ARCHIVIO ISTITUZIONALE DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

On-Policy Data-Driven Linear Quadratic Optimal Control of SISO Systems via Model Reference Adaptive Reinforcement Learning

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Bosso, A., Borghesi, M., Serrani, A., Notarstefano, G., Teel, A.R. (2025). On-Policy Data-Driven Linear Quadratic Optimal Control of SISO Systems via Model Reference Adaptive Reinforcement Learning. New York : IEEE [10.1109/cdc57313.2025.11312870].

Availability:
This version is available at: https://hdl.handle.net/11585/1042935 since: 2026-02-08

Published:

DOI: http://doi.org/10.1109/cdc57313.2025.11312870

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

07 September 2026

On-Policy Data-Driven Linear Quadratic Optimal Control of SISO Systems via Model Reference Adaptive Reinforcement Learning

Alessandro Bosso[1], Marco Borghesi[1], Andrea Serrani[1], Giuseppe Notarstefano[1], and Andrew R. Teel[2]

***Abstract*— In this paper, we address the problem of on-policy data-driven linear quadratic optimal control for continuous-time single-input single-output systems. Assuming that the plant is minimum phase and has relative degree one, we propose model reference adaptive reinforcement learning – an approach with theoretical guarantees that combines learning and model reference adaptive control. The developed algorithm features an adaptive output-feedback controller that tracks a parameter-varying reference model, whose behavior is shaped by a discrete-time optimizer. For the resulting hybrid closed-loop system, we establish semi-global boundedness of the solutions and show that, under persistency of excitation induced by a dither signal, the applied policy converges to the optimal one.**

I. Introduction

Over the years, the control community has increasingly moved from model-based to data-driven approaches. Using data as a substitute for model information has been a central theme in adaptive control [1], which aims at stabilization and tracking, and more recently in reinforcement learning (RL) [2], which focuses on optimal control. A historical overview of these trends is given in [3]. In this context, our goal is to study linear quadratic (LQ) optimal control of single-input single-output (SISO) linear time-invariant systems.

Several algorithms have been proposed for data-driven LQ optimal control in the state-feedback setting. Among these, we find off-policy approaches that compute the optimal gain offline from a dataset [4]–[6] or online while controlling the system with a suboptimal feedback law [7]. Early on-policy techniques include [8]–[10], where the data are generated under the current policy estimate, but an initial stabilizing gain is needed and the closed-loop interconnection of plant and control/learning dynamics is not fully analyzed. Other on-policy methods, such as [11], [12], provide closed-loop stability guarantees but still rely on an initial stabilizing gain. In contrast, the continuous-time value iteration approach of [13] establishes stability of the learning dynamics without such a requirement. Similarly, [14] introduces model reference adaptive reinforcement learning (MR-ARL), which combines model reference adaptive control (MRAC, see [1], [15]) for stabilization and value iteration to achieve optimality, without any need for an initial stabilizing policy. A related approach is found in [16], which also combines MRAC and RL, but without ensuring optimality as in [14].

In the output-feedback setting, [17] proposes an observer to implement and compute the optimal policy without state measurements, although without proving closed-loop stability during the learning transient. Recently, [18] applied the value iteration method of [13] within adaptive pole placement schemes for continuous-time SISO systems [1, Ch. 7], establishing closed-loop stability and optimality under persistency of excitation. However, without excitation, this scheme may suffer from the loss of stabilizability of the estimates, so their good behavior is assumed rather than enforced by design.

In this work, we extend the MR-ARL approach of [14] to continuous-time, minimum-phase, SISO systems with relative degree one. The proposed on-policy algorithm is an actor-critic architecture that does not require an initial stabilizing policy. Specifically, a continuous-time observer-based adaptive controller (actor) stabilizes the system and tracks a parameter-varying reference model, while a discrete-time optimizer (critic) adjusts the reference model by solving an LQ problem based on the observer parameters. To prevent any loss of stabilizability, we employ a non-minimal realization of the plant, known in the adaptive literature [15, Ch. 4] and revisited here in a novel state-space setting. The closed-loop system combines continuous- and discrete-time dynamics and is modeled as a hybrid dynamical system [19]. We show that: (i) the solutions remain bounded from any compact set of initial conditions if the optimizer's sampling time is sufficiently large; (ii) if a dither signal is injected to ensure persistency of excitation, the applied policy becomes optimal and the reference model converges to the optimal closed-loop behavior for the plant.

The paper is organized as follows. In Section II, we discuss the control problem. In Section III, we present the non-minimal realization. Then, Sections IV and V are dedicated to MR-ARL and its analysis. Section VI provides a numerical example and Section VII concludes the paper. Due to space constraints, some proofs are omitted.

Notation: we use $\mathbb{R}$ to denote the set of real numbers, and for any $a \in \mathbb{R}$, we let $\mathbb{R}_{\geq a} \coloneqq [a, \infty)$. I_j denotes the identity matrix of dimension j. Given a symmetric matrix M, $M \succeq 0$ denotes that it is positive semidefinite. For any square matrix M, we indicate its spectrum with $\sigma(M)$. Given vectors v_1, v_2, we often use (v_1, v_2) for $[v_1^\top \ v_2^\top]^\top$. Finally, we refer to [19] for the notation and tools of hybrid dynamical systems.

[1]A. Bosso, M. Borghesi, A. Serrani, and G. Notarstefano are with the Department of Electrical, Electronic, and Information Engineering, University of Bologna, Italy. Email: {alessandro.bosso, m.borghesi, andrea.serrani, giuseppe.notarstefano}@unibo.it

[2]A. R. Teel is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA, USA. Email: teel@ece.ucsb.edu

The research leading to these results has received funding from the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 101104404 - IMPACT4Mech.

This work is also partially funded by NextGenerationEU PNRR PRIN 2022 ECODREAM, code 202228CTKY002 - CUP J53D23000560006.

II. Problem Statement

A. On-Policy Data-Driven LQ Control of SISO Systems

Consider a continuous-time SISO LTI system of the form

$$\begin{aligned} \dot{x} &= Ax + bu \\ y &= c^\top x, \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^n$ is the state, with dimension n known, $u \in \mathbb{R}$ is the control input, $y \in \mathbb{R}$ is the measured output, and A, b, and c are unknown matrices satisfying the following assumption.

