Automatica 191 (2026) 113101

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Stability-certified on-policy data-driven LQR via recursive learning and policy gradient☆,☆☆

Lorenzo Sforni, Guido Carnevale *, Ivano Notarnicola, Giuseppe Notarstefano

Department of Electrical, Electronic and Information Engineering, Alma Mater Studiorum - Università di Bologna, Bologna, 40136, Italy

ARTICLE INFO

Article history:
Received 28 February 2024
Received in revised form 6 February 2026
Accepted 4 May 2026

Keywords:
Data-based control
Linear quadratic regulator
Numerical algorithms
Optimization-based controller synthesis

ABSTRACT

In this paper, we investigate a data-driven framework to solve Linear Quadratic Regulator (LQR) problems when the dynamics is unknown, with the additional challenge of providing stability certificates for the overall learning and control scheme. Specifically, in the proposed on-policy learning framework, the control input is applied to the actual (unknown) linear system while iteratively optimized. We propose a learning and control procedure, termed RELEARN LQR, that combines a recursive least squares method with a direct policy search based on the gradient method. The resulting scheme is analyzed by modeling it as a feedback-interconnected nonlinear dynamical system. A Lyapunov-based approach, exploiting averaging and timescale separation theories for nonlinear systems, allows us to provide formal stability guarantees for the whole interconnected scheme. The effectiveness of the proposed strategy is corroborated by numerical simulations, where RELEARN LQR is deployed on an aircraft control problem, with both static and drifting parameters.

1. Introduction

The massive availability of data in automation and robotics pushed the control community to revise the traditional model-based Optimal Control (OC) approaches toward learning-driven scenarios. In this context, the control policy is iteratively updated without an explicit knowledge of the underlying dynamical system, relying solely on the collected data. Hence, the key distinction between off-policy and on-policy methods arises from the interconnection between gathered data and the current policy. Specifically, off-policy algorithms pursue a value iteration approach, and data are, in general, independent of the current policy. Conversely, on-policy algorithms employ a policy iteration framework, evaluating the performance of the current policy using data generated under the same policy. Since the early derivations of Reinforcement Learning (RL) methods for Linear Quadratic (LQ) regulation (Bradtke et al., 1994), there has been growing interest in data-driven solutions to infinite-horizon Linear Quadratic Regulator (LQR). The recent survey (Recht, 2019) investigates connections between OC and RL.

In the off-policy context, we find iterative methods inspired by the Kleinman algorithm (Kleinman, 1968), involving either parameter identification or direct policy estimation (Krauth et al., 2019; Lopez et al., 2023; Modares et al., 2016; Pang et al., 2018, 2021; Qin et al., 2014). The recent works (Bian & Jiang, 2016; Possieri & Sassano, 2022a; Ziemann et al., 2022) propose algorithms that do not assume the existence of stabilizing initial policies. A model-free approach for discrete-time LQR based on RL is studied and developed in Kiumarsi et al. (2017). Off-policy approaches can be further distinguished between direct, where data are used directly in the policy design phase, and indirect approaches, where a preliminary identification step is performed. Direct strategies often tackle the LQR problem by exploiting Persistently Exciting (PE) data together with semi-definite programming and Linear Matrix Inequalities (LMI) approaches, as introduced in De Persis and Tesi (2019) and extended in Rotulo et al. (2020, 2022), Van Waarde et al. (2020). These LMI-based solutions also allowed for the design of control policies in case of noisy data, as explored in De Persis and Tesi (2021), Dörfler et al. (2023). Direct approaches have been deployed to address also the design of robust controllers, e.g., in Berberich et al. (2020), van Waarde et al. (2020). The recent survey (De Persis & Tesi, 2023) also includes an extension to nonlinear systems. Instead, indirect approaches are explored (Dean et al., 2019;

☆ Work supported by Fondi PNRR - Bando PE - Progetto PE11 - 3A-ITALY, "Made in Italy Circolare e Sostenibile" - Codice PE0000004, CUP: J33C22002950001 and by MOST - Sustainable Mobility National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) - MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 - D.D. 1033 17/06/2022, CN00000023).

☆☆ The material in this paper was presented at IEEE 62nd Conference on Decision and Control (CDC), Dec. 13–15, 2023, Singapore. This paper was recommended for publication in revised form by Associate Editor Alessandro Abate under the direction of Editor Florian Dorfler.

* Corresponding author.
E-mail addresses: lorenzo.sforni@unibo.it (L. Sforni), guido.carnevale@unibo.it (G. Carnevale), ivano.notarnicola@unibo.it (I. Notarnicola), giuseppe.notarstefano@unibo.it (G. Notarstefano).

https://doi.org/10.1016/j.automatica.2026.113101

Ferizbegovic et al., 2019; Mania et al., 2019). Approaches bridging the indirect and direct paradigms are proposed in Dörfler et al. (2022), Formentin and Chiuso (2018), Iannelli et al. (2020). Another successful approach in LQR, often deployed in an off-policy setting, is represented by policy-gradient methods, see the survey (Hu et al., 2023). Complete characterizations of these methods for discrete-time LQR are in Bu et al. (2019) and Fazel et al. (2018). A model-free, gradient-based, strategy is proposed in Zhang et al. (2020). In Mohammadi et al. (2021), the sample complexity and convergence properties for the continuous-time case are examined. In Mohammadi et al. (2020) the discrete-time case is considered. Sublinear regret bounds in model-free LQR are given in Abbasi-Yadkori and Szepesvári (2011), Cohen et al. (2019), while Akbari et al. (2022), Cassel et al. (2020) provide poly-logarithmic regret bounds. Sample complexity in model-free LQR is studied in Dean et al. (2020). Conversely, continuous-time on-policy techniques are proposed in Jiang and Jiang (2012), Vrabie et al. (2009). In Possieri and Sassano (2022b), stability guarantees on the learning dynamics are provided. In the discrete-time context, the on-policy setting is addressed in Kiumarsi et al. (2015) leveraging on policy iteration and value iteration approaches. In Simchowitz and Foster (2020), regret bounds for online LQR are provided. While the mentioned works about online approaches offer guarantees for (asymptotically and, some of them, probabilistically) obtaining stabilizing (possibly non-optimal) controllers, a thorough and explicit investigation into the stability properties of the closed-loop interlacing optimization, learning, and control tasks, governed by a time-varying and nonlinear dynamics, remains an open challenge.

Our main contribution is the development of a data-driven on-policy control scheme with stability certificates in LQR for unknown systems. Specifically, the estimated control policy is applied to the actual (unknown) linear system, while it is concurrently refined toward the optimal solution of the LQR problem. The proposed method, termed RELEARN LQR, short for REcursive LEARNing policy gradient for LQR, relies on the so-called *direct policy search* reformulation of the LQR problem, which is an optimization problem with the control policy gain K being the decision variable. This optimization problem, with cost function parametrized by the system matrices $(A_\star, B_\star)$, is addressed via a gradient-based method combined with an estimation procedure to deal with the missing knowledge of $(A_\star, B_\star)$. In particular, the system matrices are progressively reconstructed via a Recursive Least Squares (RLS) mechanism that iteratively processes the state-input samples obtained from the actual, closed-loop system. The on-policy nature of RELEARN LQR stems from the fact that each state-input sample is gathered by actuating the (yet non-optimal) state feedback. To ensure persistency of excitation, a probing dithering signal is also fed into the (running) closed-loop dynamics. The stability certificates for the learning and control closed-loop system are proved by resorting to Lyapunov arguments and averaging theory for two-time-scale systems. Specifically, for the whole closed-loop system consisting of the gradient update on the gain K, the RLS scheme, and the system dynamics, we show the exponential stability of a properly defined steady state, in which: (i) the feedback policy is the optimal solution of the LQR problem; (ii) the estimates of the unknown matrices are exact; and (iii) the system state oscillates about the origin with an amplitude arbitrarily tunable by setting the dither magnitude. These stability properties pave the way for characterizing algorithm effectiveness even in non-nominal conditions, where system and cost matrices change over time and/or disturbances affect the plant and the measurements. In such scenarios, the closed-loop system would adapt dynamically, restoring optimality without requiring any restart of either the optimization or the learning process. A key distinctive feature of our work is to provide a stability certificate for the overall closed-loop system that simultaneously addresses optimization, learning, and control tasks. Most existing data-driven approaches (see, e.g., De Persis and Tesi (2019)) are not on-policy, namely, they are characterized by two distinct phases in which system samples are collected and then used to find the optimal gain. Alternative, popular approaches (see, e.g., Fazel et al. (2018), Krauth et al. (2019)) are based on improving the tentative policy by performing so-called experiments of the actuated system to evaluate it. The main drawback of these approaches is that they are not online, namely, the real plant needs to be repeatedly initialized and actuated for a number of samples at each algorithm iteration. Another branch of literature relies on the certainty equivalence principle, see, e.g., Mania et al. (2019), Simchowitz and Foster (2020), which propose online strategies but focus on studying the incurred regret rather than the stability properties of the overall closed-loop system.

A preliminary short version of this paper is Sforni et al. (2023). The present article includes an improved comprehensive treatment, all the theoretical proofs, and a concrete application.

The paper unfolds as follows. Section 2 introduces the problem setup and some preliminaries. Section 3 describes RELEARN LQR and states its theoretical features. Section 4 analyzes the proposed scheme, while Section 5 presents some numerical simulations.

Notation. 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 $v_1, \dots, v_N$ is $\mathrm{col}(v_1, \dots, v_N)$. $\mathcal{B}_r(x) := \{y \in \mathbb{R}^n \mid \|y - x\| \leq r\}$ denotes the ball of radius $r > 0$ centered in $x \in \mathbb{R}^n$. $\sigma(A)$ denotes the spectrum of $A \in \mathbb{R}^{n\times n}$, while $A^\dagger$ denotes its Moore–Penrose inverse. We denote the Kronecker product by $\otimes$. We denote the concatenation of the columns of $M \in \mathbb{R}^{n\times m}$ by $\mathtt{vec}(M) := \mathrm{col}([M]_{11}, \dots, [M]_{n1}, \dots, [M]_{1m}, \dots, [M]_{nm}) \in \mathbb{R}^{nm}$, where $[M]_{ij}$ is the (i, j)th entry of M.

2. Preliminaries and problem setup

In this section we present some useful preliminaries and describe the problem we aim at investigating.

2.1. Averaging theory for two-time-scale systems

Consider the time-varying two-time-scale system

$$\chi_{t+1} = \mathcal{A}(\mathrm{z}_t)\chi_t + h(\mathrm{z}_t, t) + \epsilon g(\chi_t, \mathrm{z}_t, t) \tag{1a}$$

$$\mathrm{z}_{t+1} = \mathrm{z}_t + \epsilon f(\chi_t, \mathrm{z}_t, t), \tag{1b}$$

with $\chi_t \in \mathbb{R}^{n_\chi}$, $\mathrm{z}_t \in \mathbb{R}^{n_\mathrm{z}}$, $g : \mathbb{R}^{n_\chi} \times \mathbb{R}^{n_\mathrm{z}} \times \mathbb{N} \to \mathbb{R}^{n_\chi}$, $f : \mathbb{R}^{n_\chi} \times \mathbb{R}^{n_\mathrm{z}} \times \mathbb{N} \to \mathbb{R}^{n_\mathrm{z}}$, and $\mathcal{A} : \mathbb{R}^{n_\mathrm{z}} \to \mathbb{R}^{n_\chi \times n_\chi}$. Further, $\epsilon > 0$ is a tuning parameter that is useful to arbitrarily reduce the variations over time of subsystem (1b), which is therefore typically referred to as the *slow* subsystem, with z_t being the slow state. Coherently, subsystem (1b) is referred to as the *fast* subsystem, with χ_t being the fast state. The analysis also relies on the investigation of a time-invariant auxiliary system associated to the slow dynamics, termed the *averaged* system and defined as

$$\mathrm{z}^{\mathrm{AV}}_{t+1} = \mathrm{z}^{\mathrm{AV}}_t + \epsilon f^{\mathrm{AV}}(\mathrm{z}^{\mathrm{AV}}_t), \tag{2}$$

where $f^{\mathrm{AV}} : \mathbb{R}^{n_\mathrm{z}} \to \mathbb{R}^{n_\mathrm{z}}$ is given by

$$f^{\mathrm{AV}}(\mathrm{z}) := \lim_{T\to\infty} \frac{1}{T} \sum_{\tau=\bar{t}+1}^{\bar{t}+T} f(0, \mathrm{z}, \tau). \tag{3}$$

The averaged system effectively "neglects" the time variability of $f(\cdot, \cdot, t)$ and assumes that the fast state χ_t is at the origin. The underlying idea is that, as ϵ becomes smaller, a timescale

separation can be established between the evolution of z_t and the variations of $f(\cdot,\cdot,t)$ and χ_t so that the stability properties of (1) can be inferred from those of (2). To this end, in the following, we report some conditions that will be used in the forthcoming algorithmic analysis. First, we need the following regularity conditions on the vector fields of system (1).

Assumption 2.1. There exists r such that f, g, and h are Lipschitz continuous over $\mathcal{B}_r(0_{n_\chi+n_z})$. ■

We also assume that the origin is an equilibrium of (1).

Assumption 2.2. It holds $h(0,t) = 0$, $g(0,0,t) = 0$, and $f(0,0,t) = 0$ for all $t \in \mathbb{N}$. ■

Third, we characterize the matrix function $\mathcal{A}(z)$ of (1b).

Assumption 2.3. There exist $r, m_1, m_2 > 0$ and $a_1, a_2 \in (0,1)$ such that, for all $z \in \mathcal{B}_r(0_{n_z})$ and $t \in \mathbb{N}$, it holds

$$m_1 a_1^t \leq \|\mathcal{A}(z)^t\| \leq m_2 a_2^t.$$

Further, $\mathcal{A}$ is differentiable and it holds

$$\|\partial \mathcal{A}(z)/\partial z_i\| \leq k_a,$$

for all $i \in \{1,\dots,n_z\}$, $z \in \mathcal{B}_r(0_{n_z})$, and some $k_a > 0$. ■

Fourth, we impose that the vector field characterizing the averaged system, say $f^{\text{AV}} : \mathbb{R}^{n_z} \to \mathbb{R}^{n_z}$, is well-posed.

