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

On-Policy Data-Driven Linear Quadratic Regulator via Model Reference Adaptive Reinforcement Learning

Marco Borghesi[1], Alessandro Bosso[1], and Giuseppe Notarstefano[1]

Abstract— **In this paper, we address a data-driven linear quadratic optimal control problem in which the regulator design is performed on-policy by resorting to approaches from reinforcement learning and model reference adaptive control. In particular, a continuous-time identifier of the value function is used to generate online a reference model for the adaptive stabilizer. By introducing a suitably selected dithering signal, the resulting policy is shown to achieve asymptotic convergence to the optimal gain while the controlled plant reaches asymptotically the behavior of the optimal closed-loop system.**

I. Introduction

Optimal control is a key area of control theory aiming at designing controllers able both to achieve stabilization properties and minimize (or maximize) a desired performance index, see, e.g., [1] for a survey. While optimal control is mainly a model-based approach, it has more recently inspired the Reinforcement Learning (RL) paradigm to address problems in which the system model is uncertain or completely unknown [2], [3]. Another successful discipline in control theory is adaptive control, which has begun to deal with parameter and environmental uncertainties [4]. In this paper, we investigate a data-driven Linear-Quadratic Regulator (LQR) optimal control framework by combining tools from adaptive control and reinforcement learning.

A common reinforcement learning technique used in the control field is Policy Iteration (PI), which allows the refinement of a given feedback policy up to the optimal one. Typical assumptions of this technique are the need for an initial stabilizing policy, and the persistency of excitation of the closed-loop signals gathered to perform the learning procedure. Under both the assumptions of persistency of excitation and knowledge of an initial stabilizing policy, on-policy control techniques have been proposed in [5], [6], and [7] for the LQR problem under uncertainties. In [5] the authors perform PI on the value function, while in [6] a Q-learning approach is used. Instead, [7] considers a tracking problem and introduces a discount factor in the cost function. We stress that in these PI approaches, the policy update (improvement) is performed through discrete jumps after collecting enough data. In [8] and [9], the initial stabilizing policy is no longer required. In the first work, the authors use a gradient technique to estimate the Q-function from data. In the second, the authors design an adaptive estimator of the state matrices and provide a semi-global convergence result.

Finally, we recall that Model Reference Adaptive Control (MRAC) is a very extensively used technique from the adaptive control field that matches an unknown system dynamics to an assigned one (the reference model) with desired properties [10], [11], [12]. A recent work combining MRAC and RL is [13], where RL techniques are used to optimize a reference model based on nominal plant parameters, then MRAC is applied to assign the reference model to the actual system. However, asymptotic convergence of the input to the optimal policy for the uncertain plant is not ensured.

In this work, we address the problem of on-policy optimal control of a partially unknown linear system. Our main contribution, which we call *Model Reference Adaptive Reinforcement Learning*, is a modular architecture that interconnects approaches from the system identification, reinforcement learning, and adaptive control paradigms. Unlike other works, in this paper we design an on-policy controller where learning and control are continuously updated in a closed-loop fashion. Moreover, no assumption regarding the initial policy is required. Namely, our architecture may also be initialized to a non-stabilizing estimate of the optimal controller. Besides this, we guarantee by design the persistency of excitation conditions needed to ensure convergence towards the true system parameters and the optimal policy. By relying on different fields, our architecture achieves the following properties: (i) global boundedness of solutions, along with robust stability of the error subsystems; (ii) convergence of the policy to the optimal one; (iii) asymptotic estimation of the true system parameters.

The paper is organized as follows. In Section II, we introduce the proposed framework and provide the problem statement. In Section III, we present the main contribution of this paper, namely the *Model Reference Adaptive Reinforcement Learning* architecture. Throughout that section, we give insights from a high-level perspective. In Section IV, we present the technical results used to prove the main theorem. Finally, we show a numerical example in Section V, while Section VI concludes the paper.

II. Problem Statement

Consider a continuous-time linear time-invariant system of the form

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

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^m$ is the control input, and $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are the system matrices.

[1] This work was supported in part by the Italian Ministry of Foreign Affairs and International Cooperation, grant number BR22GR01. The authors are with the Department of Electrical, Electronic and Information Engineering, Alma Mater Studiorum - University of Bologna, Italy. E-mail: {m.borghesi, alessandro.bosso, giuseppe.notarstefano}@unibo.it.

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

For system (1), we consider the problem of finding a suitable *control policy* $\pi^\star(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ such that $u = \pi^\star(x)$ solves, for all initial conditions $x_0 \in \mathbb{R}^n$, the following linear-quadratic optimal control problem:

$$\begin{aligned} \min_{x,u} \quad & \int_0^\infty x(\tau)^\top Q x(\tau) + u(\tau)^\top R u(\tau) d\tau \qquad (2) \\ \text{subj.to} \quad & \dot{x}(t) = Ax(t) + Bu(t), \quad \forall t \in [0, \infty), \\ & x(0) = x_0, \end{aligned}$$

with $Q \geq 0$, $R > 0$ symmetric weight matrices of appropriate dimensions. From linear quadratic regulation theory [14], it is known that, under typical assumptions of stabilizability and observability of pairs (A, B) and $(\sqrt{Q}, A)$, problem (2) admits a unique optimal control policy of the form

$$\pi^\star(\cdot) : \mathbb{R}^n \to \mathbb{R}^m, \quad x \mapsto K^\star x := -R^{-1}B^\top P^\star x, \qquad (3)$$

where $P^\star = P^{\star\top} > 0$ is the solution of the Algebraic Riccati Equation (ARE):

$$A^\top P^\star + P^\star A - P^\star B R^{-1} B^\top P^\star + Q = 0. \qquad (4)$$

