MR-ARL: Model Reference Adaptive Reinforcement Learning for Robustly Stable On-Policy Data-Driven LQR

Marco Borghesi, Alessandro Bosso, Giuseppe Notarstefano

Abstract

This article introduces a novel framework for data-driven linear quadratic regulator (LQR) design. First, we introduce a reinforcement learning paradigm for on-policy data-driven LQR, where exploration and exploitation are simultaneously performed while guaranteeing robust stability of the whole closed-loop system encompassing the plant and the control/learning dynamics. Then, we propose Model Reference Adaptive Reinforcement Learning (MR-ARL), a control architecture integrating tools from reinforcement learning and model reference adaptive control. The approach stands on a variable reference model containing the currently identified value function. Then, an adaptive stabilizer is used to ensure convergence of the applied policy to the optimal one, convergence of the plant to the optimal reference model, and overall robust closed-loop stability. The proposed framework provides theoretical robustness certificates against real-world perturbations such as measurement noise, plant nonlinearities, or slowly varying parameters. The effectiveness of the proposed architecture is validated via realistic numerical simulations.

Index Terms

Data-driven control, linear quadratic regulator, model reference adaptive control, optimal control, reinforcement learning.

I. INTRODUCTION

REINFORCEMENT LEARNING (RL) is a machine learning field that emerged to perform optimization and decision-making by interacting with an environment without or with limited knowledge of its mathematical model [2], [3]. Over the past years, this field has been successfully applied to multiple domains, including computer games, biology, and economics and finance. RL has garnered the attention of the control engineering community, where it has been used to address optimal control in uncertain or model-free scenarios. Learning from system data aligns RL with principles found in adaptive control literature [4], which seeks to design dynamic controllers for regulation and tracking in the presence of model uncertainties. This work systematically investigates the connection between the fields of optimal and adaptive control, paving the way for a new RL paradigm that provides formal certificates of *robust closed-loop learning and control*, thereby leading to effective performance in real-world applications.

In particular, we focus on solving the infinite-horizon linear quadratic regulator (LQR) problem by developing an *on-policy data-driven algorithm* where data collection and optimization are done simultaneously by applying the learned policy to the actual system. The requirements of our framework are schematically presented below in Table I and later formalized in Section II. A key distinctive feature of our proposed framework is the requirement of robust asymptotic stability for the whole closed-loop system including both the learning and control dynamics. This requirement, as elucidated in the subsequent sections, encapsulates the notion that the proven learning features in nominal cases must persist in perturbed scenarios, encompassing disturbances, measurement noise, slowly varying parameters, and sample-and-hold implementations. With a priori guarantees of effective closed-loop controller implementation, our framework is particularly tailored for safety-critical applications, such as collaborative robotics and aircraft control. Motivated by the above discussion, we provide an overview of the literature pertaining to data-driven LQR, distinguishing between both so-called off-policy and on-policy approaches. Then, we recall model reference adaptive control, one of the main inspirations of the approach of this article.

TABLE I Robustly stable on-policy data-driven LQR

Plant:

$$\dot{x} = Ax + Bu, \tag{1}$$

with state $x \in \mathbb{R}^n$, input $u \in \mathbb{R}^m$, and matrices A and B.

Infinite-horizon LQR: find an optimal control policy $\pi^\star : \mathbb{R}^n \to \mathbb{R}^m$ such that $u(t) = \pi^\star(x(t))$ minimizes, for all initial conditions $x_0 \in \mathbb{R}^n$, the following cost functional along the solutions of the plant:

$$J(x_0, u(\cdot)) := \int_0^\infty x(\tau)^\top Q x(\tau) + u(\tau)^\top R u(\tau) d\tau, \tag{2}$$

with tuning matrices $Q = Q^\top \geq 0$, $R = R^\top > 0$.

Problem: with A and B partially or totally unknown, find a controller of the form

$$\begin{aligned} \dot{z} &= \varphi(x, z, d) &&\text{\textit{learning dynamics}} \\ u &= \pi(x, z, d) &&\text{\textit{applied control policy}} \end{aligned} \tag{3}$$

with $z \in \mathcal{Z} \subset \mathbb{R}^\ell$ and $d \in \mathbb{R}^q$ a dither signal, such that the following properties are achieved.

1) **Exploration:** d probes the uncertainties of A and B.
2) **Exploitation:** map $x \mapsto \pi(x, z, 0)$ converges to the optimal policy $x \mapsto \pi^\star(x)$.
3) **Robust stability:** learning is enforced through robust asymptotic stability of the closed-loop system.

Off-Policy Data-Driven LQR: Off-policy approaches involve finding the optimal policy without applying it at the same time to the actual system. In this context, we find iterative methods, often inspired by the Kleinman algorithm, involving either parameter identification or direct estimate of the policy [5]–[11]. Typically, in these methods, the stabilization of the controlled system during the evolution of the learning algorithm is circumvented by assuming an initial stabilizing policy. However, as discussed in [12], there are situations where this assumption may be unrealistic due to plant uncertainties. The algorithm [13] does not need this assumption. Finally, other relevant paradigms for off-policy approaches involve batch identification of the policy from data [14]–[19] and system-level synthesis [20]. All these approaches differ from our setup, since the exploration and exploitation phases are not performed simultaneously.

On-Policy Data-Driven LQR: As compared to off-policy approaches, the on-policy paradigm adds the significant challenge of ensuring stability of the interconnection between the plant and the control/learning algorithm. Early attempts to address this setup are [21]–[27], where the stabilizing feedback gain is updated at discrete iterations. However, the stability of the whole closed-loop system is not analyzed and an initial stabilizing policy is required, similar to off-policy approaches. A data-driven approach to compute the initial gain is presented in [28]. To the best of our knowledge, [29] is the only work in the literature that provides stability guarantees without a stabilizing initialization, although the focus is on the learning dynamics and not the overall closed-loop system.

Model Reference Adaptive Control: We finally review the literature dealing with model reference adaptive control (MRAC). The principle of this technique is to match the unknown system dynamics to a reference model with desired properties [30]–[32]. To ensure design feasibility, this stabilization technique requires the plant to satisfy constraints named *matching conditions*. A recent work combining MRAC and RL is [33], where RL techniques are used to find the optimal controller for a reference model based on nominal plant parameters. Then, MRAC is applied to assign the reference model to the real system. However, convergence to the desired policy is not proved and can only be ensured to a policy that is optimal for the reference model and not the true system.

Article Contribution: The goal of this article is to lay the foundations for a new paradigm of on-policy data-driven LQR according to the problem setting described in Table I. The main paper contribution is twofold: **(i)** introducing a novel formulation of the on-policy data-driven LQR problem where centrality is given to the stability of the whole closed-loop learning and control system; **(ii)** providing a combined

adaptive control and reinforcement learning design paradigm to address this framework.

Concerning our first contribution, we formulate the on-policy data-driven LQR problem in terms of convergence of the controller, the plant, and an exosystem (modeling the dither signal) to an asymptotically stable set. The fundamental property defining this set is that the learned policy is optimal. Additionally, the set becomes smaller as the dither amplitude is reduced. Thanks to this formulation, we ensure that asymptotic stability in the nominal scenario is preserved, practically and semiglobally, also for a broad class of perturbations, see [34, Ch. 7]. With the generality of the proposed framework, we aim to provide a solid foundation for future work in the field.

The second and main contribution of this work consists in introducing an on-policy learning and control law, termed *Model Reference Adaptive Reinforcement Learning* (MR-ARL), integrating concepts from system identification, adaptive control, and reinforcement learning paradigms. The architecture is structured as a modular actor-critic system with a time-varying reference model bridging the two modules. The actor, inspired to a MRAC architecture, guides the plant to a desired behavior set by the reference model, even in the presence of parametric uncertainties. The reference model is updated online by the critic, which leverages system identification techniques to estimate the dynamics. To impose the desired learning properties, the reference model is driven by a dither signal for which we require suitable richness properties. By relying on analysis tools related to adaptive control, differential inclusions, and singular perturbations, we prove formally that our architecture achieves the following properties for the whole closed-loop system: (i) convergence of the policy to the optimal one; (ii) asymptotic estimation of the true system parameters; (iii) uniform asymptotic stability of an attractor (tunable with the dither amplitude); (iv) robustness in the sense of semiglobal practical asymptotic stability with respect to unmodeled nonlinearities and disturbances. To the best of authors' knowledge, in the context of on-policy data-driven LQR, this algorithm is the first one possessing all these properties. Also, no assumptions are needed about the initial policy. Further, persistency of excitation, needed to ensure convergence, is not assumed a priori, but rather guaranteed by design by resorting to concepts from nonlinear adaptive systems [35]. Finally, given the inherent robustness of the proposed design framework, we ensure that the algorithm is effective in the presence of slowly varying parameters. To validate this property, our numerical simulations cover both the constant parameters case and the one with drifts.

In Section II, we provide preliminary concepts of LQR and introduce the on-policy data-driven paradigm that formalizes Table I. Then, Section III presents MR-ARL and its derivation, while Section IV provides its stability properties. The technical results for the stability analysis are given in Section V, while all related proofs are left in the Appendix. Finally, Section VI provides in-depth numerical results.

Notation: The set of real numbers is denoted by $\mathbb{R}$, while $\mathbb{R}_{\geq 0} := [0, +\infty)$. $(\cdot)^\top$ and $(\cdot)^\dagger$ denote the transpose and the Moore-Penrose pseudo-inverse of real matrices, while $\boldsymbol{S}_0^n$ (resp. $\boldsymbol{S}_+^n$) denotes the set of $n \times n$ symmetric, positive semidefinite (resp. positive definite) real matrices. $|\cdot|$ denotes the induced $2-$norm of real matrices, while $|\cdot|_F$ indicates the Frobenius norm. The notation $\dot{\xi} = f(t, \xi)$, $\xi \in C \subset \mathbb{R}^r$ represents a differential equation having flow set C, i.e., with initial state and flow constrained to C. We refer to [34] for the solution and stability notions of (hybrid) dynamical systems, covering the constrained differential equations of this article.

II. PRELIMINARIES AND PROBLEM SETUP

Following the discussion in the introduction, we now provide a rigorous formulation of the on-policy data-driven linear quadratic regulation (LQR) problem addressed in the paper.

A. *Linear Quadratic Regulation*

We start by introducing the basic concepts of LQR for system (1) and the cost functional (2). The infinite-horizon LQR problem involves finding a *control policy* $\pi^\star : \mathbb{R}^n \to \mathbb{R}^m$ such that applying $u(t) = \pi^\star(x(t))$

for all $t \in \mathbb{R}_{\geq 0}$ solves, for all initial conditions $x_0 \in \mathbb{R}^n$, the following optimal control problem:

$$\min_{u(\cdot)} J(x_0, u(\cdot)) := \int_0^\infty x(\tau)^\top Q x(\tau) + u(\tau)^\top R u(\tau) d\tau$$
$$\text{subject to:} \quad \dot{x}(t) = A x(t) + B u(t), \quad \forall t \in \mathbb{R}_{\geq 0}, \tag{4}$$
$$x(0) = x_0.$$

Under the assumption that pair (A, B) be stabilizable and $(\sqrt{Q}, A)$ be detectable, the LQR problem (4) is solved by the linear policy:

$$\pi^\star(x) := K^\star x, \qquad K^\star := -R^{-1} B^\top P^\star, \tag{5}$$

where $P^\star \in \boldsymbol{S}_0^n$ is the unique solution of the algebraic Riccati equation (ARE):

$$A^\top P^\star + P^\star A - P^\star B R^{-1} B^\top P^\star + Q = 0. \tag{6}$$

Additionally, $P^\star \in \boldsymbol{S}_+^n$ if pair $(\sqrt{Q}, A)$ is observable. We also recall that $P^\star$ specifies the *value function*, which is defined, for a given initial condition x, as the minimum of $J(x, u(\cdot))$:

$$V^\star(x) := \min_{u(\cdot)} J(x, u(\cdot)) := x^\top P^\star x. \tag{7}$$

Consider the differential Riccati equation (DRE) with flow set constrained to symmetric positive semidefinite matrices:

$$\dot{P} = A^\top P + P A - P B R^{-1} B^\top P + Q, \quad P \in \boldsymbol{S}_0^n. \tag{8}$$

Then, if (A, B) is stabilizable and $(\sqrt{Q}, A)$ is detectable, the equilibrium point $P^\star$ is uniformly globally asymptotically stable (UGAS) for the constrained differential equation (8) [36, Thm. 15]. Furthermore, if (A, B) is controllable and $(\sqrt{Q}, A)$ is observable, $P^\star$ is uniformly locally exponentially stable (ULES) [37, Thm. 4]. Formal results describing the stability properties of DRE (8) are provided in [36], [37].

B. Robustly Stable On-Policy Data-Driven LQR

Solving the LQR problem (4) involves computing the solution $P^\star$ of ARE (5), which depends on the matrices A and B of plant (1). Therefore, if A and B are unknown or only partially known, it is necessary to resort to data-driven approaches based on acquiring the data of the state-input sequences $(x(t), u(t))$ (continuously or via batches).