Assumption 2.4. The function f is piecewise continuous in t and the limit $\lim_{T\to\infty} \frac{1}{T}\sum_{\tau=\bar{t}+1}^{\bar{t}+T} f(0,z,\tau)$ exists uniformly in $\bar{t} \in \mathbb{N}$ and for all $z \in \mathcal{B}_r(0_{n_z})$. ■

Finally, we characterize the difference between f and f^{AV}.

Assumption 2.5. Consider f^{AV} as defined in (3) and let $\Delta f : \mathbb{R}^{n_z} \times \mathbb{N} \to \mathbb{R}^{n_z}$ be defined as

$$\Delta f(z,t) := f(0,z,t) - f^{\text{AV}}(z).$$

Then, there exists a nonnegative strictly decreasing function $\nu(t)$ such that $\lim_{t\to\infty} \nu(t) = 0$ and

$$\left\| \frac{1}{T} \sum_{\tau=\bar{t}+1}^{\bar{t}+T} \Delta f(z,\tau) \right\| \leq \nu(T)\,\|z\|$$
$$\left\| \frac{1}{T} \sum_{\tau=\bar{t}+1}^{\bar{t}+T} \frac{\partial \Delta f(z,\tau)}{\partial z} \right\| \leq \nu(T),$$

uniformly in $\bar{t} \in \mathbb{N}$ and for all $z \in \mathcal{B}_r(0_{n_z})$. ■

We are ready to provide a stability result for (1). Essentially, for sufficiently small ϵ, exponential stability of the origin for (2) is enough to get exponential stability of the origin for the interconnected time-varying system (1).

Theorem 2.6 (*Bai et al., 1988, Theorem 2.2.4*)**.** *Consider system (1) and let Assumptions 2.1, 2.2, 2.3, 2.4 and 2.5 hold. If there exists ϵ^{AV} such that, for all $\epsilon \in (0, \epsilon^{\text{AV}})$, the origin is exponentially stable for system (2), then there exists $\bar{\epsilon} \in (0, \epsilon^{\text{AV}})$ such that, for all $\epsilon \in (0, \bar{\epsilon})$, the origin is an exponentially stable equilibrium of system (1).* ■

2.2. On-policy data-driven LQR: Problem setup

In this paper, we focus on the LQR problem

$$\min_{\substack{x_1,x_2,\dots\\ u_0,u_1,\dots}} \ \mathbb{E}\left[\frac{1}{2}\sum_{t=0}^{\infty}\left(x_t^\top Q x_t + u_t^\top R u_t\right)\right] \tag{5a}$$

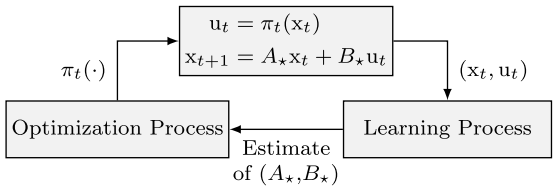

Fig. 1. Schematic representation of the stability-certified on-policy LQR setup.

$$\text{subj. to } x_{t+1} = A_\star x_t + B_\star u_t, \qquad x_0 \sim \mathcal{X}_0, \tag{5b}$$

where $x_t \in \mathbb{R}^n$ and $u_t \in \mathbb{R}^m$ denote, respectively, the state and the input of the system at time $t \in \mathbb{N}$, while $A_\star \in \mathbb{R}^{n\times n}$ and $B_\star \in \mathbb{R}^{n\times m}$ are the state and input matrices. As for the initial condition $x_0 \in \mathbb{R}^n$, we assume that it is drawn from a (known) probability distribution $\mathcal{X}_0$. Hence, the operator $\mathbb{E}[\cdot]$ denotes the expected value with respect to $\mathcal{X}_0$. The cost matrices $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$ satisfy $Q = Q^\top > 0$ and $R = R^\top > 0$. In case of detectability of $(A_\star, Q_0)$ with $Q := Q_0{}^\top Q_0$, the positive definiteness of Q can be relaxed to mere positive semidefiniteness. We characterize $(A_\star, B_\star)$ as follows.

Assumption 2.7 (*Unknown System Properties*)**.** The pair $(A_\star, B_\star)$ is unknown and controllable. ■

As it will be useful later, we collect the pair $(A_\star, B_\star)$ in a single variable $\theta_\star := [A_\star\ B_\star]^\top \in \mathbb{R}^{(n+m)\times n}$.

It is well-known that the optimal solution to problem (5) is given by a linear time-invariant policy $u_t = K_\star x_t$ with the optimal gain $K_\star \in \mathbb{R}^{m\times n}$ given by

$$K_\star = -(R + B_\star^\top P_\star B_\star)^{-1} B_\star^\top P_\star A_\star, \tag{6}$$

where $P_\star \in \mathbb{R}^{n\times n}$ solves the Discrete-time Algebraic Riccati Equation associated to problem (5), see Anderson and Moore (2007).

In this work, we are interested in devising a data-driven on-policy strategy to get a state-feedback controller solution to (5). Thus, the problem can be posed as designing a learning and control scheme that is capable of

(i) learning the optimal policy solution to problem (5),
(ii) estimating the unknown system matrices,
(iii) actuating the (real) system with the currently available state-feedback policy,

while ensuring asymptotic stability properties on the closed-loop learning and control system.

As shown in Fig. 1, our scheme iteratively computes a tentative policy, say $\pi_t(\cdot)$, to be actuated on the real, unknown system. One single sample of the system state is then collected and fed into a learning mechanism refining the matrices estimates, which, in turn, are used to improve the policy. The distinctive feature of our approach is that these steps are interwoven rather than temporally separate, as is typically done in the literature.

2.3. Model-based gradient method for LQR

Next, we recall the key ingredients for devising a model-based gradient method to address problem (5).

2.3.1. Model-based reduced problem formulation

First of all, we recall an equivalent (unconstrained) formulation of problem (5) that explicitly imposes the linear feedback structure to the optimal input and is amenable for gradient-based algorithmic solutions. Problem (5) is rewritten by substituting in the dynamics and in the cost function the input in linear feedback

form $u_t = Kx_t$, where $K \in \mathbb{R}^{m\times n}$ is to be computed. Hence, for all $t \in \mathbb{N}$, the state is uniquely determined as

$$x_t = (A_\star + B_\star K)^t x_0, \qquad x_0 \sim \mathcal{X}_0. \tag{7}$$

By using (7), assuming, without loss of generality, that $\mathcal{X}_0$ is a uniform distribution on the unit sphere, and taking the expectation on x_0, we rewrite problem (5) as

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

where $\mathcal{K} := \{K \in \mathbb{R}^{m\times n} \mid A_\star + B_\star K \text{ is Schur}\} \subseteq \mathbb{R}^{m\times n}$ is the stabilizing gains' set and $J : \mathcal{K} \times \mathbb{R}^{(n+m)\times n} \to \mathbb{R}$ reads as

$$J(K, \theta_\star) = \tag{9}$$
$$\tfrac{1}{2}\,\mathrm{Tr}\left(\sum_{t=0}^{\infty} (A_\star + B_\star K)^{t,\top}(Q + K^\top R K)(A_\star + B_\star K)^t\right).$$

This formulation highlights that (i) the overall problem actually depends on the gain K only, and, (ii) the optimal gain $K_\star$ does not depend on the initial condition x_0.

Remark 2.8. Due to the linearity of the expected value operator and the properties of the trace, the formulation in (9) holds up to a constant scaling factor for any probability distribution $\mathcal{X}_0$ with a well-defined second moment.

2.3.2. Model-based gradient method for problem (8)

The set of stabilizing gains $\mathcal{K}$ is open (Bu et al., 2020b, Lemma IV.3) and connected (Bu et al., 2020b, Lemma IV.6). Moreover, the cost function $J(\cdot, \theta_\star)$ is coercive (Bu et al., 2019, Lemma 3.7). Had the pair $(A_\star, B_\star)$ been known, the gradient descent method could have been used to solve problem (8) (see, e.g., Bu et al. (2019)). Namely, at each iteration $t \in \mathbb{N}$, an estimate K_t of $K_\star$ is maintained and iteratively updated according to

$$K_{t+1} = K_t - \gamma G(K_t, \theta_\star), \tag{10}$$

where $\gamma > 0$ is the stepsize and $G : \mathbb{R}^{m\times n} \times \mathbb{R}^{(n+m)\times n} \to \mathbb{R}^{m\times n}$ 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. By initializing $K_0 \in \mathcal{K}$ and selecting a proper stepsize γ, the optimal gain $K_\star$ is an exponentially stable equilibrium of system (10), see Bu et al. (2019, Theorem 4.6). Given $K \in \mathcal{K}$, we note that $G(K, \theta_\star)$ reads as

$$G(K, \theta_\star) = \left(RK + B_\star^\top P(A_\star + B_\star K)\right) W^c, \tag{11}$$

where $W^c, P \in \mathbb{R}^{n\times n}$ are the solutions to the equations

$$(A_\star + B_\star K) W^c (A_\star + B_\star K)^\top - W^c = -I_n$$
$$(A_\star + B_\star K)^\top P (A_\star + B_\star K) - P = -(Q + K^\top R K).$$

However, since our goal is to address the problem setup described in Section 2.2, the pair $(A_\star, B_\star)$ is unknown. Consequently, the update (10) cannot be implemented.

3. On-policy LQR for unknown systems: Concurrent learning and optimization

In this section, we present RELEARN LQR, a concurrent learning and optimization algorithm developed to solve the stability-certified on-policy LQR setup described in Section 2.2. The proposed on-policy strategy feeds the real system at each iteration t with the current feedback strategy including also an exogenous dithering signal w_t. Then, a new data sample from the system is collected and used to improve the estimates (A_t, B_t) of the unknown $(A_\star, B_\star)$ 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 according to the (approximated) gradient method. Fig. 2 shows the overall scheme.

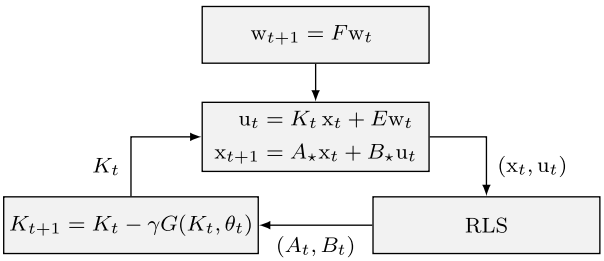

Fig. 2. Representation of the concurrent learning and optimization scheme implemented by RELEARN LQR.

Our RELEARN LQR strategy is reported in Algorithm 1, where $\theta_t \in \mathbb{R}^{(n+m)\times n}$ denotes the estimate of $\theta_\star$ at iteration t and $(A_t, B_t) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ are the corresponding estimates of $(A_\star, B_\star)$. Further, $H_t \in \mathbb{R}^{(n+m)\times(n+m)}$ and $S_t \in \mathbb{R}^{(n+m)\times n}$ are two additional states of the learning part and $\lambda \in (0, 1)$ is a forgetting factor.

Algorithm 1 RELEARN LQR

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

Data collection: generate

$$w_{t+1} = Fw_t$$
$$d_t = Ew_t$$

and actuate

$$u_t = K_t x_t + d_t$$
$$x_{t+1} = A_\star x_t + B_\star u_t$$
$$y_t = x_{t+1}^\top$$

Learning process: compute

$$H_{t+1} = \lambda H_t + \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top \tag{12a}$$
$$S_{t+1} = \lambda S_t + \begin{bmatrix} x_t \\ u_t \end{bmatrix} y_t \tag{12b}$$
$$\theta_{t+1} = \theta_t - \gamma H_t^\dagger (H_t \theta_t - S_t) \tag{12c}$$

Optimization process: update

$$K_{t+1} = K_t - \gamma G(K_t, \theta_t) \tag{13}$$

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

Data collection. Data from the controlled system (5b) are recast in an identification-oriented form described by

$$\underbrace{x_{t+1}^\top}_{y_t} = \underbrace{\begin{bmatrix} x_t^\top & u_t^\top \end{bmatrix}}_{C(x_t, u_t)^\top} \underbrace{\begin{bmatrix} A_\star & B_\star \end{bmatrix}^\top}_{\theta_\star}. \tag{14}$$

Learning process. The adopted learning strategy to compute an estimate of $\theta_\star$ relies on the interpretation of the least squares problem as an online optimization. Specifically, with the measurements (14) at hand, we consider, at each $t \in \mathbb{N}$, the online optimization problem

$$\min_{\theta\in\mathbb{R}^{(n+m)\times n}} \; \tfrac{1}{2} \sum_{\tau=0}^{t} \lambda^{t-\tau} \left\| C(x_\tau, u_\tau)^\top \theta - y_\tau \right\|^2. \tag{15}$$

We iteratively address (15) by updating a solution estimate $\theta_t \in \mathbb{R}^{(n+m)\times n}$ through a "scaled" gradient method with Newton-like

scaling matrix, namely

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

where $\mathcal{H}(\mathrm{x}_\tau, \mathrm{u}_\tau)$ and $\mathcal{S}(\mathrm{x}_\tau, \mathrm{u}_\tau, \mathrm{y}_\tau)$ are defined as

$$\mathcal{H}(\mathrm{x}_\tau, \mathrm{u}_\tau) \coloneqq C(\mathrm{x}_\tau, \mathrm{u}_\tau)C(\mathrm{x}_\tau, \mathrm{u}_\tau)^\top$$
$$\mathcal{S}(\mathrm{x}_\tau, \mathrm{u}_\tau, \mathrm{y}_\tau) \coloneqq C(\mathrm{x}_\tau, \mathrm{u}_\tau)\mathrm{y}_\tau.$$

To avoid storing the whole history of $\mathcal{H}$ and $\mathcal{S}$, we iteratively track them through the matrix states $H_t \in \mathbb{R}^{(n+m)\times(n+m)}$ and $S_t \in \mathbb{R}^{(n+m)\times n}$ giving rise to (12).

