Extremum-Seeking Policy Iteration for Data-Driven LQR

# Extremum-Seeking Policy Iteration for Data-Driven LQR

Guido Carnevale, Nicola Mimmo, Giuseppe Notarstefano

***Abstract*— In this paper, we propose a data-driven strategy to iteratively find the state feedback gain matrix solving a Linear Quadratic Regulator (LQR) problem in a model-free fashion, i.e., under unknown system and cost matrices. In our setup, we assume that, at each iteration, an oracle provides the LQR cost of the tentative policy, e.g., by running the system or a simulator. Based on this information, we develop an algorithm based on Extremum-Seeking to iteratively refine our tentative solution without any additional knowledge on the system and cost models. By using a Lyapunov-based approach exploiting averaging theory for time-varying systems, we show that the proposed algorithm exponentially converges to an arbitrarily small ball containing the optimal gain matrix. We corroborate the theoretical results by testing the proposed strategy via numerical simulations.**

## I. Introduction

This paper focuses on infinite-horizon Linear Quadratic Regulator (LQR) problems from a data-driven perspective. In this context, iterative methods based on the Kleinman algorithm [1] are given in [2]–[6]. Moreover, the paper [7] investigates an off-policy Q-learning strategy, while the works [8]–[10] develop iterative methods without assuming stabilizing properties of the initial policy. Finally, the work [11] proposes a model-free approach based on reinforcement learning.

A popular approach is given by the so-called direct methods, i.e., the off-policy strategies that use data in the policy design phase, see the pioneering works [12]–[14]. In this area, some extensions are given in [15] to deal with unknown switching linear systems, and in [16] and [17] to deal with data corrupted by noise.

Our work proposes a novel scheme inspired by the so-called policy-gradient methods, see, e.g., the related survey [18] and the works [19], [20], where the convergence properties of (policy-) gradient methods for discrete-time LQR are studied.

More in detail, since in our model-free setup the policy-gradient method is not implementable because the model is unknown, we resort to an Extremum-Seeking technique, see, e.g., the works [21]–[25].

The main contribution of the paper is the development of EXtremum-seeking Policy iteration LQR (EXP-LQR), a novel data-driven strategy based on Extremum-Seeking for LQR problems. Our method assumes the presence of an oracle providing the cost associated to the current policy. With this information at hand, we iteratively improve the policy through a mechanism based on Extremum-Seeking. We analyze the resulting algorithm by interpreting it as a nonlinear time-varying system. Specifically, we focus on the associated auxiliary dynamics named *averaged system* whose stability is investigated with a Lyapunov-based approach. Then, by relying on *averaging theory*, we leverage this result to ensure the exponential convergence of our algorithm to an arbitrarily small ball containing the optimal gain matrix. In particular, this last step relies on Theorem 1, introduced in Section II, which provides averaging-related stability results for generic discrete-time systems. To the best of the authors' knowledge, the result of Theorem 1 represents a side contribution of the work.

The paper is organized as follows. Section II introduces some preliminaries about averaging theory for discrete-time systems. Section III introduces the problem setup. Section IV describes our novel strategy and states its theoretical features. Section V is devoted to numerical simulations.

*Notation:* We say that a square matrix $M \in \mathbb{R}^{n\times n}$ is Schur if all its eigenvalues lie in the open unit disk. The identity matrix in $\mathbb{R}^{n\times n}$ is $I_n$. The vector of zeros of dimension $n$ is denoted as $0_n$. The vertical concatenation of vectors $v_1, \dots, v_N$ is $\mathrm{col}(v_1, \dots, v_N)$. Given $r > 0$ and $x \in \mathbb{R}^n$, we use $\mathcal{B}_r(x)$ to denote the closed ball of radius $r > 0$ centered in $x$, namely $\mathcal{B}_r(x) := \{y \in \mathbb{R}^n \mid \|y - x\| \leq r\}$. Given $A \in \mathbb{R}^{n\times n}$, $\mathrm{Tr}(A)$ denotes its trace. $\mathbb{R}_+$ denotes is the positive orthant in $\mathbb{R}$.

## II. Preliminaries: Averaging Theory for Discrete-Time Systems

In this preliminary part, we provide a generic stability result for discrete-time systems in the context of averaging theory. Although we will use it as an instrumental step for proving the main result of the paper, we remark that it represents a contribution per se.