However, implementing (3), (4) requires complete knowledge of the system matrices A and B. Specifically, in this work, we require that the input matrix B be available for design, whereas the state matrix A is unknown. Since we are in a data-driven scenario, we also include a dither signal w in our policy to guarantee the persistency of excitation. Therefore, this work aims to solve the following problem.

On-policy data-driven LQR scenario

Find a data-driven time-varying policy

$$(x, t) \mapsto \pi(x, t) + w$$

designed according to the following objectives:

- w is a dither signal injected to guarantee the persistency of excitation (PE) property;
- at each t, $u = \pi(x, t) + w$ is applied to the actual system (1) to learn from system trajectories;
- all the trajectories $x(t)$ of the closed loop system are bounded, forward complete, and satisfy

$$\limsup_{t\to\infty} |x(t)| \leq \alpha(\limsup_{t\to\infty} |w(t)|), \qquad (5)$$

  where α is a class $\mathcal{K}$ function;
- the policy asymptotically converges to the optimal LQR solution, i.e.,

$$\pi(x, t) \to K^\star x. \qquad (6)$$

III. Model Reference Adaptive Reinforcement Learning

For the subsequent design, it makes sense to rewrite the uncertain matrix A in vectorized form. Specifically, define $\theta_\mathrm{A} := \mathrm{vec}(A) \in \mathbb{R}^{n^2}$. Then, system (1) can be rewritten as

$$\dot{x} = (x \otimes I_n)^\top \theta_A + Bu, \qquad (7)$$

where $\otimes$ denotes the Kronecker product.

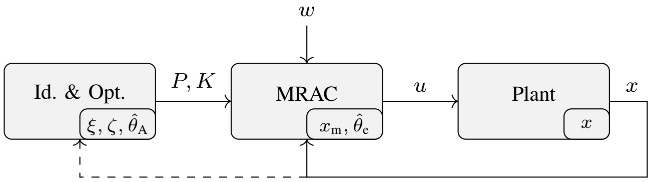

w
Id. & Opt.
$\xi, \zeta, \hat{\theta}_\mathrm{A}$
P, K
MRAC
$x_\mathrm{m}, \hat{\theta}_\mathrm{e}$
u
Plant
x
x

Fig. 1. Scheme of the *Model-Reference Adaptive Reinforcement Learning*.

The proposed algorithm is conceived as a modular architecture where we construct an online gradient estimation $\hat{\theta}_\mathrm{A}$ of θ_A, from which we compute an estimate $P(\hat{\theta}_\mathrm{A})$ of the solution $P^\star$ of ARE (4), which yields the gain $K(\hat{\theta}_\mathrm{A}) := -R^{-1}B^\top P(\hat{\theta}_\mathrm{A})$. Alone, the optimal gain estimate may be not stabilizing. For this reason, we introduce an additional adaptive feedback gain $\hat{K}_e$ following a *Model-Reference Adaptive Control* (MRAC) approach where the closed-loop system is driven by an external signal $w(t)$ having sufficient richness properties. Since the resulting MRAC is not based on a fixed model reference but on a dynamical system that is continuously tuned through the estimate $P(\hat{\theta}_\mathrm{A})$, we call this architecture *Model Reference Adaptive Reinforcement Learning*. Under the chosen policy, the closed loop dynamics of the controlled system becomes:

$$\dot{x} = (A + BK(\hat{\theta}_\mathrm{A}))x + B\hat{K}_e x + Bw, \qquad (8)$$

from which it is clear that, if $K(\hat{\theta}_\mathrm{A}) \to K^\star$ and $\hat{K}_e \to 0$, we obtain the desired behavior

$$\dot{x} = (A + BK^\star)x + Bw. \qquad (9)$$

In Fig. 1, we give a high level overview on how the controller works. The next subsections are dedicated to presenting our design, summarized in Algorithm 1. Subsequently, we describe the properties of the closed-loop system.

A. Value Function and Optimal Gain Identifier

In this subsection, we build a continuous identifier of $\hat{\theta}_A$ in (7) (equivalently, of matrix A in (1)). Then, given the estimate $\hat{A} := \mathrm{vec}^{-1}(\hat{\theta}_\mathrm{A})$, we can compute the solution $P = P^\top > 0$ of the following ARE:

$$\mathcal{R}(P, \hat{A}) := \hat{A}^\top P + P\hat{A} - PBR^{-1}B^\top P + Q = 0, \quad (16)$$

which will be denoted as $P(\hat{\theta}_\mathrm{A})$ in the following as it depends on the estimate $\hat{\theta}_\mathrm{A}$. To guarantee that the matrix $P(\hat{\theta}_\mathrm{A})$ exists, is unique, and is positive definite, we impose the following assumption.