Assumption 1. *The pair (A, b) is controllable and the pair $(c^\top, A)$ is observable.*

We associate to the model (1) the cost functional:

$$J(x_0, u(\cdot)) \coloneqq \int_0^\infty (y^2(t, x_0, u(\cdot)) + ru^2(t))\mathrm{d}t, \tag{2}$$

where $r > 0$ is a known scalar weight and $y(t, x_0, u(\cdot))$ is the output trajectory of system (1) at time t, subject to the initial condition $x_0 \in \mathbb{R}^n$ and the control sequence $u(\cdot)$. The infinite-horizon LQ optimal control problem involves finding a policy $\pi^\star : \mathbb{R}^n \to \mathbb{R}$ such that $u(t) = \pi^\star(x(t))$ minimizes the cost functional (2) for all $x_0 \in \mathbb{R}^n$.

Under Assumption 1, the optimal policy is given by

$$\pi^\star(x) \coloneqq -r^{-1} b^\top P^\star x, \tag{3}$$

where $P^\star \in \mathbb{R}^{n \times n}$ is the unique symmetric and positive definite solution of the algebraic Riccati equation (ARE):

$$A^\top P + PA - r^{-1} Pbb^\top P + cc^\top = 0. \tag{4}$$

Our goal is to obtain the optimal policy when A, b, and c are unknown and x is not available. Also, (1) may be unstable and no initial output-feedback stabilizing policy is known. To address this scenario, we consider the following problem.

On-Policy Data-Driven LQ Optimal Control Problem

Design a controller having state ς and control policy

$$u = \pi(y, \varsigma, d), \tag{5}$$

where d is a probing signal (dither), such that:

1) ***Exploration:*** the policy (5) is applied to the system (1) to learn from the data acquired online.
2) ***Exploitation:*** the policy $\pi(y, \varsigma, d)$ converges to $\pi^\star(x) + d$, with $\pi^\star(x)$ the optimal policy in (3).
3) ***Boundedness:*** given a specified compact set of initial conditions of x and ς, the resulting closed-loop signals are bounded.

B. LQ Control via Model Reference Adaptive Control

We propose to tackle the above problem with an MRAC architecture. Thus, we introduce some standard assumptions.

Assumption 2. *System (1) has relative degree one. In particular, a scalar $\beta_0 > 0$ is known such that $\beta \coloneqq c^\top b \geq \beta_0$.*

Assumption 3. *System (1) is minimum phase, i.e., all zeros of the transfer function $c^\top (sI_n - A)^{-1} b$ have strictly negative real part.*

While future work will relax Assumption 2 to address the case of arbitrary relative degree, Assumption 3 is necessary for any model reference control design.

Consider a reference model for system (1) of the form

$$\begin{aligned} \dot{x}_\mathrm{m} &= A_\mathrm{m} x_\mathrm{m} + b_\mathrm{m} u_\mathrm{r} \\ y_\mathrm{r} &= c_\mathrm{m}^\top x_\mathrm{m}, \end{aligned} \tag{6}$$

where $x_\mathrm{m} \in \mathbb{R}^\mu$ is the state, $u_\mathrm{r} \in \mathbb{R}$ and $y_\mathrm{r} \in \mathbb{R}$ are the reference input and output, while A_m, b_m, and c_m are design matrices. In classical MRAC [15], the objective is to design a controller of the plant (1) that, by processing $y(t)$, $x_\mathrm{m}(t)$, and $u_\mathrm{r}(t)$, ensures that: (i) the closed-loop signals are bounded; (ii) the error $y(t) - y_\mathrm{r}(t)$ converges asymptotically to zero.

Since this architecture solves the stabilization part of our problem, one may wonder if also learning could be achieved. It turns out that $\pi^\star$ can be learned by appropriately shaping the behavior of the reference model, instead of modifying the underlying stabilizing controller. In particular, we will show that our problem is solved if (6) is replaced by a parameter-varying system that converges to a non-minimal realization of (1). We now introduce such a realization and its properties.

III. Non-Minimal Realization of the Plant

Following classical adaptive observer design [15, Thm. 4.3], a non-minimal realization of (1) is given by

$$\begin{aligned} \dot{\xi} &= F\xi + gu \\ y &= h^\top \xi, \end{aligned} \tag{7}$$

with input and output u, $y \in \mathbb{R}$ as in (1), state $\xi \in \mathbb{R}^{2n-1}$, and matrices defined as

$$F \coloneqq \begin{bmatrix} \Lambda & 0 & \ell \\ 0 & \Lambda & 0 \\ \theta_1^\top & \theta_2^\top & \theta_y \end{bmatrix}, \quad g \coloneqq \begin{bmatrix} 0 \\ \ell \\ \beta \end{bmatrix}, \quad h \coloneqq \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \tag{8}$$

where $\Lambda \in \mathbb{R}^{(n-1)\times(n-1)}$ and $\ell \in \mathbb{R}^{n-1}$ are tuning gains described later, $\beta \coloneqq c^\top b$, and

$$\theta \coloneqq \begin{bmatrix} \theta_1^\top & \theta_2^\top & \theta_y \end{bmatrix}^\top \in \mathbb{R}^{2n-1} \tag{9}$$

are parameters matching the transfer functions of (1) and (7).

Instead of using the transfer-function framework of [15], we study the properties of (7) in a state-space setting. In particular, the next result is a reinterpretation of the non-minimal realization (7) from a point of view closely related to the original Luenberger's paper on observer design [20].

