2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

# On-Policy Data-Driven Linear Quadratic Regulator via Combined Policy Iteration and Recursive Least Squares

Lorenzo Sforni, Guido Carnevale, Ivano Notarnicola, Giuseppe Notarstefano

***Abstract*— In this paper, we address infinite-horizon Linear Quadratic Regulator (LQR) problems for unknown discrete-time systems. As an additional challenge, we address an on-policy setup in which system matrices are identified while controlling the real system with a progressively optimized policy. Specifically, we consider a time-varying control policy that, while applied to the real unknown system, is iteratively refined (based on the most updated estimate of the system matrices) towards the optimal LQR solution. The overall learning procedure combines a recursive least squares method with a direct policy search based on the gradient method. By resorting to Lyapunov-based analysis tools in combination with averaging theory for nonlinear systems, exponential stability for the closed-loop scheme can be proven. Finally, a numerical example showing the effectiveness of the considered strategy corroborates the theoretical findings.**

## I. Introduction

We address infinite-horizon discrete-time Linear Quadratic Regulator (LQR) problems in a model-free setting, i.e., when the (linear) dynamics is unknown. The model-based LQR is a cornerstone problem since it admits a closed form solution based on the well-known Riccati equation, see, e.g., classic textbooks as [1], [2]. Gradient-based methods may be also used to solve LQR as, e.g., the celebrated Anderson-Moore algorithm and Kleinman policy iteration presented in [3], [4], [5]. A renovated interest in these strategies is due to their possible use in a data-driven context. They are, in fact, early versions of policy iteration and policy gradient methods developed in reinforcement learning [2], [6]. A summary of first-order properties of the discrete-time LQR is given in [7].

In the last years, the solution of LQR has become a benchmark in the model-free control. Starting from early works in the adaptive control, [8], [9], [10], a connection with reinforcement learning has been investigated in recent years [11]. An adaptive value-iteration strategy is analyzed in [12]. In [13] robustness of the policy iteration for continuous-time systems under additive, bounded disturbances is studied. The discrete-time framework has drawn significant attention in the learning community. The work [14] proposes a safe-learning strategy for LQR via an indirect approach, i.e., the unknown dynamics is first estimated, so that the control gain is optimized on the estimated quantities. The sample complexity for model-free linear quadratic regulator is studied in [15]. A distributed instance of the model-free LQR problem is addressed in [16] via zeroth-order optimization. The paper [17] investigates an off-policy Q-learning strategy, with an additional focus on computational complexity. Due to its benchmarking role, recently, the LQR problem has been also addressed via policy-gradient methods. A model-free, gradient-based, strategy is proposed in [18]. The convergence properties of the (policy) gradient methods are thoroughly studied in [19] for discrete-time LQR. While in [20], the sample complexity and convergence properties for the continuous-time case are examined. Recent works also explored the non-asymptotic performances of model-free LQR algorithms. Sub-linear regret result is given in [21]. Poly-logarithmic regret bounds are given in [22], [23].

The main contribution of the paper is the design of a control policy that is concurrently applied to the actual (unknown) linear system while iteratively refined towards the optimal solution of a Linear Quadratic Regulator (LQR) problem. The proposed on-policy scheme relies on a suitable reformulation of the LQR problem as an optimization problem parametrized in the system matrices $(A, B)$, with the control policy gain $K$ being the decision variable. The optimization problem is then solved via a gradient-based method combined with an estimation procedure to cope with the lack of knowledge about the system matrices. Specifically, the gradient-based update is interlaced with a Recursive Least Squares (RLS) mechanism (to recover the missing information about the system matrices) elaborating trajectory samples obtained from the actual, closed-loop system, which is actuated by the (yet non-optimal) state feedback. To guarantee the persistence of excitation, the (running) closed-loop dynamics is fed by a probing dithering signal. We show exponential stability to a proper steady state, in which: $(i)$ the feedback policy is the optimal solution of the LQR, $(ii)$ the estimates of the unknown matrices are exact, and $(iii)$ the system state oscillates around the origin.

The paper is organized as follows. Section II introduces the problem setup with some preliminaries. Section III describes the proposed methodology and states its theoretical features. Section IV is devoted to sketching the proof of the stated results, while Section V provides a numerical simulation.

*Notation:* A square matrix $M \in \mathbb{R}^{n\times n}$ is Schur if all its eigenvalues lie in the open unit disk. the spectrum of $M$ is denoted as $\sigma(M)$, while its trace as $\mathrm{Tr}(M)$. $M^\dagger$ denotes the Moore-Penrose inverse of $M$. The identity matrix in $\mathbb{R}^{n\times n}$ is $I_n$. The vector of $n$ zeros is denoted as $0_n$. The vertical concatenation of $v_1, \dots, v_N$ is $\mathrm{col}(v_1, \dots, v_N)$. Given $r > 0$ and $x \in \mathbb{R}^n$, $\mathcal{B}_r(x)$ denotes the ball of radius $r > 0$ centered at $x$. We use $\otimes$ to denote the Kronecker product.

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. The corresponding author is L. Sforni lorenzo.sforni@unibo.it.

## II. Problem Setup and Preliminaries

### A. On-Policy Data-Driven LQR Problem Setup

In this paper, we consider a linear time-invariant system described by

$$\mathrm{x}_{t+1} = A\mathrm{x}_t + B\mathrm{u}_t, \qquad \mathrm{x}_0 \sim \mathcal{X}_0, \tag{1}$$

where $\mathrm{x}_t \in \mathbb{R}^n$ and $\mathrm{u}_t \in \mathbb{R}^m$ denote, respectively, 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, respectively, the state matrix and the input matrix. The initial condition $\mathrm{x}_0 \in \mathbb{R}^n$ is assumed to be drawn from a (known) probability distribution $\mathcal{X}_0$ and we enforce the following assumption about $(A, B)$.