In this work, we are interested in finding an on-policy data-driven algorithm, where data collection and learning are performed simultaneously by applying the learned policy. We now provide a novel rigorous framework to formalize this problem so that its solution guarantees, by design, desirable learning and robust stability properties.

As anticipated in the introduction, the class of controllers that we seek are described by continuous-time dynamical systems of the form

$$\begin{aligned} \dot{z} &= \varphi(x, z, d) \\ u &= \pi(x, z, d) \end{aligned} \quad z \in \mathcal{Z}, \tag{9}$$

where z is the controller state, $\mathcal{Z} \subset \mathbb{R}^\ell$ is a closed set, $d \in \mathbb{R}^q$ is a uniformly bounded signal, named hereafter *dither*, while φ and π are maps that are locally Lipschitz in their arguments. To the algorithm (9), we associate the *learning set* $\mathcal{L}$, defined as:

$$\mathcal{L} := \{z \in \mathcal{Z} : \pi(x, z, 0) = K^\star x, \ \forall x \in \mathbb{R}^n\}, \tag{10}$$

which denotes the set of all controller states such that the control policy π coincides with the optimal policy $\pi^\star$ in (5) whenever the dither d is turned off.

In (9), d is an exogenous signal that may include references for tracking, probing signals, or disturbances. To ensure well-posedness of the problem formulation, from now on we consider a general class of signals d that can be generated by an autonomous dynamical system (*exosystem*) of the form

$$\begin{aligned} \dot{w} &= s(w) \\ d &= h(w) \end{aligned} \qquad w \in \mathcal{W}, \tag{11}$$

where w is the exosystem state, $\mathcal{W} \subset \mathbb{R}^p$ is the set of admissible initial conditions $w(0)$, while s and h are locally Lipschitz maps. Since d is a bounded signal defined for all $t \in \mathbb{R}_{\geq 0}$, we suppose that the set $\mathcal{W}$ be compact and strongly forward invariant under the flow of (11).

Remark 1. *Exosystem* (11) *is not implemented in the control solution but is used to represent the class of admissible signals d. Moreover, the results of the paper hold if* (11) *is replaced by a well-posed hybrid dynamical system [34] to include discontinuous dither signals. Here, we use a continuous-time exosystem to avoid an additional notational burden.*

The closed-loop system resulting from the interconnection of exosystem (11), plant (1), and controller (9) is given by

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{x} &= Ax + B\pi(x, z, h(w)) \qquad (w, x, z) \in \mathcal{W} \times \mathbb{R}^n \times \mathcal{Z}. \\ \dot{z} &= \varphi(x, z, h(w)) \end{aligned} \tag{12}$$

We are ready to precisely state the requirements for controller (9), which include a precise stability characterization for the closed-loop system (12).

Definition 1. *We say that controller* (9) *solves the robustly stable on-policy data-driven LQR problem if the learning set $\mathcal{L}$ in* (10) *is non-empty and, for a chosen class of dither signals d generated by exosystem* (11)*, there exists a compact attractor $\mathcal{A}$, satisfying*

$$\mathcal{A} \subset \mathcal{W} \times \mathbb{R}^n \times \mathcal{L}, \tag{13}$$

that is asymptotically stable for the closed-loop system (12)*.*

We show that any algorithm satisfying Definition 1 covers all of the design requirements stated in the introduction.

- **Exploration**: choosing the dither d determines the shape and the attractivity properties of $\mathcal{A}$, thus it ensures the necessary probing to estimate the optimal policy.
- **Exploitation:** since the projection of $\mathcal{A}$ in the z direction is a subset of the learning set $\mathcal{L}$, uniform attractivity of $\mathcal{A}$ (encoded in asymptotic stability) ensures $z \to \mathcal{L}$ and, thus, correct estimation of the optimal policy.
- **Robust stability:** under the regularity properties required for the controller and assumed for the exosystem, asymptotic stability of the attractor $\mathcal{A}$ is preserved (practically and semiglobally) under a broad range of non-vanishing perturbations arising in real-world scenarios, such as disturbances, measurement noise, unmodeled dynamics, sample-and-hold implementations of the controller, and actuator dynamics.

Remark 2. *In Definition 1, we do not specify the restrictions on the knowledge of A and B to cover a broad range of applications and solutions. However, the prior knowledge on the parametric uncertainties determines the design of φ, π, and $\mathcal{Z}$. Note that, if controller* (9) *is not appropriately chosen, the learning set may be empty.*

Remark 3. *The convergence of z to the learning set $\mathcal{L}$ in Definition 1 implies that the controlled plant becomes asymptotically:*

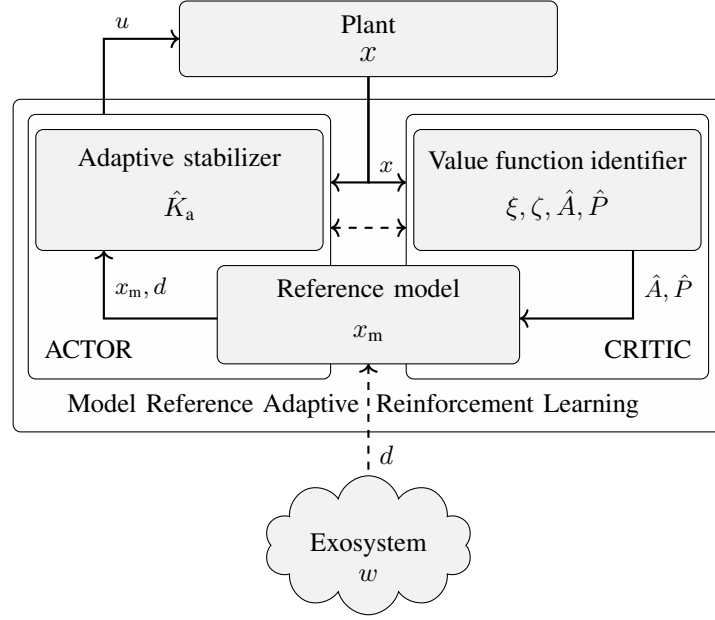$$\dot{x} = (A + BK^{\star})x + \Delta(x, z, d), \tag{14}$$

Fig. 1. Block scheme of the *Model Reference Adaptive Reinforcement Learning*.

where $\Delta(x, z, d) := B(\pi(x, z, d) - K^\star x)$ *vanishes in* $d = 0$. *Moreover, since* $\Delta(x, z, d)$ *is locally Lipschitz, it is uniformly bounded for all* (x, z, w) *in the compact attractor* $\mathcal{A}$. *As a consequence, the input-to-state stability of* (14) *implies that*

$$\limsup_{t\to\infty} |x(t)| \leq \alpha(\limsup_{t\to\infty} |d(t)|), \tag{15}$$

where α *is a class* $\mathcal{K}$ *function. In other words, the ultimate bound of* x *directly depends on the amplitude of the injected dither* d.

III. MODEL REFERENCE ADAPTIVE REINFORCEMENT LEARNING

We now present a new control and learning approach, named *Model Reference Adaptive Reinforcement Learning* (MR-ARL), which satisfies Definition 1 in a scenario of *structured uncertainties* characterized by the following assumptions.

Assumption 1. *There exists a known closed convex set* $\mathcal{C} \subset \mathbb{R}^{n\times n}$ *such that:*
1) $\mathcal{C}$ *has a non-empty interior and* $A \in \text{Int}(\mathcal{C})$;
2) $(\hat{A}, B)$ *is controllable and* $(\sqrt{Q}, \hat{A})$ *is observable for all* $\hat{A} \in \mathcal{C}$.

Remark 4. *From [29], [38], it is known that if there exists* A_0 *such that* (A_0, B) *is controllable and* $(\sqrt{Q}, A_0)$ *is observable, then there exists a scalar* $\rho > 0$ *such that* (A, B) *is controllable and* $(\sqrt{Q}, A)$ *is observable for all* A *such that* $|A - A_0| \leq \rho$. *Therefore, during implementation, the set* $\mathcal{C}$ *can be chosen as a ball centered in a nominal value* A_0 *of* A *with radius* ρ *chosen to include all possible uncertainties.*

Assumption 2. *Consider the linear map* $\mathcal{B} : K \in \mathbb{R}^{m\times n} \longmapsto BK \in \mathbb{R}^{n\times n}$, *where* B *is the input matrix in* (1). *For some known* $A_0 \in \mathcal{C}$, *it holds that*

$$A_0 - A \in \text{Im}(\mathcal{B}). \tag{16}$$

Remark 5. *Assumption 2 is an alternative formulation of the matching conditions used in the MRAC literature [30]. Notice that for any* $\hat{A}$ *such that* $\hat{A} - A_0 \in \text{Im}(\mathcal{B})$, *by* $\hat{A} - A_0 + (A_0 - A) = \hat{A} - A \in \text{Im}(\mathcal{B})$, *there exists* $K_a \in \mathbb{R}^{m\times n}$ *such that*

$$\hat{A} - A = BK_a. \tag{17}$$

Algorithm 1 MR-ARL

Initialization:
$\hat{P}(0) \in \boldsymbol{S}_0^n$, $\hat{A}(0) \in \Theta$, with Θ from Remark 5
$\lambda, \gamma, \nu, g, \mu > 0$ design gains
$d(t)$: bounded and stationary signal with each entry sufficiently rich of order $n + 1$ and uncorrelated
Repeat:
Swapping filters:

$$\dot{\xi} = -\lambda\xi + x, \qquad \dot{\zeta} = -\lambda(x + \zeta) - Bu \tag{19}$$

Identifier dynamics:

$$\dot{\hat{A}} = \underset{\hat{A} \in \mathcal{C}}{\mathrm{Proj}}\left\{-\gamma BB^\dagger \frac{\epsilon\xi^\top}{1 + \nu|\xi||\epsilon|}\right\}, \quad \epsilon := \hat{A}\xi - (x + \zeta) \tag{20}$$

Value iteration:

$$\dot{\hat{P}} = g\left(\hat{A}^\top\hat{P} + \hat{P}\hat{A} - \hat{P}BR^{-1}B^\top\hat{P} + Q\right) \tag{21}$$

Reference model:

$$\dot{x}_\mathrm{m} = (\hat{A} - BR^{-1}B^\top\hat{P})x_\mathrm{m} + Bd \tag{22}$$

Adaptive gain dynamics:

$$\dot{\hat{K}}_\mathrm{a} = -\mu B^\top\hat{P}(x - x_\mathrm{m})x^\top + B^\dagger\dot{\hat{A}} \tag{23}$$

System input:

$$u = -R^{-1}B^\top\hat{P}x + \hat{K}_\mathrm{a}x + d \tag{24}$$

Given the two sets $\mathcal{C}$ and $A_0 + \mathrm{Im}(\mathcal{B})$, we are interested in all matrices $\hat{A} \in \Theta$, where

$$\Theta := \mathcal{C} \cap (A_0 + \mathrm{Im}(\mathcal{B})) \tag{18}$$

since they can be used to build reference models.

Our controller is conceived as an *actor-critic* modular architecture where a *reference model* bridges the two parts of the design. The resulting structure is a MRAC where the reference model is continuously updated with value iteration, thus we aptly name it *Model Reference Adaptive Reinforcement Learning*. We introduce the building blocks of the design.

- **Critic:** this block performs *data-driven value function identification* to build an optimal and asymptotically stable reference model. In particular, a gradient identifier computes an estimate $\hat{A} \in \Theta$ of A that is used to obtain an estimate $\hat{P}$ of the solution $P^\star$ of ARE (6). In this respect, Assumption 1 guarantees that for any estimate $\hat{A}$ the computation of $\hat{P}$ is feasible. Then, $\hat{A}$ and the optimal gain estimate $-R^{-1}B^\top\hat{P}$ are used to build a reference model having state matrix $\hat{A} - BR^{-1}B^\top\hat{P}$. As input to the reference model, we consider a dither d with sufficient richness properties to ensure convergence to the true system parameters and to the optimal policy.
- **Actor:** this block assigns the input to the plant to *adaptively track the reference model*. During the transient, the feedback gain $-R^{-1}B^\top\hat{P}$ may be not stabilizing for the real system. For this reason, the actor introduces in the control law an additional adaptive feedback gain $\hat{K}_\mathrm{a}$ to cancel the mismatch between the estimated matrix $\hat{A}$ and the real A. Canceling such a mismatch is possible due to Assumption 2.

See Fig. 1 for a block scheme of Model Reference Adaptive Reinforcement Learning. The full description of the design is presented in Algorithm 1 and discussed in detail in the next subsections.

