a step towards addressing such challenges with a focus on the infinite-horizon LQR problem for continuous-time systems.

The globally optimal solution of the LQR problem can be obtained by solving the Riccati equation. However, computing the optimal solution becomes challenging for large-scale problems, when prior knowledge is not available, or in the presence of structural constraints on the controller. This motivates the use of direct search methods for controller synthesis. Unfortunately, the nonconvex nature of this formulation complicates the analysis of first- and second-order optimization algorithms. To make matters worse, structural constraints on the feedback gain matrix may result in a disjoint search landscape limiting the utility of conventional descent-based methods (Ackermann, 1980). Furthermore, in the model-free setting the exact model (and hence the gradient of the loss) is unknown so that only zero-order methods can be used to estimate the gradient.

Recent reference (Mohammadi et al., 2019a) showed that the gradient descent method on the state-feedback gain can indeed solve the *continuous-time* LQR problem with a linear convergence rate. While computing the exact value of gradient requires the system dynamics to be known, this convergence result motivates the use of a random search method for the model-free setting. Random search is perhaps the simplest model-free approach to RL which attempts to emulate the behavior of gradient descent using gradient approximations obtained from function values. It has been used to solve benchmark control problems such as locomotion tasks, matching state-of-the-art sample efficiency (Mania et al., 2018). However, even for the standard LQR problem, many open theoretical questions surround convergence properties and sample complexity of this method.

For the *discrete-time* LQR problem, global convergence guarantees were recently provided in Fazel et al. (2018) for gradient decent and the random search method with one-point gradient estimates. The authors established 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 $(1/\epsilon^4)\log(1/\epsilon)$. If one has access to the infinite-horizon cost values, the number of function evaluations for the random search method with one-point estimates was improved to $1/\epsilon^2$ in Malik et al. (2019). Moreover, this work showed that the use of two-point estimates reduces the number of function evaluations to $1/\epsilon$.

In this paper, we focus on the *continuous-time* LQR problem, and establish 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, we also offer a significant improvement both in terms of the required function evaluations and simulation time. Specifically, the total number of function evaluations required in our results to achieve an accuracy of ϵ is proportional to $\log(1/\epsilon)$ compared to at least $(1/\epsilon^4)\log(1/\epsilon)$ in Fazel et al. (2018) and $1/\epsilon$ in Malik et al. (2019). Similarly, the simulation time required in our results to achieve an accuracy of ϵ is proportional to $\log(1/\epsilon)$; this is in contrast to Fazel et al. (2018) which requires $\operatorname{poly}(1/\epsilon)$ simulation time and Malik et al. (2019) which assumes an infinite simulation time. We only present the main result and a sketch of the proof and refer the reader to our more recent work Mohammadi et al. (2019b) for details.

2. 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, and $x(0)$ is a random initial condition with distribution $\mathcal{D}$. The quadratic performance index 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 and the expectation is taken over $x(0)$. For a controllable pair (A, B), the solution to LQR problem (1) is given by $u(t) = -K^\star x(t) = -R^{-1}B^TP^\star x(t)$, where $P^\star$ is the unique positive definite solution to the algebraic Riccati equation, $A^T P^\star + P^\star A + Q - P^\star B R^{-1} B^T P^\star = 0$. Unfortunately, this approach is not directly applicable when the matrices A and B are not known. Exploiting the linearity of the optimal controller, we can alternatively formulate the LQR problem as a direct search for the optimal feedback gain, namely

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

where

$$f(K) \;:=\; \begin{cases} \operatorname{trace}\big((Q + K^T R K) X(K)\big), & K \in \mathcal{S} \\ \infty, & \text{otherwise.} \end{cases} \tag{2b}$$

Here, $f(K)$ determines the objective function in (1b) associated with the feedback law $u = -Kx$, $\mathcal{S} := \{K \in \mathbb{R}^{m \times n} \,|\, A - BK \text{ is Hurwitz}\}$ is the set of stabilizing feedback gains, and for any $K \in \mathcal{S}$, the matrix $X(K)$ is determined by

$$X(K) \;=\; \int_0^\infty \mathrm{e}^{(A\,-\,BK)t}\, \Omega\, \mathrm{e}^{(A\,-\,BK)^T t}\, \mathrm{d}t \tag{2c}$$

where $\Omega := \mathbb{E}[x(0)x^T(0)]$ is the covariance matrix of the zero-mean initial condition $x(0)$. Since the optimal feedback gain $K^\star$ does not depend on $x(0)$, without loss of generality, we assume $\Omega \succ 0$. In (2), K is the optimization variable and $(A, B, Q \succ 0, R \succ 0, \Omega \succ 0)$ are the problem parameters.