Let us consider the time-varying discrete-time system

$$\chi^{k+1} = \chi^k + \gamma f(\chi^k, k) \qquad \chi^0 = \chi_0, \tag{1}$$

where $\chi^k \in \mathbb{R}^n$ denotes the state, $f : \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}^n$, and $\gamma > 0$ is a tunable parameter. Let us enforce the following assumptions.

**Assumption 1.** *There exist $T \in \mathbb{N}$ and $f_{\text{AV}} : \mathbb{R}^n \to \mathbb{R}^n$ such that*

$$f_{\text{AV}}(\chi) = \frac{1}{T} \sum_{\tau=k+1}^{k+T} f(\chi, \tau), \tag{2}$$

*for all $\chi \in \mathbb{R}^n$ and $k \in \mathbb{N}$.* ■

This work was supported in part by the Italian Ministry of Foreign Affairs and International Cooperation, grant number BR22GR01. The authors are with the Department of Electrical, Electronic and Information Engineering, Alma Mater Studiorum - Università di Bologna, Bologna, 40136, Italy, e-mail: {nicola.mimmo2, guido.carnevale, giuseppe.notarstefano}@unibo.it. The corresponding author is G. Carnevale.

Assumption 1 allows for properly writing a well-posed averaged system associated to system (1), while the next one guarantees some regularity conditions on $f$ and $f_{\text{AV}}$.

**Assumption 2.** *There exists a set $\mathcal{X} \subseteq \mathbb{R}^n$ such that the functions $f(\chi, k)$, $f_{\text{AV}}(\chi)$, and their derivatives are continuous for all $k \in \mathbb{N}$ over $\mathcal{X}$.* ■

The next assumption characterizes the convergence properties of the so-called averaged system associated to (1), i.e., the auxiliary dynamics described by

$$\chi_{\text{AV}}^{k+1} = \chi_{\text{AV}}^{k} + \gamma f_{\text{AV}}(\chi_{\text{AV}}^{k}) \qquad \chi_{\text{AV}}^{0} = \chi_0. \tag{3}$$

To this end, we first introduce a continuous function $V : \mathcal{X} \to \mathbb{R}_+$ and, given any $c > 0$, its level set $\Omega_c \subset \mathbb{R}^n$ given by

$$\Omega_c := \{x \in \mathbb{R}^n \mid V(x) \leq c\}.$$

**Assumption 3.** *Consider* (3)*, there exists a differentiable and continuous function $V : \mathcal{X} \to \mathbb{R}_+$ such that (i) the level sets of $V$ are compact and (ii), for all $\chi_0$ such that $V(\chi_0) \leq c_0$ with $\Omega_{c_0} \subseteq \mathcal{X}$ and $\rho \in (0, V(\chi_0))$, there exist $\bar{\gamma}_1, a > 0$ such that, along the trajectories of* (3) *and for all $\gamma \in (0, \bar{\gamma}_1)$, it holds*

$$V(\chi_{\text{AV}}^{k}) \leq V(\chi_0) \exp(-\gamma a k), \tag{4}$$

*for all $V(\chi_{\text{AV}}^{k}) \geq \rho$.* ■

We are ready to state the following result about the original system (1). The proof has been omitted for the lack of space.

**Theorem 1.** *Consider system* (1) *and let Assumptions 1-3 hold. Then, for all $\chi_0 \in \mathcal{X}$, $c_1 > V(\chi_0)$ such that $\Omega_{c_1} \subseteq \mathcal{X}$, and $\bar{\rho} \in (0, V(\chi_0))$, there exist $\bar{\gamma}, \bar{k} > 0$ such that, for all $\gamma \in (0, \bar{\gamma})$, it holds*

$$V(\chi^k) \leq c_1, \tag{5}$$

*for all $k \in \mathbb{N}$ and*

$$V(\chi^k) \leq \bar{\rho}, \tag{6}$$

*for all $k \geq \bar{k}$.* ■

## III. Problem Setup

This section states the problem setup and recalls a model-based approach to address it.