A. Critic: Value Function Identifier

In this subsection, we build a continuous-time identifier of $P^\star$ based on the estimation of matrix A. Given the structure of system (1), we compute an estimate $\hat{A} \in \Theta$ of A by designing a swapping filter of the form (19), with $\lambda > 0$ a scalar gain for tuning the filter time constant. Using the filter states, define the *prediction error*

$$\epsilon := \hat{A}\xi - (x + \zeta), \tag{25}$$

which we can rewrite as $\epsilon = (\hat{A} - A)\xi + \tilde{\epsilon}$, where

$$\tilde{\epsilon} := A\xi - (x + \zeta), \tag{26}$$

is an error signal that is shown in Section V to converge exponentially to zero. Since the available signal ϵ converges exponentially to $(\hat{A} - A)\xi$, which contains the parameter estimation error $\hat{A} - A$, we can use the normalized projected gradient descent algorithm (20) to update the estimate $\hat{A}$. In (20), parameters $\gamma > 0$ and $\nu > 0$ are scalar gains, while the multiplicative term $BB^\dagger$ is a projection onto $\mathrm{Im}(\mathcal{B})$ [39, Sec. 5.5.4]. This projection is needed to ensure, given any initialization $\hat{A}(0) \in A_0 + \mathrm{Im}(\mathcal{B})$, that the estimate $\hat{A}(t)$ never leaves this subspace. Finally, $\mathrm{Proj}_{\hat{A} \in \mathcal{C}}\{\cdot\}$ is a Lipschitz continuous parameter projection operator, whose expression is provided, e.g., in [40, Appendix E] and depends on the shape of the set $\mathcal{C}$.

Given the estimate $\hat{A}$, we are interested in computing the matrix $\hat{P} \in \boldsymbol{S}^n_+$ that solves the ARE

$$\mathcal{R}(P, \hat{A}) := \hat{A}^\top P + P\hat{A} - PBR^{-1}B^\top P + Q = 0. \tag{27}$$

From [1], such a matrix could be obtained by computing the map $\mathcal{P}(\hat{A})$ that solves (27) for each $\hat{A}$, i.e., such that:

$$\mathcal{R}(\mathcal{P}(\hat{A}), \hat{A}) = 0, \quad \text{for all } \hat{A} \in \Theta. \tag{28}$$

For simplicity in the implementation and inspired by [29], in Algorithm 1, we compute $\hat{P}$ via the dynamical system (21), which is a DRE rescaled by the tuning gain $g > 0$. Notice that, if $\hat{A}$ is constant, then the solution $\hat{P}$ of (21) converges to $\mathcal{P}(\hat{A})$.

Remark 6. *Assumption 1 guarantees that, for each $\hat{A} \in \Theta$, $\mathcal{P}(\hat{A})$ exists, is unique, and positive definite for any $\hat{A} \in \Theta$. Although stabilizability of $(\hat{A}, B)$ would be sufficient in Assumption 1 for the solvability of $\mathcal{R}(P, \hat{A}) = 0$, controllability is essential to guarantee convergence of the identifier under sufficient richness of the dither d. Moreover, we need observability instead of simple detectability to ensure that $\mathcal{P}(\hat{A})$ is positive definite.*

Remark 7. *From Assumption 1 and the parameter projection in (20), matrix $\hat{A} - BR^{-1}B^\top\mathcal{P}(\hat{A})$ is Hurwitz by design. Therefore, if $\hat{A}$ converges to a constant matrix, (21) ensures that $\hat{A} - BR^{-1}B^\top\hat{P}$ converges to a Hurwitz matrix.*

B. Reference Model

Given the estimate $\hat{P}$ of $P^\star$, we design a reference model for system (1). The reference model has to embed all the properties required for the plant, i.e., robust stability, optimality and persistency of excitation. To these aims, consider system (22), where $x_\mathrm{m} \in \mathbb{R}^n$ is the reference model state. We embed the stability and optimality properties through $\hat{A} - BR^{-1}B^\top\hat{P}$, which is designed to converge to a Hurwitz matrix.

Remark 8. *Different from classic MRAC, the state matrix $\hat{A} - BR^{-1}B^\top\hat{P}$ of the reference model (22) is not constant but time-varying as it depends on the estimates $\hat{A}$ and $\hat{P}$. This property leads to an adaptive design where the known-plant stabilizing gains are time-varying.*

Finally, we embed the persistency of excitation properties through dither $d \in \mathbb{R}^m$, which is chosen such that it is a bounded stationary signal, whose entries are sufficiently rich of order $n + 1$ and uncorrelated.

Remark 9. *We model $d(t)$ as the output of an exosystem of the form (11). It is not necessary to actually implement the exosystem as part of the algorithm, as we show in the numerical example.*

C. Actor: Model Reference Adaptive Controller

Given the reference model (22), we design an adaptive controller for system (1). Define the tracking error $e := x - x_{\mathrm{m}}$ and compute its time derivative from (1), (22) as

$$
\begin{aligned}
\dot{e} &= Ax + Bu - (\hat{A} - BR^{-1}B^\top \hat{P})(x - e) - Bd \\
&= (\hat{A} - BR^{-1}B^\top \hat{P})e + (A - \hat{A})x + B(u + R^{-1}B^\top \hat{P}x - d).
\end{aligned}
\tag{29}
$$

To ensure that the plant (1) asymptotically copies the behavior of the reference model (22), i.e., $e(t) \to 0$, we exploit the fact that $\hat{A} \in \Theta$. In particular, from (17), for each $\hat{A}$ there exists $K_{\mathrm{a}}(\hat{A}) \in \mathbb{R}^{m \times n}$ such that

$$
\hat{A} - A = BK_{\mathrm{a}}(\hat{A}).
\tag{30}
$$

More specifically, we can ensure that map $K_{\mathrm{a}}(\hat{A})$ is smooth in $\hat{A}$ by choosing:

$$
K_{\mathrm{a}}(\hat{A}) := B^\dagger (\hat{A} - A),
\tag{31}
$$

where $B^\dagger$ denotes the Moore-Penrose pseudoinverse of B. This way, (29) becomes

$$
\dot{e} = (\hat{A} - BR^{-1}B^\top \hat{P})e + B(u + R^{-1}B^\top \hat{P}x - K_{\mathrm{a}}(\hat{A})x - d),
\tag{32}
$$

suggesting a control law of the form

$$
u := -R^{-1}B^\top \hat{P}x + K_{\mathrm{a}}(\hat{A})x + d
\tag{33}
$$

if the plant dynamics were known. However, $K_{\mathrm{a}}(\hat{A})$ is unavailable for design as it depends also on A, as highlighted in (30), thus we consider the certainty-equivalence-based adaptive controller given in (24), where $K_{\mathrm{a}}(\hat{A})$ is replaced by the adaptive gain $\hat{K}_{\mathrm{a}}$, driven by the adaptive law (23) where $\mu > 0$ is a scalar gain. The first term in the adaptive law (23) is a standard update to ensure the error e goes asymptotically to zero in a framework where the model mismatch is constant. However, since $\hat{A}$ is continuously updated by identifier (20), the second term in the update law takes into account the time-varying mismatch.

IV. MAIN RESULT

We now provide the main results of this work, where we show that Model Reference Adaptive Reinforcement Learning solves the robustly stable on-policy data-driven LQR problem as per Definition 1. The first result is given supposing to have a DRE dynamics in (21) infinitely faster than the rest of the system (*reduced-order system*), i.e., supposing $\hat{P}(t) = \mathcal{P}(A(t))$ as in (28) at each t. For this reason, we mark the results for the reduced-order system with a subscript s to highlight its slow dynamics. We then follow a singular perturbation approach to prove also Algorithm 1 solves the robustly stable on-policy data-driven LQR problem.

A. Stability Result for the Reduced-Order System

Consider the Model Reference Adaptive Reinforcement Learning with ARE implementation of $\hat{P}$ as in (28). Following the notation of Section II, the controller obtained by combining the value function identifier (19), (20), (28), reference model (22), and the adaptive stabilizer (23), (24) is in the form (9), with state

$$
z_{\mathrm{s}} := (\xi, \zeta, \hat{A}, x_{\mathrm{m}}, \hat{K}_{\mathrm{a}}) \in \mathcal{Z}_{\mathrm{s}} := \mathbb{R}^{2n} \times \Theta \times \mathbb{R}^n \times \mathbb{R}^{m \times n},
\tag{34}
$$

and output policy

$$
\pi(x, z_{\mathrm{s}}, d) := (\hat{K}_{\mathrm{a}} - BR^{-1}B^\top \mathcal{P}(\hat{A}))x + d,
\tag{35}
$$

from which it follows that the learning set $\mathcal{L}$ in (10) is non-empty and given by

$$
\mathcal{L}_{\mathrm{s}} := \{z_{\mathrm{s}} \in \mathcal{Z}_{\mathrm{s}} : \hat{K}_{\mathrm{a}} - BR^{-1}B^\top \mathcal{P}(\hat{A}) = -BR^{-1}B^\top P^\star\}.
\tag{36}
$$

In particular, we recall that our algorithm aims at reaching the learning set by ensuring $\hat{A} \to A$ (hence $\hat{P} \equiv \mathcal{P}(\hat{A}) \to P^\star$ by continuity of the map $\mathcal{P}(\hat{A})$) and $\hat{K}_a \to 0$. The following result shows that, with $\gamma > 0$ sufficiently small, the Model Reference Adaptive Reinforcement Learning with ARE implementation of $\hat{P}$ solves the robustly stable on-policy data-driven LQR problem.

Theorem 1. *Consider the closed-loop system given by the interconnection of plant* (1) *and the controller of Algorithm 1, with $\hat{P}(t) = \mathcal{P}(\hat{A}(t))$ for all t and $\mathcal{P}(\hat{A})$ satisfying* (28). *Let the stationary dither d be generated by an exosystem of the form* (11) *and let its entries be sufficiently rich of order $n + 1$ and uncorrelated. Then, there exists $\gamma^\star > 0$ such that, for all $\gamma \in (0, \gamma^\star]$, there exists a compact set $\mathcal{A}_s$ satisfying*

$$\mathcal{A}_s \subset \{(w, x, z_s) \in \mathcal{W} \times \mathbb{R}^n \times \mathcal{L}_s : \hat{A} = A, \hat{K}_a = 0, x = x_m, \epsilon = 0\} \tag{37}$$

that is uniformly globally asymptotically stable.

B. Stability Result for MR-ARL

Consider the Model Reference Adaptive Reinforcement Learning (MR-ARL) algorithm with DRE implementation of $\hat{P}$ as in (21) (Algorithm 1). Following the notation of Section II, the controller obtained by combining the value function identifier (19), (20), (21), reference model (22), and the adaptive stabilizer (23), (24) is in the form (9), with state

$$z := (z_s, \hat{P}) \in \mathcal{Z} := \mathcal{Z}_s \times \boldsymbol{S}_0^n, \tag{38}$$

where z_s and $\mathcal{Z}_s$ are given in (34). The output policy then becomes

$$\pi(x, z, d) := (\hat{K}_a - BR^{-1}B^\top \hat{P})x + d, \tag{39}$$

and the learning set $\mathcal{L}$ is given by

$$\mathcal{L} := \{z \in \mathcal{Z} : \hat{K}_a - BR^{-1}B^\top \hat{P} = -BR^{-1}B^\top P^\star\}. \tag{40}$$

In this case, the learning set is reached via $\hat{A} \to A$ (hence $\hat{P} \to P^\star$ by asymptotic stability of the DRE (8)) and $\hat{K}_a \to 0$. The next result, which is the main result of this work, shows that with $\gamma > 0$ sufficiently small and $g > 0$ sufficiently large, the Model Reference Adaptive Reinforcement Learning in Algorithm 1 solves the robustly stable on-policy data-driven LQR problem.

Theorem 2. *Consider the closed-loop system given by the interconnection of plant* (1) *and the controller of Algorithm 1. Pick d and γ as in Theorem 1. Then, the compact set*

$$\mathcal{A} := \mathcal{A}_s \times P^\star \subset \{(w, x, z) \in \mathcal{W} \times \mathbb{R}^n \times \mathcal{L} : \hat{A} = A, \hat{K}_a = 0, x = x_m, \epsilon = 0, P = P^\star\} \tag{41}$$

is semiglobally uniformly asymptotically stable in the tuning parameter $g > 0$, where $\mathcal{A}_s$ is given in (37) *and g is the one in* (21). *Namely, for any compact set $\mathcal{K} \subset \mathcal{W} \times \mathbb{R}^n \times \mathcal{Z}$ of initial conditions for the closed-loop system, there exists $g > 0$ such that $\mathcal{A}$ is uniformly asymptotically stable with domain of attraction containing $\mathcal{K}$.*

V. ALGORITHM ANALYSIS

In the following, we will only study the properties of the reduced-order version of the algorithm, i.e., with $\hat{P}(t) = \mathcal{P}(A(t))$ for all t and $\mathcal{P}(\hat{A})$ satisfying (28). The second result, i.e., the stability of Algorithm 1 (implementing the DRE), is obtained by invoking singular perturbations techniques.

A. Error Dynamics

We begin the analysis by presenting the closed-loop dynamics in error coordinates, which is used to provide the technical results of the following subsections.

1) Identifier Dynamics: Consider the error coordinate $\tilde{\epsilon}$ in (26), which can be written as

$$\tilde{\epsilon} := A\xi - (x + \zeta). \tag{42}$$

Then, from (1), (19), it holds that

$$\begin{aligned}
\dot{\tilde{\epsilon}} &= A(-\lambda\xi + x) - (Ax + Bu - \lambda(x + \zeta) - Bu) \\
&= -\lambda(A\xi - (x + \zeta)) = -\lambda\tilde{\epsilon},
\end{aligned} \tag{43}$$

which ensures that the prediction error $\epsilon := \hat{A}\xi - (x + \zeta) = (\hat{A} - A)\xi + \tilde{\epsilon}$ converges to $(\hat{A} - A)\xi$ exponentially.