Optimization process. The estimate θ_t is concurrently used in the update of the gain K_t, replacing the unavailable $\theta_\star$ into (10) giving rise to (13). To ensure sufficiently informative data, we equip our feedback policy with an additive dithering signal $\mathrm{d}_t \in \mathbb{R}^m$, namely

$$\mathrm{u}_t = K_t\mathrm{x}_t + \mathrm{d}_t, \tag{16}$$

where $\mathrm{d}_t \in \mathbb{R}^m$ is the output of an exogenous, discrete-time oscillator dynamics (see, e.g., Turner (2003)) described by

$$\mathrm{w}_{t+1} = F\mathrm{w}_t \tag{17a}$$
$$\mathrm{d}_t = E\mathrm{w}_t, \tag{17b}$$

where $\mathrm{w}_t \in \mathbb{R}^{n_\mathrm{w}}$, with $n_\mathrm{w} \geq n + m$, is the state, while $F \in \mathbb{R}^{n_\mathrm{w}\times n_\mathrm{w}}$ and $E \in \mathbb{R}^{m\times n_\mathrm{w}}$ are state and output matrices. The design requirements of (17) are as follows.

Assumption 3.1 (*Persistency of Excitation*)**.** The signal w_t is *persistently exciting*, while d_t is *sufficiently rich of order* $(n+1)$, i.e., there exist $\alpha_1, \alpha_2, t_\mathrm{w}, \mathbf{t}_\mathrm{d} > 0$ such that, if $\mathrm{w}_0 \neq 0_{n_\mathrm{w}}$, then, for all $\bar{t} \in \mathbb{N}$, it holds

$$\alpha_1 I_{n_\mathrm{w}} \leq \sum_{\tau=\bar{t}+1}^{\bar{t}+t_\mathrm{w}} \mathrm{w}_\tau\mathrm{w}_\tau^\top \leq \alpha_2 I_{n_\mathrm{w}}, \tag{18a}$$

$$\operatorname{rank}\left(\begin{bmatrix} \mathrm{d}_{\bar{t}} & \mathrm{d}_{\bar{t}+1} & \dots & \mathrm{d}_{\bar{t}+\mathbf{t}_\mathrm{d}-n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{d}_{\bar{t}+n} & \mathrm{d}_{\bar{t}+n+1} & \dots & \mathrm{d}_{\bar{t}+\mathbf{t}_\mathrm{d}-1} \end{bmatrix}\right) = m(n+1), \tag{18b}$$

Further, the eigenvalues of F lie on the unit circle. ■

In finite-time windows, the property (18b) is known as *persistency of excitation of order* $(n+1)$ (see De Persis and Tesi (2019), Willems et al. (2005)), while we used the notion of *sufficient richness* (Bai & Sastry, 1985).

Remark 3.2. A possible way to build F and E to verify Assumption 3.1 is as follows. First, we set $q \coloneqq n+1$, $n_\mathrm{w} = 2q$, and $F \coloneqq \mathrm{blkdiag}(F_1, \dots, F_q)$, where, for all $i \in \{1, \dots, q\}$, $F_i \in \mathbb{R}^{2\times 2}$ is defined as

$$F_i \coloneqq \begin{bmatrix} \cos(\omega_i) & \sin(\omega_i) \\ -\sin(\omega_i) & \cos(\omega_i) \end{bmatrix},$$

for given ω_i such that $\omega_i = 2\omega_{i-1}$ for all $i \in \{2, \dots, q\}$. By choosing an initial condition w_0 that satisfies

$$([\mathrm{w}_0]_{2i-1})^2 + ([\mathrm{w}_0]_{2i})^2 \neq 0, \quad i \in \{1, \dots, q\},$$

the chosen structure of F guarantees (18a) according to Padoan et al. (2017, Thm. 2). As for (18b), it is achieved by selecting E such that $[E^\top\ (EF)^\top\ \dots\ (EF^n)^\top]$ is nonsingular. ■

The closed-loop system resulting from Algorithm 1 is

$$\mathrm{w}_{t+1} = F\mathrm{w}_t \tag{19a}$$
$$\mathrm{x}_{t+1} = (A_\star + B_\star K_t)\mathrm{x}_t + B_\star E\mathrm{w}_t \tag{19b}$$
$$H_{t+1} = \lambda H_t + \begin{bmatrix} \mathrm{x}_t \\ K_t\mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}\begin{bmatrix} \mathrm{x}_t \\ K_t\mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}^\top \tag{19c}$$
$$S_{t+1} = \lambda S_t + \begin{bmatrix} \mathrm{x}_t \\ K_t\mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}\begin{bmatrix} \mathrm{x}_t \\ K_t\mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}^\top \theta_\star \tag{19d}$$
$$\theta_{t+1} = \theta_t - \gamma H_t^\dagger (H_t\theta_t - S_t) \tag{19e}$$
$$K_{t+1} = K_t - \gamma G(K_t, \theta_t), \tag{19f}$$

in which we use the expressions of y_t (cf. (14)) and u_t (cf. (16)). In order to establish the stability properties of the closed-loop system (19), let us introduce the sets $\mathcal{S} \coloneqq R^{n_\mathrm{w}} \times \mathbb{R}^n \times \mathbb{R}^{(n+m)\times(n+m)} \times \mathbb{R}^{(n+m)\times n} \times \mathbb{R}^{(n+m)\times n} \times \mathcal{K}$ and $\mathcal{S}_{SS}(\Pi_1, \Pi_2, \Pi_3) \coloneqq \{(w, x, H, S, \theta, K) \in \mathcal{S} \mid \mathrm{w} \neq 0_{n_\mathrm{w}}, x = \Pi_1\mathrm{w}, H = \mathrm{v}_H(\Pi_2, \mathrm{w}), S = \mathrm{v}_S(\Pi_3, \mathrm{w}), (\theta, K) = (\theta_\star, K_\star)\}$, where $\mathrm{v}_H : \mathbb{R}^{(n+m)^2\times n_\mathrm{w}^2} \times \mathbb{R}^{n_\mathrm{w}} \to \mathbb{R}^{(n+m)\times(n+m)}$ and $\mathrm{v}_S : \mathbb{R}^{(n+m)m\times n_\mathrm{w}^2} \times \mathbb{R}^{n_\mathrm{w}} \to \mathbb{R}^{(n+m)\times n}$ are defined as

$$\mathrm{v}_H(\Pi_1, \mathrm{w}) \coloneqq \mathtt{unvec}\left(\Pi_1\mathtt{vec}\left(\mathrm{w}\mathrm{w}^\top\right)\right) \tag{20a}$$
$$\mathrm{v}_S(\Pi_2, \mathrm{w}) \coloneqq \mathtt{unvec}\left(\Pi_2\mathtt{vec}\left(\mathrm{w}\mathrm{w}^\top\right)\right). \tag{20b}$$

Theorem 3.3. *Let Assumptions 2.7 and 3.1 hold. Then, there exist* $\Pi_\mathrm{x} \in \mathbb{R}^{n\times n_\mathrm{w}}$, $\Pi_H \in \mathbb{R}^{(n+m)^2\times n_\mathrm{w}^2}$, $\Pi_S \in \mathbb{R}^{(n+m)m\times n_\mathrm{w}^2}$, *and* $\bar{\gamma} > 0$ *such that, for all* $\gamma \in (0, \bar{\gamma})$, *the set* $\mathcal{S}_{SS}(\Pi_\mathrm{x}, \Pi_H, \Pi_S)$ *is exponentially stable for system (19).* ■

The proof of Theorem 3.3 is provided in Section 4.3.

Besides stability, Theorem 3.3 ensures exponential convergence of $(\mathrm{x}_t, \theta_t, K_t)$ toward $(\Pi_\mathrm{x}\mathrm{w}_t, \theta_\star, K_\star)$. Namely, for some $a_1, a_2 > 0$, along the trajectories of (19), it holds

$$\|\mathrm{col}(\mathrm{x}_t - \Pi_\mathrm{x}\mathrm{w}_t, \theta_t - \theta_\star, K_t - K_\star)\| \leq a_1\exp(-a_2 t), \tag{21}$$

for all $t \in \mathbb{N}$. Hence, our method asymptotically reconstructs $(A_\star, B_\star)$ and $K_\star$ with linear rate. Moreover, we recall that w_t follows an oscillating dynamics (cf. Assumption 3.1) such that $\|\mathrm{w}_t\| = \|\mathrm{w}_0\|$ for all $t \in \mathbb{N}$. Then, by (21) and for all $\rho > \|\Pi_\mathrm{x}\mathrm{w}_0\|$, the ball $\mathcal{B}_\rho(0_n)$ is exponentially attractive for (19b). Since we can arbitrarily reduce w_0, this implies that the origin of (19b) is practically exponentially stable, provided that the other states lie in their steady-state locus.

Remark 3.4. From the proof of Theorem 3.3 (which uses Proposition 4.3 proved in Appendix C), one can see that the initial condition (θ_0, K_0) must lie in a neighborhood of $(\theta_\star, K_\star)$. Hence, the initialization requirements for RELEARN LQR in Theorem 3.3 are more stringent than those in existing results in the literature, which typically only require an initial stabilizing controller for the true model $(A_\star, B_\star)$ (see Fazel et al. (2018), Hu et al. (2023), Krauth et al. (2019), Lopez et al. (2023), Modares et al. (2016), Pang et al. (2021), Qin et al. (2014), Zhang et al. (2020)). These works, however, only prove convergence to a neighborhood of $K^\star$ and do not analyze the stability of the closed-loop system arising from the interaction between the real system and the proposed algorithm. By contrast, Theorem 3.3 guarantees *exact* convergence to the optimal gain $K^\star$, in addition to stability of the closed-loop system (19). ■

Remark 3.5. In realistic scenarios, an approximate (controllable) model of the system is typically available and can be used as (A_0, B_0). One can then compute a stabilizing controller K_0 for the pair (A_0, B_0). For instance, K_0 can be obtained as the solution of the discrete-time algebraic Riccati equation (6) with (A_0, B_0) in place of the unknown $(A_\star, B_\star)$. ■

Remark 3.6. The proof of Theorem 3.3 exploits system theory tools based on averaging theory for two-time-scale systems (cf. Theorem 2.6 in Section 2.1) and, thus, introduces an auxiliary system called the *averaged system*. Such auxiliary system involves modified averaged dynamics of θ_t and K_t (see Section 4.2) and is shown to have an exponentially stable equilibrium in its origin. Further, we are also able to show that K_t remains a stabilizing gain for $(A_\star, B_\star)$ at all $t \in \mathbb{N}$ (see Appendix C). Such stabilizing property combined with closeness between trajectories of the averaged and original systems (imposed through γ, see, e.g., Bai et al. (1988)), allows for concluding that K_t, generated by (19), stabilizes $(A_\star, B_\star)$ for all $t \in \mathbb{N}$. ■

4. Stability analysis

In this section, we perform the stability analysis of the closed-loop system (19). First, we rewrite it in suitable error coordinates. Second, we resort to the averaging theory to prove the exponential stability of the origin for the averaged system associated to the error dynamics. This result is then exploited to prove Theorem 3.3.

4.1. Closed-loop dynamics in error coordinates

As a preliminary step, we express system (19) into suitable error coordinates. First, we consider vectorized versions of the matrix updates in (19c)–(19d). To this end, let $H^{\mathrm{vc}} \in \mathbb{R}^{(n+m)^2}$ and $S^{\mathrm{vc}} \in \mathbb{R}^{(n+m)n}$ be defined as

$$\begin{bmatrix} H \\ S \end{bmatrix} \longmapsto \begin{bmatrix} H^{\mathrm{vc}} \\ S^{\mathrm{vc}} \end{bmatrix} := \begin{bmatrix} \mathrm{vec}\,(H) \\ \mathrm{vec}\,(S) \end{bmatrix}. \tag{22}$$

Therefore, (19c)–(19d) can be recast as

$$H^{\mathrm{vc}}_{t+1} = \lambda H^{\mathrm{vc}}_t + \mathrm{vec}\left(\begin{bmatrix} \mathrm{x}_t \\ K_t \mathrm{x}_t + E\mathrm{w}_t \end{bmatrix} \begin{bmatrix} \mathrm{x}_t \\ K_t \mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}^\top \right) \tag{23a}$$

$$S^{\mathrm{vc}}_{t+1} = \lambda S^{\mathrm{vc}}_t + \mathrm{vec}\left(\begin{bmatrix} \mathrm{x}_t \\ K_t \mathrm{x}_t + E\mathrm{w}_t \end{bmatrix} \begin{bmatrix} \mathrm{x}_t \\ K_t \mathrm{x}_t + E\mathrm{w}_t \end{bmatrix}^\top \theta_\star \right). \tag{23b}$$

Next, we will inspect (23) together with (19b) to provide the steady-state locus (see, e.g., Isidori (2017, Ch. 12) for a formal definition) when the system is fed with the signal w_t, which evolves according to (19a). To this end, set $n_\chi := n + (n+m)^2 + (n+m)n$ and let $\chi \in \mathbb{R}^{n_\chi}$ be defined as

$$\chi := \mathrm{col}(\mathrm{x}, H^{\mathrm{vc}}, S^{\mathrm{vc}}).$$

Then, by using (23), the dynamics in (19a)–(19d) can be compactly expressed in the new coordinates as

$$\mathrm{w}_{t+1} = F\mathrm{w}_t \tag{24a}$$

$$\chi_{t+1} = \mathcal{A}_K(K_t)\chi_t + \phi(\chi_t, K_t, \mathrm{w}_t), \tag{24b}$$

where we introduced $\mathcal{A}_K : \mathbb{R}^{m\times n} \to \mathbb{R}^{n_\chi \times n_\chi}$ and $\phi : \mathbb{R}^{n_\chi} \times \mathbb{R}^{m\times n} \times \mathbb{R}^{n_\mathrm{w}} \to \mathbb{R}^{n_\chi}$ be defined as

$$\mathcal{A}_K(K) := \begin{bmatrix} A_\star + B_\star K & 0 \\ 0 & \lambda I_{(n+m)(2n+m)} \end{bmatrix} \tag{25a}$$

$$\phi(\chi, K, \mathrm{w}) := \begin{bmatrix} B_\star E\mathrm{w} \\ \mathrm{vec}\left(\begin{bmatrix} \chi_1 \\ K\chi_1 + E\mathrm{w} \end{bmatrix} \begin{bmatrix} \chi_1 \\ K\chi_1 + E\mathrm{w} \end{bmatrix}^\top \right) \\ \mathrm{vec}\left(\begin{bmatrix} \chi_1 \\ K\chi_1 + E\mathrm{w} \end{bmatrix} \begin{bmatrix} \chi_1 \\ K\chi_1 + E\mathrm{w} \end{bmatrix}^\top \theta_\star \right) \end{bmatrix}, \tag{25b}$$

in which $\chi_1 \in \mathbb{R}^n$ denotes the first n components of χ.