The formulation of the LQR problem given by (2) has been studied for both continuous-time (Anderson and Moore, 1990; Mohammadi et al., 2019a) and discrete-time systems (Fazel et al., 2018; Bu et al., 2019). In this paper, we study the convergence of a local-search method based on random sampling (Mania et al., 2018; Recht, 2019) for solving problem (2) that does not require knowledge of system matrices A and B. At each iteration in Algorithm 1, we form an empirical approximation $\overline{\nabla} f(K)$ to the gradient of the objective function via simulation of system (1a) for randomly perturbed feedback gains $K \pm U_i$ with $i = 1, \ldots, N$ and update 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 use system matrices A and B in (1a); only access to a simulation engine is required.

3. Main result

Our analysis of the convergence of the random search method (RS) exploits two key properties of the LQR objective function f: smoothness and the Polyak-Łojasiewicz (PL) condition over its sublevel sets $\mathcal{S}(a) := \{K \in \mathcal{S} | f(K) \le a\}$, where a is a positive scalar. In particular, for all feedback gains K and K' such that the line segment between them belongs to $\mathcal{S}(a)$, the function f satisfies

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

PL condition: $\quad f(K) \,-\, f(K^\star) \;\le\; \dfrac{1}{2\mu_f(a)}\, \|\nabla f(K)\|_F^2$

where the smoothness and PL parameters $L_f(a)$ and $\mu_f(a)$ are positive rational functions of a. This result holds for both continuous-time (Mohammadi et al., 2019a; Fatkhullin and Polyak, 2020) and discrete-time (Fazel et al., 2018; Bu et al., 2019) LQR problems. We also make the following assumption on the statistical properties of the initial condition.