Define $\tilde{A} := \hat{A} - A$. Then, from (26), (43), we can rewrite the identifier dynamics (19), (20), (25) in error coordinates as the following cascaded system

$$\begin{aligned}
\dot{\tilde{\epsilon}} &= -\lambda\tilde{\epsilon} \\
\dot{\hat{A}} &= \Proj_{\hat{A} \in \mathcal{C}} \left\{ -\gamma BB^\dagger \frac{\tilde{A}\xi\xi^\top + \tilde{\epsilon}\xi^\top}{1 + \nu|\xi||\epsilon|} \right\},
\end{aligned} \tag{44}$$

driven by $\xi(t)$, solution of the filter

$$\dot{\xi} = -\lambda\xi + x. \tag{45}$$

Remark 10. *To ensure $\hat{A}(t) \to A$, it is known from the adaptive control literature that vector $\xi(t)$ must be a persistently exciting (PE) signal [31]. However, notice that $\xi(t)$ is a filtered version of $x(t)$, which is generated in closed-loop by interconnecting the plant and the controller. For this reason, special care will be dedicated to its analysis.*

2) Reference Model Dynamics: From (28), when $\hat{P} = \mathcal{P}(\hat{A})$, system (22) can be written highlighting the dependence on the estimate $\hat{A}$ of the identifier:

$$\dot{x}_\mathrm{m} = (\hat{A} - BR^{-1}B^\top\mathcal{P}(\hat{A}))x_\mathrm{m} + Bd, \tag{46}$$

where from (20), (28), the pointwise-in-time value of $\mathcal{P}(\hat{A})$ is provided implicitly as the solution of a parameter-varying ARE. By [41, Thm. 4.1], $\mathcal{P}(\hat{A})$ is an analytic function of $\hat{A}$, being all matrices of ARE $\mathcal{R}(P, \hat{A}) = 0$ in (27) analytic functions of $\hat{A} \in \Theta$. From this fact, matrix $\hat{A} - BR^{-1}B^\top\mathcal{P}(\hat{A})$ is Hurwitz and an analytic function of $\hat{A}$.

3) Adaptive Tracking Dynamics: We conclude this overview by studying the interconnection of the error dynamics (32) and the adaptive controller (23), (24). We define $\tilde{K}_\mathrm{a} := \hat{K}_\mathrm{a} - K_\mathrm{a}(\hat{A})$. By choosing (24) as input for (29), we obtain:

$$\dot{e} = (\hat{A} - BR^{-1}B^\top\mathcal{P}(\hat{A}))e + B(\hat{K}_\mathrm{a}x - K(\hat{A})x) \tag{47}$$

By choosing expression (31) for $K_\mathrm{a}(\hat{A})$, we can explicitly calculate the variation in time of $K_\mathrm{a}(\hat{A})$ due to the movement of $\hat{A}$. This is out of the standard framework of model reference adaptive control, and thus particular attention is required. We can calculate the time derivative of $K_\mathrm{a}(\hat{A})$ by deriving (31):

$$\dot{K}_\mathrm{a} = B^\dagger \dot{\hat{A}}. \tag{48}$$

Since both B and $\dot{\hat{A}}$ are known, we can use their knowledge to implement adaptive law (23), which takes into account this drift. Given equations (23) and (48), the induced dynamics for $\tilde{K}_\mathrm{a}$ is:

$$\begin{aligned}
\dot{\tilde{K}}_\mathrm{a} &= \dot{\hat{K}}_\mathrm{a} - \dot{K}_\mathrm{a} \\
&= -\mu B^\top\mathcal{P}(\hat{A})(x - x_\mathrm{m})x^\top + B^\dagger\dot{\hat{A}} - B^\dagger\dot{\hat{A}} \\
&= -\mu B^\top\mathcal{P}(\hat{A})ex^\top.
\end{aligned} \tag{49}$$

B. Global Boundedness of Solutions

We now show boundedness and forward completeness of the solutions of the closed-loop system obtained from the interconnection of the identifier dynamics (44), (45), the reference model (46), and the adaptive error system (47), (49). The overall analysis entails proving uniform bounds on the solutions of the main involved subsystems, then combining the results using arguments similar to [40, Thm. 6.3] (see the proof of Proposition 1). To increase readability, we leave the proofs of the technical lemmas in the Appendix.

We begin by showing uniform boundedness of $\hat{A}$ and $\dot{\hat{A}}$.

Lemma 1. *Let the maximal interval of solutions of* (44), (45), (46), (47), (49) *be* $[0, t_f)$. *Then, it holds that*

I) $\tilde{\epsilon}(\cdot), \tilde{A}(\cdot)$ are uniformly bounded in the interval $[0, t_f)$
II) $\hat{A}(t) \in \Theta$ for all $t \in [0, t_f)$
Furthermore, if $t_f = \infty$, the origin $(\tilde{\epsilon}, \tilde{A}) = 0$ of system (44), driven by input $\xi(t)$, is uniformly globally stable (UGS).

Lemma 2. *Let the maximal interval of solutions of* (44), (45), (46), (47), (49) *be* $[0, t_f)$. *Then, it holds that*

$$|\dot{\hat{A}}(t)| \leq \gamma, \qquad \forall t \in [0, t_f). \tag{50}$$

Remark 11. *The above results hold even if the input $\xi(t)$ of the identifier escapes to infinity as $t \to t_f$.*

Although the overall boundedness analysis entails also the study of $\xi(t)$, system (45) ISS with respect to input $x(t)$, thus its behavior will be analyzed directly in Proposition 1.

Then, we show that the reference model (46) is bounded as long as $|\dot{\hat{A}}(t)|$ is sufficiently small.

Lemma 3. *Let the maximal interval of solutions of* (44), (45), (46), (47), (49) *be* $[0, t_f)$. *There exists $\gamma_b^\star > 0$ such that, if $|\dot{\hat{A}}(t)| \leq \gamma_b^\star$ for all $t \in [0, t_f)$, then $x_m(\cdot)$ is uniformly bounded over the interval $[0, t_f)$. Furthermore, if $t_f = \infty$, then the reference model (46) with input $d(t)$ is input-to-state stable.*

Next, we provide a statement for system (47), (49).

Lemma 4. *Let the maximal interval of solutions of* (44), (45), (46), (47), (49) *be* $[0, t_f)$. *Pick $\gamma_b^\star > 0$ from Lemma 3 and let $|\dot{\hat{A}}(t)| \leq \gamma_b^\star$ for all $t \in [0, t_f)$. Then, signals $e(\cdot), \tilde{K}_a(\cdot)$ are uniformly bounded in the interval $[0, t_f)$. Furthermore, if $t_f = \infty$, the origin $(e, \tilde{K}_a) = 0$ of system (47), (49), with input $\hat{A}(t)$, is UGS.*

Finally, we combine the previous results to obtain that solutions are globally bounded and forward complete.

Proposition 1. *Consider the closed-loop system obtained from the interconnection of the identifier dynamics* (44), (45), *the reference model* (46), *and the adaptive error system* (47), (49). *Pick $\gamma_b^\star$ from Lemma 2. If $\gamma \in (0, \gamma_b^\star]$, then the closed-loop solutions are bounded and forward complete.*

Proof. Suppose that the maximal interval of existence of the solution of (44), (45), (46), (47), and (49) is $[0, t_f)$. Then, from Lemma 1, $\tilde{A}(\cdot)$ and $\tilde{\epsilon}(\cdot)$ are uniformly bounded. From Lemma 2, $|\dot{\hat{A}}(\cdot)|$ is uniformly bounded by γ. Consider any $\gamma \in (0, \gamma_b^\star]$, then Lemmas 3 and 4 ensure that $x_m(\cdot)$, $e(\cdot)$, and $\tilde{K}_a(\cdot)$ are uniformly bounded, thus also $\xi(\cdot)$ is uniformly bounded from (45) and standard ISS results.

We have thus shown that all signals of the closed-loop system are bounded, with bounds that do not depend on t_f. By contradiction, we conclude that $t_f = \infty$, thus the solutions are forward complete. Namely, if t_f were finite, the solutions would leave any compact set as $t \to t_f$, contradicting the independence of the bounds on t_f [40, Thm. 6.3]. □

C. Exponential Convergence to the Optimal Policy

We now focus on the uniform asymptotic stability properties of the closed-loop system (44), (45), (46), (47), (49). First, we show that $x_{\mathrm{m}}(t)$ is persistently exciting as long as $|\hat{A}|$ is sufficiently small.

Lemma 5. *Let the entries of stationary input d be sufficiently rich of order $n+1$ and uncorrelated. There exists $\gamma_{PE}^{\star} \in (0, \gamma_b^{\star}]$, with $\gamma_b^{\star}$ from Proposition 1, such that, for all $\gamma \in (0, \gamma_{PE}^{\star}]$, the solutions $x_{\mathrm{m}}(t)$ of the reference model (46) are persistently exciting (PE).*

Next, we provide a direct consequence of Lemma 5 for the adaptive error dynamics (47), (49).

Lemma 6. *Let the hypotheses of Lemma 5 hold and let $\gamma \in (0, \gamma_{PE}^{\star}]$, where $\gamma_{PE}^{\star}$ is given in Lemma 5. Then, the origin $(e, \tilde{K}_a) = 0$ of system (47), (49) is uniformly globally asymptotically stable (UGAS) and uniformly locally exponentially stable (ULES).*

Now that we have established that every solution $e(t)$ converges exponentially to zero, uniformly from compact sets of initial conditions, we can conclude the convergence analysis by studying the identifier dynamics (44).

Lemma 7. *Let the hypotheses of Lemma 5 hold and let $\gamma \in (0, \gamma_{PE}^{\star}]$, where $\gamma_{PE}^{\star}$ is given in Lemma 5. Then, the origin $(\tilde{\epsilon}, \tilde{A}) = 0$ of system (44), with input $\xi(t)$, is uniformly globally exponentially stable (UGES).*

D. Proof of the Main Results

1) Proof of Theorem 1: Pick $\gamma^{\star} := \min\{\gamma_b^{\star}, \gamma_{PE}^{\star}\} = \gamma_{PE}^{\star}$, where $\gamma_b^{\star}$ is from Proposition 1 and $\gamma_{PE}^{\star}$ is the one of Lemma 5. Then, if $\gamma \leq \gamma^{\star}$, the closed-loop solutions are bounded and forward complete. Moreover, x_{m} is PE. The remainder of the proof involves showing the existence of a UGAS attractor using the concept of ω-limit set of a set, see [34, Def. 6.23].

By Lemmas 6 and 7, from any compact set of initial conditions, it holds that $\hat{A} \to A, \tilde{\epsilon} \to 0, e \to 0, \tilde{K}_a \to 0$ exponentially. Moreover, by Lemma 3, the model reference subsystem (46) is ISS with uniformly bounded input $d(t)$, in particular we have that:

$$|x_{\mathrm{m}}| \geq X_{\mathrm{m}} \implies \dot{V}_{\mathrm{m}} \leq -\frac{q}{4}|x_{\mathrm{m}}|^2, \tag{51}$$

where X_{m} can be found in (88) and depends on $\|d(\cdot)\|_{\infty}$, and V_{m} in (80) is an ISS Lyapunov function for the reference model. Consider the ξ subsystem in (19). It holds that

$$|\xi| \geq \frac{2|x|}{\lambda} \implies \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}|\xi|^2\right) \leq -\frac{\lambda}{2}|\xi|^2. \tag{52}$$

Denote $\Xi := \frac{2}{\lambda}X_{\mathrm{m}}$ and define the compact set

$$\mathcal{K}_s^{\star} := \{(w, x, z_{\mathrm{s}}) \in \mathcal{W} \times \mathbb{R}^n \times \mathcal{Z}_{\mathrm{s}} : \hat{A} = A, \tilde{\epsilon} = 0, e = 0, |x_{\mathrm{m}}| \leq X_{\mathrm{m}}, |\xi| \leq \Xi\} \subset \mathcal{W} \times \mathbb{R}^n \times \mathcal{L}_{\mathrm{s}}, \tag{53}$$

where $\mathcal{L}_{\mathrm{s}}$ is the learning set given in (36). Consider a set of initial conditions $\mathcal{K}_s := \mathcal{K}_s^{\star} + c\mathbb{B}$, with $c > 0$ arbitrary, and note that the solutions are empty if they start outside $\mathcal{W} \times \mathbb{R}^n \times \mathcal{Z}_{\mathrm{s}}$. We now prove that $\mathcal{K}_s^{\star}$ is uniformly attractive from $\mathcal{K}_s$. By the above-mentioned properties for the subsystems $\tilde{A}, \tilde{\epsilon}, e, \tilde{K}_a, x_{\mathrm{m}}$, there exists $T' > 0$ such that, for any $\varepsilon > 0$, it holds that

$$|\tilde{A}(t)| \leq \varepsilon, \quad |\tilde{\epsilon}(t)| \leq \varepsilon, \quad |e(t)| \leq \min\left(\varepsilon, \frac{\lambda}{2}\frac{\varepsilon}{3}\right), \quad |x_{\mathrm{m}}(t)| \leq X_{\mathrm{m}} + \min\left(\varepsilon, \frac{\lambda}{2}\frac{\varepsilon}{3}\right). \tag{54}$$

for all $t \geq T'$, from which it holds also that

$$\frac{2|x(t)|}{\lambda} \leq \frac{2}{\lambda}(|x_{\mathrm{m}}(t)| + |e(t)|)$$
$$\leq \Xi + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \leq \Xi + \frac{2}{3}\varepsilon. \tag{55}$$

Thus, from (52), there exists $T \geq T'$ such that $|\xi(t)| \leq \Xi + \varepsilon$ for all $t \geq T$.