System (24) together with the exosystem (19a) is a cascade whose steady-state locus can be characterized by the nonlinear map $\chi^{\mathrm{ss}} : \mathbb{R}^{n_\mathrm{w}} \to \mathbb{R}^{n_\chi}$ defined as

$$\chi^{\mathrm{ss}}(\mathrm{w}) := \begin{bmatrix} \Pi_\mathrm{x}\mathrm{w} \\ \begin{bmatrix} \Pi_H & 0 \\ 0 & \Pi_S \end{bmatrix} \mathrm{vec}\left(\mathrm{w}\mathrm{w}^\top\right) \end{bmatrix}, \tag{26}$$

where Π_x, Π_H, and Π_S are those referred in Theorem 3.3 (see (A.2) and (A.6) in Appendix A for their explicit definition). Formally, the following lemma holds true.

Lemma 4.1. *Let the assumptions of Theorem 3.3 hold true. Consider the map χ^{ss} defined in (26), the feedback gain $K_\star$ solving (5), the matrix F as in (17), and the functions $\mathcal{A}_K$ and ϕ defined in (25). Then, it holds*

$$\chi^{\mathrm{ss}}(F\mathrm{w}) = \mathcal{A}_K(K_\star)\chi^{\mathrm{ss}}(\mathrm{w}) + \phi(\chi^{\mathrm{ss}}(\mathrm{w}), K_\star, \mathrm{w}), \tag{27}$$

for all $\mathrm{w} \in \mathbb{R}^{n_\mathrm{w}}$. Moreover, it holds

$$(\theta_\star^\top \otimes I_{n+m})\Pi_H = \Pi_S. \quad ■ \tag{28}$$

The proof of Lemma 4.1 is provided in Appendix A.

Lemma 4.1 ensures that $\mathrm{col}(\chi^{\mathrm{ss}}(\mathrm{w}), \theta_\star, K_\star)$ is the steady-state locus of the overall closed-loop system (19). In this regard, we also include condition (28) since it allows us to show that $\theta_\star$ is an equilibrium of (19e) restricted to the case in which H_t and S_t lie in the steady-state locus. Indeed, when $\chi_t = \chi^{\mathrm{ss}}(\mathrm{w}_t)$, system (19e) reduces to

$$\begin{aligned} \theta_{t+1}\Big|_{\chi_t=\chi^{\mathrm{ss}}(\mathrm{w}_t)} &= \theta_t - \gamma\,(H_t\theta_t - S_t)\Big|_{\chi_t=\chi^{\mathrm{ss}}(\mathrm{w}_t)} \\ &= \theta_t - \gamma \mathrm{v}_S(\Pi_S, \mathrm{w}_t)\theta_t \\ &= \theta_t - \gamma\,\mathrm{unvec}\left(\left(\theta_\star^\top \otimes I_{n+m}\right)\Pi_H \mathrm{vec}\left(\mathrm{w}_t\mathrm{w}_t^\top\right)\right) \\ &= \theta_t - \gamma \mathrm{v}_H(\Pi_H, \mathrm{w}_t)\,(\theta_t - \theta_\star)\,, \end{aligned}$$

where we use a vectorization operator property[1] and the definitions of v_H and v_S given in (20). As for the equilibrium of (19f) when the other states lie on the steady-state locus, it turns out to be $K_\star$ since $G(K_\star, \theta_\star) = 0$.

Before proceeding, let us collect also the remaining states in (19) in $\mathrm{z} \in \mathbb{R}^{n_\mathrm{z}}$, with $n_\mathrm{z} := (n + 2\,m) \times n$, defined as

$$\mathrm{z} := \mathrm{col}(K, \theta).$$

With Lemma 4.1 at hand, let us introduce the error coordinates $\tilde\chi \in \mathbb{R}^{n_\chi}$ and $\tilde{\mathrm{z}} \in \mathbb{R}^{n_\mathrm{z}}$ defined as

$$\begin{bmatrix} \mathrm{w} \\ \chi \\ \mathrm{z} \end{bmatrix} \longmapsto \begin{bmatrix} \mathrm{w} \\ \tilde\chi \\ \tilde{\mathrm{z}} \end{bmatrix} := \begin{bmatrix} \mathrm{w} \\ \begin{bmatrix} I_n & 0 \\ 0 & \gamma I_{(n+m)(2n+m)} \end{bmatrix} (\chi - \chi^{\mathrm{ss}}(\mathrm{w})) \\ \mathrm{z} - \begin{bmatrix} K_\star \\ \theta_\star \end{bmatrix} \end{bmatrix}. \tag{29}$$

For notational convenience, we will sometimes refer to the components of $\tilde{\mathrm{z}}$ as $\mathrm{col}(\tilde K, \tilde\theta)$. Finally, the closed-loop dynamics (19) in the new coordinates (29) reads as

$$\tilde\chi_{t+1} = \mathcal{A}(\tilde{\mathrm{z}}_t)\tilde\chi_t + h(\tilde{\mathrm{z}}_t, t) + \gamma g(\tilde\chi_t, \tilde{\mathrm{z}}_t, t) \tag{30a}$$

$$\tilde{\mathrm{z}}_{t+1} = \tilde{\mathrm{z}}_t + \gamma f(\tilde\chi_t, \tilde{\mathrm{z}}_t, t), \tag{30b}$$

where $\mathcal{A}(\tilde{\mathrm{z}}) := \mathcal{A}_K(\tilde{\mathrm{z}}_1 + K_\star)$ (cf. (25a)) and we introduced $h : \mathbb{R}^{n_\mathrm{z}} \times \mathbb{N} \to \mathbb{R}^{n_\chi}$, $g : \mathbb{R}^{n_\chi} \times \mathbb{R}^{n_\mathrm{z}} \times \mathbb{N} \to \mathbb{R}^{n_\chi}$, and $f : \mathbb{R}^{n_\chi} \times \mathbb{R}^{n_\mathrm{z}} \times \mathbb{N} \to \mathbb{R}^{n_\mathrm{z}}$ defined respectively as

$$h(\tilde{\mathrm{z}}, t) := \begin{bmatrix} B_\star \tilde{\mathrm{z}}_1 \Pi_\mathrm{x}\mathrm{w}_t \\ 0_{(n+m)(2n+m)} \end{bmatrix} \tag{31a}$$

[1] Given any two matrices $X_1 \in \mathbb{R}^{n_1\times n_2}$ and $X_2 \in \mathbb{R}^{n_2\times n_3}$, it holds $\mathrm{vec}\,(X_1X_2) = (X_2^\top \otimes I_{n_1})\mathrm{vec}\,(X_1)$.

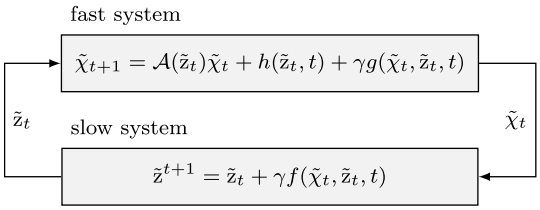

Fig. 3. Block diagram describing system (30).

$$g(\tilde{\chi}, \tilde{z}, t) \tag{31b}$$
$$:= \begin{bmatrix} 0_n \\ \phi_2(\tilde{\chi} + \chi^{ss}(w_t), \tilde{z}_1 + K_\star, w_t) - \phi_2(\chi^{ss}(w_t), K_\star, w_t) \end{bmatrix}$$
$$f(\tilde{\chi}, \tilde{z}, t) := \begin{bmatrix} f_1(\tilde{\chi}, \tilde{z}, t) \\ f_2(\tilde{\chi}, \tilde{z}, t) \end{bmatrix}, \tag{31c}$$

with

$$f_1(\tilde{\chi}, \tilde{z}, t) := -G(\tilde{z}_1 + K_\star, \tilde{z}_2 + \theta_\star) \tag{31d}$$
$$f_2(\tilde{\chi}, \tilde{z}, t) := -(\mathtt{unvec}\,(\tilde{\chi}_2) + H_t^{ss})^\dagger \Big((\mathtt{unvec}\,(\tilde{\chi}_2) + H_t^{ss})\tilde{z}_2 + \mathtt{unvec}\,(\tilde{\chi}_2 - \tilde{\chi}_3)\,\theta_\star \Big), \tag{31e}$$

where, for the sake of readability, in (31) we used the shorthands $\tilde{\chi} := \mathrm{col}(\tilde{\chi}_1, \tilde{\chi}_2, \tilde{\chi}_3)$ and $\tilde{z} = \mathrm{col}(\tilde{z}_1, \tilde{z}_2)$, we partitioned $\phi(\chi, K, w) := \mathrm{col}(\phi_1(\chi, K, w), \phi_2(\chi, K, w))$ (cf. (25b)), we defined $H_t^{ss} \in \mathbb{R}^{(n+m)\times(n+m)}$ as

$$H_t^{ss} := v_H(\Pi_H, w_t), \tag{32}$$

which represents the steady-state value of H_t (see (20a) for the definition of v_H) and we introduced the error coordinates $\tilde{H} \in \mathbb{R}^{(n+m)\times(n+m)}$ and $\tilde{S} \in \mathbb{R}^{(n+m)\times n}$, as

$$\begin{bmatrix} w \\ H^{vc} \\ S^{vc} \end{bmatrix} \longmapsto \begin{bmatrix} w \\ \tilde{H} \\ \tilde{S} \end{bmatrix} := \begin{bmatrix} w \\ \mathtt{unvec}\,(H^{vc} - v_H(\Pi_H, w)) \\ \mathtt{unvec}\,\left(S^{vc} - \Pi_S \mathtt{vec}\,(ww^\top)\right) \end{bmatrix}. \tag{33}$$

We point out that with this transformation we obtained a dynamical system with two-time scales as the one described in Section 2.1 (cf. system (1)). As customary in this context, we distinguish between (i) the fast dynamics (30a) with state $\tilde{\chi}$, and (ii) the slow one (30b) with state $\tilde{z}$. Fig. 3 shows the mentioned interconnected structure of system (30).

In this reformulation, the effect of the exogenous signal w_t is embedded in the time dependency of h, g, and f. Finally, by definition of h, g, and f (cf. (31)) and since $G(K_\star, \theta_\star) = 0$, for all $t \in \mathbb{N}$, we have

$$h(0, t) = 0, \qquad g(0, 0, t) = 0, \qquad f(0, 0, t) = 0. \tag{34}$$

4.2. Averaged system analysis

Next, we carry out the stability analysis of the time-varying system (30) by using the averaging and timescale separation theories (cf. Section 2.1). System (30) enjoys a two-time-scale structure. Hence, we can study (30) by only investigating an auxiliary system typically termed the *averaged system* (see Section 2.1). The latter is obtained by considering the slow dynamics (30b) in which the fast state is frozen to its equilibrium ($\tilde{\chi}_t = 0$ for all $t \in \mathbb{N}$) and the vector field describing the dynamics is averaged with respect to time. The following result is instrumental to properly write the averaged system.

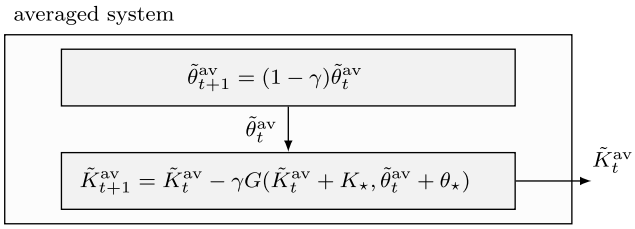

Fig. 4. Block diagram of (36) with $\tilde{z}^{av}_t = \mathrm{col}(\tilde{K}^{av}{}_t, \tilde{\theta}^{av}{}_t)$.

Lemma 4.2. *Let the assumptions of Theorem 3.3 hold true. Consider f defined in (31c). Then, it holds*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{\tau=\bar{t}+1}^{\bar{t}+T} f(0, \tilde{z}, \tau) = -\begin{bmatrix} G(\tilde{K} + K_\star, \tilde{\theta} + \theta_\star) \\ \tilde{\theta} \end{bmatrix} \tag{35}$$

uniformly in $\bar{t} \in \mathbb{N}$ and for all $\tilde{z} = \mathrm{col}(\tilde{K}, \tilde{\theta}) \in \mathbb{R}^{n_z}$. ■

The proof of Lemma 4.2 is given in Appendix B.

Lemma 4.2 provides a suitable approximation of the dynamics of $\tilde{z}$ in (30b) when (i) the convergence of the fast state $\tilde{\chi}$ to its equilibrium has already occurred and (ii) by averaging over time t the vector field $f(0, \tilde{z}, t)$. Specifically, under this approximation, Lemma 4.2 ensures that the two components of the driving term of the dynamics of $\tilde{z}$ are given by (i) a proportional term $-\gamma\tilde{\theta}$ and (ii) an approximated version of the correct gradient $G(\tilde{K} + K_\star, \theta_\star)$. Next, we will leverage averaging theory to prove the stability of the origin for system (30).

Once the averaged vector field has been characterized in Lemma 4.2, we can introduce $f^{av} : \mathbb{R}^{n_z} \to \mathbb{R}^{n_z}$ given by

$$f^{av}(\tilde{z}) := \lim_{T\to\infty} \tfrac{1}{T} \sum_{\tau=\bar{t}+1}^{\bar{t}+T} f(0, \tilde{z}, t).$$

Then, the averaged system associated to (30) reads as

$$\tilde{z}^{av}_{t+1} = \tilde{z}^{av}_t + \gamma f^{av}(\tilde{z}^{av}_t), \tag{36}$$

with state $\tilde{z}^{av}_t := \mathrm{col}(\tilde{K}^{av}{}_t, \tilde{\theta}^{av}{}_t) \in \mathbb{R}^{n_z}$. Expanding the expression of f^{av} (cf. (35)), the dynamics in (36) results in a cascade as depicted in Fig. 4.

For the sake of compactness, we also introduce the (averaged) estimates $A^{av}_t \in \mathbb{R}^{n\times n}$ and $B^{av}_t \in \mathbb{R}^{n\times m}$ of A and B, defined as

$$[A^{av}_t \; B^{av}_t]^\top := \tilde{\theta}^{av}{}_t + \theta_\star. \tag{37}$$

We establish exponential stability of the origin for (36).