*Assumption 2.1:* The pair of system matrices $(A, B)$ is controllable and unknown. ■

We consider an infinite horizon LQR problem

$$\min_{\substack{\mathrm{x}_1,\mathrm{x}_2,\dots,\\ \mathrm{u}_0,\mathrm{u}_1,\dots}} \ \tfrac{1}{2}\sum_{t=0}^{\infty}\Big(\mathrm{x}_t{}^\top Q\mathrm{x}_t + \mathrm{u}_t{}^\top R\mathrm{u}_t\Big) \tag{2a}$$

$$\text{subj. to } \mathrm{x}_{t+1} = A\mathrm{x}_t + B\mathrm{u}_t, \qquad \mathrm{x}_0 \sim \mathcal{X}_0, \tag{2b}$$

where the cost matrices $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$ are such that $Q = Q^\top > 0$ and $R = R^\top > 0$. It is well-known that, when $(A, B)$ are knwon the optimal solution is given by a linear time-invariant policy $\mathrm{u}_t = K^\star\mathrm{x}_t$ 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 (2), see [1].

In this paper, we are interested in devising a data-driven feedback policy for (1). The main innovation of the proposed approach is that we *concurrently*

(i) learn the unknown dynamics;
(ii) improve the policy toward a solution of (2);
(iii) actuate the (real) system with our tentative optimal state-feedback policy.

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

*1) Model-based reduced problem formulation:* We recall an equivalent (unconstrained) formulation of Problem (2), which explicitly imposes the linear feedback structure to the optimal input. That is, letting $K \in \mathbb{R}^{m\times n}$, the problem is rewritten by explicitly substituting in the cost function the feedback input

$$\mathrm{u}_t = K\mathrm{x}_t.$$

Such a formulation highlights that $(i)$ the overall cost actually depends on the gain $K$ only, and, $(ii)$ the optimal gain $K^\star$ does not depend on the initial condition $\mathrm{x}_0$. First of all, for any gain $K$, the original (open-loop) dynamics (1) admits the closed-loop formulation $\mathrm{x}_{t+1} = (A + BK)\mathrm{x}_t$. So that, for all $t \geq 0$, the state is uniquely determined as

$$\mathrm{x}_t = (A + BK)^t\mathrm{x}_0, \qquad \mathrm{x}_0 \sim \mathcal{X}_0. \tag{3}$$

Hence, Problem (2) can be compactly written as

$$\min_K \ \tfrac{1}{2}\mathrm{x}_0^\top\Big(\sum_{t=0}^{\infty}(A + BK)^{t\top}(Q + K^\top RK)(A + BK)^t\Big)\mathrm{x}_0$$

that holds for all initial conditions $\mathrm{x}_0$. Averaging on the initial condition we obtain

$$\min_K \tfrac{1}{2}\operatorname{Tr}\Big(\sum_{t=0}^{\infty}(A + BK)^{t\top}(Q + K^\top RK)(A + BK)^t\Sigma_0\Big)$$

where $\Sigma_0 := \mathbb{E}[\mathrm{x}_0\mathrm{x}_0^\top]$, with $\mathbb{E}[\cdot]$ denoting the expected value with respect the distribution $\mathcal{X}_0$. Without loss of generality, we consider $\mathcal{X}_0$ to be a uniform distribution about the unit sphere and, so we can finally write the problem as

$$\min_{K\in\mathcal{D}} \ J(K, \theta^\star), \tag{4}$$

where the cost function $J : \mathcal{D} \times \mathbb{R}^{(n+m)\times n} \to \mathbb{R}$ is

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

the parameter $\theta^\star$ collects $A$ and $B$ as $\theta^\star := \begin{bmatrix}A & B\end{bmatrix}^\top \in \mathbb{R}^{(n+m)\times n}$, and the set $\mathcal{D} \subset \mathbb{R}^{m\times n} := \{K \in \mathbb{R}^{m\times n} \mid J(K, \theta^\star) < \infty\}$ is the domain of $J$, i.e., the set over which $J$ is well-defined. It is possible to show that the set of stabilizing gains $\mathcal{S} := \{K \in \mathbb{R}^{m\times n} \mid A + BK \text{ is Schur}\} \subseteq \mathbb{R}^{m\times n}$ coincides with the interior of $\mathcal{D}$ [7, Lemma 3.2]. Clearly, the optimal gain $K^\star$ must belong to $\mathcal{D}$.

*2) Model-based gradient method for* (4)*:* If the matrices $(A, B)$ were known, a gradient descent method could be used to solve Problem (4). (see, e.g., [7]). Namely, at each iteration $t \in \mathbb{N}$, we maintain a solution estimate $K_t$ and we update it according to

$$K_{t+1} = K_t - \gamma G(K_t, \theta^\star), \tag{5}$$

where $\gamma > 0$ is the stepsize, while $G : \mathbb{R}^{m\times n}\times\mathbb{R}^{(n+m)\times n} \to \mathbb{R}^{m\times n}$ is the gradient of $J$ with respect to $K$ evaluated at $(K_t, \theta^\star)$ when $\mathbb{R}^{m\times n}$ is equipped with the Frobenius inner product. It is possible to show that, by initializing $K_t$ into $\mathcal{S}$ and selecting a proper stepsize $\gamma$, the optimal gain $K^\star$ is an exponentially stable equilibrium of the dynamical system (5) [7, Th. 4.6]. The procedure to evaluate $G(K_t, \theta^\star)$ reads as follows:

(i) solve for $W_t^c \in \mathbb{R}^{n\times n}$ and $P_t \in \mathbb{R}^{n\times n}$ the Lyapunov-type equations

$$\begin{aligned}&(A + BK_t)W_t^c(A + BK_t)^\top - W_t^c = -I_n,\\ &(A + BK_t)^\top P_t(A + BK_t) - P_t = -(Q + K_t{}^\top RK_t),\end{aligned} \tag{6a}$$

(ii) compute $G(\mathrm{x}_t, \theta^\star)$ as

$$G(K_t, \theta^\star) = \big(RK_t + B^\top P_t(A + BK_t)\big)\,W_t^c. \tag{6b}$$

Notice that we want to solve Problem (4) without resorting to the knowledge of $\theta^\star$, i.e., when $(A, B)$ are unknown (cf. Assumption 2.1). Therefore, in our framework it is not possible to implement update (5).