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 $\tau$, 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 $x_i$ from distribution $\mathcal{D}$.
        – For $j \in \{1, 2\}$, simulate system (1a) up to time $\tau$ with the feedback gain $K_{i,j}$ and initial
        condition $x(0) = x_i$ to form $\hat{f}_{i,j} = \int_0^\tau (x^T(t) Q x(t) + u^T(t) R u(t))\, \mathrm{d}t$.
    end for
Output: The gradient estimate $\overline{\nabla} f(K) := \frac{1}{2rN} \sum_{i=1}^{N} \left(\hat{f}_{i,1} - \hat{f}_{i,2}\right) U_i$.
```

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, i.e., for a random vector $v \in \mathbb{R}^n$ distributed according to $\mathcal{D}$, $\mathbb{E}[v_i] = 0$ and $\|v_i\|_{\psi_2} \le \kappa$, for some constant κ and $i = 1, \ldots, n$.*

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*

$$\tau \;\ge\; \theta'(a) \log(1/\epsilon) \quad \text{and} \quad N \;\ge\; c\left(1 + \beta^4 \kappa^4\, \theta(a) \log^6 N\right) n$$

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 $K^0 \in \mathcal{S}(a)$, random search 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

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

iterations. This holds with probability at least $1 - c'k(n^{-\beta} + N^{-\beta} + N\mathrm{e}^{-\frac{n}{8}} + \mathrm{e}^{-c'N})$. Here, $\omega(a) := c''\left(\sqrt{m} + \beta\kappa^2\theta(a)\sqrt{mn}\log n\right)^2$, 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 (Fazel et al., 2018; Malik et al., 2019) that require $O(1/\epsilon)$ function evaluations and $\mathrm{poly}(1/\epsilon)$ simulation time.

4. Proof sketch

The random search method in (RS) updates the iterates using gradient estimates obtained via Algorithm 1. For any stabilizing feedback gain $K \in \mathcal{S}$, we have (Lin et al., 2013),

$$\nabla f(K) \;=\; 2\left(R K - B^T P(K)\right) X(K) \tag{3}$$

where $X(K)$ is given by (2c) and

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

Although nonconvex, the smoothness and PL property of the objective function were utilized to prove linear convergence of gradient descent for *continuous-time* systems (Mohammadi et al., 2019a). We note that similar analysis was used to show convergence of gradient descent for LQR problem for *discrete-time* systems (Fazel et al., 2018).

In the random search method, we do not have access to the gradient and $\overline{\nabla} f(K)$ is a biased estimate of $\nabla f(K)$. According to Fazel et al. (2018), achieving $\|\overline{\nabla} f(K) - \nabla f(K)\|_F \leq \epsilon$ may take $N = \Omega(1/\epsilon^4)$ samples. Because of this poor sample complexity, we take an alternative route and give up on the objective of controlling the gradient estimation error. By exploiting the problem structure, we show that with a linear number of samples $N = \tilde{O}(n)$, where n is the number of states, the estimate $\overline{\nabla} f(K)$ concentrates with *high probability* when projected to the direction of $\nabla f(K)$.

In what follows, we first control the bias by establishing 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,

$$\|\overline{\nabla} f(K) \;-\; \widehat{\nabla} f(K)\|_F \;\leq\; \epsilon \tag{5}$$

where the definition of $\widehat{\nabla} f(K)$ is given in Eq. (7). We then show that by choosing a large number of samples N, our unbiased estimate $\widehat{\nabla} f(K)$ becomes highly correlated with the gradient. In particular, we show that the following two events

$$\mathsf{M}_1 \;:=\; \left\{\left\langle \widehat{\nabla} f(K), \nabla f(K)\right\rangle \geq \mu_1 \|\nabla f(K)\|_F^2\right\},\; \mathsf{M}_2 \;:=\; \left\{\|\widehat{\nabla} f(K)\|_F^2 \leq \mu_2 \|\nabla f(K)\|_F^2\right\} \tag{6}$$

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

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

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

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

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

4.1. 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 given by Eq. (5). To simplify our presentation, for any $K \in \mathbb{R}^{m\times n}$, we define the closed-loop Lyapunov operator $\mathcal{A}_K : \mathbb{S}^n \to \mathbb{S}^n$ as

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

where $\mathbb{S}^n \subset \mathbb{R}^{n\times n}$ 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 and we have $X(K) = \mathcal{A}_K^{-1}(-\Omega)$ and $P(K) = (\mathcal{A}_K^*)^{-1}(-K^T RK - Q)$.

For any $\tau \geq 0$ and $x_0 \in \mathbb{R}^n$, let us define the τ-truncated version of the LQR objective function

$$f_{x_0,\tau}(K) \;:=\; \int_0^\tau \big(x^T(t)Qx(t) + u^T(t)Ru(t)\big)\,\mathrm{d}t$$

associated with system (1a) with the initial condition $x(0) = x_0$ and feedback law $u = -Kx$. 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. (3))

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

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 $\mathbb{E}[x_0 x_0^T] = \Omega$ the linearity of $\mathcal{A}_K$ implies

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

Next we define the following three estimates of the gradient

$$\begin{aligned}
\overline{\nabla} f(K) \;&:=\; \frac{1}{2rN}\sum_{i=1}^N \left(f_{x_i,\tau}(K+rU_i) - f_{x_i,\tau}(K-rU_i)\right) U_i \\
\widetilde{\nabla} f(K) \;&:=\; \frac{1}{2rN}\sum_{i=1}^N \left(f_{x_i}(K+rU_i) - f_{x_i}(K-rU_i)\right) U_i \qquad (7)\\
\widehat{\nabla} f(K) \;&:=\; \frac{1}{N}\sum_{i=1}^N \langle \nabla f_{x_i}(K), U_i\rangle\, U_i.