### A. Data-Driven LQR Problem Setup

In this paper, we focus on the LQR problem

$$\min_{\substack{x_1, x_2, \dots, \\ u_0, u_1, \dots}} \quad \mathbb{E}\left[\frac{1}{2}\sum_{t=0}^{\infty}\left({x_t}^\top Q x_t + {u_t}^\top R u_t\right)\right] \tag{7a}$$

$$\text{subj. to } x_{t+1} = Ax_t + Bu_t, \qquad x^0 \sim \mathcal{X}^0, \tag{7b}$$

where $x_t \in \mathbb{R}^n$ and $u_t \in \mathbb{R}^m$ denote the state and the input of the system at time $t \in \mathbb{N}$, $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$ represent the state and the input matrices, while $Q \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{m \times m}$ are the cost matrices. As for the initial condition $x^0 \in \mathbb{R}^n$, we assume that it is drawn from a probability distribution $\mathcal{X}^0$. The operator $\mathbb{E}[\cdot]$ denotes the expected value with respect to $\mathcal{X}^0$. We require the following properties on the pairs $(A, B)$ and $(Q, R)$.

**Assumption 4** (Properties of the system and cost matrices)**.** *The pair $(A, B)$ is controllable, while the cost matrices $Q$ and $R$ are both symmetric and positive definite, i.e., $Q = Q^\top \succ 0$ and $R = R^\top \succ 0$.* ■

While Assumption 4 is customary in the field of LQR, the next assumption makes challenging our setup.

**Assumption 5** (Unknown System and Cost Matrices)**.** *The pairs $(A, B)$ and $(Q, R)$ are unknown.* ■

Under the properties enforced in Assumption 4, when $(A, B)$ and $(Q, R)$ are known, the optimal solution to problem (7) is ruled by a linear time-invariant policy $u_t = K^\star x_t$ with $K^\star \in \mathbb{R}^{m \times n}$ given by

$$K^\star = -(R + B^\top P^\star B)^{-1} B^\top P^\star A,$$

where $P^\star \in \mathbb{R}^{n \times n}$ solves the Discrete-time Algebraic Riccati Equation associated to problem (7), see [26].

In this paper, we are interested in devising a data-driven strategy to iteratively obtain a state-feedback controller solution to (7) without the knowledge of $(A, B)$ and $(Q, R)$ (cf. Assumption 5).

### B. Model-based Gradient Method for LQR

Next, we recall a model-based gradient method to address problem (7). Let us introduce $\mathcal{K}$ to denote the set of stabilizing gains, namely

$$\mathcal{K} := \{K \in \mathbb{R}^{m \times n} \mid A + BK \text{ is Schur}\}.$$

As it will become useful later, we also introduce $r_{\mathcal{K}} > 0$ to denote the radius of the largest ball contained in $\mathcal{K}$. By assuming that $\mathcal{X}^0$ is a uniform distribution about the unitary-radius sphere and considering the state-feedback control $u_t = Kx_t$, with $K \in \mathbb{R}^{m \times n}$ such that $(A + BK)$ is Schur, it is possible to recast (see, e.g., [19]) problem (7) as

$$\min_{K \in \mathcal{K}} \; J(K), \tag{8}$$

where the cost function $J : \mathcal{K} \to \mathbb{R}$ is given by

$$J(K) := \tfrac{1}{2} \operatorname{Tr} \sum_{t=0}^{\infty} (A + BK)^{t,\top} (Q + K^\top R K)(A + BK)^t.$$

We remark that the choice on $\mathcal{X}^0$ implies that $\mathcal{K}$ coincides with the domain of $J$. Being the set of stabilizing gains $\mathcal{K}$ open [27, Lemma IV.3] and connected [27, Lemma IV.6], one could use the gradient descent method to solve problem (8) (see, e.g., [19]). Namely, at each iteration $k \in \mathbb{N}$, an estimate $K^k$ of $K^\star$ is maintained and iteratively updated according to

$$K^{k+1} = K^k - \gamma G(K^k), \tag{9}$$

where $\gamma > 0$ is the step size, while, when $\mathbb{R}^{m \times n}$ is equipped with the Frobenius inner product, $G : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ is the gradient of $J$ with respect to $K$ evaluated at $K^k$. If $K^0 \in \mathcal{K}$ and the step size $\gamma$ is sufficiently small, it is possible to

prove that the optimal gain $K^\star$ is an exponentially stable equilibrium of system (9), see [19, Theorem 4.6]. However, in the unknown scenario formalized by Assumption 5, the update law in (9) cannot be implemented. Indeed, the gradient $G(K^k)$ reads as

$$G(K^k) = \left(RK^k + B^\top P^k(A+BK^k)\right) W_c^k,$$

where $W_c^k \in \mathbb{R}^{n\times n}$ and $P^k \in \mathbb{R}^{n\times n}$ solve the equations

$$(A+BK^k)W_c^k(A+BK^k)^\top - W_c^k = -I_n$$
$$(A+BK^k)^\top P^k(A+BK^k) - P^k = -(Q+{K^k}^\top RK^k).$$

In our setup, it is not possible to compute $G(K^k)$ and implement (9) because it requires the knowledge of the pairs $(A,B)$ and $(Q,R)$ that are unknown (cf. Assumption 5). Our idea is to employ an Extremum-Seeking mechanism to compensate for this lack of knowledge.