Lemma 1. *For any controllable pair (Λ, ℓ) such that no eigenvalue of Λ is a zero of the plant transfer function $c^\top (sI_n - A)^{-1} b$, there exist parameters θ and a similarity transformation $(\eta, x) = T\xi$ such that system (7) becomes:*

$$\begin{bmatrix} \dot{\eta} \\ \dot{x} \end{bmatrix} = \begin{bmatrix} \Lambda & \ell c^\top \\ 0 & A \end{bmatrix} \begin{bmatrix} \eta \\ x \end{bmatrix} + \begin{bmatrix} 0 \\ b \end{bmatrix} u, \qquad y = \begin{bmatrix} 0 & c^\top \end{bmatrix} \begin{bmatrix} \eta \\ x \end{bmatrix}. \tag{10}$$

Proof: Under Assumption 2, there exists a similarity transformation such that system (1) can be rewritten in normal form [21, Ch. 2]:

$$\dot{z} = A_z z + b_z y, \qquad \dot{y} = c_z^\top z + \alpha y + \beta u, \tag{11}$$

where $z \in \mathbb{R}^{n-1}$. From the PBH test, (A_z, b_z) is controllable and $(c_z^\top, A_z)$ is observable. Also, A_z is Hurwitz by Assumption 3 and $\sigma(\Lambda) \cap \sigma(A_z) = \emptyset$ by hypothesis. Let $X \in \mathbb{R}^{(n-1)\times(n-1)}$ be the unique solution of the following Sylvester equation:

$$\Lambda X - X A_z = \ell c_z^\top \beta^{-1}. \tag{12}$$

From the proof of [22, Lemma 1.5.6], X is nonsingular since (Λ, ℓ) is controllable and $(c_z^\top, A_z)$ is observable. Define:

$$\chi \coloneqq Xz + \ell\beta^{-1} y, \tag{13}$$

and rewrite (11) in the coordinates (χ, y):

$$\dot{\chi} = \Lambda\chi + b_\chi y + \ell u, \qquad \dot{y} = \theta_\chi^\top \chi + \theta_y y + \beta u, \tag{14}$$

where $b_\chi \coloneqq Xb_z + (\alpha I_{n-1} - \Lambda)\ell\beta^{-1}$, $\theta_\chi^\top \coloneqq c_z^\top X^{-1}$, and $\theta_y \coloneqq \alpha - c_z^\top X^{-1}\ell\beta^{-1}$. If the plant parameters were known, a reduced-order observer of z could be obtained by copying the dynamics of χ in (14) and using the transformation (13).

We now show that ξ can be transformed as follows:

$$\begin{bmatrix}\eta \\ \chi \\ y\end{bmatrix} = \begin{bmatrix} I_{n-1} & 0 & 0 \\ Y & I_{n-1} & 0 \\ 0 & 0 & 1\end{bmatrix} \xi, \tag{15}$$

where $Y \in \mathbb{R}^{(n-1)\times(n-1)}$ is a suitable matrix, while χ and y satisfy (14). Using (7), (14), we compute the derivative on both sides of (15) to obtain, after some simplifications,

$$\Lambda Y = Y\Lambda, \qquad Y\ell = b_\chi \tag{16a}$$

$$\theta_\chi^\top \begin{bmatrix} Y & I_{n-1}\end{bmatrix} = \begin{bmatrix}\theta_1^\top & \theta_2^\top\end{bmatrix}. \tag{16b}$$

From [23, Appendix II], (16a) has a unique solution because (Λ, ℓ) is controllable. Then, once Y is found, (θ_1, θ_2) is computed from (16b). The existence of T follows by combining (13), (15), and the transformation from (z, y) to x. □

With Lemma 1, we obtain the following key properties.

Lemma 2. *Consider Λ, ℓ, and θ from Lemma 1, and let Λ be Hurwitz. Then, the pair (F, g) is controllable and the pair $(h^\top, F)$ is detectable.*

Lemma 3. *Consider Λ, ℓ, θ, and T from Lemma 1, and let Λ be Hurwitz. Then, the ARE obtained from (7) and (2):*

$$F^\top Q + QF - r^{-1} Q g g^\top Q + hh^\top = 0, \tag{17}$$

admits a unique symmetric, positive semidefinite solution:

$$Q^\star \coloneqq T^\top \begin{bmatrix} 0 & 0 \\ 0 & P^\star\end{bmatrix} T, \tag{18}$$

where $P^\star$ is the matrix in (3). Thus, the optimal policy for (7) with cost (2) is $-r^{-1} g^\top Q^\star \xi = -r^{-1} b^\top P^\star x = \pi^\star(x)$.

IV. Model Reference Adaptive Reinforcement Learning

The proposed scheme, named *model reference adaptive reinforcement learning* (MR-ARL) for SISO systems, is summarized in Algorithm 1 and illustrated in Fig. 1 in the next page. In particular, we consider an actor-critic architecture combining MRAC and a parameter-varying reference model whose updates are triggered in discrete time by the clock (19). We now describe the building blocks of MR-ARL.

Algorithm 1 MR-ARL for SISO Systems

Initialization:

Tuning: $\tau_\mathrm{s} > 0$, (Λ, ℓ) as in Lemma 1 with Λ Hurwitz, $\lambda > 0$, $k > 0$, $\gamma > 0$, $\epsilon > 0$, $\beta_0 > 0$ from Assumption 2.
Inputs: dither d (continuous, bounded, stationary, sufficiently rich of order $3n$ [1, Def. 5.2.3]), plant output y.
States: $\tau \in [0, \tau_\mathrm{s}]$, $\xi_\mathrm{m} \in \mathbb{R}^{2n-1}$, $\hat{\zeta} \in \mathbb{R}^{2n-2}$, $\hat{y} \in \mathbb{R}$, $(\hat{\theta}_\mathrm{a}, \hat{\beta}_\mathrm{a}) \in \mathbb{R}^{2n-1} \times \mathbb{R}_{\geq\beta_0}$, $(\hat{\theta}_\mathrm{c}, \hat{\beta}_\mathrm{c}) \in \mathcal{R}_\epsilon$, with $\mathcal{R}_\epsilon$ in (29).