Assumption 1. *A compact, convex set $\mathcal{K}_A \subset \mathbb{R}^{n^2}$ is known such that $\theta_A \in \mathrm{Int}(\mathcal{K}_A)$ and $(\mathrm{vec}^{-1}(\hat{\theta}_A), B)$ is controllable and $(\sqrt{Q}, \mathrm{vec}^{-1}(\hat{\theta}_A))$ is observable for all $\hat{\theta}_A \in \mathcal{K}_A$.*

It is known from linear quadratic regulator theory that stabilizability of pair $(\hat{A}, B) = (\mathrm{vec}^{-1}(\hat{\theta}_\mathrm{A}), B)$ would be sufficient in Assumption 1 for the solvability of $\mathcal{R}(P, \hat{A}) = 0$ in (16). However, as shown in the following, controllability is required to ensure convergence of the identifier under sufficient richness of the external signal $w(t)$. Moreover, we need observability instead of simple detectability to ensure the solution of (16) is positive definite. From an

Algorithm 1 Model Reference Adaptive RL Policy

```
Require: $\mathcal{K}_\mathrm{A}$ as in Assumption 1
Require: $\gamma, \lambda, \mu, \nu > 0$ design gains
Ensure: $\gamma$ is small enough
Require: $w(t)$ : bounded stationary signal, whose each entry
    is sufficiently rich of order $n+1$ and uncorrelated
    Swapping Filters:
        $\dot{\xi} = -\lambda\xi + x, \qquad \dot{\zeta} = -\lambda(x+\zeta) - Bu,$   (10)
    Identifier Dynamics:
        $\dot{\hat{\theta}}_\mathrm{A} = \operatorname*{Proj}_{\hat{\theta}_\mathrm{A}\in\mathcal{K}_\mathrm{A}} \left\{ -\gamma \frac{(\xi\otimes I_n)\epsilon}{1+\nu|\xi|^2} \right\},$
        with: $\epsilon := (\xi\otimes I_n)^\top \hat{\theta}_A - (x+\zeta),$   (11)
    Optimizer:
        Find $P(\hat{\theta}_\mathrm{A})$ sol. of $\mathcal{R}(\mathrm{vec}^{-1}(\hat{\theta}_\mathrm{A}), P) = 0$ in (16)
        $K(\hat{\theta}_\mathrm{A}) = -R^{-1}B^\top P(\hat{\theta}_\mathrm{A})$   (12)
        $A_\mathrm{cl}(\hat{\theta}_\mathrm{A}) = \hat{A} + BK(\hat{\theta}_\mathrm{A}),$
    Reference Model Dynamics:
        $\dot{x}_\mathrm{m} = A_\mathrm{cl}(\hat{\theta}_\mathrm{A})x_\mathrm{m} + Bw,$   (13)
    Adaptive Gains Dynamics:
        $\dot{\hat{\theta}}_\mathrm{e} = -\mu(x\otimes I_m)B^\top P(\hat{\theta}_\mathrm{A})(x-x_m) + (I_n\otimes B)^\dagger \dot{\hat{\theta}}_\mathrm{A},$   (14)
    System Input:
        $u = K(\hat{\theta}_\mathrm{A})x + (x\otimes I_m)^\top \hat{\theta}_\mathrm{e} + w.$   (15)
```

implementation point of view, solving (16) at each time instant may cause computation overhead. In this paper we do not address this problem and we leave its solution to future work. Given the structure of system (7), we compute an estimate $\hat{\theta}_\mathrm{A}$ of θ_A, thus an estimate $\hat{A}$ of A, by designing a swapping filter of the form (10), with $\lambda > 0$ a scalar gain for tuning, and defining the *prediction error*

$$\epsilon := (\xi\otimes I_n)^\top \hat{\theta}_\mathrm{A} - (x+\zeta). \tag{17}$$

In particular, we can write $\epsilon = (\xi\otimes I_n)^\top(\hat{\theta}_\mathrm{A} - \theta_\mathrm{A}) + \tilde{\epsilon}$, where

$$\tilde{\epsilon} := (\xi\otimes I_n)^\top \theta_\mathrm{A} - (x+\zeta), \tag{18}$$

is an error signal that is shown in Section IV to converge exponentially to zero. This way, we can use a normalized projected gradient algorithm to update the estimate $\hat{\theta}_\mathrm{A}$, with dynamics given in (11). Parameters $\gamma > 0$ and $\nu > 0$ are scalar gains, while $\mathrm{Proj}_{\hat{\theta}_\mathrm{A}\in\mathcal{K}_\mathrm{A}}\{\cdot\}$ is a Lipschitz continuous parameter projection operator whose expression is provided in [15, Appendix E] for a generic compact convex set $\mathcal{K}_\mathrm{A}$. Finally, given the estimate $\hat{\theta}_\mathrm{A}$, matrices $P(\hat{\theta}_\mathrm{A})$, $K(\hat{\theta}_\mathrm{A})$ and $A_\mathrm{cl}(\hat{\theta}_\mathrm{A})$ are computed in (12).