## IV. EXP-LQR: ALGORITHM DESCRIPTION AND CONVERGENCE PROPERTIES

In this section, we present EXP-LQR, i.e., a novel method to solve problem (7) without model knowledge. Our algorithmic idea is to mimic the (model-based) gradient descent update through an Extremum-Seeking scheme. To this end, at each iteration $k$, we perturb a given policy gain $K^k$ to get $K^k + \delta \mathrm{D}^k$, where $\delta > 0$ is an amplitude parameter and $\mathrm{D}^k \in \mathbb{R}^{m\times n}$ is the so-called dither matrix, whose $(i,j)$-th element $D_{ij}^k$ is

$$D_{ij}^k := \sin\left(\frac{2\pi k}{T_{ij}} + \phi_{ij}\right),$$

where $T_{ij} \in \mathbb{N}$ and $\phi_{ij} \in \mathbb{R}$ are the period and the phase of component $(i,j)$, respectively. We force the orthonormality of the dither in the following assumption.

**Assumption 6.** *Let $T \in \mathbb{N}$ be the least common multiple of all periods $T_1, \dots, T_{nm}$. Then, it holds*

$$\sum_{k=1}^{T} \sin\left(\tfrac{2\pi k}{T_p} + \phi_p\right) = 0 \tag{10a}$$
$$\sum_{k=1}^{T} \sin\left(\tfrac{2\pi k}{T_p} + \phi_p\right)\sin\left(\tfrac{2\pi k}{T_q} + \phi_q\right) = \frac{T}{2} \tag{10b}$$
$$\sum_{k=1}^{T} \sin\left(\tfrac{2\pi k}{T_p} + \phi_p\right)\sin\left(\tfrac{2\pi k}{T_p} + \phi_p\right)\sin\left(\tfrac{2\pi k}{T_r} + \phi_r\right) = 0, \tag{10c}$$

*for all $p,q,r \in \{1,\dots,nm\}$ such that $p \neq q$, $q \neq r$, and $p \neq r$.* ■

EXP-LQR follows the steps proposed in Algorithm 1. More in detail, we suppose a simulation phase in which the feedback control law $u_t = (K^k + \delta \mathrm{D}^k)x_t$ is implemented into an oracle providing the corresponding $J(K^k + \delta \mathrm{D}^k)$. This scenario may occur, for example, when a simulator of a complex system is available, but the analytical knowledge of the dynamics being implemented for the simulations is unavailable because confidential. With $J(K^k + \delta \mathrm{D}^k)$ at hand, we perform the algorithm iteration detailed in (11). Specifically, the variable $\mathrm{z}^k \in \mathbb{R}$ filters the variation of $J(K^k+\delta \mathrm{D}^k)$ (see (11a)), while the update of the gain matrix $K^k$ follows the extremum-seeking update (11b).

**Algorithm 1** EXP-LQR

**Initialization:** $x^0 \in \mathbb{R}^n$, $K^0 \in \mathcal{K}$, $\mathrm{z}^0 \in \mathbb{R}$.

**for** $k = 0, 1, 2 \dots$ **do**

**Simulation phase**

Set the controller $u_t = (K^k + \delta \mathrm{D}^k)x_t$

Simulate $x_{t+1} = Ax_t + Bu_t$ retrieving $J(K^k+\delta \mathrm{D}^k)$

**Optimization phase:**

$$\mathrm{z}^{k+1} = \mathrm{z}^k + \gamma(J(K^k + \delta \mathrm{D}^k) - \mathrm{z}^k) \tag{11a}$$
$$K^{k+1} = K^k - \gamma \frac{2(J(K^k + \delta \mathrm{D}^k) - \mathrm{z}^k)\mathrm{D}^k}{\delta} \tag{11b}$$

**end for**

A block scheme of the strategy is depicted in Fig. 1.