Clock:

$$\begin{cases} \dot{\tau} = 1, & \tau \in [0, \tau_\mathrm{s}] \\ \tau^+ = 0, & \tau = \tau_\mathrm{s}. \end{cases} \tag{19}$$

Reference model (continuous time):

Reference dynamics:

$$\dot{\xi}_\mathrm{m} = \hat{F}(\hat{\theta}_\mathrm{c})\xi_\mathrm{m} + \hat{g}(\hat{\beta}_\mathrm{c}) u_\mathrm{r}, \qquad y_\mathrm{r} = h^\top \xi_\mathrm{m}, \tag{20}$$

with $\hat{F}$, $\hat{g}$, and h defined in (26).
Reference input:

$$u_\mathrm{r} = -r^{-1}\hat{g}^\top(\hat{\beta}_\mathrm{c})\hat{Q}\xi_\mathrm{m} + d. \tag{21}$$

Actor (continuous time):

Adaptive observer:

$$\begin{aligned} \dot{\hat{\zeta}} &= \begin{bmatrix}\Lambda & 0 \\ 0 & \Lambda\end{bmatrix}\hat{\zeta} + \begin{bmatrix}\ell & 0 \\ 0 & \ell\end{bmatrix}\begin{bmatrix} y \\ u\end{bmatrix} \\ \dot{\hat{y}} &= \phi_\mathrm{a}^\top \hat{\theta}_\mathrm{a} + \hat{\beta}_\mathrm{a} u - \lambda(\hat{y} - y), \qquad \phi_\mathrm{a} \coloneqq (\hat{\zeta}, y) \\ \dot{\hat{\theta}}_\mathrm{a} &= -\gamma\phi_\mathrm{a}(\hat{y} - y), \quad \dot{\hat{\beta}}_\mathrm{a} = \mathrm{Proj}_{\geq\beta_0}\{-\gamma u(\hat{y} - y)\}, \end{aligned} \tag{22}$$

with $\mathrm{Proj}_{\geq\beta_0}\{\cdot\}$ defined in (28).
Control input:

$$u = \hat{\beta}_\mathrm{a}^{-1}(\xi_\mathrm{m}^\top \hat{\theta}_\mathrm{c} - \phi_\mathrm{a}^\top \hat{\theta}_\mathrm{a} + \hat{\beta}_\mathrm{c} u_\mathrm{r} - k(\hat{y} - y_\mathrm{r})). \tag{23}$$

Critic (discrete time):

Critic update during jumps:

$$(\hat{\theta}_\mathrm{c}, \hat{\beta}_\mathrm{c})^+ = \begin{cases} (\hat{\theta}_\mathrm{a}, \hat{\beta}_\mathrm{a}), & \text{if } |\det \hat{R}(\hat{\theta}_\mathrm{a}, \hat{\beta}_\mathrm{a})| \geq \epsilon \\ (\hat{\theta}_\mathrm{c}, \hat{\beta}_\mathrm{c}), & \text{otherwise,} \end{cases} \tag{24}$$

with $\hat{R}$ defined in (29).
Algebraic Riccati equation: find $\hat{Q} = \hat{Q}^\top \succeq 0$ such that

$$\hat{F}^\top(\hat{\theta}_\mathrm{c})\hat{Q} + \hat{Q}\hat{F}(\hat{\theta}_\mathrm{c}) - r^{-1}\hat{Q}\hat{g}(\hat{\beta}_\mathrm{c})\hat{g}^\top(\hat{\beta}_\mathrm{c})\hat{Q} + hh^\top = 0. \tag{25}$$

Reference model: This continuous-time block specifies the target behavior for the controlled plant (1). In particular, the matrices of system (20) are defined as

$$\hat{F} : \hat{\theta} \mapsto \left[\begin{array}{ccc} \Lambda & 0 & \ell \\ 0 & \Lambda & 0 \\ \hline & \hat{\theta}^\top & \end{array}\right], \quad \hat{g} : \hat{\beta} \mapsto \begin{bmatrix} 0 \\ \ell \\ \hat{\beta}\end{bmatrix}, \quad h \coloneqq \begin{bmatrix} 0 \\ 0 \\ 1\end{bmatrix}. \tag{26}$$

Thus, (20) has the same structure of (7), with (θ, β) replaced by the estimates $(\hat{\theta}_\mathrm{c}, \hat{\beta}_\mathrm{c})$ received from the critic. Moreover, the reference input (21) is the sum of a dither signal d and a feedback law that yields the following closed-loop matrix,

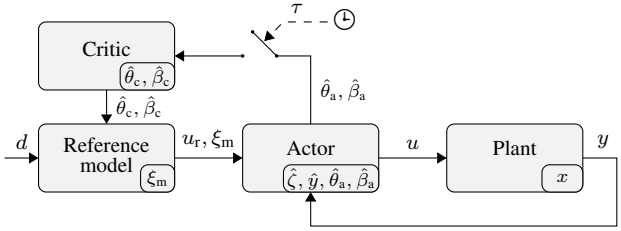

Fig. 1. Interconnection of MR-ARL with the controlled plant.

whose properties are determined by the critic:

$$F_{\text{cl}}(\hat{\theta}_c, \hat{\beta}_c) \coloneqq \hat{F}(\hat{\theta}_c) - r^{-1}\hat{g}(\hat{\beta}_c)\hat{g}^\top(\hat{\beta}_c)\hat{Q}(\hat{\theta}_c, \hat{\beta}_c). \quad (27)$$

Actor: This continuous-time block is an adaptive controller based on MRAC design that ensures that $y(t) - y_r(t)$ converges to zero. The scheme comprises the adaptive observer (22), with parameter estimates $(\hat{\theta}_a, \hat{\beta}_a)$, and the control input (23). In (22), $\text{Proj}_{\geq\beta_0}\{\cdot\}$ is a projection operator that enforces the constraint $\hat{\beta}_a \geq \beta_0$ along flows:

$$\text{Proj}_{\geq\beta_0}\{v\} \coloneqq \begin{cases} v, & (\hat{\beta}_a > \beta_0) \vee ((\hat{\beta}_a = \beta_0) \wedge (v \geq 0)) \\ 0, & (\hat{\beta}_a = \beta_0) \wedge (v < 0). \end{cases} \quad (28)$$

Critic: This discrete-time module is used for updating the reference model according to the estimates generated by the actor. In particular, the critic samples the weights $(\hat{\theta}_a, \hat{\beta}_a)$ whenever they belong to the set $\mathcal{R}_\epsilon$ defined as

$$\begin{aligned} \mathcal{R}_\epsilon &\coloneqq \{(\hat{\theta}, \hat{\beta}) \in \mathbb{R}^{2n} : |\det \hat{R}(\hat{\theta}, \hat{\beta})| \geq \epsilon\} \\ \hat{R}(\hat{\theta}, \hat{\beta}) &\coloneqq \begin{bmatrix} \hat{g}(\hat{\beta}) & \hat{F}(\hat{\theta})\hat{g}(\hat{\beta}) & \cdots & \hat{F}^{2n-2}(\hat{\theta})\hat{g}(\hat{\beta}) \end{bmatrix}, \end{aligned} \quad (29)$$

i.e., the estimated non-minimal plant is controllable with a margin ϵ. Otherwise, the block freezes the values of $(\hat{\theta}_c, \hat{\beta}_c)$ until the next jump. The weights are then used in (25) to compute an estimate $\hat{Q}$ of the matrix $Q^\star$ in (18). As a result, F_{cl} in (27) is Hurwitz at all times.