For compactness of notation, denote $\boldsymbol{x}_{\mathrm{s}} := (w, x, z_{\mathrm{s}})$. The arguments above have proved that $\mathcal{K}_s^\star$ is uniformly attractive from $\mathcal{K}_s$. Namely, for any $\rho > 0$, there exists $T_\rho \geq 0$ such that $|\phi(t, \boldsymbol{x}_{\mathrm{s}})|_{\mathcal{K}_s^\star} \leq \rho$, for all $t \geq T_\rho$ and $\boldsymbol{x}_{\mathrm{s}} \in \mathcal{K}_{\mathrm{s}}$, where $\phi(t, \boldsymbol{x}_{\mathrm{s}})$ is the solution at time t of the closed-loop system having initial condition $\boldsymbol{x}_{\mathrm{s}}$.

Denote with $\mathcal{A}_s := \Omega(\mathcal{K}_{\mathrm{s}})$ the ω-limit set of $\mathcal{K}_{\mathrm{s}}$. We want to prove that $\mathcal{A}_s \subset \mathcal{K}_s^\star$. We do it by contradiction, i.e., we suppose that $\mathcal{A}_s \subset \mathcal{K}_s^\star$ is false. Under this hypothesis, there exists $\bar{\boldsymbol{x}}_{\mathrm{s}} \in \mathcal{A}_s$ and $\rho > 0$ such that $|\bar{\boldsymbol{x}}_{\mathrm{s}}|_{\mathcal{K}_s^\star} \geq 3\rho$. By definition [34, Def. 6.23], the ω-limit set of $\mathcal{K}_{\mathrm{s}}$ is the set of all points $\boldsymbol{x}_{\mathrm{s}}$ such that there exist sequences $\boldsymbol{x}_{\mathrm{s},n} \in \mathcal{K}_{\mathrm{s}}$, $t_n \geq 0$ such that $\lim_{n \to \infty} t_n = \infty$ and $\lim_{n \to \infty} \phi(t_n, \boldsymbol{x}_{\mathrm{s},n}) = \boldsymbol{x}_{\mathrm{s}}$. Therefore, by definition of limit, there exists $\bar{n} \in \mathbb{N}$ such that

$$|\phi(t_n, \boldsymbol{x}_{\mathrm{s},n}) - \bar{\boldsymbol{x}}_{\mathrm{s}}| \leq \rho, \quad \forall n \geq \bar{n}. \tag{56}$$

Pick any subsequence $\boldsymbol{x}_{\mathrm{s},n_i}$, t_{n_i} such that, for $n_i \geq \bar{n}$, then $t_{n_i} \geq T_\rho$, where T_ρ derives from the uniform attractivity of $\mathcal{K}_s^\star$ (see above). We have thus proved that, for $n_i \geq \bar{n}$, $|\phi(t_{n_i}, \boldsymbol{x}_{\mathrm{s},n_i}) - \bar{\boldsymbol{x}}_{\mathrm{s}}| \leq \rho$, thus $|\phi(t_{n_i}, \boldsymbol{x}_{\mathrm{s},n_i})|_{\mathcal{K}_s^\star} \geq 2\rho$, and at the same time $|\phi(t_{n_i}, \boldsymbol{x}_{\mathrm{s},n_i})|_{\mathcal{K}_s^\star} \leq \rho$ by uniform attractivity of $\mathcal{K}_s^\star$. This is a contradiction, hence necessarily $\mathcal{A}_s \subset \mathcal{K}_s^\star$.

To summarize the previous results, we have thus proved that the solutions are globally bounded and forward complete and

$$\mathcal{A}_s := \Omega(\mathcal{K}_{\mathrm{s}}) \subset \mathcal{K}_s^\star \subset \mathrm{Int}(\mathcal{K}_{\mathrm{s}}) \subset \mathcal{K}_{\mathrm{s}}. \tag{57}$$

By [34, Corollary 7.7], $\mathcal{A}_s = \Omega(\mathcal{K}_{\mathrm{s}})$ is asymptotically stable, with domain of attraction containing $\mathcal{K}_{\mathrm{s}}$. Since $\mathcal{K}_{\mathrm{s}}$ can be chosen arbitrarily large due to Proposition 1, we conclude UGAS of $\mathcal{A}_s$.

2) Proof of Theorem 2: Consider the reduced-order system with state $\boldsymbol{x}_{\mathrm{s}} := (w, x, z_{\mathrm{s}})$ and the boundary layer system with state $\hat{P}$. Define the indicator functions

$$\omega_s(\boldsymbol{x}_{\mathrm{s}}) := \begin{cases} |\boldsymbol{x}_{\mathrm{s}}|_{\mathcal{A}_s} & \boldsymbol{x}_{\mathrm{s}} \in \mathcal{W} \times \mathbb{R}^n \times \mathcal{Z}_{\mathrm{s}} \\ \infty & \text{elsewhere} \end{cases}$$
$$\omega_f(\boldsymbol{x}_{\mathrm{s}}, \hat{P}) := \begin{cases} |\hat{P} - \mathcal{P}(\hat{A})| & (\boldsymbol{x}_{\mathrm{s}}, \hat{P}) \in \mathcal{W} \times \mathbb{R}^n \times \mathcal{Z} \\ \infty & \text{elsewhere.} \end{cases} \tag{58}$$

By Theorem 1, the reduced-order system satisfies

$$\omega_s(\boldsymbol{x}_{\mathrm{s}}(t)) \leq \beta_s(\omega_s(\boldsymbol{x}_{\mathrm{s}}(0)), t), \tag{59}$$

where β_s is a class $\mathcal{KL}$ function. Moreover, by the DRE properties [37, Thm. 4], the boundary-layer system $d\hat{P}/d\tau(\tau) = \mathcal{R}(\hat{P}(\tau), \hat{A})$, with $\hat{A} \in \Theta$ constant and $\tau := gt$, satisfies

$$\omega_f(\boldsymbol{x}_{\mathrm{s}}, \hat{P}(\tau)) \leq \beta_f(\omega_f(\boldsymbol{x}_{\mathrm{s}}, \hat{P}(0)), \tau), \tag{60}$$

where β_f is a class $\mathcal{KL}$ function. From [42, Thm. 1] (Assumptions $1, 3, 4, 7, 8$ can be verified), from any compact set of initial conditions $\mathcal{K} \subset \mathcal{W} \times \mathbb{R}^n \times \mathcal{Z}$ and for any $\delta > 0$, there exists $g^\star > 0$ such that, for all $g \geq g^\star$, the solutions are forward complete and satisfy:

$$\omega_s(\boldsymbol{x}_{\mathrm{s}}(t)) \leq \beta_s(\omega_s(\boldsymbol{x}_{\mathrm{s}}(0)), t) + \delta$$
$$\omega_f(\boldsymbol{x}_{\mathrm{s}}(t), \hat{P}(t)) \leq \beta_f(\omega_f(\boldsymbol{x}_{\mathrm{s}}(0), \hat{P}(0)), gt) + \delta. \tag{61}$$

In particular, choose $\mathcal{K} := \mathcal{A}_s \times P^\star + c\mathbb{B}$, with $c > 0$ arbitrary. Reference model dynamics (22) can be rewritten as:

$$\begin{aligned} \dot{x}_{\mathrm{m}} &= (\hat{A} - BR^{-1}B^\top \hat{P})x_{\mathrm{m}} + Bd \\ &= (A - BR^{-1}B^\top P^\star)x_{\mathrm{m}} + Bd + (\tilde{A} - BR^{-1}B^\top(\hat{P} - P^\star))x_{\mathrm{m}} \\ &= (A - BR^{-1}B^\top P^\star)x_{\mathrm{m}} + Bd + (\tilde{A} - BR^{-1}B^\top(\hat{P} - \mathcal{P}(\hat{A}) + \mathcal{P}(\hat{A}) - \mathcal{P}(A)))x_{\mathrm{m}}. \end{aligned} \tag{62}$$

Notice that x_m is PE if $\hat{P} = \mathcal{P}(\hat{A})$ and $\tilde{A} = 0$. Furthermore, since $\mathcal{P}(\hat{A})$ is an analytic function of $\hat{A}$, x_m is PE by [43, Lemma 6.1.2] if $|\hat{P} - \mathcal{P}(\hat{A})|$ and $|\tilde{A}|$ are sufficiently small, because the solutions of (62) are sufficiently close to those with $\hat{P} = \mathcal{P}(\hat{A})$ and $\tilde{A} = 0$. Moreover, also x and ξ are PE if x_m is PE and $|e|$ is sufficiently small. Choose $\delta > 0$ such that the conditions $\omega_s(\boldsymbol{x}_\mathrm{s}) \leq 2\delta$ and $\omega_f(\boldsymbol{x}_\mathrm{s}, \hat{P}) \leq 2\delta$ imply that x_m, x, and ξ are PE. Then, pick $g \geq g^\star$, where $g^\star$ is obtained from the considered $\mathcal{K}$ and δ. From (61), the closed-loop solutions converge in finite time T to a compact set satisfying $\omega_s(\boldsymbol{x}_\mathrm{s}) \leq 2\delta$ and $\omega_f(\boldsymbol{x}_\mathrm{s}, \hat{P}) \leq 2\delta$. Then, for $t \geq T$, $\tilde{\epsilon} \to 0, \hat{A} \to A$ exponentially from Lemma 7 since ξ is PE. From the local exponential stability of the DRE [37, Thm. 4], it follows that $\hat{P} \to P^\star$ exponentially. By Lemma 6 and $\hat{P} \to \mathcal{P}(\hat{A})$, we conclude that $e \to 0, \hat{K}_\mathrm{a} \to 0$ exponentially. As a consequence, the same arguments of Theorem 1 (omitted here to avoid repetition) can be used to show that the compact set

$$
\begin{aligned}
\mathcal{K}^\star &:= \{(\boldsymbol{x}_\mathrm{s}, \hat{P}) \in \mathcal{W} \times \mathbb{R}^n \times \mathcal{Z} : \hat{A} = A, \tilde{\epsilon} = 0, e = 0, |x_\mathrm{m}| \leq X_\mathrm{m}, |\xi| \leq \Xi, \hat{P} = P^\star\} \\
&= \mathcal{K}_s^\star \times P^\star,
\end{aligned}
\tag{63}
$$

is uniformly attractive from $\mathcal{K}$, with $\mathcal{K}_s^\star$ given in (53) . The same steps as in Theorem 1 allow to prove that $\mathcal{A} := \Omega(\mathcal{K}) \subset \mathcal{K}^\star \subset \mathrm{Int}(\mathcal{K}) \subset \mathcal{K}$, thus $\mathcal{A}$ is uniformly asymptotically stable with domain of attraction containing $\mathcal{K}$, and since $\mathcal{K}$ can be chosen arbitrarily large we can conclude semiglobal uniform asymptotic stability of $\mathcal{A}$.

Finally, we want to prove that $\mathcal{A} = \mathcal{A}_s \times P^\star$, where $\mathcal{A}_s = \Omega(\mathcal{K}_s^\star)$. In $\mathcal{A}$, it holds that $\hat{P} = P^\star$, $\hat{A} = A$ and $\tilde{\epsilon} = 0$, from which it holds that $\hat{P} = \mathcal{P}(\hat{A}) = \mathcal{P}(A) = P^\star$ for all points in this set. For this reason, in $\mathcal{A}$, the vector field of Algorithm 1 coincides with the vector field of the reduced-order system with $\hat{A} = A$ and $\tilde{\epsilon} = 0$. Since the vector fields coincide, we have that in this set solutions $\boldsymbol{x}(t)$ of Algorithm 1 can be written as $\boldsymbol{x}(t) = \boldsymbol{x}_s(t) \times P^\star$, where $\boldsymbol{x}_s(t)$ is the solution of the reduced-order system having the same initial conditions.

From (63) and $\mathcal{A} \subset \mathcal{K}^\star$, it follows that $\mathcal{A} = \Omega(\mathcal{K}) = \Omega(\mathcal{K}^\star) = \Omega(\mathcal{K}_s^\star \times P^\star) = \mathcal{A}' \times P^\star$. Since for the slow states of Algorithm 1 the solutions coincide with those of the reduced-order system, it follows that $\mathcal{A}' = \mathcal{A}_s$. As a consequence, $\mathcal{A} = \mathcal{A}_s \times P^\star$.

VI. Numerical Analysis: Control of a Doubly Fed Induction Motor

In this section, we propose two numerical examples to show the effectiveness of Model Reference Adaptive Reinforcement Learning. In the first example, we consider the model of a doubly fed induction motor (DFIM) at constant speed with unknown rotor and stator resistances. In the second example, we test the robustness of the proposed algorithm by considering a DFIM with slowly time-varying unknown resistances, due to the motor heating up, and rotor acceleration.