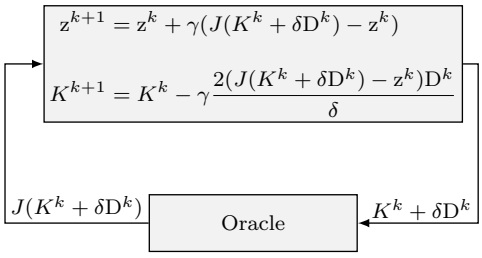


Fig. 1. Graphical representation of Algorithm 1.

Next, we provide the main result of the paper, i.e., the practical stability properties of system (11).

**Theorem 2.** *Consider* (11) *and let Assumptions 4, 5, and 6 hold. Then, for all $(\mathrm{z}^0, K^0) \in \mathbb{R} \times \mathcal{K}$ and $\bar{r} \in (0, r_{\mathcal{K}})$, there exist $\bar{\gamma}, \bar{\delta}, \bar{k} > 0$, such that, for all $\gamma \in (0, \bar{\gamma})$ and $\delta \in (0, \bar{\delta})$, the trajectories of* (11) *are bounded and it holds*

$$\left\|K^k - K^\star\right\| \leq \bar{r}, \tag{12}$$

*for all $k \geq \bar{k}$.* ■

The proof of Theorem 2 is based on three main building blocks. The first block regards the gradient approximation property of the mechanism incorporated in our algorithm, see Lemma 1. The second block is the stability analysis of the averaged system associated to system (11), see Lemma 2. The third block is the stability result provided in Theorem 1 for generic time-varying discrete-time systems. For the sake of space, the proofs of Lemma 2 and Theorem 1 will be given in a forthcoming document. However, in Section V-B, assuming the availability of both of these results, we provide the proof of Theorem 2.

## V. Stability Analysis

In this section, we perform the stability analysis of system (11). First, in Section V-A, by resorting to averaging theory, we characterize the convergence properties of the so-called *averaged system* associated to (11). In Section V-B, we use the stability properties of the averaged system to analyze the ones of the original time-varying system (11) and conclude the proof. Assumptions 4, 5, and 6 are valid throughout the entire section.

### A. Averaged System

The *averaged system* associated to (11) is an auxiliary system derived by averaging the vector field of (11) over a time horizon of length $T$ (see Assumption 6). To construct this system, we exploit [28, Lemma 1].

**Lemma 1** ([28])**.** *There exists $\ell : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ such that, for all $K \in \mathcal{K}$, $\mathrm{z} \in \mathbb{R}$, and $k \in \mathbb{N}$, it holds*

$$\frac{2}{\delta T}\sum_{\tau=k+1}^{k+T}\left(J(K+\delta \mathrm{D}^\tau)-\mathrm{z}\right)\mathrm{D}^\tau = G(K)+\delta^2\ell(K). \quad (13)$$

*Moreover, given any compact set $\mathcal{S} \subset \mathbb{R}^{m\times n}$, if $\delta \in (0,1]$, there exists $\beta_{\mathcal{S}} > 0$ such that*

$$\|\ell(K)\| \le \beta_{\mathcal{S}}, \quad (14)$$

*for all $K \in \mathcal{S}$.* ■

By applying Lemma 1 and the frequencies' property (10a), the averaged system associate to (11) reads as

$$\mathrm{z}_{\text{AV}}^{k+1} = \mathrm{z}_{\text{AV}}^{k} + \gamma\left(J_{\text{AV}}(K_{\text{AV}}^k,\delta) - \mathrm{z}_{\text{AV}}^k\right) \quad (15a)$$

$$K_{\text{AV}}^{k+1} = K_{\text{AV}}^k - \gamma G(K_{\text{AV}}^k) - \gamma\delta^2\ell(K_{\text{AV}}^k), \quad (15b)$$

where the function $J_{\text{AV}} : \mathcal{K}\times\mathbb{R} \to \mathbb{R}$ is defined as

$$J_{\text{AV}}(K,\delta) = \frac{1}{T}\sum_{\tau=k+1}^{k+T} J(K+\delta\mathrm{D}^\tau).$$

The next lemma ensures that $(J_{\text{AV}}(K^\star,\delta),K^\star)$ is a practically exponentially stable equilibrium point of system (15). To provide this result, we need to introduce the candidate Lyapunov function $V : \mathbb{R}\times\mathcal{K} \to \mathbb{R}_+$ defined as

$$V(\mathrm{z},K) := \frac{1}{2}\|\mathrm{z}\|^2 + \lambda\left(J(K)-J(K^\star)\right), \quad (16)$$

with $\lambda \ge 1$.