Given (Λ, ℓ) from Algorithm 1, consider θ and T from Lemma 1. It is worth noting that, if: (i) the parameter estimates satisfy $(\hat{\theta}_a, \hat{\beta}_a) = (\hat{\theta}_c, \hat{\beta}_c) = (\theta, \beta)$; and (ii) the observer tracks the plant and the reference model, i.e., $(\hat{\zeta}, \hat{y}) = \xi_m$ and $(\hat{\eta}, \hat{x}) \coloneqq T(\hat{\zeta}, \hat{y})$ is such that $\hat{x} = x$, then the control input in (23) becomes, due to Lemma 3:

$$u = -r^{-1}g^\top Q^\star \xi_m + d = -r^{-1}b^\top P^\star x + d. \quad (30)$$

In the next section, we will show semi-global boundedness of the closed-loop solutions and, under persistency of excitation induced by d, convergence to these conditions.

V. Algorithm Analysis

We study the closed-loop system by reformulating it as a well-posed hybrid inclusion. First, we present its building blocks based on Lemma 1 and its proof.

A. Dither Dynamics

To aid with the analysis, it is convenient to model d as the output of an autonomous system (exosystem) of the form

$$\begin{aligned} \dot{w} &\in S(w), \qquad w \in \mathcal{W} \\ d &= \Delta(w), \end{aligned} \quad (31)$$

where $\mathcal{W} \subset \mathbb{R}^{n_w}$ is a compact set, $\Delta : \mathcal{W} \to \mathbb{R}$ is a continuous function, and $S : \mathcal{W} \rightrightarrows \mathbb{R}^{n_w}$ is an outer semicontinuous and locally bounded set-valued map with non-empty, convex values on $\mathcal{W}$. Also, we assume that every solution of (31) is forward complete. Note that (31) can generate a wide range of bounded signals beyond sums of sinusoids.

B. Stabilization Dynamics

Observer estimation error: Consider θ from Lemma 1. Here, we study the performance of the observer (22) in reconstructing the plant state x and the parameters (θ, β). Instead of using x, we consider the plant in the coordinates (χ, y) as in (14). Using (15), define the estimation errors:

$$\begin{aligned} \tilde{\chi} &\coloneqq \begin{bmatrix} Y & I_{n-1} \end{bmatrix} \hat{\zeta} - \chi, & \tilde{y} &\coloneqq \hat{y} - y \\ \tilde{\theta} &\coloneqq \hat{\theta}_a - \theta, & \tilde{\beta} &\coloneqq \hat{\beta}_a - \beta. \end{aligned} \quad (32)$$

Also, define the set-valued map $\mathcal{P} : \mathbb{R}_{\geq\beta_0} \times \mathbb{R} \rightrightarrows \mathbb{R}$ as

$$\mathcal{P}(\hat{\beta}_a, v) \coloneqq \begin{cases} v, & (\hat{\beta}_a > \beta_0) \vee ((\hat{\beta}_a = \beta_0) \wedge (v \geq 0)) \\ [v, 0] & (\hat{\beta}_a = \beta_0) \wedge (v \leq 0), \end{cases} \quad (33)$$

which is a regularized version of the projection operator $\text{Proj}_{\geq\beta_0}\{\cdot\}$ in (28). Combining (14), (16a), (22), and (33), we obtain the differential inclusion:

$$\begin{aligned} \begin{bmatrix} \dot{\tilde{\chi}} \\ \dot{\tilde{y}} \end{bmatrix} &= \begin{bmatrix} \Lambda & 0 \\ \theta_\chi^\top & -\lambda \end{bmatrix} \begin{bmatrix} \tilde{\chi} \\ \tilde{y} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} (\phi_a^\top \tilde{\theta} + u\tilde{\beta}) \\ \dot{\tilde{\theta}} &= -\gamma\phi_a\tilde{y}, \qquad \dot{\tilde{\beta}} \in \mathcal{P}(\tilde{\beta} + \beta, -\gamma u\tilde{y}), \end{aligned} \quad (34)$$

with $(\tilde{\chi}, \tilde{y}, \tilde{\theta}, \tilde{\beta}) \in \mathbb{R}^{3n-1} \times \mathbb{R}_{\geq\beta_0-\beta}$.

Tracking error: We are now interested in studying how $(\hat{\zeta}, \hat{y})$ in (22) tracks the reference state ξ_m of (20). Recalling (15), consider the following mismatch errors:

$$\tilde{\eta} \coloneqq \begin{bmatrix} I_{n-1} & 0 \end{bmatrix} \hat{\zeta} - \begin{bmatrix} I_{n-1} & 0 & 0 \end{bmatrix} \xi_m, \quad e \coloneqq \hat{y} - y_r. \quad (35)$$

Using (20), (22), and (23), we obtain

$$\begin{bmatrix} \dot{\tilde{\eta}} \\ \dot{e} \end{bmatrix} = \begin{bmatrix} \Lambda & \ell \\ 0 & -k \end{bmatrix} \begin{bmatrix} \tilde{\eta} \\ e \end{bmatrix} - \begin{bmatrix} \ell \\ \lambda \end{bmatrix} \tilde{y}, \quad (36)$$

with $(\tilde{\eta}, e) \in \mathbb{R}^n$.

Finally, we denote with $\boldsymbol{x}_s \coloneqq (\tilde{\chi}, \tilde{y}, \tilde{\theta}, \tilde{\beta}, \tilde{\eta}, e)$ the overall state in error coordinates and by $\mathcal{X}_s \coloneqq \mathbb{R}^{3n-1} \times \mathbb{R}_{\geq\beta_0-\beta} \times \mathbb{R}^n$ the flow set of the stabilization dynamics (34), (36).