## III. On-policy LQR for Unknown Systems: Concurrent Learning and Optimization

In this section, we present the concurrent learning and optimization algorithm developed to solve Problem (2) under Assumption 2.1. The proposed on-policy strategy feeds the real system dynamics at each iteration $t$ with the current feedback input including an additive exogenous dithering signal $\mathrm{w}_t$. Then, a new sample data from the system is collected and used to progressively improve the estimates $(A_t, B_t)$ of the unknown $(A, B)$ via a learning process inspired by Recursive Least Squares (RLS). In turn, $(A_t, B_t)$ is used to refine the feedback gain $K_t$ for (2), and the procedure is repeated. The overall scheme is shown in Fig. 1.

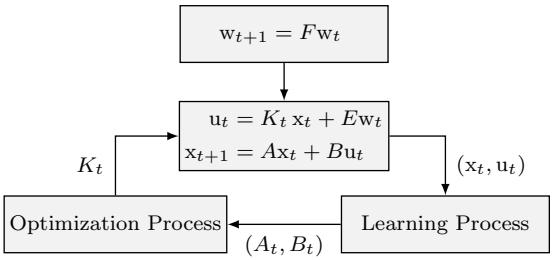


Fig. 1. Representation of the concurrent learning and optimization scheme.

The overall strategy is reported in Algorithm 1 where, for notational convenience, we denote as $\theta_t \in \mathbb{R}^{(n+m)\times n}$ the estimate of $\theta^\star$ at iteration $t \in \mathbb{N}$ and, consistently, $A_t \in \mathbb{R}^{n\times n}$ and $B_t \in \mathbb{R}^{n\times m}$ are the corresponding estimates of $A$ and $B$. Moreover, $H_t$ and $S_t$ denote two additional states of the learning process, $\lambda \in (0, 1)$ is a forgetting factor, while $\gamma$ is the stepsize as in (5).

**Algorithm 1** On-policy LQR for Unknown Systems

**Initialization:** $\mathrm{x}_0 \in \mathbb{R}^n$, $H_0 \in \mathbb{R}^{(n+m)\times(n+m)}$, $S_0 \in \mathbb{R}^{(n+m)\times n}$, $\theta_0 \in \mathbb{R}^{(n+n)\times n}$, $K_0 \in \mathbb{R}^{m\times n}$ and $\mathrm{w}_0 \in \mathbb{R}^{n_\mathrm{w}}$.

**for** $t = 0, 1, 2 \ldots$ **do**

**Data collection**

$$\begin{aligned} \mathrm{w}_{t+1} &= F\mathrm{w}_t \\ \mathrm{u}_t &= K_t\mathrm{x}_t + E\mathrm{w}_t \\ \mathrm{x}_{t+1} &= A\mathrm{x}_t + B\mathrm{u}_t \\ \mathrm{y}_t &= \mathrm{x}_{t+1}^\top \end{aligned}$$

**Learning process**

$$H_{t+1} = \lambda H_t + \begin{bmatrix}\mathrm{x}_t \\ \mathrm{u}_t\end{bmatrix}\begin{bmatrix}\mathrm{x}_t \\ \mathrm{u}_t\end{bmatrix}^\top \tag{7a}$$

$$S_{t+1} = \lambda S_t + \begin{bmatrix}\mathrm{x}_t \\ \mathrm{u}_t\end{bmatrix}\mathrm{y}_t \tag{7b}$$

$$\theta_{t+1} = \theta_t - \gamma H_t^\dagger\left(H_t\theta_t - S_t\right). \tag{7c}$$

**Optimization process**

$$K_{t+1} = K_t - \gamma G(K_t, \theta_t). \tag{8}$$

Next, we detail the main steps of the proposed algorithm.

*Data collection:* Data from the controlled system (1) are recast in an identification-oriented fashion given by

$$\underbrace{\mathrm{x}_{t+1}^\top}_{\mathrm{y}_t} = \underbrace{\begin{bmatrix}\mathrm{x}_t^\top & \mathrm{u}_t^\top\end{bmatrix}}_{C(\mathrm{x}_t,\mathrm{u}_t)^\top}\underbrace{\begin{bmatrix}A^\top \\ B^\top\end{bmatrix}}_{\theta^\star}. \tag{9}$$

*Learning process:* The adopted learning strategy to compute an estimate of $\theta^\star$ relies on the interpretation of the least squares problem in terms of online optimization. Specifically, with the measurements (9) at hand, we pose, at each $t \in \mathbb{N}$, the following online optimization problem

$$\min_{\theta\in\mathbb{R}^{(n+m)\times n}} \tfrac{1}{2}\sum_{\tau=0}^{t}\lambda^{t-\tau}\left\|C(\mathrm{x}_\tau,\mathrm{u}_\tau)^\top\theta - \mathrm{y}_\tau\right\|^2. \tag{10}$$