A. Example 1: Constant Parameters

A DFIM at constant speed can be modeled [44] with a linear system in the form of (1) with state

$$
x = (i_{1u}, i_{1v}, i_{2u}, i_{2v}) \in \mathbb{R}^4,
\tag{64}
$$

where i_{1u}, i_{1v} are the stator currents and i_{2u}, i_{2v} are the rotor currents. The input is

$$
u = (u_{1u}, u_{1v}, u_{2u}, u_{2v}) \in \mathbb{R}^4,
\tag{65}
$$

where u_{1u}, u_{1v} are the stator voltages and u_{2u}, u_{2v}, the rotor voltages. System matrices are defined as

$$
A = \frac{1}{\bar{\bar{L}}} \begin{bmatrix} -L_2 R_1 & -\alpha + \beta & L_m R_2 & \beta_2 \\ \alpha - \beta & -L_2 R_1 & -\beta_2 & -L_m R_2 \\ L_m R_1 & -\beta_1 & -L_1 R_2 & -\alpha - \beta_{12} \\ \beta_1 & L_m R_1 & \alpha + \beta_{12} & -L_1 R_2 \end{bmatrix}, \qquad B = \frac{1}{\bar{\bar{L}}} \begin{bmatrix} L_2 & 0 & -L_m & 0 \\ 0 & L_2 & 0 & -L_m \\ -L_m & 0 & L_1 & 0 \\ 0 & -L_m & 0 & L_1 \end{bmatrix}, \tag{66}
$$

TABLE II
PHYSICAL PARAMETERS OF THE MOTOR.

Parameter	Value	Parameter	Value
L_1 [H]	0.02645	R_1 [Ω]	0.036
L_2 [H]	0.0264	R_2 [Ω]	0.038
L_m [H]	0.0257	ω_0 [rad/s]	$2\pi70.8$
p	3	ω_r [rad/s]	$2\pi62$

TABLE III
UNCERTAINTY PARAMETERS FOR EXAMPLE 1.

Parameter	Value	Parameter	Value
$\bar{R}_1$ [Ω]	0.03	r_1 [Ω]	0.01
$\bar{R}_2$ [Ω]	0.03	r_2 [Ω]	0.01
ρ	20		

where

$$\bar{L} := L_1 L_2 - L_m^2, \qquad \alpha := \bar{L}\omega_0, \qquad \beta := L_m^2 \omega_r$$
$$\beta_{12} := L_1 L_2 \omega_r, \qquad \beta_1 := L_1 L_m \omega_r, \qquad \beta_2 := L_2 L_m \omega_r. \tag{67}$$

Parameters R_1, R_2 are the stator and rotor resistances, while L_1, L_2, L_m are the stator and rotor auto-inductances and the mutual inductance, respectively. Finally, ω_r and ω_0 are the electrical angular speeds of the rotor and the rotating reference frame, which we suppose constant.

Remark 12. *We suppose to have uncertainties on the parameters R_1 and R_2. This makes the matrix A uncertain in half of its entries. In this example B is such that $\mathrm{Im}(\mathcal{B}) = \mathbb{R}^{n \times n}$, so Assumption 2 is fulfilled for any $A_0 \in \mathbb{R}^{n \times n}$.*

Denote the true resistances as R_1, R_2. We model our uncertainties specifying nominal values $\bar{R}_1, \bar{R}_2$ and radiuses $r_1, r_2 > 0$ such that

$$R_1 \in [\bar{R}_1 - r_1, \bar{R}_1 + r_1]$$
$$R_2 \in [\bar{R}_2 - r_2, \bar{R}_2 + r_2]. \tag{68}$$

Next, we define $\mathcal{C}$ as a ball about the nominal $\bar{A}$ (i.e., having the structure (66) with resistances $\bar{R}_1$ and $\bar{R}_2$) containing all possible parameter variation, i.e.,

$$\mathcal{C} := \{\hat{A} \in \mathbb{R}^{n \times n} : |\hat{A} - \bar{A}|_F \leq \rho\} \tag{69}$$

with $\rho > 0$ big enough. We report in Table II the physical parameters of the motor. In Table III, we specify the values used for the uncertainties and the desired performances. The dither $d(t)$ is designed, on each entry $d_i(t)$, according to

$$d_i(t) = 10 \sum_{j=1}^{4} \mathrm{sawtooth}(2\omega_s i j t), \quad i \in \{1, 2, 3, 4\} \tag{70}$$

where sawtooth($\cdot$) is a triangular wave of unitary amplitude and $\omega_s = 0.2$ rad/s. In Fig. 2-(a), we show the difference between the estimate $\hat{A}(t)$ and the true matrix A. Next, in Fig. 2-(b), we show how the error between the optimal feedback gain $K^\star$ and the overall applied feedback gain $-R^{-1}B^\top \hat{P} + \hat{K}_a$ approaches zero, thus controlling in an optimal way the system. In Fig. 3, we show for completeness the error between the reference model and the real system, which reaches a small amplitude in a few seconds.
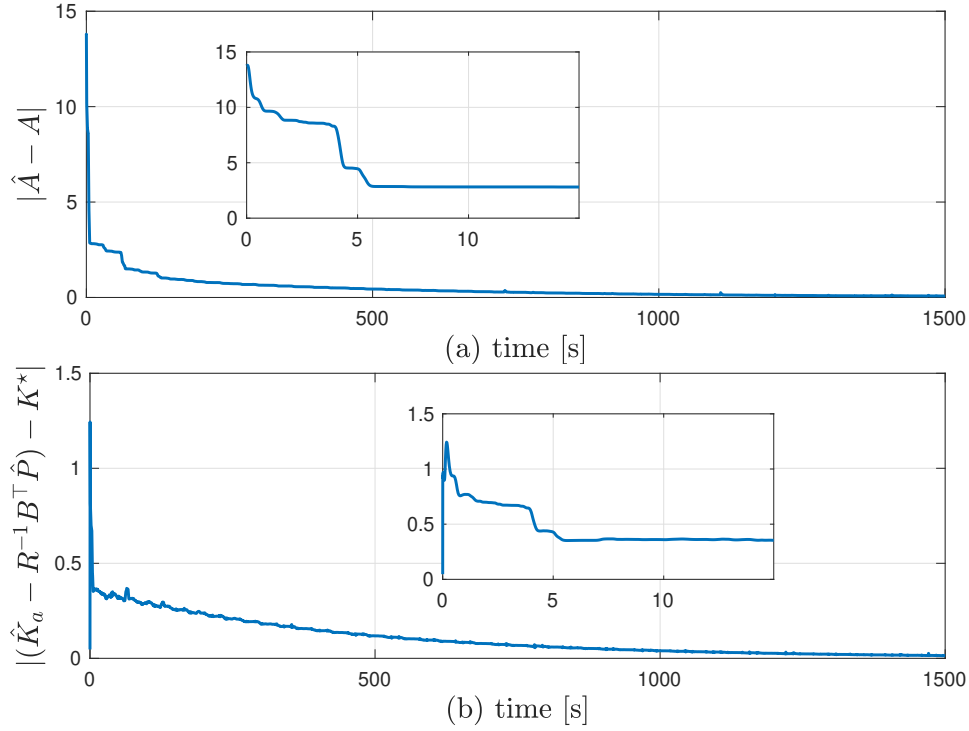
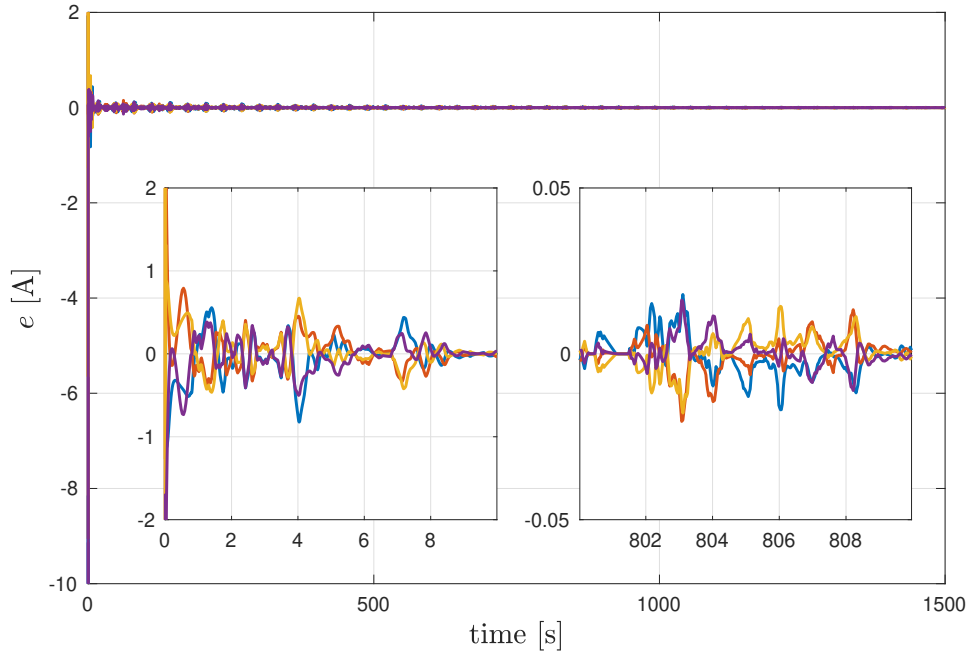Fig. 2. Convergence to true A and to optimal gain $K^\star$.

Fig. 3. Tracking error between plant and reference model. Different colors stand for different components of e.

B. Example 2: Drifting Parameters and Variable Speed

In this example, we apply perturbations to the DFIM with model given in (66) to test the robustness of MR-ARL. We consider two perturbations to the nominal model occurring together: the first one is a time-varying resistance due to motor heating up, while the second one is a time-varying rotor speed due to load changes. We model both disturbances with sigmoid functions and we report them in the plots. The temperature disturbance lasts for about 600 s and brings the temperature from 20 °C to 100 °C, i.e., $\Delta T = 80$ °C. The speed disturbance is a total increase of speed of $2\pi20$ rad/s occurring in about 60 s.

TABLE IV
UNCERTAINTY PARAMETERS FOR EXAMPLE 2.

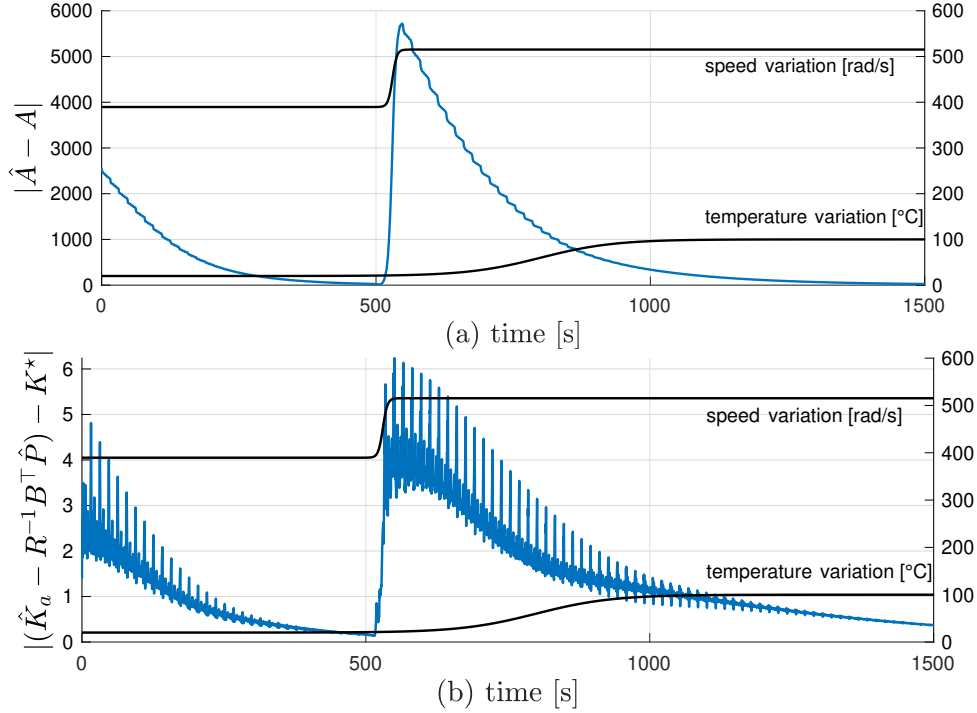| Parameter	Value	Parameter	Value
R_1 [Ω]	0.2	r_1 [Ω]	0.18
$\bar{R}_2$ [Ω]	0.2	r_2 [Ω]	0.18
$\bar{\omega}_r$ [rad/s]	$2\pi 70$	r_ω [rad/s]	$2\pi 15$
ρ	4830		

Fig. 4. Convergence to true $A(t)$ and to optimal gain $K^\star(t)$.

We model the dependence of resistances on temperature with

$$R_i(\Delta T) = R_i + \alpha \Delta T, \qquad i \in \{1, 2\}, \tag{71}$$

where $\alpha_{\mathrm{CU}} = 4.041 \times 10^{-3} \ \Omega/°\mathrm{C}$ is the temperature coefficient of resistance of the copper.