\end{aligned}$$

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 $\overline{\nabla} f(K)$ of 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 $\widetilde{\nabla} f(K)$ 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 $\|U\|_F \leq \sqrt{mn}$. In this paper, 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 *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, $\widetilde{\nabla} f(K)$ is well defined since the feedback gains $K + rU_i$ 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 $\widetilde{\nabla} f(K)$ through the use of the triangle inequality

$$\|\widehat{\nabla} f(K) \,-\, \overline{\nabla} f(K)\|_F \;\leq\; \|\widetilde{\nabla} f(K) \,-\, \overline{\nabla} f(K)\|_F \,+\, \|\widehat{\nabla} f(K) \,-\, \widetilde{\nabla} f(K)\|_F. \tag{8}$$

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 $r \leq r(a)$, 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 (8) 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 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 $\nabla f(K)$

To show that the events M_i in (6) hold with high probability, we exploit the problem structure to isolate 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 $X_{x_0} = -\mathcal{A}_K^{-1}(x_0 x_0^T)$. This allows us to write

$$\widehat{\nabla} f(K) \;=\; \frac{1}{N} \sum_{i\,=\,1}^{N} \langle E X_{x_i}, U_i \rangle\, U_i \;=\; \frac{1}{N} \sum_{i\,=\,1}^{N} \langle E(X_{x_i} - X), U_i \rangle\, U_i \;+\; \frac{1}{N} \sum_{i\,=\,1}^{N} \langle \nabla f(K), U_i \rangle\, U_i.$$

It is now easy to verify that $\mathbb{E}[\langle E(X_{x_i} - X), U_i \rangle\, U_i] = 0$ and $\mathbb{E}[\langle \nabla f(K), U_i \rangle\, U_i] = \nabla f(K)$. We next present our key technical result in Proposition 3 that allows us to show that with enough samples $N = \tilde{O}(n)$, the inner product of the zero-mean term 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 in (6). The analysis for the event M_2 follows similar steps; see Mohammadi et al. (2019b) for details.

Proposition 3 *Let $X_1, \ldots, X_N \in \mathbb{R}^{n \times n}$ be i.i.d. random matrices 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* $\delta, \beta > 0$,

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

if $N \geq C\beta^4\kappa^4(n/\delta^2)\log^6 N\left(\|\mathcal{M}^*\|_2 + \|\mathcal{M}^*\|_S\right)^2$, *where* $\|\cdot\|_S$ *denotes the spectral induced norm.*

5. Computational experiments

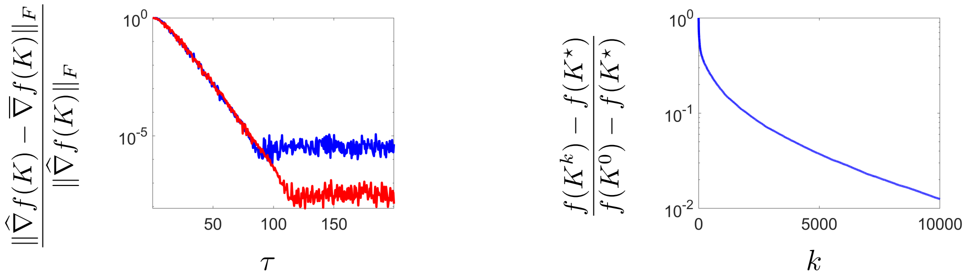
We consider a mass-spring-damper system with $s = 10$ masses, where we set all spring and damping constants as well as masses to unity. In state-space representation (1a), the state vector $x = [\, p^T \;\; v^T \,]^T$ contains the position and velocity of masses and the dynamic and input matrices are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

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

where 0 and I are $s \times s$ zero and identity matrices, and T is a Toeplitz matrix with 2 on the main diagonal and -1 on the first upper and lower sub-diagonals. 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.

Figure 1(*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 $K = 0$ and two values of smoothing parameter $r = 10^{-4}$ (blue) and $r = 10^{-5}$ (red). 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 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.

Figure 1(*b*) illustrates the convergence curve of the random search method (RS) with stepsize $\alpha = 10^{-4}$, where the parameters of Algorithm 1 are $r = 10^{-5}$ and $\tau = 10$. This figure demonstrates linear convergence for (RS), as established in Theorem 1.



(*a*) (*b*)

Figure 1: (a) The bias in gradient estimation as a function of the simulation time τ; (b) the convergence curve of the random search method (RS).

References

Y. Abbasi-Yadkori, N. Lazic, and C. Szepesvári. Model-free linear quadratic control via reduction to expert prediction. In *Proc. Mach. Learn. Res.*, volume 89, pages 3108–3117. PMLR, 2019.

J. Ackermann. Parameter space design of robust control systems. *IEEE Trans. Automat. Control*, 25 (6):1058–1072, 1980.

B. D. O. Anderson and J. B. Moore. *Optimal Control; Linear Quadratic Methods*. Prentice Hall, New York, NY, 1990.

D. Bertsekas. Approximate policy iteration: A survey and some new methods. *J. Control Theory Appl.*, 9(3):310–335, 2011.

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

S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu. On the sample complexity of the linear quadratic regulator. *Found. Comput. Math.*, pages 1–47, 2017.

I. Fatkhullin and B. Polyak. Optimizing static linear feedback: gradient method. 2020. arXiv:2004.09875.

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*, pages 1467–1476, 2018.

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*, 58(9):2426–2431, 2013.

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 *AISTATS: Conference on AI and Statistics*, 2019.

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

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.

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*, pages 7474–7479, Nice, France, 2019a.

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*, 2019b. submitted; also arXiv:1912.11899.

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.*, pages 7559–7566, 2018.

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

M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Proc. Mach. Learn. Res.*, pages 439–473, 2018.