Proposition 4.3. *Let the assumptions of Theorem 3.3 hold true. Consider the averaged system (36). Then, there exists $\bar{\gamma}^{av} > 0$ such that, for all $\gamma \in (0, \bar{\gamma}^{av})$, the origin of (36) is exponentially stable.* ■

The proof of Proposition 4.3 is given in Appendix C.

Once this result has been posed, we can proceed with the proof of Theorem 3.3 in the next subsection.

4.3. Proof of Theorem 3.3

We will use Theorem 2.6 given in Section 2.1 to guarantee the exponential stability of the origin for (30). Specifically, in order to apply Theorem 2.6, we need to verify

(i) the exponential stability of the origin for the averaged system (36);
(ii) the Lipschitz continuity of the vector field of the original system (30) (cf. Assumption 2.1);

(iii) that the origin is an equilibrium point of the original system (30) (cf. Assumption 2.2);
(iv) that the matrix function $\mathcal{A}(\tilde{z})$ satisfies Assumption 2.3;
(v) that the difference between the vector fields f and f^{av} of the original (30b) and (36), respectively, satisfies

$$\left\| \frac{1}{T} \sum_{\tau=\bar{t}+1}^{\bar{t}+T} \Delta f(\tilde{z}, \tau) \right\| \leq \nu(T) \left\| \tilde{z} \right\| \tag{38a}$$

$$\left\| \frac{1}{T} \sum_{\tau=\bar{t}+1}^{\bar{t}+T} \frac{\partial \Delta f(\tilde{z}, \tau)}{\partial \tilde{z}} \right\| \leq \nu(T), \tag{38b}$$

for all $\tau \in \mathbb{N}$, where $\Delta f(\tilde{z}, \tau) := f(0, \tilde{z}, \tau) - f^{\text{av}}(\tilde{z})$ and $\nu(t)$ is a nonnegative strictly decreasing function with the property $\nu(t) \to 0$ as $t \to \infty$ (see Assumption 2.5).

Condition (i) follows from Proposition 4.3. Condition (ii) is satisfied by using the quantities in (C.3) in Appendix C and the invertibility of H_t^{ss} (cf. Lemma B.1 in Appendix C) to find the required Lipschitz constants of the vector field of (30). Condition (iii) is verified by (34). As for condition (iv), we note that $\mathcal{A}(\tilde{z})$ is Schur for all $\tilde{K} \in \mathbb{R}^{m \times n}$ (the first component of $\tilde{z}$, see its definition in (29)) such that $(\tilde{K}+K_\star) \in \mathcal{K}$ by definition of $\mathcal{K}$. Hence, condition (iv) is verified with the largest ball contained in $\mathcal{K}$ and centered in $K_\star$. Finally, to check condition (v) (cf. (38)), we use the definitions of f (cf. (31c)) and f^{av} (cf. (35)) to write

$$\Delta f(\tilde{z}, t) = \begin{bmatrix} 0 \\ (H_t^{\text{ss}})^\dagger H_t^{\text{ss}} \tilde{\theta} - \tilde{\theta} \end{bmatrix} \overset{(a)}{=} 0, \tag{39}$$

where in (a) we used the fact that H_t^{ss} is actually a square invertible matrix for all $t \in \mathbb{N}$ (cf. Lemma B.1 in Appendix C). Therefore, the conditions in (38) are satisfied and, thus, the proof follows by Theorem 2.6.

5. Numerical simulations

In this section, we numerically test our strategy. We consider the model of the longitudinal dynamics of a highly maneuverable aircraft linearized at an altitude of 3000 [ft] and a velocity of 0.6 [Mach], see Kapasouris et al. (1990). The resulting linear continuous-time time-invariant dynamics reads as

$$\dot{x} = \begin{bmatrix} -0.0151 & -60.5651 & 0 & -32.174 \\ -0.0001 & -1.3411 & 0.9929 & 0 \\ 0.00018 & 43.2541 & -0.86939 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} x + \begin{bmatrix} -2.516 & -13.136 \\ -0.1689 & -0.2514 \\ -17.251 & -1.5766 \\ 0 & 0 \end{bmatrix} u, \tag{40}$$

where $x \in \mathbb{R}^4$ contains the forward velocity, the attack angle, the pitch rate and the pitch angle, while $u \in \mathbb{R}^2$ contains the elevator and flaperon angles. The discrete-time system matrices $A_\star$ and $B_\star$ are obtained from the continuous-time ones in (40) with a Forward-Euler discretization with sampling time $T_s = 0.05$ [s]. The cost matrices $Q \in \mathbb{R}^{4\times4}$ and $R \in \mathbb{R}^{2\times2}$ are randomly generated ensuring that $Q = Q^\top \geq 0$ and $R = R^\top > 0$. As for the design of F and E, we use the procedure outlined in Remark 3.2 by choosing $\omega_1 = 1/5$, $\omega_{i+1} = \omega_i$ if i is odd and $\omega_i = 2\omega_{i-2}$ if i is even for all $i \in \{2, \dots, n_{\text{w}}\}$. Further, we set $\gamma = 5 \cdot 10^{-4}$ and $\text{w}_0 = [0.01\ 0\ \dots\ 0.01\ 0]^\top$. Finally, the initial condition x_0 is sampled from a normal distribution with mean value 10 for each state.

5.1. Aircraft control

We consider system (40) and run RELEARN LQR with the exogenous signal generated via the procedure detailed above.

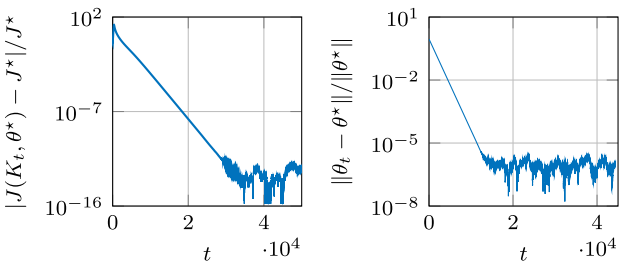

Fig. 5. Evolution of the (left) cost error $|J(K_t, \theta_t) - J^\star|/J^\star$ and (right) estimation error $\|\theta_t - \theta_\star\| / \|\theta_\star\|$.

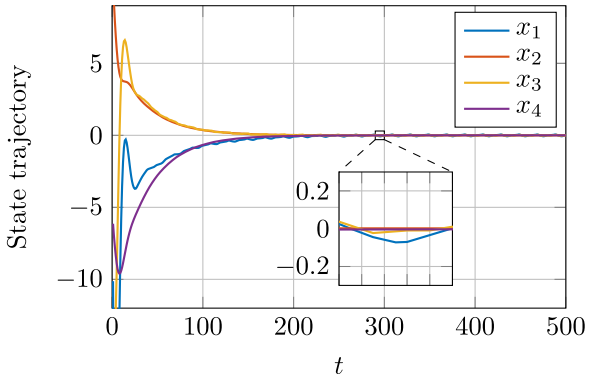

Fig. 6. State trajectory of the closed-loop system.

Fig. 5 shows the evolution of the normalized (left) cost error $|J(K_t, \theta_t) - J^\star|/J^\star$ (with $J^\star := J(K_\star, \theta_\star)$ and $\theta_\star := [A_\star \ B_\star]^\top$) and (right) estimation error $\|\theta_t - \theta_\star\| / \|\theta_\star\|$. The convergence to the optimal cost $J^\star$ and true parameters $\theta_\star$ is achieved. Finally, Fig. 6 shows the closed-loop system trajectory. After a transient, the states oscillate about the origin due to d_t.

5.2. Aircraft control with drifting parameters

To better highlight the capabilities of our algorithm, we also consider the case where the system matrices $A_\star$, $B_\star$, slowly change over time. The new time-varying state and input matrices are denoted as $A_\star^t$ and $B_\star^t$, respectively. More in detail, the time-varying system matrices $A_\star^t$ and $B_\star^t$ smoothly evolve from $A_\star$ and $B_\star$ toward a new pair of matrices A_+ and B_+, according to the update law

$$\begin{bmatrix} A_\star^t & B_\star^t \end{bmatrix} = (1 - \sigma(t)) \begin{bmatrix} A_\star & B_\star \end{bmatrix} + \sigma(t) \begin{bmatrix} A_+ & B_+ \end{bmatrix},$$

for all $t \in \mathbb{N}$, with $\sigma(t)$ being the sigmoid function $\sigma(t) = 1/(1 + \exp((t - t^{\text{mid}})/\alpha))$, where $\alpha \in \mathbb{R}$ determines the transition width and $t^{\text{mid}} \in \mathbb{N}$ defines the transition center. We select $t^{\text{mid}} = 1.5 \cdot 10^5$, $\alpha = 5 \cdot 10^3$, while the entries θ_{ij}^+ of $[A_+ \ B_+] \in \mathbb{R}^{n \times nm}$ are randomly generated as

$$\theta_{ij}^+ = \begin{cases} \theta_{ij} & \text{if } \theta_{ij} = 0 \\ \theta_{ij} + \sigma \theta_{ij}^A & \text{otherwise,} \end{cases}$$

for all $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, nm\}$, where θ_{ij}^A is a random variable normally distributed and $\sigma = 0.1$ is a scale factor. Fig. 7 compares $J(K_t, \theta_{\star,t})$ and $J_t^\star := J(K_{\star,t}, \theta_{\star,t})$, where $\theta_{\star,t} := [A_\star^t \ B_\star^t]^\top$ and $K_{\star,t}$ is the corresponding optimal gain. Finally, Fig. 8 shows the evolution of the normalized (left) cost error $|J(K_t, \theta_{\star,t}) - J_t^\star|/J_t^\star$ and (right) estimation error $\|\theta_t - \theta_{\star,t}\| / \|\theta_{\star,t}\|$. In both cases, asymptotic tracking of $J_t^\star$ and $\theta_{\star,t}$ is achieved. As one may expect, in the neighborhood of the inflection point $t \approx t^{\text{mid}}$, both errors increase. However, we note the adaptability

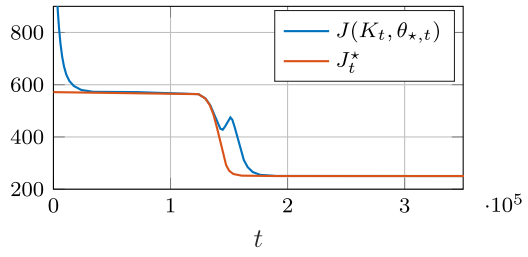

Fig. 7. Comparison between $J(K_t, \theta_{\star,t})$ and $J_t^\star$.

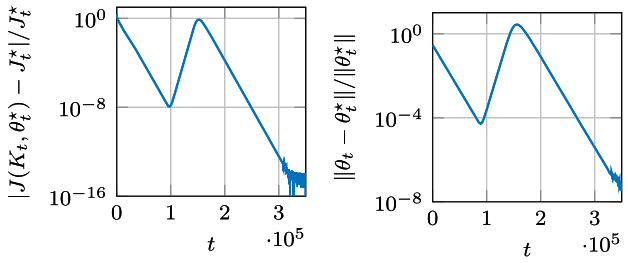

Fig. 8. Evolution of the (left) cost error $|J(K_t, \theta_{\star,t}) - J_t^\star|/J_t^\star$ and (right) estimation error $\|\theta_t - \theta_{\star,t}\| / \|\theta_{\star,t}\|$.

of our policy as it quickly recovers convergence toward $J_t^\star$ and $\theta_{\star,t}$.

6. Conclusions

In this paper, we addressed infinite-horizon LQR problems with unknown system matrices. We proposed a method mixing the identification of the unknown matrices with the optimization of the feedback policy. We proved exponential convergence of the overall closed-loop system to the optimal steady-state associated to the optimal gain and the exact matrices by using tools from Lyapunov-based analysis and averaging theory. Although our analysis considers a time-invariant plant without disturbances, our procedure is inherently applicable to more challenging scenarios in which, e.g., the system and cost matrices vary over time and disturbances affect the plant. Thus, our work paves the way for stability certificates also in these more complex settings. Future research will focus on extending the proposed framework to stochastic settings and on studying the connections between regret and stability analysis.