We aim to solve (10) through an iterative algorithm that progressively refines a solution estimate $\theta_t \in \mathbb{R}^{(n+m)\times n}$. In particular, the estimate $\theta_t$ can be updated according to a Newton-like method. A plain application of the method to problem (10) would give the iteration

$$\begin{aligned}\theta_{t+1} = \theta_t - \gamma&\left(\sum_{\tau=0}^{t}\lambda^{t-\tau}\mathcal{H}(\mathrm{x}_\tau,\mathrm{u}_\tau)\right)^\dagger \\ &\times\left(\sum_{\tau=0}^{t}\lambda^{t-\tau}\left(\mathcal{H}(\mathrm{x}_\tau,\mathrm{u}_\tau)\theta_t - \mathcal{S}(\mathrm{x}_\mathrm{t},\mathrm{u}_\tau)\right)\right),\end{aligned}$$

where $\mathcal{H} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{(n+m)\times(n+m)}$ and $\mathcal{S} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{(n+m)\times n}$ are defined as

$$\begin{aligned}\mathcal{H}(\mathrm{x}_\tau,\mathrm{u}_\tau) &:= C(\mathrm{x}_\tau,\mathrm{u}_\tau)C(\mathrm{x}_\tau,\mathrm{u}_\tau)^\top \\ \mathcal{S}(\mathrm{x}_\tau,\mathrm{u}_\tau) &:= C(\mathrm{x}_\tau,\mathrm{u}_\tau)\mathrm{y}_\tau.\end{aligned}$$

To overcome the issue of storing the whole history of $\mathcal{H}(\cdot,\cdot)$ and $\mathcal{S}(\cdot,\cdot)$, we keep track of them through the states $H_t \in \mathbb{R}^{(n+m)\times(n+m)}$ and $S_t \in \mathbb{R}^{(n+m)\times n}$ giving rise to (7).

*Optimization process:* The estimate $\theta_t$ is concurrently exploited in the update of the feedback gain $K_t$. That is, the unavailable $\theta^\star$ of (5) is replaced by $\theta_t$ giving rise to (8).

To ensure sufficiently informative data, we equip our feedback policy with an additive dithering signal $\mathrm{d}_t \in \mathbb{R}^m$. Namely, we implement

$$\mathrm{u}_t = K_t\mathrm{x}_t + \mathrm{d}_t, \tag{11}$$

where $\mathrm{d}_t$ is the output of an exogenous system evolving according to a marginally stable linear discrete-time oscillator dynamics (see, e.g., [24]) described by

$$\mathrm{w}_{t+1} = F\mathrm{w}_t \tag{12a}$$

$$\mathrm{d}_t = E\mathrm{w}_t, \tag{12b}$$

where $\mathrm{w}_t \in \mathbb{R}^{n_\mathrm{w}}$, with $n_\mathrm{w} \geq n + m$, is the state of the exogenous system having $F \in \mathbb{R}^{n_\mathrm{w}\times n_\mathrm{w}}$ and $E \in \mathbb{R}^{n_\mathrm{w}}$ as state and output matrix, respectively. The matrix $F$ is a degree of freedom to properly shape the oscillation frequency of $\mathrm{w}_t$. The following assumption formalizes the requirements for the design of the exogenous system (12).

*Assumption 3.1:* There exist $\alpha_1, \alpha_2, t_{\mathrm{w}} > 0$ such that, if $\mathrm{w}_0 \neq 0_{n_{\mathrm{w}}}$, then

$$\alpha_1 I_{n_{\mathrm{w}}} \leq \sum_{\tau=\bar{t}+1}^{\bar{t}+t_{\mathrm{w}}} \mathrm{w}_\tau \mathrm{w}_\tau^\top \leq \alpha_2 I_{n_{\mathrm{w}}}, \tag{13}$$

for all $\bar{t} \in \mathbb{N}$. Moreover, it holds

$$\operatorname{rank}\left(\begin{bmatrix} \mathrm{d}_0 & \mathrm{d}_1 & \dots & \mathrm{d}_{\mathbf{t}_{\mathrm{d}}-n-1} \\ \mathrm{d}_1 & \mathrm{d}_1 & \dots & \mathrm{d}_{\mathbf{t}_{\mathrm{d}}-n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{d}_n & \mathrm{d}_{n+1} & \dots & \mathrm{d}_{\mathbf{t}_{\mathrm{d}}-1} \end{bmatrix}\right) = m(n+1), \tag{14}$$

for some $\mathbf{t}_{\mathrm{d}} > 0$. ■

Property (13) is usually referred to as *persistency of excitation* of the signal $\mathrm{w}_t$, see, e.g., [25].

The resulting closed-loop dynamics is

$$\mathrm{w}_{t+1} = F\mathrm{w}_t \tag{15a}$$
$$\mathrm{x}_{t+1} = (A + BK_t)\mathrm{x}_t + BE\mathrm{w}_t \tag{15b}$$
$$H_{t+1} = \lambda H_t + \begin{bmatrix} \mathrm{x}_t \\ K_t\mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}\begin{bmatrix} \mathrm{x}_t \\ K_t\mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}^\top \tag{15c}$$
$$S_{t+1} = \lambda S_t + \begin{bmatrix} \mathrm{x}_t \\ K_t\mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}\begin{bmatrix} \mathrm{x}_t \\ K_t\mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}^\top \theta^\star \tag{15d}$$
$$\theta_{t+1} = \theta_t - \gamma H_t^\dagger \left(H_t\theta_t - S_t\right) \tag{15e}$$
$$K_{t+1} = K_t - \gamma G(K_t, \theta_t), \tag{15f}$$

in which we have used the explicit expressions for $\mathrm{y}_t$ (cf. (9)) and $\mathrm{u}_t$ (cf. (11)). In order to state our (local) exponential stability result, we introduce the set $\mathcal{B}_{r^\star}(K^\star) \subset \mathbb{R}^{m\times n}$ being the neighborhood of $K^\star$ such that $A + BK$ is Schur for all $K \in \mathcal{B}_r(K^\star)$. Indeed, we notice that being the matrix $A + BK^\star$ Schur, such a set must exist by continuity. We can now provide the main result of the paper, i.e., the convergence properties of (15).