Remark 1. *In general, the feedback gain $K(\hat{\theta}_A)$ does not make the closed loop matrix $A + BK(\hat{\theta}_A)$ Hurwitz at all times. However, from Assumption 1 and the parameter projection in* (11)*, it is such that $A_{cl}(\hat{\theta}_A)$ in* (12) *is Hurwitz.*

B. Optimal Model Reference Adaptive Control

Given the estimate $P(\hat{\theta}_\mathrm{A}) = P(\hat{\theta}_\mathrm{A})^\top > 0$, we design an adaptive controller for system (1) with two fundamental roles: (i) to ensure global boundedness of solutions once interconnected with identifier (10), (11), (12); (ii) to impose some form of PE to the system trajectories. To these aims, consider a continuous input $w(t) \in \mathbb{R}^m$, then introduce the *reference model* (13), where $x_\mathrm{m} \in \mathbb{R}^n$ is the reference model state, $A_\mathrm{cl}(\hat{\theta}_\mathrm{A})$ is given in (12) and is Hurwitz by design (Remark 1), and B is the same as in (1) and (7).

Remark 2. *Different from classic MRAC, the state matrix $A_{cl}(\hat{\theta}_A)$ of reference model* (13) *is not a constant matrix but a time-varying one as it depends on the estimate $\hat{A}(t)$. This property leads to an adaptive design where the known-plant stabilizing gains are time-varying.*

Given the reference model (13), we define the tracking error $e := x - x_\mathrm{m}$, whose dynamics are computed from (1), (13) as

$$\begin{aligned} \dot{e} &= Ax + Bu - A_\mathrm{cl}(\hat{\theta}_\mathrm{A})(x-e) - Bw \\ &= A_\mathrm{cl}(\hat{\theta}_\mathrm{A})e + (A - A_\mathrm{cl}(\hat{\theta}_\mathrm{A}))x + B(u-w) \\ &= A_\mathrm{cl}(\hat{\theta}_\mathrm{A})e + (A-\hat{A})x + B(u - K(\hat{\theta}_\mathrm{A})x - w). \end{aligned} \tag{19}$$

To ensure that the plant (1) copies the behavior of the reference model (13), namely, $e(t) \to 0$, we need the following matching condition, typical of full-state feedback adaptive control design [12, §3.4.2].

Assumption 2. *For all $\hat{\theta}_A \in \mathcal{K}_A$, with $\hat{A} := \mathrm{vec}^{-1}(\hat{\theta}_A)$, there exists $K_e = K_e(\hat{\theta}_A)$ such that*

$$\hat{A} - A = BK_e. \tag{20}$$

Equivalently, in vectorized form,

$$\hat{\theta}_A - \theta_A = (I_n\otimes B)\theta_e, \qquad \theta_e := \mathrm{vec}(K_e). \tag{21}$$

Since K_e depends on the estimate $\hat{\theta}_\mathrm{A}$, it will be denoted as $K_e(\hat{\theta}_\mathrm{A})$ in the following. Under Assumption 2, (19) becomes

$$\dot{e} = A_\mathrm{cl}(\hat{\theta}_\mathrm{A})e + B(u - K(\hat{\theta}_\mathrm{A})x - K_e(\hat{\theta}_\mathrm{A})x - w), \tag{22}$$

suggesting a control law of the form $u := K(\hat{\theta}_\mathrm{A})x + K_e(\hat{\theta}_\mathrm{A})x + w = K(\hat{\theta}_\mathrm{A})x + (x\otimes I_m)^\top\theta_e(\hat{\theta}_\mathrm{A}) + w$ if the plant dynamics were known. However, $K_e(\hat{\theta}_\mathrm{A})$ is unavailable for design as it depends also on A, as highlighted in (20), thus we consider the certainty-equivalence-based adaptive controller given in (15), where $\theta_e(\hat{\theta}_\mathrm{A})$ is replaced by the adaptive gain $\hat{\theta}_\mathrm{e}$, driven by the adaptive law (14) where $\mu > 0$ is a scalar gain and $(I_n\otimes B)^\dagger$ is the pseudo-inverse of $I_n \otimes B$. The first term in the adaptive law (14) 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 (11), the second term in the update law takes into account the time-varying mismatch.

C. Main Result

We are now ready to state the main result of this work, providing formal guarantees for the proposed control architecture and ensuring that the requirements of the on-policy data-driven scenario of Section II are achieved.

Theorem 1. *Let Assumptions 1 and 2 hold. Consider the closed-loop system given by the interconnection of plant* (1)*, swapping filters* (10)*, identifier* (11), (12)*, reference model* (13)*, and adaptive controller* (14), (15)*, with bounded and stationary input $w(t)$. For all $E \geq 0$, there exists $\gamma^\star = \gamma^\star(E) > 0$ such that, if:*

- *the identifier gain γ is chosen such that $\gamma \in (0, \gamma^\star]$;*
- *the initial conditions satisfy $|\tilde{\epsilon}(0)| \leq E$;*
- *the entries of $w(t)$ are sufficiently rich of order $n+1$ and uncorrelated;*

then, the closed-loop solutions:

1) *are bounded, forward complete and satisfy*

$$\limsup_{t\to\infty} |x(t)| \leq \beta \limsup_{t\to\infty} |w(t)|, \tag{23}$$

where β is a positive scalar;

2) *satisfy:*

$$\lim_{t\to\infty} (\hat{\theta}_A(t), e(t), \hat{\theta}_e(t)) = (\theta_A, 0, 0), \tag{24}$$

where the convergence in (24) *is exponential, with uniform bounds for any given compact set of initial conditions such that $|\tilde{\epsilon}(0)| \leq E$;*

With (24), the right-hand side of (8) converges exponentially to $(A + BK^\star)x + Bw$, reducing controlled system (8) to the desired structure of (9). (24) states also convergence of the gradient estimation $\hat{A}$ to the true state matrix A.