**Lemma 2.** *Consider* (15)*. Then, for all $(\mathrm{z}_{\text{AV}}^0, K_{\text{AV}}^0) \in \mathbb{R}\times\mathcal{K}$, $\rho > 0$, there exist $\bar\gamma_1, \bar\delta_1, a > 0$ and $\bar\lambda \ge 1$ such that, for all $\gamma \in (0,\bar\gamma_1)$, $\delta \in (0,\bar\delta_1)$, and $\lambda \ge \bar\lambda$, it holds*

$$V(\mathrm{z}_{\text{AV}}^k - J_{\text{AV}}(K_{\text{AV}}^k,\delta), K_{\text{AV}}^k) \le \exp(-\gamma a k)V(\mathrm{z}_{\text{AV}}^0 - J_{\text{AV}}(K_{\text{AV}}^0,\delta), K_{\text{AV}}^0), \quad (17)$$

*for all $V(\mathrm{z}_{\text{AV}}^k, K_{\text{AV}}^k) \ge \rho$.* ■

The proof of Lemma 2 will be provided in a forthcoming document.

### B. Sketch of Proof of Theorem 2

The proof relies on the application to system (11) of Theorem 1 (cf. Section II). In particular, we need to choose the parameters $\bar\rho, c_1 > 0$ and show that system (11) satisfies the conditions required in Assumptions 1, 2, and 3, namely, the ones due to apply Theorem 1. By looking at the statement of Theorem 1 and given the desired final radius $\bar r$, we start by designing the parameters $c_1$ and $\bar\rho$. By [29, Lemma 3.8], there exists $\sigma > 0$ such that

$$\sigma\|K-K^\star\| \le J(K)-J(K^\star), \quad (18)$$

for all $K \in \mathcal{K}$. Therefore, we set $c_1 > V(\mathrm{z}^0 - J_{\text{AV}}(K^0,\delta),K^0)$ and $\bar\rho \in (0,\bar r/\sigma)$. Assumptions 1, 2, and 3 are verified as detailed hereafter. First, Assumption 1 is trivially satisfied by the dither signal design (cf. Assumption 6). Second, we need to guarantee the continuity of the EXP-LQR system (11) and the corresponding averaged system (15) required in Assumption 2. By looking at the dynamics of (11) and (15), the continuity is guaranteed over the set $\{(\mathrm{z},K) \in \mathbb{R}\times\mathcal{K} \mid K+\delta\mathrm{D}^k \in \mathcal{K} \text{ for all } k \in \mathbb{N}\}$. By looking at the definition of the function $V$ (cf. (16)) and since $\lambda \ge 1$, we are able to guarantee that $J(K)-J(K^\star) \le c_1$ for all $(\mathrm{z},K)$ such that $V(\mathrm{z}-J_{\text{AV}}(K,\delta),K) \le c_1$. Then, by using the fact that $\mathcal{K}$ is open [27, Lemma IV.3] and $\mathrm{D}^k$ is bounded for all $k \in \mathbb{N}$, we guarantee the existence of $\bar\delta_2 > 0$ such that $K+\delta\mathrm{D}^k \in \mathcal{K}$ for all $K$ such that $J(K)-J(K^\star) \le c_1$, $\delta \in (0,\bar\delta_2)$, and $k \in \mathbb{N}$. Third, Lemma 2 ensures the convergence properties of $V$ along the trajectories of the averaged system (15) required in Assumption 3 provides an upper bound $\bar\delta_1$ for $\delta$. Hence, we set $\delta \in (0,\bar\delta)$ with $\bar\delta := \min\{\bar\delta_1,\bar\delta_2\}$ and apply Theorem 1, thus guaranteeing

$$V(\mathrm{z}^k - J(K^k), K^k) \le \bar\rho, \quad (19)$$

for all $k \ge \bar k$. The proof of (12) follows by combining $J(K)-J(K^\star) \le V(\mathrm{z}-J(K),K)$, (18), (19), and the choice of $\bar\rho$.