*Theorem 3.2:* Let Assumptions 2.1, and 3.1 hold. Consider system (15), then for each $(\mathrm{w}_0, \mathrm{x}_0, H_0, S_0, \theta_0, K_0) \in \mathbb{R}^n \times \mathbb{R}^{(n+m)\times(n+m)} \times \mathbb{R}^{(n+m)\times n} \times \mathbb{R}^{(n+m)\times n} \times \mathcal{B}_{r^\star}(K^\star)$ such that $\mathrm{w}_0 \neq 0$, $A_0 + B_0K_0$ is Schur, there exist $\Pi_{\mathrm{x}} \in \mathbb{R}^{n\times n_{\mathrm{w}}}$, $\Pi_H \in \mathbb{R}^{(n+m)^2\times n_{\mathrm{w}}^2}$, $\Pi_S \in \mathbb{R}^{(n+m)m\times n_{\mathrm{w}}^2}$, $a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, \bar{\gamma} > 0$ such that, it holds

$$\|\mathrm{x}_t - \Pi_{\mathrm{x}}\mathrm{w}_t\| \leq a_1 \|\mathrm{x}_0 - \Pi_{\mathrm{x}}\mathrm{w}_0\| \exp(-a_2 t) \tag{16a}$$
$$\left\|H_t - \mathtt{unvec}\left(\Pi_H \mathtt{vec}\left(\mathrm{w}_t\mathrm{w}_t^\top\right)\right)\right\| \leq a_3 \left\|H_0 - \mathtt{unvec}\left(\Pi_H \mathtt{vec}\left(\mathrm{w}_0\mathrm{w}_0^\top\right)\right)\right\| \exp(-a_4 t) \tag{16b}$$
$$\left\|S_t - \mathtt{unvec}\left(\Pi_S \mathtt{vec}\left(\mathrm{w}_t\mathrm{w}_t^\top\right)\right)\right\| \leq a_5 \left\|S_0 - \mathtt{unvec}\left(\Pi_S \mathtt{vec}\left(\mathrm{w}_0\mathrm{w}_0^\top\right)\right)\right\| \exp(-a_6 t) \tag{16c}$$
$$\left\|\begin{bmatrix} \theta_t \\ \tilde{K}_t \end{bmatrix}\right\| \leq a_7 \left\|\begin{bmatrix} \theta_0 \\ K_0 \end{bmatrix}\right\| \exp(-a_8 t), \tag{16d}$$

for any $\gamma \in (0, \bar{\gamma})$. ■

For a sketch of proof of Theorem 3.2 see Section IV.

Notice that the initialization in Theorem 3.2 does not necessarily require the knowledge of $(A, B)$. Indeed, one can compute a stabilizing controller $K_0$ in a data-based fashion, see, e.g., [26] and the discussion in [17]. The result (16a) of Theorem 3.2 ensures that $\Pi_{\mathrm{x}}\mathrm{w}_t$ is an exponentially practically stable equilibrium for (15b).

## IV. Sketch of Stability Analysis

In order to perform the analysis, we preliminarily write the vectorized version of the matrix updates in (15c) and (15d). To this end, let us introduce $H^{\mathrm{vc}} \in \mathbb{R}^{(n+m)^2}$ and $S^{\mathrm{vc}} \in \mathbb{R}^{(n+m)n}$ defined as

$$\begin{cases} H_t \\ S_t \end{cases} \longmapsto \begin{cases} H_t^{\mathrm{vc}} := \mathtt{vec}\left(H_t\right) \\ S_t^{\mathrm{vc}} := \mathtt{vec}\left(S_t\right). \end{cases} \tag{17}$$

It allows us to write

$$\mathrm{w}_{t+1} = F\mathrm{w}_t \tag{18a}$$
$$\mathrm{x}_{t+1} = (A + BK_t)\mathrm{x}_t + BE\mathrm{w}_t \tag{18b}$$
$$H_{t+1}^{\mathrm{vc}} = \lambda H_t^{\mathrm{vc}} + \mathtt{vec}\left(\begin{bmatrix} \mathrm{x}_t \\ K_t\mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}\begin{bmatrix} \mathrm{x}_t \\ K_t\mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}^\top\right) \tag{18c}$$
$$S_{t+1}^{\mathrm{vc}} = \lambda S_t^{\mathrm{vc}} + \mathtt{vec}\left(\begin{bmatrix} \mathrm{x}_t \\ K_t\mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}\begin{bmatrix} \mathrm{x}_t \\ K_t\mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}^\top \theta^\star\right) \tag{18d}$$
$$\theta_{t+1} = \theta_t - \gamma H_t^\dagger \left(H_t\theta_t - S_t\right) \tag{18e}$$
$$K_{t+1} = K_t - \gamma G(K_t, \theta_t) \tag{18f}$$

where we abuse of notation in leaving the unvectorized version of the states $H_t$ and $S_t$ in (18e) and (18f).

Next, we provide the steady-state locus (see, e.g., [27, Ch. 12]) for the dynamical system (18). To this end, let $\chi := \mathrm{col}(\mathrm{x}, H^{\mathrm{vc}}, S^{\mathrm{vc}}, \mathtt{vec}\left(\theta\right), \mathtt{vec}\left(K\right))$ and recast (18) in a more compact way as the following cascade

$$\mathrm{w}_{t+1} = F\mathrm{w}_t \tag{19a}$$
$$\chi_{t+1} = \phi(\chi_t, \mathrm{w}_t), \tag{19b}$$

where $\phi$ properly collects all the vectorized updates in (18). Moreover, we also introduce the nonlinear map $\mathrm{w} \mapsto \chi^{\mathrm{ss}}(\mathrm{w})$ defined as

$$\chi^{\mathrm{ss}}(\mathrm{w}) := \begin{bmatrix} \Pi_{\mathrm{x}}\mathrm{w} \\ \Pi_H \mathtt{vec}\left(\mathrm{w}\mathrm{w}^\top\right) \\ \Pi_S \mathtt{vec}\left(\mathrm{w}\mathrm{w}^\top\right) \\ \mathtt{vec}\left(\theta^\star\right) \\ \mathtt{vec}\left(K^\star\right) \end{bmatrix}, \tag{20}$$

where $\Pi_{\mathrm{x}}$, $\Pi_H$, and $\Pi_S$ are defined as in the statement Theorem 3.2. The following lemma holds.