Remark 3. *In Theorem 1, we study the closed-loop solutions in a semi-global sense with respect to the initial conditions $\tilde{\epsilon}(0)$, with $\tilde{\epsilon}$ as in* (18)*. This result is not restrictive because*

$$\xi(0) = 0, \zeta(0) = -x(0) \implies \tilde{\epsilon}(0) = 0, \tag{25}$$

which implies $E = 0$, leading to $\gamma \in (0, \gamma^\star(0)]$.

IV. Closed-Loop Stability Analysis

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 (18), which can be written as

$$\tilde{\epsilon} := (\xi \otimes I_n)^\top \theta_{\rm A} - (x + \zeta) = A\xi - (x + \zeta). \tag{26}$$

Then, from (1), (10), it holds that

$$\begin{aligned} \dot{\tilde{\epsilon}} &= A(-\lambda\xi + x) - (Ax + Bu - \lambda(x + \zeta) - Bu) \\ &= -\lambda(A\xi - (x + \zeta)) = -\lambda\tilde{\epsilon}, \end{aligned} \tag{27}$$

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.

Notice that the properties of the Kronecker product imply

$$(\xi \otimes I_n)(\xi \otimes I_n)^\top = (\xi \otimes I_n)(\xi^\top \otimes I_n) = (\xi\xi^\top) \otimes I_n. \tag{28}$$

Define $\tilde{\theta}_A := \hat{\theta}_{\rm A} - \theta_A$, then from (18), (27), (28) we can rewrite the identifier dynamics (10), (11), (17) in error coordinates as the following cascaded system

$$\begin{aligned} \dot{\tilde{\epsilon}} &= -\lambda\tilde{\epsilon} \\ \dot{\tilde{\theta}}_A &= \underset{\theta_A - \tilde{\theta}_A \in \mathcal{K}_{\rm A}}{\operatorname{Proj}} \left\{ -\gamma \frac{((\xi\xi^\top) \otimes I_n)\tilde{\theta}_A + (\xi \otimes I_n)\tilde{\epsilon}}{1 + \nu|\xi|^2} \right\}, \end{aligned} \tag{29}$$

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

$$\dot{\xi} = -\lambda\xi + x. \tag{30}$$

Remark 4. *To ensure $\hat{\theta}_A(t) \to \theta_{\rm A}$, equivalently, $\hat{A}(t) \to A$, it is known from the adaptive control literature that vector $x(t)$ must be a persistently exciting (PE) signal [11]. However, notice that $x(t)$ is generated in closed loop by interconnecting the plant and the controller, so special care will be dedicated to its analysis.*

2) Reference model dynamics: From (12), system (13) can be written highlighting the dependence on the estimate $\hat{\theta}_{\rm A}$ of the identifier:

$$\dot{x}_{\rm m} = \underbrace{(\mathrm{vec}^{-1}(\hat{\theta}_{\rm A}) - BR^{-1}B^\top P(\hat{\theta}_{\rm A}))}_{A_{\rm cl}(\hat{\theta}_{\rm A})} x_{\rm m} + Bw, \tag{31}$$

where from (11), (16), the pointwise-in-time value of $P(\hat{\theta}_{\rm A})$ is provided implicitly as the solution of a parameter-varying ARE. By [16, Thm. 4.1], $P(\hat{\theta}_{\rm A})$ is an analytic function of $\hat{\theta}_{\rm A}$, being all matrices of ARE $\mathcal{R}(P, \hat{A}) = 0$ in (16) analytic functions of $\hat{\theta}_{\rm A} \in \mathcal{K}_{\rm A}$. From this fact, matrix $A_{\rm cl}(\hat{\theta}_{\rm A})$ is Hurwitz and an analytic function of $\hat{\theta}_{\rm A}$.