## VI. Numerical Simulations

In this section, we numerically test the effectiveness of EXP-LQR. We generate the scenario outlined in Section III by setting $n = 4$, $m = 2$, and by randomly generating the system and cost matrices imposing the controllability of $(A,B)$ and the positive definiteness and symmetry for $Q$ and $R$. Specifically, we randomly generate the matrix $A$ such that its eigenvalues have absolute values in the interval $[0,10]$, the matrix $B$ with components randomly extracted from the interval $[0,1]$ with uniform probability, and the cost matrices $Q$ and $R$ such that their eigenvalues lie in the interval $(0,10)$. As for the algorithm parameters, we empirically tune $\gamma = 10^{-7}$ and $\delta = 10^{-2}$. As for the dither frequencies, we ordered the pairs $(i,j) \in \{1,\dots,n\}\times\{1,\dots,m\}$ with indices $p = 1,\dots,nm$ and chosen $T_p = 11^{-(p-1)/2}$ and $\phi_p = 0$ for $p$ odd, while $T_p = T_{p-1}$ and $\phi_p = \pi/2$ for $p$ even. This choice ensures we satisfy Assumption 6 with $T = 11$. Fig. 2 shows the evolution of the cost error $J(K^k)-J(K^\star)$ in logarithmic scale, while Fig. 3 shows the evolution of $J(K^k)$ and $J(K^\star)$. As predicted by Theorem 2,

Fig. 2 and 3 show that Algorithm 1 asymptotically converges in a neighborhood of the optimal gain $K^\star$.

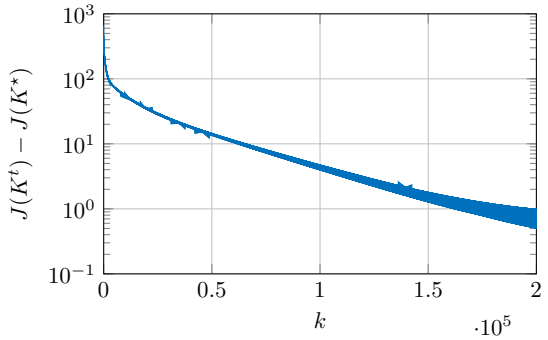


Fig. 2. Evolution of the cost error $J(K^k) - J(K^\star)$.

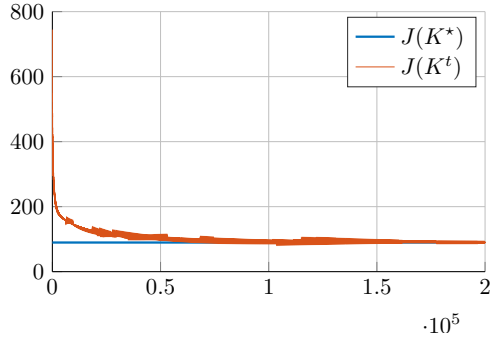


Fig. 3. Comparison between $J(K^k)$ and $J(K^\star)$.

## VII. Conclusions

We proposed a data-driven method able to iteratively find the state feedback gain matrix solving a Linear Quadratic Regulator (LQR) problem without any knowledge of the system and cost matrices. Given an oracle able to evaluate the current cost, our method refines its estimate according to a mechanism based on Extremum-Seeking. We analyzed the resulting time-varying algorithm by exploiting system theory tools based on Lyapunov stability and averaging theory. Specifically, we guaranteed that our algorithm exponentially converges to an arbitrarily small ball containing the optimal gain matrix. We numerically tested the proposed solution.

## References


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

[2] B. Pang, T. Bian, and Z.-P. Jiang, "Robust policy iteration for continuous-time linear quadratic regulation," *IEEE Transactions on Automatic Control*, vol. 67, no. 1, pp. 504–511, 2021.

[3] C. Qin, H. Zhang, and Y. Luo, "Online optimal tracking control of continuous-time linear systems with unknown dynamics by using adaptive dynamic programming," *International Journal of Control*, vol. 87, no. 5, pp. 1000–1009, 2014.

[4] K. Krauth, S. Tu, and B. Recht, "Finite-time analysis of approximate policy iteration for the linear quadratic regulator," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[5] H. Modares, F. L. Lewis, and Z.-P. Jiang, "Optimal output-feedback control of unknown continuous-time linear systems using off-policy reinforcement learning," *IEEE Transactions on Cybernetics*, vol. 46, no. 11, pp. 2401–2410, 2016.