*Lemma 4.1:* Consider the dynamical system (19) and the map $\chi^{\mathrm{ss}}$ defined in (20). Then, it holds

$$\chi^{\mathrm{ss}}(F\mathrm{w}) = \phi(\chi^{\mathrm{ss}}(\mathrm{w}), \mathrm{w}),$$

for all $\mathrm{w} \in \mathbb{R}^{n_{\mathrm{w}}}$. ■

The proof of Lemma 4.1 is omitted for the sake of space.

With the previous result at hand, let us introduce the error coordinates $\tilde{\mathrm{x}}_t \in \mathbb{R}^n$, $\tilde{H}_t^{\mathrm{vc}} \in \mathbb{R}^{(n+m)\times(n+m)}$, $\tilde{S}_t^{\mathrm{vc}} \in$

$\mathbb{R}^{(n+m)\times n}$, $\tilde{\theta}_t \in \mathbb{R}^{(n+m)\times n}$, and $\tilde{K}_t \in \mathbb{R}^{m\times n}$ defined via the following change of coordinates

$$\begin{cases} \mathrm{w}_t \\ \mathrm{x}_t \\ H_t^{\mathrm{vc}} \\ S_t^{\mathrm{vc}} \\ \theta_t \\ K_t \end{cases} \longmapsto \begin{cases} \mathrm{w}_t \\ \tilde{\mathrm{x}}_t := \mathrm{x}_t - \Pi_{\mathrm{x}} \mathrm{w}_t \\ \tilde{H}_t^{\mathrm{vc}} := H_t^{\mathrm{vc}} - \Pi_H \mathtt{vec}\left(\mathrm{w}_t \mathrm{w}_t^\top\right) \\ \tilde{S}_t^{\mathrm{vc}} := S_t^{\mathrm{vc}} - \Pi_S \mathtt{vec}\left(\mathrm{w}_t \mathrm{w}_t^\top\right) \\ \tilde{\theta}_t := \theta_t - \theta^\star \\ \tilde{K}_t := K_t - K^\star. \end{cases} \tag{21}$$

The dynamics (18) can be expressed in error coordinates as

$$\mathrm{w}_{t+1} = F \mathrm{w}_t \tag{22a}$$
$$\tilde{\mathrm{x}}_{t+1} = (A + B\tilde{K}_t + BK^\star)\tilde{\mathrm{x}}_t + B\tilde{K}_t \Pi_{\mathrm{x}} \mathrm{w}_t \tag{22b}$$
$$\tilde{H}_{t+1}^{\mathrm{vc}} = \lambda \tilde{H}_t^{\mathrm{vc}} + \ell(\tilde{\mathrm{x}}_t, \tilde{K}_t, \mathrm{w}_t) \tag{22c}$$
$$\tilde{S}_{t+1}^{\mathrm{vc}} = \lambda \tilde{S}_t^{\mathrm{vc}} + \ell(\tilde{\mathrm{x}}_t, \tilde{K}_t, \mathrm{w}_t)\theta^\star \tag{22d}$$
$$\tilde{\theta}_{t+1} = \tilde{\theta}_t - \gamma(\tilde{H}_t + H_t^{\mathrm{ss}})^\dagger((\tilde{H}_t + H_t^{\mathrm{ss}})\tilde{\theta}_t + (\tilde{H}_t - \tilde{S}_t)\theta^\star) \tag{22e}$$
$$\tilde{K}_{t+1} = \tilde{K}_t - \gamma G(\tilde{K}_t + K^\star, \tilde{\theta}_t + \theta^\star), \tag{22f}$$

where, for the sake of readability, we introduced $\tilde{H}_t, H_t^{\mathrm{ss}} \in \mathbb{R}^{(n+m)\times(n+m)}$ and $\tilde{S}_t \in \mathbb{R}^{(n+m)\times n}$ given by

$$\tilde{H}_t := \mathtt{unvec}\left(\tilde{H}_t^{\mathrm{vc}}\right) \tag{23a}$$
$$H_t^{\mathrm{ss}} := \mathtt{unvec}\left(\Pi_H \mathtt{vec}\left(\mathrm{w}_t \mathrm{w}_t^\top\right)\right) \tag{23b}$$
$$\tilde{S}_t := \mathtt{unvec}\left(\tilde{S}_t^{\mathrm{vc}}\right), \tag{23c}$$

while $\ell : \mathbb{R}^n \times \mathbb{R}^{m\times n} \times \mathbb{R}^{n_{\mathrm{w}}}$ is defined as

$$\ell(\tilde{\mathrm{x}}, \tilde{K}, \mathrm{w}) := -\mathtt{vec}\left(M \mathrm{w}\mathrm{w}^\top M^\top\right) + \mathtt{vec}\left(\begin{bmatrix} \tilde{\mathrm{x}} + \Pi_{\mathrm{x}}\mathrm{w} \\ (\tilde{K}+K^\star)(\tilde{\mathrm{x}}+\Pi_{\mathrm{x}}\mathrm{w}) + E\mathrm{w}_t \end{bmatrix}\begin{bmatrix} \tilde{\mathrm{x}} + \Pi_{\mathrm{x}}\mathrm{w} \\ (\tilde{K}+K^\star)(\tilde{\mathrm{x}}+\Pi_{\mathrm{x}}\mathrm{w}) + E\mathrm{w} \end{bmatrix}^\top\right).$$