We set new nominal $\bar{R}_1, \bar{R}_2, \bar{\omega}_r$ with associated range r_1, r_2, r_ω (reported in Table IV) to consider these uncertainties. We recalculate $\mathcal{C}$ as in the previous example. Finally, we leave the dither as in (70).

Remark 13. *Due to the parameter variations, the plant becomes a slowly time-varying system. Consistently with the theoretical result, due to the "small" variations, the stability properties of Theorem 2 are practically preserved and recovered when the variations vanish.*

In Fig. 4-(a), we show the difference between the estimate $\hat{A}(t)$ and the true time-varying matrix $A(t)$. Notice that as soon as the speed disturbance ends, the gradient estimator is able to adapt and recover convergence of the estimation to a small ball about the true parameters. Next, in Fig. 4-(b), we show how the data-driven feedback gain approaches the optimal one. Since in this simulation we have a LTV plant, we calculate at each time instant the optimal gain $K^\star(t)$ by solving an LQR problem with constant $A(t)$. The importance of the adaptive controller action is particularly clear in presence of the speed disturbance, where the estimated matrix is far from the true one and thus the optimal action is likely to be destabilizing.

Finally, we show in Fig. 5 how the error between the reference model and the real plant is kept bounded also in the presence of these disturbances.
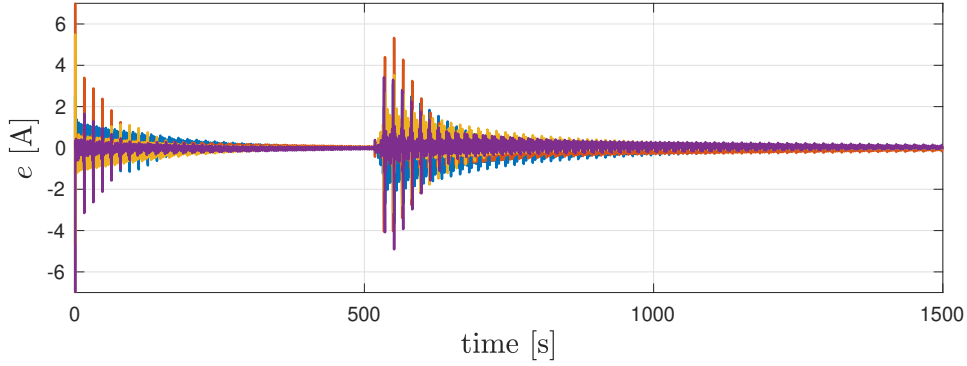
Fig. 5. Tracking error between plant and reference model. Different colors stand for different components of e.

VII. Conclusions

In this paper, we have addressed the problem of data-driven optimal control of partially unknown linear systems. First, we have proposed a framework that formalizes a robustlty stable on-policy data-driven LQR problem in which optimality of the learned strategy is obtained while guaranteeing robust stability of the whole learning and control closed-loop system. Next, we have proposed a new solution to this problem consisting in the combination of model reference adaptive control and reinforcement learning. As main result, we showed that our design has a semiglobally uniformly asymptotically stable attractor where the plant follows the optimal reference model. To demonstrate the effectiveness of the solution, we tested it in the control of a doubly fed induction motor. The results show that our solution is also able to manage non-vanishing perturbations typical of real-world applications.

VIII. Appendix

A. Proofs

Proof of Lemma 1. At first, to simplify the expressions for the Lyapunov function, we introduce the following vectorized coordinates:

$$\hat{\theta}_A := \mathrm{vec}(\hat{A}) \in \mathbb{R}^{n^2}, \qquad \tilde{\theta}_A := \mathrm{vec}(\tilde{A}) \in \mathbb{R}^{n^2}. \tag{72}$$

It can be verified the following relation holds:

$$\mathrm{vec}(BB^\dagger \epsilon \xi^\top) = \bar{B}(\xi \otimes I_n)\epsilon, \qquad \bar{B} := (I_n \otimes BB^\dagger), \tag{73}$$

where $\bar{B}$ defines a projection onto $\mathrm{Im}(I_n \otimes B) \subset \mathbb{R}^{n^2}$. Notice that, for any $\hat{\theta}_A \in \mathrm{Im}(I_n \otimes B)$ and $\tau \in \mathbb{R}^{n^2}$, then $\tilde{\theta}_A \in \mathrm{Im}(I_n \otimes B)$ and, since the scalar product of orthogonal vectors is zero and by idempotence of the projection,

$$\tilde{\theta}_A^\top \tau = \tilde{\theta}_A^\top(\tau_\| + \tau_\perp) = \tilde{\theta}_A^\top \tau_\| + 0 = \tilde{\theta}_A^\top \bar{B}\tau_\| = \tilde{\theta}_A^\top \bar{B}(\tau_\| + \tau_\perp) = \tilde{\theta}_A^\top \bar{B}\tau, \tag{74}$$

where $\tau_\| \in \mathrm{Im}(I_n \otimes B)$ and $\tau_\perp \in (I_n \otimes B)^\perp$. We rewrite (20) by using the vectorized coordinates defined above:

$$\dot{\hat{\theta}}_A = \underset{\mathrm{vec}^{-1}\hat{\theta}_A \in \mathcal{C}}{\mathrm{Proj}} \left\{ -\gamma \bar{B}\frac{(\xi \otimes I_n)\epsilon}{1 + \nu|\xi||\epsilon|} \right\}. \tag{75}$$

The computations from here are similar to [40, Lemma 6.1] but we report them for the reader's convenience. Define

$$V_A(\tilde{\epsilon}, \tilde{\theta}_A) := \frac{1}{\lambda}|\tilde{\epsilon}|^2 + \frac{1}{2\gamma}|\tilde{\theta}_A|^2, \tag{76}$$

which is positive definite with respect to $(0,0)$ and radially unbounded. Note that $\epsilon = (\xi \otimes I_n)^\top \tilde{\theta}_A + \tilde{\epsilon}$. Then, using (74) and [40, Lemma E.1] to treat the projection operator $\text{Proj}_{\hat{A} \in \mathcal{C}}\{\cdot\}$, the derivative of V_A along the solutions of (44) is

$$
\begin{aligned}
\dot{V}_A &= -2|\tilde{\epsilon}|^2 + \tilde{\theta}_A^\top \underset{\text{vec}^{-1}\,\hat{\theta}_A \in \mathcal{C}}{\text{Proj}} \left\{ -\bar{B}\frac{(\xi \otimes I_n)\epsilon}{1 + \nu|\xi||\epsilon|} \right\} \\
&\leq -2|\tilde{\epsilon}|^2 - \frac{\tilde{\theta}_A^\top \bar{B}(\xi \otimes I_n)\epsilon}{1 + \nu|\xi||\epsilon|} = -2|\tilde{\epsilon}|^2 - \frac{\tilde{\theta}_A^\top (\xi \otimes I_n)\epsilon}{1 + \nu|\xi||\epsilon|} \\
&\leq -2|\tilde{\epsilon}|^2 - \frac{|\epsilon|^2 - \tilde{\epsilon}^\top \epsilon}{1 + \nu|\xi|^2} \\
&\leq -2|\tilde{\epsilon}|^2 - \frac{1}{4}\frac{|\epsilon|^2}{(1 + \nu|\xi|^2)^2} + \frac{\tilde{\epsilon}^\top \epsilon}{1 + \nu|\xi|^2} - \frac{3}{4}\frac{|\epsilon|^2}{1 + \nu|\xi|^2} \\
&= -|\tilde{\epsilon}|^2 - \left(\frac{1}{2}\frac{|\epsilon|}{1 + \nu|\xi|^2} - \tilde{\epsilon} \right)^2 - \frac{3}{4}\frac{|\epsilon|^2}{1 + \nu|\xi|^2} \leq 0
\end{aligned}
\tag{77}
$$

implying that $(\tilde{\epsilon}(t), \tilde{\theta}_A(t))$ is contained for all $t \in [0, t_f)$ in a compact sublevel set of V_A. We conclude the proof by recalling [40, Lemma E.1] to ensure $\hat{A}(t) \in \Theta$, for all its domain of existence. $\qquad\square$

Proof of Lemma 2. Using [40, Lemma E.1] to treat the projection operator $\text{Proj}_{\hat{A} \in \mathcal{C}}\{\cdot\}$ and the fact that $|\bar{B}| = 1$ due to (73), we can bound $|\dot{\hat{A}}|$ as follows:

$$
|\dot{\hat{A}}| \leq |\dot{\hat{A}}|_F = |\dot{\hat{\theta}}_A| \leq \gamma|\bar{B}|\frac{|(\xi \otimes I_n)||\epsilon|}{1 + \nu|\xi||\epsilon|} \leq \frac{\gamma|\xi||\epsilon|}{1 + \nu|\xi||\epsilon|} \leq \gamma.
\tag{78}
$$

$\qquad\square$

Proof of Lemma 3. Function $\mathcal{P}(\hat{A})$ being continuous and Θ a compact set, there exist scalars $p_{\min}, p_{\max} > 0$ such that

$$
p_{\min} I_n \leq \mathcal{P}(\hat{A}) \leq p_{\max} I_n, \quad \forall \hat{A} \in \Theta.
\tag{79}
$$

Then, define the Lyapunov function

$$
V_m(x_m, t) := x_m^\top \mathcal{P}(\hat{A}) x_m
\tag{80}
$$

which is positive definite and radially unbounded, and whose derivative along the solutions of (22) is given by:

$$
\dot{V}_m = x_m^\top \left(\mathcal{P}(\hat{A})A_{cl}(\hat{A}) + A_{cl}(\hat{A})^\top \mathcal{P}(\hat{A}) \right) x_m + x_m^\top \left(\frac{\partial \mathcal{P}(\hat{A})}{\partial \hat{A}} \odot \dot{\hat{A}} \right) x_m + 2x_m^\top \mathcal{P}(\hat{A})Bd,
\tag{81}
$$

where

$$
A_{cl}(\hat{A}) := \hat{A} - BR^{-1}B^\top \mathcal{P}(\hat{A})
\tag{82}
$$

and the product $\odot$ is defined as

$$
\frac{\partial \mathcal{P}(\hat{A})}{\partial \hat{A}} \odot \dot{\hat{A}} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial \mathcal{P}(\hat{A})}{\partial [\hat{A}]_{ij}} [\dot{\hat{A}}]_{ij},
\tag{83}
$$

with $[\hat{A}]_{ij}$ the i-th row and j-th column entry of matrix $\hat{A}$. Since $\hat{P} = \mathcal{P}(\hat{A})$ solves at each time instant ARE (27), it holds that:

$$
A_{cl}(\hat{A})\mathcal{P}(\hat{A}) + \mathcal{P}(\hat{A})A_{cl}(\hat{A}) = \underbrace{-Q - K(\hat{A})^\top R K(\hat{A})}_{=:-\bar{Q}(\hat{A})},
\tag{84}
$$

where from Assumption 1 and Θ being compact, $\bar{Q}(\hat{A}) \geq q > 0$ for all $\hat{A} \in \Theta$, with q defined as

$$q := \min_{\hat{A} \in \Theta} \lambda_{\min}\left(-Q - \mathcal{P}(\hat{A})BRB^\top\mathcal{P}(\hat{A})\right), \tag{85}$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a matrix. Define $c := \max_{i,j \in \{1,\ldots,n\}}\left\{\max_{\hat{A} \in \Theta}\left|\frac{\partial\mathcal{P}(\hat{A})}{\partial[\hat{A}]_{ij}}\right|\right\}$, then we obtain

$$\left|\frac{\partial\mathcal{P}(\hat{A})}{\partial\hat{A}} \odot \dot{\hat{A}}\right| = \left|\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial\mathcal{P}(\hat{A})}{\partial[\hat{A}]_{ij}}[\dot{\hat{A}}]_{ij}\right| \tag{86}$$

$$\leq c\sum_{i=1}^{n}\left(\sum_{j=1}^{n}|[\dot{\hat{A}}]_{ij}|\right) \leq cn\max_{1 \leq i \leq n}\sum_{j=1}^{n}|[\dot{\hat{A}}]_{ij}| \leq cn^{\frac{3}{2}}|\dot{\hat{A}}|.$$

By letting $|\dot{\hat{A}}| \leq \gamma_b^\star := q/(2cn^{\frac{3}{2}})$, (81) becomes

$$\dot{V}_{\mathrm{m}} = -x_{\mathrm{m}}^\top\left(\bar{Q}(\hat{A}) - \frac{\partial\mathcal{P}(\hat{A})}{\partial\hat{A}} \odot \dot{\hat{A}}\right)x_{\mathrm{m}} + 2x_{\mathrm{m}}^\top\mathcal{P}(\hat{A})Bd$$

$$\leq -(q - c\rho|\dot{\hat{A}}|)|x_{\mathrm{m}}|^2 + 2p_{\max}|B||x_{\mathrm{m}}||d| \tag{87}$$

$$\leq -\frac{q}{2}|x_{\mathrm{m}}|\left(|x_{\mathrm{m}}| - \frac{4p_{\max}|B||d|}{q}\right).$$