[6] B. Pang, T. Bian, and Z.-P. Jiang, "Data-driven finite-horizon optimal control for linear time-varying discrete-time systems," in *2018 IEEE Conference on Decision and Control (CDC)*, pp. 861–866, IEEE, 2018.

[7] V. G. Lopez, M. Alsalti, and M. A. Müller, "Efficient off-policy Q-learning for data-based discrete-time LQR problems," *IEEE Transactions on Automatic Control*, 2023.

[8] C. Possieri and M. Sassano, "Q-learning for continuous-time linear systems: A data-driven implementation of the Kleinman algorithm," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 10, pp. 6487–6497, 2022.

[9] T. Bian and Z.-P. Jiang, "Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design," *Automatica*, vol. 71, pp. 348–360, 2016.

[10] I. Ziemann, A. TSIAMis, H. Sandberg, and N. Matni, "How are policy gradient methods affected by the limits of control?," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 5992–5999, IEEE, 2022.

[11] B. Kiumarsi, F. L. Lewis, and Z.-P. Jiang, "$H_\infty$ control of linear discrete-time systems: Off-policy reinforcement learning," *Automatica*, vol. 78, pp. 144–152, 2017.

[12] C. De Persis and P. Tesi, "Formulas for data-driven control: Stabilization, optimality, and robustness," *IEEE Transactions on Automatic Control*, vol. 65, no. 3, pp. 909–924, 2019.

[13] H. J. Van Waarde, J. Eising, H. L. Trentelman, and M. K. Camlibel, "Data informativity: a new perspective on data-driven analysis and control," *IEEE Transactions on Automatic Control*, vol. 65, no. 11, pp. 4753–4768, 2020.

[14] M. Rotulo, C. De Persis, and P. Tesi, "Data-driven linear quadratic regulation via semidefinite programming," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 3995–4000, 2020.

[15] M. Rotulo, C. De Persis, and P. Tesi, "Online learning of data-driven controllers for unknown switched linear systems," *Automatica*, vol. 145, p. 110519, 2022.

[16] C. De Persis and P. Tesi, "Low-complexity learning of linear quadratic regulators from noisy data," *Automatica*, vol. 128, p. 109548, 2021.

[17] F. Dörfler, P. Tesi, and C. De Persis, "On the certainty-equivalence approach to direct data-driven LQR design," *IEEE Transactions on Automatic Control*, 2023.

[18] B. Hu, K. Zhang, N. Li, M. Mesbahi, M. Fazel, and T. Başar, "Toward a theoretical foundation of policy optimization for learning control policies," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 6, pp. 123–158, 2023.

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

[20] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *International conference on machine learning*, pp. 1467–1476, PMLR, 2018.

[21] B. Wittenmark and A. Urquhart, "Adaptive extremal control," in *Proceedings of 1995 34th IEEE Conference on Decision and Control*, vol. 2, pp. 1639–1644, IEEE, 1995.

[22] A. R. Teel and D. Popovic, "Solving smooth and nonsmooth multivariable extremum seeking problems by the methods of nonlinear programming," in *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, vol. 3, pp. 2394–2399, IEEE, 2001.

[23] K. B. Ariyur and M. Krstic, *Real-time optimization by extremum-seeking control*. John Wiley & Sons, 2003.

[24] M. Krstić and H.-H. Wang, "Stability of extremum seeking feedback for general nonlinear dynamic systems," *Automatica*, vol. 36, no. 4, pp. 595–601, 2000.

[25] Y. Tan, D. Nešić, and I. Mareels, "On non-local stability properties of extremum seeking control," *Automatica*, vol. 42, no. 6, pp. 889–903, 2006.

[26] B. D. Anderson and J. B. Moore, *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

[27] J. Bu, A. Mesbahi, and M. Mesbahi, "On topological properties of the set of stabilizing feedback gains," *IEEE Transactions on Automatic Control*, vol. 66, no. 2, pp. 730–744, 2020.

[28] N. Mimmo, G. Carnevale, A. Testa, and G. Notarstefano, "Extremum seeking tracking for derivative-free distributed optimization," *arXiv preprint arXiv:2110.04234*, 2021.

[29] J. Bu, A. Mesbahi, and M. Mesbahi, "LQR via first order flows," in *2020 American Control Conference (ACC)*, pp. 4683–4688, IEEE, 2020.