Next we analyze the time-varying system (22) leveraging on averaging theory tools (see, e.g., [28]). First of all, we need to recast (22) in a dynamical system with two time scales. Specifically, let the new states $\xi_t$ and $\mathrm{z}_t$ be defined as

$$\xi_t := \begin{bmatrix} \tilde{\mathrm{x}}_t \\ \gamma \tilde{H}_t^{\mathrm{vc}} \\ \gamma \tilde{S}_t^{\mathrm{vc}} \end{bmatrix}, \quad \mathrm{z}_t := \begin{bmatrix} \tilde{\theta}_t \\ \tilde{K}_t \end{bmatrix}.$$

As we will see in the next, $\xi_t$ highlights the state of the fast dynamic embedded into (15), while, on the contrary, $\mathrm{z}_t$ represents the state of its slow dynamics. Indeed, we can now reformulate (15) as a two-time-scale dynamical system described by

$$\xi_{t+1} = \mathcal{A}(\mathrm{z}_t)\xi_t + h(\mathrm{z}_t, t) + \gamma g(\xi_t, \mathrm{z}_t, t) \tag{24a}$$
$$\mathrm{z}_{t+1} = \mathrm{z}_t + \gamma f(\xi_t, \mathrm{z}_t, t), \tag{24b}$$

where

$$\mathcal{A}(\mathrm{z}) := \begin{bmatrix} A + BK^\star + B\tilde{K} & 0 & 0 \\ 0 & \lambda I & 0 \\ 0 & 0 & \lambda I \end{bmatrix}$$

$$h(\mathrm{z}, t) := \begin{bmatrix} B\tilde{K}\Pi_{\mathrm{x}}\mathrm{w}_t \\ 0 \\ 0 \end{bmatrix}, \quad g(\xi, \mathrm{z}, t) := \begin{bmatrix} 0 \\ \ell(\tilde{\mathrm{x}}, \tilde{K}, t) \\ \ell(\tilde{\mathrm{x}}, \tilde{K}, t)\theta^\star \end{bmatrix}$$

$$f(\xi, \mathrm{z}, t) := -\begin{bmatrix} (\tilde{H} + H_t^{\mathrm{ss}})^\dagger \left((\tilde{H} + H_t^{\mathrm{ss}})\tilde{\theta} + (\tilde{H}_t - \tilde{S}_t)\theta^\star\right) \\ G(\tilde{K} + K^\star, \tilde{\theta} + \theta^\star) \end{bmatrix},$$

in which, with a slight abuse of notation, we maintained a hybrid set of coordinates with $(\xi, \mathrm{z})$ on the left-hand side and $(\tilde{\mathrm{x}}, \tilde{H}, \tilde{S}, \tilde{\theta}, \tilde{K})$ on the right-hand one. The next result explicitly provides the averaged system associated to (24).

*Lemma 4.2:* Consider (24). Then, the associated averaged system reads as

$$\mathrm{z}_{t+1}^{\mathrm{AV}} = \mathrm{z}_t^{\mathrm{AV}} + \gamma f^{\mathrm{AV}}(\mathrm{z}_t^{\mathrm{AV}}), \tag{25}$$

where

$$f^{\mathrm{AV}}(\mathrm{z}) := \lim_{T\to\infty} \frac{1}{T} \sum_{\tau=\bar{t}+1}^{\bar{t}+T} f(0, \mathrm{z}, t)$$

exists uniformly in $\bar{t} \in \mathbb{N}$ and for any $\mathrm{z}$ and

$$f^{\mathrm{AV}}(\mathrm{z}) = \begin{bmatrix} -\tilde{\theta} \\ -G(\tilde{K} + K^\star, \tilde{\theta} + \theta^\star) \end{bmatrix},$$

where, with a slight abuse of notation, we used $\mathrm{z} = \mathrm{col}(\tilde{K}, \tilde{\theta})$. ■

The proof of Lemma 4.2 is omitted for the sake of space. This lemma shows that the averaged system (25) is a cascade system given by $(i)$ an autonomous linear dynamics (with state matrix $(1-\gamma)I$) related to the estimate of $\theta^\star$, and $(ii)$ a perturbed gradient method that becomes the standard update (5) when $\bar{\theta} = 0$. Therefore, by relying on the structure of the cascade system (25) and the differentiability properties of the gradient map $G$, [7], it is possible to show that, for sufficiently small values of $\gamma$, the origin is semi-globally exponentially stable for (25). In turn, this result allows us to exploit results from averaging theory (cf. [28, Th.2.2.4]) to show the exponential stability of the origin for (22) thus proving the result of Theorem 3.2.

## V. Numerical Simulations

In this section, we perform some numerical simulations to corroborate our theoretical findings. We randomly generated system matrices $A \in \mathbb{R}^{3\times3}$ and $B \in \mathbb{R}^{3\times2}$ and the cost matrices $Q \in \mathbb{R}^{3\times3}$ and $R \in \mathbb{R}^{2\times2}$. In the problem instance considered in these simulations, we have

$$A = \begin{bmatrix} -0.53 & 0.42 & -0.44 \\ 0.42 & -0.56 & -0.65 \\ -0.44 & -0.65 & 0.35 \end{bmatrix} \qquad B = \begin{bmatrix} 0.43 & -0.82 \\ 0.53 & -0.78 \\ 0.26 & -0.40 \end{bmatrix}$$

and

$$Q = \begin{bmatrix} 6.12 & 1.72 & 0.53 \\ 1.72 & 6.86 & 1.72 \\ 0.53 & 1.72 & 5.73 \end{bmatrix} \qquad R = \begin{bmatrix} 1.15 & -0.23 \\ -0.23 & 3.62 \end{bmatrix}.$$

We empirically tune $\gamma = 0.01$ and $\|\mathrm{w}_0\| = 1$. Fig. 2 shows the evolution of the normalized cost error $|J(K_t, \theta^\star) - J^\star|/J^\star$, with $J^\star := J(K^\star, \theta^\star)$ in logarithmic scale. Finally, Fig. 3 shows the evolution of the normalized estimation error $\|\theta_t - \theta^\star\| / \|\theta^\star\|$ in logarithmic scale. Notice that, in both cases, convergence to the optimal cost $J^\star$ and true parameters $\theta^\star$ is achieved.