Therefore,

$$|x_{\mathrm{m}}| \geq \frac{8p_{\max}|B||d|}{q} \implies \dot{V}_{\mathrm{m}} \leq -\frac{q}{4}|x_{\mathrm{m}}|^2, \tag{88}$$

which concludes the statement. $\qquad\square$

Proof of Lemma 4. At first, to simplify the expressions for the Lyapunov function, we introduce the following vectorized coordinates:

$$\hat{\theta}_{\mathrm{a}} := \mathrm{vec}(\hat{K}_{\mathrm{a}}) \in \mathbb{R}^{mn}, \qquad \tilde{\theta}_{\mathrm{a}} := \mathrm{vec}(\tilde{K}_{\mathrm{a}}) \in \mathbb{R}^{mn}. \tag{89}$$

We rewrite dynamics (47) and (49) by using the vectorized coordinates above defined:

$$\dot{e} = A_{\mathrm{cl}}(\hat{A})e + B\hat{K}_{\mathrm{a}}x - BK_{\mathrm{a}}(\hat{A})x$$
$$= A_{\mathrm{cl}}(\hat{A})e + B(x \otimes I_m)^\top\tilde{\theta}_{\mathrm{a}}, \tag{90}$$
$$\dot{\tilde{\theta}}_{\mathrm{a}} = -\mu(x \otimes I_m)B^\top\mathcal{P}(\hat{A})e.$$

Consider the Lyapunov function

$$V_e(e, \tilde{\theta}_{\mathrm{a}}, t) := e^\top\mathcal{P}(\hat{A})e + \frac{1}{\mu}|\tilde{\theta}_{\mathrm{a}}|^2, \tag{91}$$

which is positive definite and radially unbounded. The time derivative of V_e along the trajectories of (90) is given by

$$\dot{V}_e = e^\top\left(\mathcal{P}(\hat{A})A_{\mathrm{cl}}(\hat{A}) + A_{\mathrm{cl}}(\hat{A})^\top\mathcal{P}(\hat{A}) + \frac{\partial\mathcal{P}(\hat{A})}{\partial\hat{A}} \odot \dot{\hat{A}}\right)e$$

$$+ 2e^\top\mathcal{P}(\hat{A})B(x \otimes I_m)^\top\tilde{\theta}_{\mathrm{a}} - \frac{2}{\mu}\tilde{\theta}_{\mathrm{a}}^\top(\mu(x \otimes I_m)B^\top\mathcal{P}(\hat{A})e) \tag{92}$$

$$= -e^\top\left(\bar{Q}(\hat{A}) - \frac{\partial\mathcal{P}(\hat{A})}{\partial\hat{A}} \odot \dot{\hat{A}}\right)e \leq -\frac{q}{2}|e|^2 \leq 0,$$

where $A_{\text{cl}}(\hat{A})$ is defined in (82), $\bar{Q}(\hat{A})$ is given in (84), and q is found in (85). We have ensured that $(e(t), \tilde{\theta}_{\text{a}}(t))$ is contained for all $t \in [0, t_f)$ in a compact sublevel set of V_e, thus concluding the proof. $\quad\square$

Proof of Lemma 5. For all $\hat{A} \in \Theta$, pair $(A_{\text{cl}}(\hat{A}), B)$ in (46), with $A_{\text{cl}}(\hat{A})$ given in (82), is controllable because $(\hat{A}, B)$ is controllable from Assumption (1). Additionally, the origin of system $\dot{x}_{\text{m}} = A_{\text{cl}}(\hat{A})x_{\text{m}}$ is UGES from Lemma 3.

From classical results on PE [31, §5.6.4], if $\dot{\hat{A}} = 0$ then $x_{\text{m}}(t)$ is PE, i.e, there exist $T > 0$, $\alpha > 0$ such that

$$\int_t^{t+T} x_{\text{m}}(s)x_{\text{m}}(s)^\top \mathrm{d}s \geq \alpha I_n, \quad \forall t \geq 0. \tag{93}$$

By [45, Thm. 6.1], there exist a constant scalar $\eta > 0$, such that, if

$$|A_{\text{cl}}(\hat{A}(s)) - A_{\text{cl}}(\hat{A}(\tau))| \leq \eta, \quad \forall s, \tau \in [t, t+T], \tag{94}$$

for all $t \geq 0$, then $x_{\text{m}}(t)$ is PE also when $\dot{\hat{A}} \neq 0$. Recall that $A_{\text{cl}}(\hat{A}) := \hat{A} - BR^{-1}B^\top \mathcal{P}(\hat{A})$ is an analytic function of $\hat{A}$ [41, Thm. 4.1]. Thus, from the mean-value theorem and similar computations to (86), we obtain:

$$|A_{\text{cl}}(\hat{A}(s)) - A_{\text{cl}}(\hat{A}(\tau))| = \left| (s - \tau)\frac{\partial A_{\text{cl}}(\hat{A})}{\partial \hat{A}} \odot \dot{\hat{A}}(\varsigma) \right|$$
$$\leq |s - \tau| cn^{\frac{3}{2}} |\dot{\hat{A}}(\varsigma)|, \tag{95}$$

where $\varsigma \in [s, \tau]$, $c := \max_{i,j \in \{1,\ldots,n\}} \left\{ \max_{\hat{A} \in \Theta} \left| \frac{\partial A_{\text{cl}}(\hat{A})}{\partial [\hat{A}]_{ij}} \right| \right\}$. From the fact that $|\dot{\hat{A}}(\cdot)| \leq \gamma$, we conclude that for $\gamma_{PE}^\star := \eta/(Tcn^{\frac{3}{2}})$, if $\gamma \in (0, \gamma_{PE}^\star]$, then bound η in (94) is enforced and thus $x_{\text{m}}(t)$ is PE. $\quad\square$

Proof of Lemma 6. From Lemma 4 and the solutions being forward complete, it holds that the origin $(e, \tilde{\theta}_{\text{a}}) = 0$ of system (90) is UGS. Note that the regressor in (49) is given by $x(t)$. Therefore, if $x(t)$ is uniformly PE (u-PE) as in [35, Def. 5], then UGAS and ULES of $(e, \tilde{\theta}_{\text{a}}) = 0$ follows from [35, Thm. 1 and 2]. To prove u-PE of $x(t)$, note that $x(t) = x_{\text{m}}(t) + e(t)$, where $x_{\text{m}}(t)$ is PE from Lemma 5. Therefore, we conclude u-PE of $x(t)$ from [35, Prop. 2]. $\quad\square$

Proof of Lemma 7. From Lemma 1 and UGES of the $\tilde{\epsilon}$ subsystem, we only need to prove UGES of system (44) with $\tilde{\epsilon} = 0$, which we write here in vectorized coordinates:

$$\dot{\tilde{\theta}}_A = \operatorname*{Proj}_{\operatorname{vec}^{-1}\hat{\theta}_A \in \mathcal{C}} \left\{ -\gamma \bar{B} \frac{(\xi \otimes I_n)(\xi \otimes I_n)^\top \tilde{\theta}_A}{1 + \nu|\xi||\tilde{A}\xi|} \right\}. \tag{96}$$

Since the directions where learning happens are unchanged by the projection operator and by $\bar{B}$, we are interested in studying regressor $\bar{\xi}(t) := \frac{\xi(t) \otimes I_n}{\sqrt{1 + \nu|\xi(t)||\tilde{A}(t)\xi(t)|}}$ in order to prove our result. Given a small enough gain γ, it holds from Lemma 5 that $x_m(t)$ is PE, while $e(t) \to 0$ exponentially fast from Lemma 6. From (45), $\xi(t)$ is a filtered version of the PE signal $x_{\text{m}}(t) + e(t)$, thus $\xi(t)$ is PE [31, Lemma. 4.8.3]. Since all signals are bounded and $(\xi \otimes I_n)(\xi \otimes I_n)^\top = (\xi\xi^\top) \otimes I_n$, PE of $\xi(t)$ implies that

$$\int_t^{t+T} \bar{\xi}(s)\bar{\xi}(s)^\top \mathrm{d}s \geq \int_t^{t+T} \frac{(\xi(s)\xi(s)^\top) \otimes I_n}{1 + \xi_M^2 \tilde{A}_M} \mathrm{d}s \geq \alpha I_{n^2}, \tag{97}$$

for some $T, \alpha > 0$ and all $t \in \mathbb{R}_{\geq 0}$, with $\xi_M := \sup_{t \in \mathbb{R}} |\xi(t)|$, $\tilde{A}_M := \sup_{t \in \mathbb{R}} |\tilde{A}(t)|$, thus $\bar{\xi}(t)$ is PE. From [31, Thm. 8.5.6], we conclude that $\tilde{A} = 0$ is UGES. $\quad\square$

References

[1] M. Borghesi, A. Bosso, and G. Notarstefano, "On-policy data-driven linear quadratic regulator via model reference adaptive reinforcement learning," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 32–37.

[2] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2042–2062, 2017.

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

[4] A. M. Annaswamy and A. L. Fradkov, "A historical perspective of adaptive control and learning," *Annual Reviews in Control*, vol. 52, pp. 18–41, 2021.

[5] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.

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

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

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

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

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

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

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

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

[14] F. Dörfler, P. Tesi, and C. De Persis, "On the role of regularization in direct data-driven LQR control," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 1091–1098.

[15] F. Celi, G. Baggio, and F. Pasqualetti, "Closed-form estimates of the LQR gain from finite data," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 4016–4021.

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

[17] G. R. G. da Silva, A. S. Bazanella, C. Lorenzini, and L. Campestrini, "Data-driven LQR control design," *IEEE control systems letters*, vol. 3, no. 1, pp. 180–185, 2018.

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

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

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

[21] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *Proceedings of 1994 American Control Conference-ACC'94*, vol. 3. IEEE, 1994, pp. 3475–3479.

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

[23] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, 2009.

[24] H. Modares and F. L. Lewis, "Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning," *IEEE Transactions on Automatic Control*, vol. 59, no. 11, pp. 3051–3056, 2014.

[25] S. A. A. Rizvi and Z. Lin, "Output feedback reinforcement Q-learning control for the discrete-time linear quadratic regulator problem," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 1311–1316.

[26] ——, "Reinforcement learning-based linear quadratic regulation of continuous-time systems using dynamic output feedback," *IEEE Transactions on Cybernetics*, vol. 50, no. 11, pp. 4670–4679, 2019.

[27] B. Kiumarsi, F. L. Lewis, M.-B. Naghibi-Sistani, and A. Karimpour, "Optimal tracking control of unknown discrete-time linear systems using input-output measured data," *IEEE Transactions on Cybernetics*, vol. 45, no. 12, pp. 2770–2779, 2015.

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

[29] ——, "Value iteration for continuous-time linear time-invariant systems," *IEEE Transactions on Automatic Control*, 2022.

[30] G. Tao, "Multivariable adaptive control: A survey," *Automatica*, vol. 50, no. 11, pp. 2737–2764, 2014.

[31] P. A. Ioannou and J. Sun, *Robust adaptive control*. PTR Prentice-Hall Upper Saddle River, NJ, 1996, vol. 1.

[32] K. S. Narendra and A. M. Annaswamy, *Stable adaptive systems*. Courier Corporation, 2012.

[33] A. Guha and A. M. Annaswamy, "Online policies for real-time control using MRAC-RL," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 1808–1813.

[34] R. Goebel, R. G. Sanfelice, and A. R. Teel, *Hybrid Dynamical Systems: Modeling Stability, and Robustness*. Princeton University Press, Princeton, NJ, 2012.

[35] E. Panteley, A. Loria, and A. Teel, "Relaxed persistency of excitation for uniform asymptotic stability," *IEEE Transactions on Automatic Control*, vol. 46, no. 12, pp. 1874–1886, 2001.

[36] V. Kučera, "A review of the matrix Riccati equation," *Kybernetika*, vol. 9, no. 1, pp. 42–61, 1973.

[37] R. S. Bucy, "Global theory of the Riccati equation," *Journal of computer and system sciences*, vol. 1, no. 4, pp. 349–361, 1967.

[38] L. Menini, C. Possieri, and A. Tornambè, "Algebraic analysis of the structural properties of parametric linear time-invariant systems," *IET Control Theory & Applications*, vol. 14, no. 20, pp. 3568–3579, 2020.

[39] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU press, 2013.

[40] M. Krstic, P. V. Kokotovic, and I. Kanellakopoulos, *Nonlinear and adaptive control design*. John Wiley & Sons, Inc., 1995.

[41] A. C. Ran and L. Rodman, "On parameter dependence of solutions of algebraic Riccati equations," *Mathematics of Control, Signals and Systems*, vol. 1, pp. 269–284, 1988.

[42] A. R. Teel, L. Moreau, and D. Nesic, "A unified framework for input-to-state stability in systems with two time scales," *IEEE Transactions on Automatic Control*, vol. 48, no. 9, pp. 1526–1544, 2003.

[43] S. Sastry, M. Bodson, and J. F. Bartram, "Adaptive control: stability, convergence, and robustness," 1990.

[44] W. Leonhard, *Control of electrical drives*. Springer Science & Business Media, 2001.

[45] I. M. Mareels and M. Gevers, "Persistency of excitation criteria for linear, multivariable, time-varying systems," *Mathematics of Control, Signals and Systems*, vol. 1, pp. 203–226, 1988.