C. Optimization Dynamics

We include the remaining states in the optimization dynamics. Define $\boldsymbol{x}_o \coloneqq (\tau, \xi_m, z, \hat{\theta}_c, \hat{\beta}_c)$, $\mathcal{X}_o \coloneqq [0, \tau_s] \times \mathbb{R}^{3n-2} \times \mathcal{R}_\epsilon$, and let $C_o \coloneqq \mathcal{X}_o$ and $D_o \coloneqq \{\boldsymbol{x}_o \in \mathcal{X}_o : \tau = \tau_s\}$. Then, the interconnection of the clock (19), the reference model (20), the reference input (21), the critic (24), (25), and

the z-subsystem of (11) is given by following hybrid system:

$$\begin{cases} \dot{\tau} = 1 \\ \dot{\xi}_{\mathrm{m}} = F_{\mathrm{cl}}(\hat{\theta}_{\mathrm{c}}, \hat{\beta}_{\mathrm{c}})\xi_{\mathrm{m}} + \hat{g}(\hat{\beta}_{\mathrm{c}})d \\ \dot{z} = A_z z + b_z(h^\top \xi_{\mathrm{m}} + e - \tilde{y}) \qquad \boldsymbol{x}_{\mathrm{o}} \in C_{\mathrm{o}} \\ (\dot{\hat{\theta}}_{\mathrm{c}}, \dot{\hat{\beta}}_{\mathrm{c}}) = 0 \end{cases} \tag{37}$$

$$\begin{cases} \tau^+ = 0 \\ \xi_{\mathrm{m}}^+ = \xi_{\mathrm{m}} \\ z^+ = z \qquad \boldsymbol{x}_{\mathrm{o}} \in D_{\mathrm{o}}, \\ (\hat{\theta}_{\mathrm{c}}, \hat{\beta}_{\mathrm{c}})^+ \in \begin{cases} (\hat{\theta}_{\mathrm{a}}, \hat{\beta}_{\mathrm{a}}), & |\det \hat{R}(\hat{\theta}_{\mathrm{a}}, \hat{\beta}_{\mathrm{a}})| \geq \epsilon \\ (\hat{\theta}_{\mathrm{c}}, \hat{\beta}_{\mathrm{c}}), & |\det \hat{R}(\hat{\theta}_{\mathrm{a}}, \hat{\beta}_{\mathrm{a}})| \leq \epsilon \end{cases} \end{cases}$$

where $F_{\mathrm{cl}}(\hat{\theta}_{\mathrm{c}}, \hat{\beta}_{\mathrm{c}})$ has been defined in (27).

D. Main Results

The interconnection of the dither dynamics (31), the stabilization dynamics (34), (36), and the optimization dynamics (37) can be compactly rewritten as follows:

$$\begin{cases} \dot{w} \in S(w) \\ \dot{\boldsymbol{x}}_{\mathrm{s}} \in \mathcal{F}_{\mathrm{s}}(\Delta(w), \boldsymbol{x}_{\mathrm{s}}, \boldsymbol{x}_{\mathrm{o}}) \\ \dot{\boldsymbol{x}}_{\mathrm{o}} = \mathcal{F}_{\mathrm{o}}(\Delta(w), \boldsymbol{x}_{\mathrm{s}}, \boldsymbol{x}_{\mathrm{o}}) \end{cases} \quad \begin{bmatrix} w \\ \boldsymbol{x}_{\mathrm{s}} \\ \boldsymbol{x}_{\mathrm{o}} \end{bmatrix} \in \mathcal{W} \times \mathcal{X}_{\mathrm{s}} \times C_{\mathrm{o}}$$

$$\begin{cases} w^+ = w \\ \boldsymbol{x}_{\mathrm{s}}^+ = \boldsymbol{x}_{\mathrm{s}} \\ \boldsymbol{x}_{\mathrm{o}}^+ \in \mathcal{G}(\boldsymbol{x}_{\mathrm{s}}, \boldsymbol{x}_{\mathrm{o}}) \end{cases} \quad \begin{bmatrix} w \\ \boldsymbol{x}_{\mathrm{s}} \\ \boldsymbol{x}_{\mathrm{o}} \end{bmatrix} \in \mathcal{W} \times \mathcal{X}_{\mathrm{s}} \times D_{\mathrm{o}}. \tag{38}$$

By inspecting the defined sets and maps, it can be verified that (38) satisfies the hybrid basic assumptions [19, As. 6.5].

We provide a useful result for the stabilization dynamics.

Lemma 4. *There exist a continuously differentiable function $V_{\mathrm{s}} : \mathcal{X}_{\mathrm{s}} \to \mathbb{R}_{\geq 0}$, two positive scalars ν_1, ν_2, and a continuous function $U : \mathcal{W} \times \mathcal{X}_{\mathrm{s}} \times C_{\mathrm{o}} \to \mathbb{R}_{\geq 0}$ such that*

$$\nu_1 |\boldsymbol{x}_{\mathrm{s}}|^2 \leq V_s(\boldsymbol{x}_{\mathrm{s}}) \leq \nu_2 |\boldsymbol{x}_{\mathrm{s}}|^2, \qquad \forall \boldsymbol{x}_{\mathrm{s}} \in \mathcal{X}_{\mathrm{s}}, \tag{39}$$

and, for all $(w, \boldsymbol{x}_{\mathrm{s}}, \boldsymbol{x}_{\mathrm{o}}) \in \mathcal{W} \times \mathcal{X}_{\mathrm{s}} \times C_{\mathrm{o}}$,

$$\max_{f \in \mathcal{F}_{\mathrm{s}}(\Delta(w), \boldsymbol{x}_{\mathrm{s}}, \boldsymbol{x}_{\mathrm{o}})} \langle \nabla V_{\mathrm{s}}(\boldsymbol{x}_{\mathrm{s}}), f \rangle \leq -U(w, \boldsymbol{x}_{\mathrm{s}}, \boldsymbol{x}_{\mathrm{o}}), \tag{40}$$

where $U(w, \boldsymbol{x}_{\mathrm{s}}, \boldsymbol{x}_{\mathrm{o}}) = 0$ if and only if $(\tilde{\chi}, \tilde{y}, \tilde{\eta}, e) = 0$.