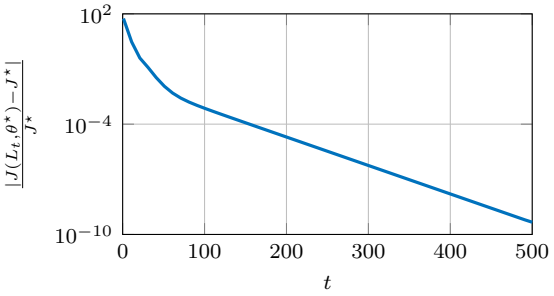


Fig. 2. Evolution of the normalized cost error $|J(K_t, \theta^\star) - J^\star|/J^\star$ with respect to iterations $t$.

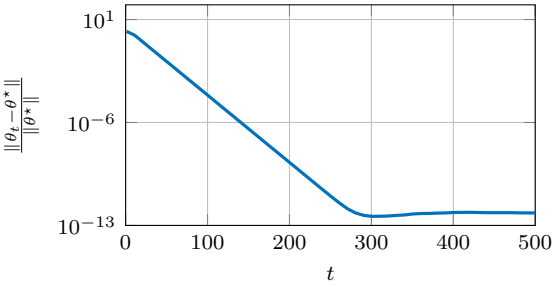


Fig. 3. Evolution of the normalized estimation error $\mathrm{x}\|\theta_t - \theta^\star\| \,/\, \|\theta^\star\|$ with respect to iterations $t$.

## VI. Conclusions

In this paper, we addressed infinite-horizon LQR problems with unknown state-input matrices. Specifically, we propose a procedure mixing the identification phase of the unknown matrices with the optimization of the feedback policy. We design an iterative algorithm combining a Recursive Least Squares scheme (elaborating samples from the closed-loop system persistently excited by a dithering signal) with the gradient method. We proved exponential convergence of the overall procedure to the optimal steady-state associated to the optimal gain and the exact matrices by using tools from Lyapunov-based analysis tools in combination with averaging theory for nonlinear systems.

## References


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

[2] D. P. Bertsekas *et al.*, "Dynamic programming and optimal control II," *Belmont, MA: Athena Scientific*, 2011.

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

[4] W. Levine and M. Athans, "On the determination of the optimal constant output feedback gains for linear multivariable systems," *IEEE Transactions on Automatic control*, vol. 15, no. 1, pp. 44–48, 1970.

[5] T. Rautert and E. W. Sachs, "Computational design of optimal output feedback controllers," *SIAM Journal on Optimization*, vol. 7, no. 3, pp. 837–852, 1997.

[6] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[7] 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.

[8] T. L. Lai and C. Z. Wei, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *The Annals of Statistics*, vol. 10, no. 1, pp. 154–166, 1982.

[9] M. C. Campi and P. Kumar, "Adaptive linear quadratic gaussian control: the cost-biased approach revisited," *SIAM Journal on Control and Optimization*, vol. 36, no. 6, pp. 1890–1907, 1998.

[10] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012.

[11] B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 253–279, 2019.

[12] C. Possieri and M. Sassano, "Value iteration for continuous-time linear time-invariant systems," *IEEE Transactions on Automatic Control*, 2022.

[13] 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.

[14] S. Dean, S. Tu, N. Matni, and B. Recht, "Safely learning to control the constrained linear quadratic regulator," in *IEEE American Control Conference (ACC)*, pp. 5582–5588, 2019.

[15] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Foundations of Computational Mathematics*, vol. 20, no. 4, pp. 633–679, 2020.

[16] L. Furieri, Y. Zheng, and M. Kamgarpour, "Learning the globally optimal distributed LQ regulator," in *Conference on Learning for Dynamics and Control*, vol. 120 of *Proceedings of Machine Learning Research*, pp. 287–297, PMLR, 2020.

[17] 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.

[18] K. Zhang, B. Hu, and T. Basar, "Policy optimization for $H_2$ linear control with $H_\infty$ robustness guarantee: Implicit regularization and global convergence," in *Learning for Dynamics and Control*, pp. 179–190, PMLR, 2020.

[19] 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.

[20] 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 Transactions on Automatic Control*, vol. 67, no. 5, pp. 2435–2450, 2021.

[21] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 1–26, JMLR Workshop and Conference Proceedings, 2011.

[22] A. Cassel, A. Cohen, and T. Koren, "Logarithmic regret for learning linear quadratic regulators efficiently," in *International Conference on Machine Learning*, pp. 1328–1337, PMLR, 2020.

[23] M. Akbari, B. Gharesifard, and T. Linder, "Achieving logarithmic regret via hints in online learning of noisy lqr systems," in *IEEE 61st Conference on Decision and Control (CDC)*, pp. 4700–4705, 2022.

[24] C. S. Turner, "Recursive discrete-time sinusoidal oscillators," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 103–111, 2003.

[25] E.-W. Bai and S. S. Sastry, "Persistency of excitation, sufficient richness and parameter convergence in discrete time adaptive control," *Systems & control letters*, vol. 6, no. 3, pp. 153–163, 1985.

[26] 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.

[27] A. Isidori, *Lectures in feedback design for multivariable systems*. Springer, 2017.

[28] E.-W. Bai, L.-C. Fu, and S. S. Sastry, "Averaging analysis for discrete time and sampled data adaptive systems," *IEEE Transactions on Circuits and Systems*, vol. 35, no. 2, pp. 137–148, 1988.