3) Adaptive tracking dynamics: We conclude this overview by studying the interconnection of the error dynamics (22) and the adaptive controller (14), (15). We may rewrite (22) with the same notation as in (13) as

$$\dot{e} = A_{\rm cl}(\hat{\theta}_{\rm A})e + B(u - K(\hat{\theta}_{\rm A})x - (x \otimes I_m)^\top \theta_e(\hat{\theta}_{\rm A}) - w).$$

Thus, we define $\tilde{\theta}_{\rm e} := \hat{\theta}_{\rm e} - \theta_e(\hat{\theta}_{\rm A})$. By choosing (15) as input for (19), we obtain:

$$\begin{aligned} \dot{e} &= A_{\rm cl}(\hat{\theta}_{\rm A})e + B(x \otimes I_m)^\top \hat{\theta}_{\rm e} - B(x \otimes I_n)^\top \theta_e(\hat{\theta}_{\rm A}) \\ &= A_{\rm cl}(\hat{\theta}_{\rm A})e + B(x \otimes I_m)^\top \tilde{\theta}_{\rm e}. \end{aligned} \tag{32}$$

We are interested now in an explicit expression for map $\theta_e(\hat{\theta}_{\rm A})$. Since we suppose the matching condition (21) has at least a solution, all of them can be written as

$$\theta_e(\hat{\theta}_{\rm A}) = (I_n \otimes B)^\dagger(\hat{\theta}_{\rm A} - \theta_A) + v, \quad v \in \ker(I_n \otimes B), \tag{33}$$

which then allows us to choose $\theta_e(\hat{\theta}_{\rm A}) := (I_n \otimes B)^\dagger(\hat{\theta}_{\rm A} - \theta_{\rm A})$. From equations (14) and (33), the dynamics of the adaptive gain error is given by:

$$\begin{aligned} \dot{\tilde{\theta}}_{\rm e} = \dot{\hat{\theta}}_{\rm e} - \frac{\partial\theta_e}{\partial\hat{\theta}_{\rm A}}(\hat{\theta}_{\rm A})\dot{\hat{\theta}}_{\rm A} &= \dot{\hat{\theta}}_{\rm e} - (I_n \otimes B)^\dagger \dot{\hat{\theta}}_{\rm A} = \\ &= -\mu(x \otimes I_m)B^\top P(\hat{\theta}_{\rm A})e + (I_n \otimes B)^\dagger \dot{\hat{\theta}}_{\rm A} \\ &\quad - (I_n \otimes B)^\dagger \dot{\hat{\theta}}_{\rm A} = -\mu(x \otimes I_m)B^\top P(\hat{\theta}_{\rm A})e, \end{aligned} \tag{34}$$

so that we obtain the overall adaptive error system:

$$\begin{aligned}\dot{e} &= A_{\mathrm{cl}}(\hat{\theta}_{\mathrm{A}})e + B(x \otimes I_m)^\top \tilde{\theta}_{\mathrm{e}} \\ \dot{\tilde{\theta}}_{\mathrm{e}} &= -\mu(x \otimes I_m)B^\top P(\hat{\theta}_{\mathrm{A}})e,\end{aligned} \tag{35}$$

written in the usual form of adaptive systems. Note that the system matrices depend on the time-varying parameter $\hat{\theta}_{\mathrm{A}}$.

B. Boundedness of Solutions

We show boundedness and forward completeness of the solutions of the closed-loop system obtained from the interconnection of the identifier dynamics (29), (30), the reference model (13), and the adaptive error system (35). The overall analysis entails proving uniform bounds on the solutions of the involved subsystems, then combining the results using arguments similar to [15, Thm. 6.3]. We begin by showing uniform boundedness of $\hat{\theta}_{\mathrm{A}}$ and $\dot{\hat{\theta}}_{\mathrm{A}}$.

Lemma 1. *Let the maximal interval of solutions of* (29), (30), (31), (35) *be* $[0, t_f)$. *Then, it holds that* $\tilde{\epsilon}(\cdot), \tilde{\theta}_A(\cdot) \in \mathcal{L}_\infty[0, t_f)$ *and* $\hat{\theta}_A(t) \in \mathcal{K}_A$ *for all* $t \in [0, t_f)$. *Furthermore, if* $t_f = \infty$, *the origin* $(\tilde{\epsilon}, \tilde{\theta}_A) = 0$ *of system* (29), *driven by input* $\xi(t)$, *is uniformly globally stable (UGS).*

Remark 5. *Since* $\mathcal{K}_A$ *is compact, we know there exists a bound on the maximum model mismatch, here defined as*

$$\rho := \max_{\hat{\theta}_A \in \mathcal{K}_A} |\hat{\theta}_A - \theta_A|. \tag{36}$$

Lemma 2. *Let the maximal interval of solutions of* (29), (30), (31), (35) *be* $[0, t_f)$. *Then, it holds that*

$$|\dot{\hat{\theta}}_A(t)| \leq \gamma(\rho + |\tilde{\epsilon}(0)|), \qquad \forall t \in [0, t_f), \tag{37}$$

where ρ *is given in Remark 5 and* γ *is the gain in* (11).

The above results hold even if the input $\xi(t)$ of the identifier escapes to infinity as $t \to t_f$. Note that $\xi(t)$ is bounded if $x(t)$ is bounded since system (30) is ISS. Then, we show that the reference model (13) is bounded as long as $|\dot{\hat{\theta}}_{\mathrm{A}}(t)|$ is sufficiently small.

Lemma 3. *Let the maximal interval of solutions of* (29), (30), (31), (35) *be* $[0, t_f)$. *There exists* $\delta > 0$ *such that, if* $|\dot{\hat{\theta}}_A(t)| \leq \delta$ *for all* $t \in [0, t_f)$, *then* $x_{\mathrm{m}}(\cdot) \in \mathcal{L}_\infty[0, t_f)$. *Furthermore, if* $t_f = \infty$, *then system* (13) *with input* $w(t)$ *is ISS, with* $\limsup_{t\to\infty} |x_{\mathrm{m}}(t)| \leq \beta \limsup_{t\to\infty} |w(t)|$, *for a positive scalar* β.