Appendix A. Proof of Lemma 4.1

Since (27) is obtained by setting $K_t = K_\star$ in (24) (which collects (19a), (19b), and (23)), we study (19a)–(19b) in the manifold in which $K_t = K_\star$, namely

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

$$\mathrm{x}_{t+1} = (A_\star + B_\star K_\star)\mathrm{x}_t + B_\star E\mathrm{w}_t. \tag{A.1b}$$

The steady-state locus of the cascade system (A.1) is $\mathrm{col}(\mathrm{w}_t, \mathrm{x}_t) = \mathrm{col}(I_{n_\mathrm{w}}, \Pi_\mathrm{x})\mathrm{w}_t$, with $\Pi_\mathrm{x} \in \mathbb{R}^{n\times n_\mathrm{w}}$ such that

$$\Pi_\mathrm{x} F = (A_\star + B_\star K_\star)\Pi_\mathrm{x} + B_\star E. \tag{A.2}$$

Since $\sigma(F) \cap \sigma(A_\star + B_\star K_\star) = \emptyset$, Π_x exists and is unique. Then, we study (23) restricted to the manifold in which $(\mathrm{x}_t, K_t) = (\Pi_\mathrm{x}\mathrm{w}_t, K_\star)$. Let $M \in \mathbb{R}^{(n+m)\times(n+m)}$ be

$$M := \begin{bmatrix} \Pi_\mathrm{x}^\top & (K_\star\Pi_\mathrm{x} + E)^\top \end{bmatrix}^\top, \tag{A.3}$$

then it holds

$$\mathtt{vec}\left(\mathrm{w}_{t+1}\mathrm{w}_{t+1}^\top\right) = \mathtt{vec}\left(F\mathrm{w}_t\mathrm{w}_t^\top F^\top\right) \tag{A.4a}$$

$$H_{t+1}^{\mathrm{vc}} = \lambda H_t^{\mathrm{vc}} + \mathtt{vec}\left(M\mathrm{w}_t\mathrm{w}_t^\top M^\top\right) \tag{A.4b}$$

$$S_{t+1}^{\mathrm{vc}} = \lambda S_t^{\mathrm{vc}} + \mathtt{vec}\left(M\mathrm{w}_t\mathrm{w}_t^\top M^\top\theta_\star\right), \tag{A.4c}$$

where (A.4a) comes from the vectorization of (19a). By using the vectorization properties,[2] we rewrite (A.4) as

$$\mathtt{vec}\left(\mathrm{w}_{t+1}\mathrm{w}_{t+1}^\top\right) = (F\otimes F)\mathtt{vec}\left(\mathrm{w}_t\mathrm{w}_t^\top\right) \tag{A.5a}$$

$$H_{t+1}^{\mathrm{vc}} = \lambda H_t^{\mathrm{vc}} + (M\otimes M)\mathtt{vec}\left(\mathrm{w}_t\mathrm{w}_t^\top\right) \tag{A.5b}$$

$$S_{t+1}^{\mathrm{vc}} = \lambda S_t^{\mathrm{vc}} + (\theta_\star{}^\top M\otimes M)\mathtt{vec}\left(\mathrm{w}_t\mathrm{w}_t^\top\right). \tag{A.5c}$$

We note that (A.5) describes a cascade of linear systems with states $\mathtt{vec}\left(\mathrm{w}_t\mathrm{w}_t^\top\right)$ and $(H_t^{\mathrm{vc}}, S_t^{\mathrm{vc}})$. Hence, it is well known that its steady-state can be characterized by resorting to a Sylvester equation. To this end, let $\Pi_H \in \mathbb{R}^{(n+m)^2\times n_\mathrm{w}^2}$ and $\Pi_S \in \mathbb{R}^{(n+m)n\times n_\mathrm{w}^2}$ solve

$$\Pi_H(F\otimes F) = \lambda\Pi_H + M\otimes M \tag{A.6a}$$

$$\Pi_S(F\otimes F) = \lambda\Pi_S + (\theta_\star{}^\top M)\otimes M. \tag{A.6b}$$

Since $\sigma(F\otimes F)\cap\sigma(\lambda I) = \emptyset$, Π_S and Π_H exist and are unique (Bhatia & Rosenthal, 1997). The proof of (27) follows by the definition of χ^{ss} (cf. (26)) and plugging (A.2) and (A.6) into (24).

To show (28), we multiply (A.6a) by $\theta_\star{}^\top \otimes I_{n+m}$ and get

$$\begin{aligned}(\theta_\star{}^\top \otimes I_{n+m})\Pi_H(F\otimes F - \lambda I) &= (\theta_\star{}^\top \otimes I_{n+m})(M\otimes M)\\ &\overset{(a)}{=} (\theta_\star{}^\top M)\otimes M\\ &\overset{(b)}{=} \Pi_S(F\otimes F - \lambda I),\end{aligned} \tag{A.7}$$

where (a) uses a property of the Kronecker operator,[3] while (b) follows from (A.6b). The proof is complete.

Appendix B. Proof of Lemma 4.2

To prove Lemma 4.2, we need the following result.

Lemma B.1. *Let the assumptions of Theorem 3.3 hold true. Then, H_t^{ss} (cf. (32)) is invertible for all $t \in \mathbb{N}$.*

Proof. We prove the invertibility of H_t^{ss} by studying the evolution of H_t. Its dynamics (19c) restricted to the manifold in which $\mathrm{x}_t = \Pi_\mathrm{x}\mathrm{w}_t$ and $K_t = K_\star$ reads as

$$H_{t+1} = \lambda H_t + M\mathrm{w}_t\mathrm{w}_t^\top M^\top, \tag{B.1}$$

with M as in (A.3). The explicit solution of (B.1) is

$$H_t = \lambda^t H_0 + M\underbrace{\left(\sum_{\tau=0}^{t-1}\lambda^{t-1-\tau}\mathrm{w}_\tau\mathrm{w}_\tau^\top\right)}_{=:\mathcal{W}_t}M^\top. \tag{B.2}$$

Since $\lambda \in (0,1)$, the free evolution $\lambda^t H_0$ of (B.2) asymptotically vanishes and does not affect the invertibility of H_t. Thus, we focus on the forced response $M\mathcal{W}_t M^\top$ only. Since $\alpha_1 I_{n_\mathrm{w}} \le \sum_{\tau=\bar{t}+1}^{\bar{t}+t_\mathrm{w}}\mathrm{w}_\tau\mathrm{w}_\tau^\top \le \alpha_2 I_{n_\mathrm{w}}$ for all $\bar{t} \in \mathbb{N}$ (cf. Assumption 3.1), we have $\mathcal{W}_t > 0$ for all $t \ge t_\mathrm{w}$ by Johnstone et al. (1982, Lemma 1). Let us consider the Cholesky decomposition of $\mathcal{W}_t$ given by $\mathcal{W}_t = \mathcal{C}_t\mathcal{C}_t^\top$, with $\mathcal{C}_t \in \mathbb{R}^{n_\mathrm{w}\times n_\mathrm{w}}$ invertible. Then,[4] for all $t \ge t_\mathrm{w}$, we can write

$$\mathrm{rank}\left(M\mathcal{W}_t M^\top\right) = \mathrm{rank}(M\mathcal{C}_t) \overset{(a)}{=} \mathrm{rank}(M), \tag{B.3}$$

[2] Given any $X_1 \in \mathbb{R}^{n_1\times n_2}$, $X_2 \in \mathbb{R}^{n_2\times n_3}$, and $X_3 \in \mathbb{R}^{n_3\times n_4}$, it holds $\mathtt{vec}(X_1X_2X_3) = (X_3^\top \otimes X_1)\mathtt{vec}(X_2)$.

[3] Given any $X_1 \in \mathbb{R}^{n_1\times n_2}$, $X_2 \in \mathbb{R}^{n_3\times n_4}$, $X_3 \in \mathbb{R}^{n_2\times n_5}$, and $X_4 \in \mathbb{R}^{n_4\times n_6}$, it holds $(X_1\otimes X_2)(X_3\otimes X_4) = (X_1X_3)\otimes(X_2X_4)$.

[4] $\mathrm{rank}(XX^\top) = \mathrm{rank}(X) = \mathrm{rank}(X^\top)$ for all $X \in \mathbb{R}^{n\times m}$.

where (a) uses full-rankness of $\mathcal{C}_t$ and a rank property.[5] To compute rank(M), we study system (19a)–(19b) restricted to the manifold in which $K_t = K_\star$, namely

$$\mathrm{w}_{t+1} = F\mathrm{w}_t \tag{B.4a}$$

$$\mathrm{x}_{t+1} = (A_\star + B_\star K_\star)\mathrm{x}_t + B_\star E\mathrm{w}_t. \tag{B.4b}$$

Recalling that $\mathrm{d}_t = E\mathrm{w}_t$ satisfies condition (18b) (cf. Assumption 3.1) and that $(A_\star, B_\star)$ is controllable (cf. Assumption 2.7), we can invoke (Willems et al., 2005, Cor. 2) to claim that

$$\operatorname{rank}\left(\begin{bmatrix} \mathrm{x}_0 & \dots & \mathrm{x}_{\mathbf{t}_\mathrm{d}-1} \\ \mathrm{d}_0 & \dots & \mathrm{d}_{\mathbf{t}_\mathrm{d}-1} \end{bmatrix}\right) = n + m, \tag{B.5}$$

for all $(\mathrm{x}_0, \mathrm{d}_0) \in \mathbb{R}^n \times \mathbb{R}^m$. When the initial condition of (B.4b) lies in the steady-state locus (cf. (A.2)) (i.e., $\mathrm{x}_0 = \Pi_\mathrm{x}\mathrm{w}_0$), the condition in (B.5) simplifies to

$$\operatorname{rank}\left(M\begin{bmatrix} \mathrm{w}_0 & \dots & \mathrm{w}_{\mathbf{t}_\mathrm{d}-1} \end{bmatrix}\right) = n + m, \tag{B.6}$$

which implies[6] rank$(M) \geq n + m$. Being rank$(M) \leq n + m$ by construction, the last inequality yields rank$(M) = n + m$, which combined with (B.3), leads to

$$\operatorname{rank}\left(M\mathcal{W}_t M^\top\right) = n + m, \tag{B.7}$$

for all $t \geq t_\mathrm{w}$. To study rank$(M\mathcal{W}_t M^\top)$, we note that, since $\lambda \in (0, 1)$, $M\mathcal{W}_t M^\top$ exponentially converges to H_t^{ss}. Thus, by continuity, there exists $t_\infty \geq t_\mathrm{w}$ such that

$$\operatorname{rank}(H_t^{\mathrm{ss}}) = n + m,$$

for all $t \geq t_\infty$. Finally, being H_t^{ss} a static function of the periodic signal w_t, then H_t^{ss} is periodic as well so that its full-rankness is independent of t. Thus, it must be that rank$(H_t^{\mathrm{ss}}) = n + m$ for all $t \in \mathbb{N}$, and the proof follows. ■

Now, let us label the two components of f^{av} as

$$\begin{bmatrix} f_1^{\mathrm{av}}(\tilde{\mathrm{z}})^\top & f_2^{\mathrm{av}}(\tilde{\mathrm{z}})^\top \end{bmatrix}^\top := \lim_{T\to\infty} \tfrac{1}{T} \sum_{\tau=\bar{t}+1}^{\bar{t}+T} f(0, \tilde{\mathrm{z}}, \tau).$$

The first block $f_1^{\mathrm{av}}(\tilde{\mathrm{z}})$ of f (cf. (31d)) does not depend on t and, thus, it trivially coincides with f_1, namely

$$f_1^{\mathrm{av}}(\tilde{\mathrm{z}}) := -G(\tilde{K} + K_\star, \tilde{\theta} + \theta_\star),$$

with $\tilde{\mathrm{z}} = \operatorname{col}(\tilde{K}, \tilde{\theta})$. As for $f_2^{\mathrm{av}}(\tilde{\mathrm{z}})$, with $\tilde{\chi} = 0$, it holds $f_2^{\mathrm{av}}(\tilde{\mathrm{z}}) = -\lim_{T\to\infty} \frac{1}{T} \sum_{\tau=\bar{t}+1}^{\bar{t}+T} (H^{\mathrm{ss}}{}_\tau)^\dagger H^{\mathrm{ss}}{}_\tau \tilde{\theta} \overset{(a)}{=} -\tilde{\theta}$, where the last equality uses Lemma B.1.

Appendix C. Proof of Proposition 4.3

The proof resorts to a suitable Lyapunov candidate function whose increment along trajectories of system (36) will allow us to claim exponential stability of the origin.

To ease the notation, we use $\tilde{\mathrm{z}}^{\mathrm{av}} := \operatorname{col}(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}})$. By Bu et al. (2019, Lemma 3.12), J is gradient dominated, namely, it holds

$$J(K, \theta_\star) - J(K_\star, \theta_\star) \leq \mu \left\| G(K, \theta_\star) \right\|^2, \tag{C.1}$$

for all $K \in \mathcal{K}$ and some $\mu > 0$. We now consider the Lyapunov function $V : \mathcal{K} \times \mathbb{R}^{(n+m)\times n} \to \mathbb{R}$ defined as

$$V(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) := \kappa(J(\tilde{K}^{\mathrm{av}} + K_\star, \theta_\star) - J(K_\star, \theta_\star)) + \tfrac{1}{2}\|\tilde{\theta}^{\mathrm{av}}\|^2, \tag{C.2}$$

where $\kappa > 0$ will be set later. Since $(A_\star, B_\star)$ is controllable (cf. Assumption 2.7) and $A_\star + B_\star K_\star$ is Schur, by continuity (see, e.g., Klamka (2008, Thm. 10) and Horn and Johnson (2012, Thm. 6.3.12)), there exists a neighborhood, say it $Z \subseteq \mathcal{K} \times \mathbb{R}^{(n+m)\times n}$, of $(K_\star, \theta_\star)$ such that (i) (A, B) is controllable and (ii) $A + BK$ is Schur for every pair $(K, \theta) := (K, [A\ B]^\top) \in Z$. Let $r^{\mathrm{M}} > 0$ be the radius of the largest ball centered in $(K_\star, \theta_\star)$ and contained in Z. Then, let us arbitrarily choose radius $\tilde{r}^{\mathrm{av}} \in (0, r^{\mathrm{M}})$ and use it to define the constants

$$\beta_0 := \max_{(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) \in \mathcal{B}_{\tilde{r}^{\mathrm{av}}}(0_{n_\mathrm{z}})} \{\|G(\tilde{K}^{\mathrm{av}} + K_\star, \tilde{\theta}^{\mathrm{av}} + \theta_\star)\|\} \tag{C.3a}$$

$$\beta_1 := \max_{(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) \in \mathcal{B}_{\tilde{r}^{\mathrm{av}}}(0_{n_\mathrm{z}})} \left\{ \left\| \tfrac{\partial G(\tilde{K}^{\mathrm{av}} + K_\star, \tilde{\theta}^{\mathrm{av}} + \theta_\star)}{\partial \tilde{K}^{\mathrm{av}}} \right\| \right\} \tag{C.3b}$$

$$\beta_2 := \max_{(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) \in \mathcal{B}_{\tilde{r}^{\mathrm{av}}}(0_{n_\mathrm{z}})} \left\{ \left\| \tfrac{\partial G(\tilde{K}^{\mathrm{av}} + K_\star, \tilde{\theta}^{\mathrm{av}} + \theta_\star)}{\partial \tilde{\theta}^{\mathrm{av}}} \right\| \right\}. \tag{C.3c}$$

Then, we use the decomposition $\tilde{\theta}^{\mathrm{av}} := [\tilde{\theta}_A^{\mathrm{av}}\ \tilde{\theta}_B^{\mathrm{av}}]^\top$ to write

$$(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) \in \mathcal{B}_{\tilde{r}^{\mathrm{av}}}(0_{n_\mathrm{z}})$$
$$\implies (\tilde{\theta}_A^{\mathrm{av}} + A_\star) + (\tilde{\theta}_B^{\mathrm{av}} + B_\star)(\tilde{K}^{\mathrm{av}} + K_\star) \text{ is Schur.} \tag{C.4}$$