With Lemma 4, we obtain the first main result of this paper, which ensures the boundedness requirement of Section II.

Theorem 1. *For any compact set $\mathcal{K} \subset \mathcal{W} \times \mathcal{X}_{\mathrm{s}} \times \mathcal{X}_{\mathrm{o}}$, there exists a minimum sampling time $\tau_{\mathrm{s}}^\star > 0$ such that, for all $\tau_{\mathrm{s}} \geq \tau_{\mathrm{s}}^\star$, all maximal solutions of (38) initialized in $\mathcal{K}$ are precompact, i.e., they are bounded and forward complete.*

Remark 1. *Note that if $(w, \boldsymbol{x}_{\mathrm{s}}, \boldsymbol{x}_{\mathrm{o}})$ is bounded, then x and all the states in Algorithm 1 are bounded. This fact is obtained by using (32), (35), and recalling (13).*

From Lemma 4, we also obtain the following technical result, which is a straightforward application of [19, Cor. 8.4].

Lemma 5. *Let ψ be any precompact solution of (38), and let $\mathcal{S}$ be such that[1] $\overline{\mathrm{rge}\, \psi} \subset \mathcal{S}$. Also, define $V(w, \boldsymbol{x}_{\mathrm{s}}, \boldsymbol{x}_{\mathrm{o}}) \coloneqq V_{\mathrm{s}}(\boldsymbol{x}_{\mathrm{s}})$, with V_{s} given in Lemma 4. Then, for some $\rho \in V(\mathcal{S})$, ψ approaches the non-empty set that is the largest weakly invariant subset of:*

$$\mathcal{E} \coloneqq V^{-1}(\rho) \cap \mathcal{S} \cap \{(w, \boldsymbol{x}_{\mathrm{s}}, \boldsymbol{x}_{\mathrm{o}}) \in \mathcal{W} \times \mathcal{X}_{\mathrm{s}} \times \mathcal{X}_{\mathrm{o}} : (\tilde{\chi}, \tilde{y}, \tilde{\eta}, e) = 0\}. \tag{41}$$

On the set $\mathcal{E}$, the solutions of (38) satisfy the clock dynamics (19) and the following differential inclusion

$$\begin{aligned} \dot{w} &\in S(w) \\ \dot{\xi}_{\mathrm{m}} &= F_{\mathrm{cl}}(\hat{\theta}_{\mathrm{c}}, \hat{\beta}_{\mathrm{c}})\xi_{\mathrm{m}} + \hat{g}(\hat{\beta}_{\mathrm{c}})\Delta(w) \\ \dot{z} &= A_z z + b_z h^\top \xi_{\mathrm{m}} \\ (\dot{\hat{\theta}}_{\mathrm{c}}, \dot{\hat{\beta}}_{\mathrm{c}}) &= 0, \qquad (\dot{\tilde{\theta}}, \dot{\tilde{\beta}}) = 0, \end{aligned} \tag{42}$$

with $(w, \xi_{\mathrm{m}}, z, \hat{\theta}_{\mathrm{c}}, \hat{\beta}_{\mathrm{c}}, \tilde{\theta}, \tilde{\beta}) \in \mathcal{W} \times \mathbb{R}^{3n-2} \times \mathcal{R}_\epsilon \times \mathbb{R}^{2n-1} \times \mathbb{R}_{\geq \beta_0 - \beta}$, and the following constraint, derived from the right-hand side of $\dot{\tilde{y}}$ in (34):

$$\phi_{\mathrm{a}}^\top \tilde{\theta} + u\tilde{\beta} = 0. \tag{43}$$

From (42), (43), we derive the next convergence result.

Theorem 2. *Suppose that, for any solution of the system (42), there exist positive scalars δ, ϱ such that:*

$$\int_t^{t+\delta} \bar{\phi}(s)\bar{\phi}^\top(s)\mathrm{d}s \succeq \varrho I_{2n}, \quad \bar{\phi} \coloneqq \begin{bmatrix} \eta_{\mathrm{m}}^\top & z^\top & y_{\mathrm{r}} & \dot{y}_{\mathrm{r}} \end{bmatrix}^\top, \tag{44}$$

for all $t \in \mathbb{R}_{\geq 0}$, where $\eta_{\mathrm{m}} \coloneqq [I_{n-1} \ 0 \ 0]\xi_{\mathrm{m}}$. Then, every precompact solution of system (38) satisfies

$$\lim_{t+j \to \infty} (\tilde{\theta}(t, j), \tilde{\beta}(t, j)) = 0. \tag{45}$$

Additionally, if $\epsilon > 0$ in (37) is sufficiently small,

$$\lim_{t+j \to \infty} (\hat{\theta}_{\mathrm{c}}(t, j), \hat{\beta}_{\mathrm{c}}(t, j)) = (\theta, \beta), \tag{46}$$

and u converges to (30), thus the LQ problem is solved.

Remark 2. *Due to space limitations, we only briefly discuss how to obtain (44) with d. Similar to Lemma 2, the system*

$$\begin{bmatrix} \dot{\eta}_{\mathrm{m}} \\ \dot{z} \\ \dot{y}_{\mathrm{r}} \end{bmatrix} = \begin{bmatrix} \Lambda & 0 & \ell \\ 0 & A_z & b_z \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \eta_{\mathrm{m}} \\ z \\ y_{\mathrm{r}} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} v \tag{47}$$

is controllable. Therefore, by [1, §5.6.3], condition (44) holds if the input $v = \dot{y}_{\mathrm{r}}$ is sufficiently rich of order $2n$, e.g., if it consists of n sinusoids at distinct frequencies. From the proof of [15, Thm. 4.3], we obtain that the transfer function of system (20) has $n-1$ zeros determined by $\hat{\theta}_{\mathrm{c}}$ and $n-1$ stable pole-zero cancellations. Since the zero dynamics are invariant under state feedback, the transfer function from d to y_{r} has at most $n-1$ zeros on the imaginary axis, which become n considering $\dot{y}_{\mathrm{r}}$ as an output. To avoid any blocking effect, we then require that d be rich of order $3n$.