Next, we provide a statement for system (35).

Lemma 4. *Let the maximal interval of solutions of* (29), (30), (31), (35) *be* $[0, t_f)$. *Pick* $\delta > 0$ *from Lemma 3 and let* $|\dot{\hat{\theta}}_A(t)| \leq \delta$ *for all* $t \in [0, t_f)$. *Then, it holds that* $e(\cdot), \tilde{\theta}_e(\cdot) \in \mathcal{L}_\infty[0, t_f)$. *Furthermore, if* $t_f = \infty$, *the origin* $(e, \tilde{\theta}_A) = 0$ *of subsystem* (32), (34) *with input* $\hat{\theta}_A(t)$ *is UGS.*

Finally, we combine the previous boundedness results.

Proposition 1. *Consider the closed-loop system obtained from the interconnection of the identifier dynamics* (29), (30), *the reference model* (31), *and the adaptive error system* (35). *For any* $E > 0$, *let* γ_b *be defined as*

$$\gamma_b := \delta/(\rho + E), \tag{38}$$

where δ *is given by Lemma 3 and* ρ *is found in Remark 5. If* $|\tilde{\epsilon}(0)| \leq E$ *and* $\gamma \in (0, \gamma_b]$, *then the closed-loop solutions are bounded and forward complete.*

C. Exponential Convergence to the Optimal Policy

We now focus on the uniform asymptotic stability properties of the closed-loop system (29), (30), (31), (35). Firstly, we show that $x_{\mathrm{m}}(t)$ is persistently exciting as long as $|\dot{\hat{\theta}}_{\mathrm{A}}|$ is sufficiently small.

Lemma 5. *Let the entries of input* w *be sufficiently rich of order* $n+1$ *and uncorrelated. For any* $E > 0$ *such that* $|\tilde{\epsilon}(0)| \leq E$, *there exists* $\gamma^\star \in (0, \gamma_b]$, *with* γ_b *from Proposition 1, such that, for all* $\gamma \in (0, \gamma^\star]$, *the solutions* $x_{\mathrm{m}}(t)$ *of the reference model* (31) *are persistently exciting.*

Next, we provide a direct consequence of Lemma 5 for the adaptive error dynamics (32), (34).

Lemma 6. *Let the hypotheses of Lemma 5 hold. Given* $E > 0$ *such that* $|\tilde{\epsilon}(0)| \leq E$, *let* $\gamma \in (0, \gamma^\star]$, *where* $\gamma^\star$ *is given in Lemma 5. Then, the origin* $(e, \tilde{\theta}_e) = 0$ *of system* (32), (34) *is uniformly globally asymptotically stable (UGAS) and uniformly locally exponentially stable (ULES).*

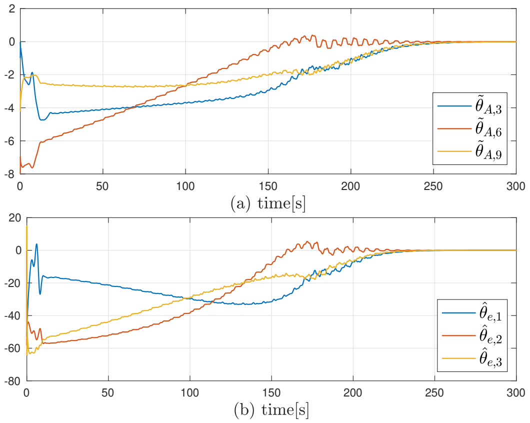
Now that we have established that every solution $e(t)$ converges exponentially to zero, uniformly from compact sets of initial conditions, we can conclude the convergence analysis by studying the identifier dynamics (29).

Lemma 7. *Let the hypotheses of Lemma 5 hold. Given* $E > 0$ *such that* $|\tilde{\epsilon}(0)| \leq E$, *let* $\gamma \in (0, \gamma^\star]$, *where* $\gamma^\star$ *is given in Lemma 5. Then, for any compact set of initial conditions of the MRAC states* $(x_{\mathrm{m}}, e, \tilde{\theta}_e)$ *and of the filter state* ξ, *the origin* $(\tilde{\epsilon}, \tilde{\theta}_A) = 0$ *of system* (29) *is uniformly globally exponentially stable (UGES).*

Remark 6. *The uniform convergence in Lemma 7 holds only once the compact set of initial conditions of the MRAC and filter states is fixed. In fact, larger initial conditions imply that signal* $x(t)$ *may cause a slower convergence rate.*