By combining Bu et al. (2019, Proposition 3.10) (and similar arguments) with (C.4), it can be shown that the functions $G(\tilde{K}^{\mathrm{av}} + K_\star, \tilde{\theta}^{\mathrm{av}} + \theta_\star)$, $\partial G(\cdot + K_\star, \cdot + \theta_\star)/\partial \tilde{K}^{\mathrm{av}}$, and $\partial G(\cdot + K_\star, \cdot + \theta_\star)/\partial \tilde{\theta}^{\mathrm{av}}$ are continuous in $\mathcal{B}_{\tilde{r}^{\mathrm{av}}}(0_{n_\mathrm{z}})$. In turn, since $\mathcal{B}_{\tilde{r}^{\mathrm{av}}}(0_{n_\mathrm{z}})$ is compact by definition, this implies that β_0, β_1, and β_2 are finite. With this result at hand, we show that, for a proper bound on γ, the set $\mathcal{B}_{r^{\mathrm{av}}}(0_{n_\mathrm{z}})$ is forward invariant for (36) for some $r^{\mathrm{av}} \in (0, \tilde{r}^{\mathrm{av}}]$ that we will determine later. To this end, let us consider $(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) \in \mathcal{B}_{\tilde{r}^{\mathrm{av}}}(0_{n_\mathrm{z}})$ and proceed by induction. Let $(\tilde{K}^{\mathrm{av}}{}_+, \tilde{\theta}^{\mathrm{av}}{}_+) \in \mathbb{R}^{m\times n} \times \mathbb{R}^{(n+m)\times n}$ such that $\operatorname{col}(\tilde{K}^{\mathrm{av}}{}_+, \tilde{\theta}^{\mathrm{av}}{}_+) := \operatorname{col}(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) + \gamma f^{\mathrm{av}}(\operatorname{col}(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}))$. Then, the increment of V along trajectories of (36) is given by

$$\begin{aligned}
&\Delta V(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) := V(\tilde{K}^{\mathrm{av}}{}_+, \tilde{\theta}^{\mathrm{av}}{}_+) - V(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) \\
&= \kappa(J(\tilde{K}^{\mathrm{av}}{}_+ + K_\star, \theta_\star) - J(\tilde{K}^{\mathrm{av}} + K_\star, \theta_\star)) - \gamma\left(1 - \tfrac{\gamma}{2}\right)\|\tilde{\theta}^{\mathrm{av}}\|^2 \\
&\overset{(a)}{=} \kappa J(\tilde{K}^{\mathrm{av}} + K_\star - \gamma G(\tilde{K}^{\mathrm{av}} + K_\star, \tilde{\theta}^{\mathrm{av}} + \theta_\star), \theta_\star) \\
&\quad - \kappa J(\tilde{K}^{\mathrm{av}} + K_\star - \gamma G(\tilde{K}^{\mathrm{av}} + K_\star, \theta_\star), \theta_\star) \\
&\quad + \kappa J(\tilde{K}^{\mathrm{av}} + K_\star - \gamma G(\tilde{K}^{\mathrm{av}} + K_\star, \theta_\star), \theta_\star) \\
&\quad - \kappa J(\tilde{K}^{\mathrm{av}} + K_\star, \theta_\star) - \gamma\,(1 - \gamma/2)\,\|\tilde{\theta}^{\mathrm{av}}\|^2,
\end{aligned} \tag{C.5}$$

where in (a) we add $\pm\kappa J(\tilde{K}^{\mathrm{av}} + K_\star - \gamma G(\tilde{K}^{\mathrm{av}} + K_\star, \theta_\star), \theta_\star)$. By using β_0 (cf. (C.3a)), we note that

$$\|\tilde{K}^{\mathrm{av}} - \gamma G(\tilde{K}^{\mathrm{av}} + K_\star, \tilde{\theta}^{\mathrm{av}} + \theta_\star)\| \leq \tilde{r}^{\mathrm{av}} + \gamma\beta_0, \tag{C.6}$$

for all $(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) \in \mathcal{B}_{\tilde{r}^{\mathrm{av}}}(0_{n_\mathrm{z}})$. Hence, the bound (C.6) ensures $(\tilde{K}^{\mathrm{av}} - \gamma G(\tilde{K}^{\mathrm{av}} + K_\star, \tilde{\theta}^{\mathrm{av}} + \theta_\star), \tilde{\theta}^{\mathrm{av}}) \in \mathcal{B}_{r^{\mathrm{M}}}(0_{n_\mathrm{z}})$ for all $(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) \in \mathcal{B}_{\tilde{r}^{\mathrm{av}}}(0_{n_\mathrm{z}})$ and $\gamma \in (0, \bar{\gamma}_0{}^{\mathrm{av}})$, where $\bar{\gamma}_0{}^{\mathrm{av}} := (r^{\mathrm{M}} - \tilde{r}^{\mathrm{av}})/\beta_0$ Then, let $\beta_3, \beta_4 > 0$ be defined as

$$\beta_3 := \max_{(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) \in \mathcal{B}_{r^{\mathrm{M}}}(0_{n_\mathrm{z}})} \left\{ \left\| \tfrac{\partial G(\tilde{K}^{\mathrm{av}} + K_\star, \tilde{\theta}^{\mathrm{av}} + \theta_\star)}{\partial \tilde{K}^{\mathrm{av}}} \right\| \right\} \tag{C.7a}$$

$$\beta_4 := \max_{(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) \in \mathcal{B}_{r^{\mathrm{M}}}(0_{n_\mathrm{z}})} \left\{ \left\| \tfrac{\partial G(\tilde{K}^{\mathrm{av}} + K_\star, \tilde{\theta}^{\mathrm{av}} + \theta_\star)}{\partial \tilde{\theta}^{\mathrm{av}}} \right\| \right\}. \tag{C.7b}$$

By using (C.7a) and expanding J about $(\tilde{K}^{\mathrm{av}}{}_t + K_\star, \theta_\star)$ at $(\tilde{K}^{\mathrm{av}}{}_t + K_\star - \gamma G(\tilde{K}^{\mathrm{av}}{}_t + K_\star, \theta_\star), \theta_\star)$, we bound (C.5) as

$$\begin{aligned}
&\Delta V(\tilde{K}^{\mathrm{av}}, \tilde{\theta}^{\mathrm{av}}) \leq \kappa J(\tilde{K}^{\mathrm{av}} + K_\star - \gamma G(\tilde{K}^{\mathrm{av}} + K_\star, \tilde{\theta}^{\mathrm{av}} + \theta_\star), \theta_\star) \\
&- \kappa J(\tilde{K}^{\mathrm{av}} + K_\star - \gamma G(\tilde{K}^{\mathrm{av}} + K_\star, \theta_\star), \theta_\star) \\
&- \gamma\kappa(1 - \gamma\tfrac{\beta_3}{2})\|G(\tilde{K}^{\mathrm{av}} + K_\star, \theta_\star)\|^2 - \gamma\left(1 - \tfrac{\gamma}{2}\right)\|\tilde{\theta}^{\mathrm{av}}\|^2.
\end{aligned} \tag{C.8}$$

Now, we handle the first two terms in (C.8). We expand J about $(\tilde{K}^{\mathrm{av}} + K_\star, \theta_\star)$ evaluated at $(\tilde{K}^{\mathrm{av}} + K_\star - \gamma G(\tilde{K}^{\mathrm{av}} + K_\star, \tilde{\theta}^{\mathrm{av}} + \theta_\star), \theta_\star)$ and $(\tilde{K}^{\mathrm{av}} + K_\star - \gamma G(\tilde{K}^{\mathrm{av}} + K_\star, \theta_\star), \theta_\star)$, use (C.7a), the Cauchy–Schwarz inequality, and get

$$J(\tilde{K}^{\mathrm{av}} + K_\star - \gamma G(\tilde{K}^{\mathrm{av}} + K_\star, \tilde{\theta}^{\mathrm{av}} + \theta_\star), \theta_\star)$$

[5] Given $X_1 \in \mathbb{R}^{n_1\times n_2}$ and $X_2 \in \mathbb{R}^{n_2\times n_3}$ it holds rank(X_1X_2) = rank(X_1) if rank$(X_2) = n_2$.

[6] Given $X_1 \in \mathbb{R}^{n_1\times n_2}$ and $X_2 \in \mathbb{R}^{n_2\times n_3}$, it holds rank$(X_1X_2) \leq \min\{\operatorname{rank}(X_1), \operatorname{rank}(X_2)\}$.

$$
\begin{aligned}
&- J(\tilde{K}^{\text{AV}} + K_\star - \gamma G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star), \theta_\star) \\
&\leq \gamma \|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\| \\
&\qquad \cdot \|G(\tilde{K}^{\text{AV}} + K_\star, \tilde{\theta}^{\text{AV}} + \theta_\star) - G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\| \\
&\quad + \tfrac{\gamma^2 \beta_3}{2}(\|G(\tilde{K}^{\text{AV}} + K_\star, \tilde{\theta}^{\text{AV}} + \theta_\star)\|^2 + \|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\|^2) \\
&\overset{(a)}{\leq} \gamma \beta_4 \|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\| \|\tilde{\theta}^{\text{AV}}\| \\
&\quad + \tfrac{\gamma^2 \beta_3}{2} \|G(\tilde{K}^{\text{AV}} + K_\star, \tilde{\theta}^{\text{AV}} + \theta_\star) \pm G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\|^2 \\
&\quad + \tfrac{\gamma^2 \beta_3}{2} \|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\|^2 \\
&\overset{(b)}{\leq} \gamma \beta_4 \|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\| \|\tilde{\theta}^{\text{AV}}\| + \gamma^2 \beta_3 \beta_4 \|\tilde{\theta}^{\text{AV}}\|^2 \\
&\quad + \gamma^2 \beta_3 \|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\|^2 + \tfrac{\gamma^2 \beta_3}{2} \|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\|^2, \qquad \text{(C.9)}
\end{aligned}
$$

where in (a) we use (C.7b) and add $\pm G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)$ inside the norm of the second term, while (b) uses (C.7b) and a standard property of $\|\cdot\|$.[7] By plugging the bound in (C.9) into (C.8), we get

$$
\begin{aligned}
\Delta V(\tilde{K}^{\text{AV}}, \tilde{\theta}^{\text{AV}}) \leq & -\gamma\kappa(1 - \gamma \tfrac{3\beta_3}{2}) \|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\|^2 \\
& + \gamma\kappa\beta_4 \|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\| \|\tilde{\theta}^{\text{AV}}\| \\
& - \gamma(1 - \gamma \tfrac{1 + \kappa\beta_3\beta_4^2}{2}) \|\tilde{\theta}^{\text{AV}}\|^2, \qquad \text{(C.10)}
\end{aligned}
$$

Let us arbitrarily choose $\nu_1, \nu_2 \in (0, 1)$. Then, for all $\gamma \in (0, \bar{\gamma}^{\text{AV}})$ with $\bar{\gamma}^{\text{AV}} := \min\{\bar{\gamma}_0{}^{\text{AV}}, \frac{2\nu_1}{3\beta_3}, 2\nu_2\}$, we further bound (C.10) as

$$
\begin{aligned}
&\Delta V(\tilde{K}^{\text{AV}}{}_t, \tilde{\theta}^{\text{AV}}{}_t) \leq -\gamma\kappa\nu_1 \|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\|^2 \\
&+ \gamma\kappa\beta_4 \|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\| \|\tilde{\theta}^{\text{AV}}\| - \gamma(\nu_2 - \kappa\beta_3\beta_4^2) \|\tilde{\theta}^{\text{AV}}\|^2 \\
&\overset{(a)}{=} -\gamma \begin{bmatrix} \|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\| \\ \|\tilde{\theta}^{\text{AV}}\| \end{bmatrix}^\top U(\kappa) \begin{bmatrix} \|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\| \\ \|\tilde{\theta}^{\text{AV}}\| \end{bmatrix}, \qquad \text{(C.11)}
\end{aligned}
$$

where in (a) we introduce the matrix

$$
U(\kappa) := \begin{bmatrix} \kappa\nu_1 & -\frac{\kappa\beta_4}{2} \\ -\frac{\kappa\beta_4}{2} & \nu_2 - \kappa\beta_3\beta_4^2 \end{bmatrix}.
$$

We set $\kappa \in (0, 4\nu_1\nu_2/\beta_4^2(1 + 4\nu_1\beta_3))$. By Sylvester criterion this implies $U(\kappa) > 0$. Then, by denoting with $\eta > 0$ the smallest eigenvalue of $U(\kappa)$, we bound (C.11) as

$$
\begin{aligned}
&\Delta V(\tilde{K}^{\text{AV}}, \tilde{\theta}^{\text{AV}}) \leq -\gamma\eta(\|G(\tilde{K}^{\text{AV}} + K_\star, \theta_\star)\|^2 + \|\tilde{\theta}^{\text{AV}}\|^2) \\
&\overset{(a)}{\leq} -\gamma \tfrac{\eta}{\mu}(J(\tilde{K}^{\text{AV}} + K_\star, \theta_\star) - J(K_\star, \theta_\star)) - \gamma\eta\|\tilde{\theta}^{\text{AV}}\|^2 \\
&\overset{(b)}{\leq} -\gamma\eta \min\{1/\mu\kappa, 1\} V(\tilde{K}^{\text{AV}}, \tilde{\theta}^{\text{AV}}), \qquad \text{(C.12)}
\end{aligned}
$$

where (a) uses (C.1) and (b) uses (C.2). Moreover, by Bu et al. (2020a, Lemma 3.8) and the Lipschitz continuity of G in $\mathcal{B}_{r^{\text{AV}}}(0_{n_z})$ (see (C.3b)), there exist $c_2 \geq c_1 > 0$ such that

$$
c_1\|\tilde{K}^{\text{AV}}\|^2 \leq J(\tilde{K}^{\text{AV}} + K_\star, \theta_\star) - J(K_\star, \theta_\star) \leq c_2\|\tilde{K}^{\text{AV}}\|^2,
$$

for all $(\tilde{\theta}^{\text{AV}}, \tilde{K}^{\text{AV}}) \in \mathcal{B}_{\tilde{r}^{\text{AV}}}(0_{n_z})$, which, combined with the definition of V (cf. (C.2)), implies

$$
c_3\|(\tilde{K}^{\text{AV}}, \tilde{\theta}^{\text{AV}})\|^2 \leq V(\tilde{K}^{\text{AV}}, \tilde{\theta}^{\text{AV}}) \leq c_4\|(\tilde{K}^{\text{AV}}, \tilde{\theta}^{\text{AV}})\|^2, \qquad \text{(C.13)}
$$

for all $(\tilde{\theta}^{\text{AV}}, \tilde{K}^{\text{AV}}) \in \mathcal{B}_{\tilde{r}^{\text{AV}}}(0_{n_z})$, where $c_3 := \min\{c_1\kappa, 1/2\}$ and $c_4 := \max\{c_2\kappa, 1/2\}$. Then, by monotonicity of V (cf. (C.12)) and its quadratic bounds (cf. (C.13)), we get

$$
\tilde{z}_+ \in \mathcal{B}_{\sqrt{c_4/c_3}\|\tilde{z}\|}(0_{n_z}), \qquad \text{(C.14)}
$$

for all $\tilde{z} \in \mathcal{B}_{\tilde{r}^{\text{AV}}}(0_{n_z})$, where $\tilde{z}_+$ is the update along (36). By observing (C.14) and setting $r^{\text{AV}} := \sqrt{c_3/c_4}\tilde{r}^{\text{AV}}$, we get invariance of $\mathcal{B}_{r^{\text{AV}}}(0_{n_z})$ for system (36). Once this invariance is achieved, we use (C.12) and (C.13) to claim exponential stability of the origin for system (36) (cf. Haddad and Chellaboina (2011, Theorem 13.2)) and the proof follows.