[1] $\overline{\mathrm{rge}\, \psi}$ denotes the closure of the range of the hybrid arc ψ.

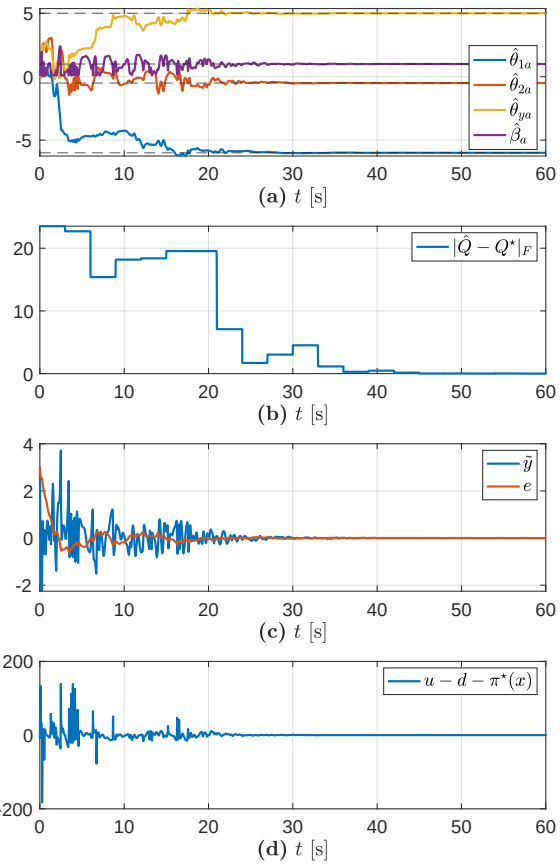

Fig. 2. Simulation run of MR-ARL applied to system (48). In the plot (a), the plant parameters are reported as dashed lines.

VI. Numerical Results

We test MR-ARL on a system having input-output model

$$G(s) = \frac{s+1}{(s-1)(s-2)}, \tag{48}$$

which we represent with a state-space realization in controller canonical form ($n = 2$). Our goal is to find the optimal controller associated with (48) and (2), with weight $r = 1$.

The tuning parameters are $\Lambda = -2$, $\ell = 2$, $\tau_s = 3$, $\lambda = k = \gamma = \epsilon = 1$, $\beta_0 = 0.1$, while we let $d(t) = 5(\sin(t) + \sin(5t) + \sin(10t))$, which is sufficiently rich of order $6 = 3n$.

Using the proof of Lemma 1, the "true parameters" of the non-minimal realization (7), used for analysis, are $\theta = (-6, -0.5, 5)$ and $\beta = 1$. In Fig. 2, we show a simulation run of MR-ARL with $\hat{\theta}_c(0,0) = (-2,1,2)$, $\hat{\beta}_c(0,0) = 3$, $\hat{\beta}_a(0,0) = 0.3$, $\hat{y}(0,0) = 3$, and all other states initialized to 0. In plot (a), we see that $(\hat{\theta}_a, \hat{\beta}_a)$ converges to (θ, β). In plot (b), we use the Frobenius norm $|\tilde{Q}|_F := \sqrt{\mathrm{Tr}(\tilde{Q}^\top \tilde{Q})}$ to show that $\hat{Q}$ converges to $Q^\star$. In plot (c), we provide the output estimation and mismatch errors, while in plot (d) we show that the applied policy converges to the optimal one.

VII. Conclusion

We proposed an algorithm for the LQ optimal control of SISO minimum-phase systems, combining an adaptive controller, a parameter-varying reference model, and an optimizer. For the resulting hybrid closed-loop dynamics, we established semi-global boundedness of the solutions and provided persistency of excitation conditions ensuring convergence to the optimal control policy. Future work will aim at extending the approach to systems with arbitrary relative degree and to multi-input multi-output systems.

References

[1] P. A. Ioannou and J. Sun, *Robust Adaptive Control*. New York, NY: Dover, 2012.

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 2018.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral reinforcement learning for continuous-time input-affine nonlinear systems with simultaneous invariant explorations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 916–932, 2014.

[12] L. Sforni, G. Carnevale, I. Notarnicola, and G. Notarstefano, "Stability-certified on-policy data-driven LQR via recursive learning and policy gradient," *arXiv preprint arXiv:2403.05367*, 2024.

[13] C. Possieri and M. Sassano, "Value iteration for continuous-time linear time-invariant systems," *IEEE Transactions on Automatic Control*, vol. 68, no. 5, pp. 3070–3077, 2022.

[14] M. Borghesi, A. Bosso, and G. Notarstefano, "MR-ARL: Model reference adaptive reinforcement learning for robustly stable on-policy data-driven LQR," *arXiv preprint arXiv:2402.14483*, 2024.

[15] K. S. Narendra and A. M. Annaswamy, *Stable Adaptive Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[16] A. M. Annaswamy, A. Guha, Y. Cui, S. Tang, P. A. Fisher, and J. E. Gaudio, "Integration of adaptive control and reinforcement learning for real-time control and learning," *IEEE Transactions on Automatic Control*, vol. 68, no. 12, pp. 7740–7755, 2023.

[17] S. A. A. Rizvi and Z. Lin, "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.

[18] C. Possieri, "Value iteration for linear quadratic optimal control of single-input single-output systems via output feedback," *IEEE Control Systems Letters*, 2024.

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

[20] D. G. Luenberger, "Observing the state of a linear system," *IEEE Transactions on Military Electronics*, vol. 8, no. 2, pp. 74–80, 1964.

[21] A. Isidori, *Lectures in Feedback Design for Multivariable Systems*. Switzerland: Springer International Publishing, 2017.

[22] A. Isidori, L. Marconi, and A. Serrani, *Robust Autonomous Guidance: an Internal Model Approach*. London: Springer-Verlag, 2003.

[23] A. Bosso, M. Borghesi, A. Iannelli, G. Notarstefano, and A. R. Teel, "Data-driven control of continuous-time LTI systems via non-minimal realizations," *arXiv preprint arXiv:2505.22505*, 2025.