D. Sketch of the Proof of Theorem 1

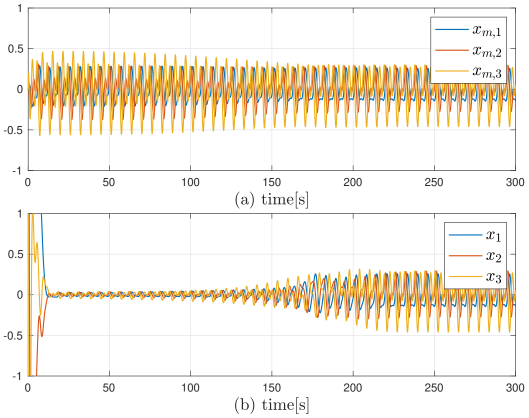
The proof involves studying the closed-loop system in the error coordinates defined in Section IV-A. Point 1 is ensured in Section IV-B by choosing $\gamma \in (0, \gamma_b]$, with γ_b defined in Proposition 1, which is proved by combining Lemmas 1, 2, 3, and 4. Point 2 is then derived from the statements of Section IV-C. Choose $\gamma \in (0, \gamma^\star]$, with $\gamma^\star \in (0, \gamma_b]$ given in Lemma 5. Then, the UGAS and ULES result for the adaptive error dynamics in Lemma 6 ensures that $(e(t), \hat{\theta}_{\mathrm{e}}(t) - \theta_{\mathrm{e}}(\hat{\theta}_{\mathrm{A}}(t)))$ converges exponentially to zero, uniformly from any compact set of initial conditions. Similarly, the result for the identifier dynamics in Lemma 7 implies exponential convergence of $\hat{\theta}_{\mathrm{A}}(t) - \theta_A$ to zero, uniformly in compact sets of initial conditions. Thus, to obtain (24), it is sufficient to recall from



(a) time[s]

(b) time[s]

Fig. 2. Parameter estimation errors and adaptive gain.



(a) time[s]

(b) time[s]

Fig. 3. Behavior of the reference model and the controlled system.

(21) that if $\hat{\theta}_\mathrm{A}(t) \to \theta_A$, then $\theta_e(\hat{\theta}_\mathrm{A}(t)) \to 0$. Finally, (23) follows from the ISS result of Lemma 3, combined with $e(t) \to 0$.

V. Numerical Example

We provide a numerical example to illustrate the effectiveness of the *Model Reference Adaptive Reinforcement Learning*. In order to satisfy Assumptions 1 and 2, we consider matrices A, B in controllability canonical form:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2 & 3 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{39}$$

We choose as weights for the LQ problem (2) the matrices $Q = 10I_3$, $R = 0.1$, while we consider the dither signal $w(t) = \sum_{i=1}^{2} \sin(2\pi i t)$. For $\hat{\theta}_\mathrm{A}$, we restrict our search in

$$\mathcal{K}_\mathrm{A} = \left\{ \hat{\theta}_\mathrm{A} \in \mathbb{R}^9 : \mathrm{vec}^{-1}(\hat{\theta}_\mathrm{A}) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \alpha_2 & \alpha_1 & \alpha_0 \end{bmatrix}, \right. \tag{40}$$
$$\left. \begin{bmatrix} \alpha_2 + 2 & \alpha_1 - 3 & \alpha_0 - 1 \end{bmatrix}^\top \in r\mathbb{B}^3 \right\},$$

where $r = 50$, and $\mathbb{B}^3$ denotes the closed unit ball in $\mathbb{R}^3$. To verify that the proposed algorithm does not require initialization to a stabilizing gain, we pick an initial parameter $\hat{\theta}_\mathrm{A}(0)$ whose associated gain $K(\hat{\theta}_\mathrm{A}(0))$ is not stabilizing for the true system. In Fig. 2-(a), we show that the parameter error $\tilde{\theta}_\mathrm{A}$ converges to zero, implying that $P(\hat{\theta}_\mathrm{A})$ converges to the optimal $P^\star$. In Figure 2-(b), we provide the adaptive gain $\hat{K}_e$ used to stabilize the plant even in presence of a non-stabilizing gain $K(\hat{\theta}_\mathrm{A})$. Once the estimation of $P^\star$ is complete, $\hat{K}_e$ becomes zero, leaving only the optimal gain as a feedback controller. Finally, we include in Fig. 3 the behavior of the reference model and the controlled plant.

VI. Conclusions

We addressed the problem of optimal control of partially unknown linear systems by proposing an algorithm combining MRAC, continuous-time identification, and LQR. Under matching conditions typical of the MRAC literature, we ensured the boundedness of solutions and, by injection of a sufficiently rich dither signal, convergence to the optimal control policy and the true system parameters. This way, the plant converges to an optimal reference model. The proposed controller does not require a priori hypotheses on PE properties of the closed-loop system trajectories or knowledge of an initial stabilizing policy. Future works on this subject will be dedicated to addressing the scenario with uncertain input matrix B and generalizing the approach to other classes of systems.

References

[1] R. Sargent, "Optimal control," *Journal of Computational and Applied Mathematics*, vol. 124, no. 1-2, pp. 361–371, 2000.

[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] 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.

[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 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.

[8] K. G. Vamvoudakis, "Q-learning for continuous-time linear systems: A model-free infinite horizon optimal control approach," *Systems & Control Letters*, vol. 100, pp. 14–20, 2017.

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

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

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

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

[13] 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.

[14] R. E. Kalman *et al.*, "Contributions to the theory of optimal control," *Bol. soc. mat. mexicana*, vol. 5, no. 2, pp. 102–119, 1960.

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

[16] 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.