[7] Given any $v_1, v_2 \in \mathbb{R}^n$, it holds $\|v_1 - v_2\|^2 \leq 2\|v_1\|^2 + 2\|v_2\|^2$.

References

Abbasi-Yadkori, Y., & Szepesvári, C. (2011). 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.

Akbari, M., Gharesifard, B., & Linder, T. (2022). Achieving logarithmic regret via hints in online learning of noisy LQR systems. In *IEEE 61st conference on decision and control* (pp. 4700–4705).

Anderson, B. D., & Moore, J. B. (2007). *Optimal control: linear quadratic methods*. Courier Corporation.

Bai, E.-W., Fu, L.-C., & Sastry, S. S. (1988). Averaging analysis for discrete time and sampled data adaptive systems. *IEEE Transactions on Circuits and Systems*, *35*(2), 137–148.

Bai, E.-W., & Sastry, S. S. (1985). Persistency of excitation, sufficient richness and parameter convergence in discrete time adaptive control. *Systems & Control Letters*, *6*(3), 153–163.

Berberich, J., Koch, A., Scherer, C. W., & Allgöwer, F. (2020). Robust data-driven state-feedback design. In *IEEE American control conference* (pp. 1532–1538).

Bhatia, R., & Rosenthal, P. (1997). How and why to solve the operator equation AX- XB=Y. *Bulletin of the London Mathematical Society*, *29*(1), 1–21.

Bian, T., & Jiang, Z.-P. (2016). Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design. *Automatica*, *71*, 348–360.

Bradtke, S. J., Ydstie, B. E., & Barto, A. G. (1994). Adaptive linear quadratic control using policy iteration. In *IEEE American control conference*: *Vol. 3*, (pp. 3475–3479).

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

Bu, J., Mesbahi, A., & Mesbahi, M. (2020a). LQR via first order flows. In *2020 American control conference* (pp. 4683–4688). IEEE.

Bu, J., Mesbahi, A., & Mesbahi, M. (2020b). On topological properties of the set of stabilizing feedback gains. *IEEE Transactions on Automatic Control*, *66*(2), 730–744.

Cassel, A., Cohen, A., & Koren, T. (2020). Logarithmic regret for learning linear quadratic regulators efficiently. In *International conference on machine learning* (pp. 1328–1337). PMLR.

Cohen, A., Koren, T., & Mansour, Y. (2019). Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International conference on machine learning* (pp. 1300–1309). PMLR.

De Persis, C., & Tesi, P. (2019). Formulas for data-driven control: Stabilization, optimality, and robustness. *IEEE Transactions on Automatic Control*, *65*(3), 909–924.

De Persis, C., & Tesi, P. (2021). Low-complexity learning of linear quadratic regulators from noisy data. *Automatica*, *128*, Article 109548.

De Persis, C., & Tesi, P. (2023). Learning controllers for nonlinear systems from data. *Annual Reviews in Control*, Article 100915.

Dean, S., Mania, H., Matni, N., Recht, B., & Tu, S. (2020). On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, *20*(4), 633–679.

Dean, S., Tu, S., Matni, N., & Recht, B. (2019). Safely learning to control the constrained linear quadratic regulator. In *IEEE American control conference* (pp. 5582–5588).

Dörfler, F., Coulson, J., & Markovsky, I. (2022). Bridging direct and indirect data-driven control formulations via regularizations and relaxations. *IEEE Transactions on Automatic Control*, *68*(2), 883–897.

Dörfler, F., Tesi, P., & De Persis, C. (2023). On the certainty-equivalence approach to direct data-driven LQR design. *IEEE Transactions on Automatic Control*.

Fazel, M., Ge, R., Kakade, S., & Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning* (pp. 1467–1476). PMLR.

Ferizbegovic, M., Umenberger, J., Hjalmarsson, H., & Schön, T. B. (2019). Learning robust LQ-controllers using application oriented exploration. *IEEE Control Systems Letters*, *4*(1), 19–24.

Formentin, S., & Chiuso, A. (2018). CoRe: Control-oriented regularization for system identification. In *IEEE conference on decision and control* (pp. 2253–2258).

Haddad, W. M., & Chellaboina, V. (2011). Nonlinear dynamical systems and control. In *Nonlinear dynamical systems and control*. Princeton University Press.

Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press.

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

Iannelli, A., Khosravi, M., & Smith, R. S. (2020). Structured exploration in the finite horizon linear quadratic dual control problem. *IFAC-PapersOnLine*, *53*(2), 959–964.

Isidori, A. (2017). *Lectures in feedback design for multivariable systems*. Springer.
Jiang, Y., & Jiang, Z.-P. (2012). Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, *48*(10), 2699–2704.
Johnstone, R. M., Johnson Jr, C. R., Bitmead, R. R., & Anderson, B. D. (1982). Exponential convergence of recursive least squares with exponential forgetting factor. *Systems & Control Letters*, *2*(2), 77–82.
Kapasouris, P., Athans, M., & Stein, G. (1990). Design of feedback control systems for unstable plants with saturating actuators. In *Proc. IFAC symp. on nonlinear control system design* (pp. 302–307). Pergamon Press.
Kiumarsi, B., Lewis, F. L., & Jiang, Z.-P. (2017). H_∞ control of linear discrete-time systems: Off-policy reinforcement learning. *Automatica*, *78*, 144–152.
Kiumarsi, B., Lewis, F. L., Naghibi-Sistani, M.-B., & Karimpour, A. (2015). Optimal tracking control of unknown discrete-time linear systems using input-output measured data. *IEEE Transactions on Cybernetics*, *45*(12), 2770–2779.
Klamka, J. (2008). Controllability of dynamical systems. *Mathematica Applicanda*, *36*(50/09).
Kleinman, D. (1968). On an iterative technique for Riccati equation computations. *IEEE Transactions on Automatic Control*, *13*(1), 114–115.
Krauth, K., Tu, S., & Recht, B. (2019). Finite-time analysis of approximate policy iteration for the linear quadratic regulator. *Advances in Neural Information Processing Systems*, *32*.
Lopez, V. G., Alsalti, M., & Müller, M. A. (2023). Efficient off-policy Q-learning for data-based discrete-time LQR problems. *IEEE Transactions on Automatic Control*
Mania, H., Tu, S., & Recht, B. (2019). Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, *32*.
Modares, H., Lewis, F. L., & Jiang, Z.-P. (2016). Optimal output-feedback control of unknown continuous-time linear systems using off-policy reinforcement learning. *IEEE Transactions on Cybernetics*, *46*(11), 2401–2410.
Mohammadi, H., Soltanolkotabi, M., & Jovanović, M. R. (2020). On the linear convergence of random search for discrete-time LQR. *IEEE Control Systems Letters*, *5*(3), 989–994.
Mohammadi, H., Zare, A., Soltanolkotabi, M., & Jovanović, M. R. (2021). Convergence and sample complexity of gradient methods for the model-free linear–quadratic regulator problem. *IEEE Transactions on Automatic Control*, *67*(5), 2435–2450.
Padoan, A., Scarciotti, G., & Astolfi, A. (2017). A geometric characterization of the persistence of excitation condition for the solutions of autonomous systems. *IEEE Transactions on Automatic Control*, *62*(11), 5666–5677.
Pang, B., Bian, T., & Jiang, Z.-P. (2018). Data-driven finite-horizon optimal control for linear time-varying discrete-time systems. In *2018 IEEE conference on decision and control* (pp. 861–866). IEEE.
Pang, B., Bian, T., & Jiang, Z.-P. (2021). Robust policy iteration for continuous-time linear quadratic regulation. *IEEE Transactions on Automatic Control*, *67*(1), 504–511.
Possieri, C., & Sassano, M. (2022a). Q-learning for continuous-time linear systems: A data-driven implementation of the kleinman algorithm. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *52*(10), 6487–6497.
Possieri, C., & Sassano, M. (2022b). Value iteration for continuous-time linear time-invariant systems. *IEEE Transactions on Automatic Control*, *68*(5), 3070–3077.
Qin, C., Zhang, H., & Luo, Y. (2014). Online optimal tracking control of continuous-time linear systems with unknown dynamics by using adaptive dynamic programming. *International Journal of Control*, *87*(5), 1000–1009.
Recht, B. (2019). A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, *2*, 253–279.
Rotulo, M., De Persis, C., & Tesi, P. (2020). Data-driven linear quadratic regulation via semidefinite programming. *IFAC-PapersOnLine*, *53*(2), 3995–4000.
Rotulo, M., De Persis, C., & Tesi, P. (2022). Online learning of data-driven controllers for unknown switched linear systems. *Automatica*, *145*, Article 110519.
Sforni, L., Carnevale, G., Notarnicola, I., & Notarstefano, G. (2023). On-policy data-driven linear quadratic regulator via combined policy iteration and recursive least squares. In *IEEE 62nd conference on decision and control* (pp. 5047–5052).
Simchowitz, M., & Foster, D. (2020). Naive exploration is optimal for online LQR. In H. D. III, & A. Singh (Eds.), *Proceedings of machine learning research*: *Vol. 119*, *Proceedings of the 37th international conference on machine learning* (pp. 8937–8948). PMLR.
Turner, C. S. (2003). Recursive discrete-time sinusoidal oscillators. *IEEE Signal Processing Magazine*, *20*(3), 103–111.
Van Waarde, H. J., Eising, J., Trentelman, H. L., & Camlibel, M. K. (2020). Data informativity: a new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, *65*(11), 4753–4768.
Vrabie, D., Pastravanu, O., Abu-Khalaf, M., & Lewis, F. L. (2009). Adaptive optimal control for continuous-time linear systems based on policy iteration. *Automatica*, *45*(2), 477–484.
van Waarde, H. J., Camlibel, M. K., & Mesbahi, M. (2020). From noisy data to feedback controllers: Nonconservative design via a matrix S-lemma. *IEEE Transactions on Automatic Control*, *67*(1), 162–175.
Willems, J. C., Rapisarda, P., Markovsky, I., & De Moor, B. L. (2005). A note on persistency of excitation. *Systems & Control Letters*, *54*(4), 325–329.
Zhang, K., Hu, B., & Basar, T. (2020). Policy optimization for H_2 linear control with H_∞ robustness guarantee: Implicit regularization and global convergence. In *Learning for dynamics and control* (pp. 179–190). PMLR.
Ziemann, I., Tsiamis, A., Sandberg, H., & Matni, N. (2022). How are policy gradient methods affected by the limits of control? In *IEEE 61st conference on decision and control* (pp. 5992–5999).

Lorenzo Sforni received the Ph.D. degree in "Biomedical, Electrical, and Systems Engineering" from the Alma Mater Studiorum Università di Bologna in 2024. From the same institution, he was awarded a Master's degree with honours in 2020. He has been visiting researcher at the California Institute of Technology in 2023. His research interests include nonlinear optimal control, optimization algorithms and reinforcement learning.

Guido Carnevale is a Junior Assistant Professor in the Department of Electrical, Electronic, and Information Engineering G. Marconi at Alma Mater Studiorum Università di Bologna. He received the M.Sc. degree "summa cum laude" in Automation Engineering from the University of Bologna, in 2019, and the Ph.D. degree "summa cum laude" in "Biomedical, Electrical, and Systems Engineering" from the same university. He was a visiting scholar at the University of Oxford in 2022. His research interests include distributed optimization and games over networks, robotics, and optimal control.

Ivano Notarnicola received the M.Sc. in Computer Engineering and the Ph.D. degree in Engineering of Complex Systems from the University of Salento, Lecce (Italy) in 2014 and 2018, respectively. From 2018 to 2020, he was a Post-doctoral fellow with the Department of Electrical Engineering at the University of Bologna (Italy) where he is currently senior assistant professor. He received the 2021 IEEE Transactions on Control of Network Systems Outstanding Paper Award from the IEEE CSS. He is an Associate Editor of the IEEE CSS Conference Editorial Board. His ongoing research focuses on distributed optimization, optimal control algorithms and system theory for optimization algorithms in learning and network systems.

Giuseppe Notarstefano is a Professor in the Department of Electrical, Electronic, and Information Engineering G. Marconi at Alma Mater Studiorum Università di Bologna. He was Associate Professor and Assistant Professor, Ricercatore, at the Università del Salento, Lecce, Italy. He received the Laurea degree "summa cum laude" in Electronics Engineering from the Università di Pisa in 2003 and the Ph.D. degree in Automation and Operation Research from the Università di Padova in 2007. He has been visiting scholar at the University of Stuttgart, University of California Santa Barbara and University of Colorado Boulder. His research interests include distributed optimization and learning, cooperative and distributed robotics, nonlinear optimal control and learning, and trajectory optimization and maneuvering of autonomous vehicles. He has served as an Associate Editor for IEEE Transactions on Automatic Control, IEEE Transactions on Control Systems Technology and IEEE Control Systems Letters. He has been part of the Conference Editorial Board of IEEE Control Systems Society and EUCA. He was a recipient of an ERC Starting Grant 2014.