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

An Observer-Based Reinforcement Learning Solution for Model-Following Problems

Mohammed I. Abouheaf[1], Kyriakos G. Vamvoudakis[2], Mohammad A. Mayyas[1], and Hashim A. Hashim[3]

***Abstract*— This paper introduces a novel model-free solution for a multi-objective model-following control problem, utilizing an observer-based adaptive learning approach. The goal is to regulate model-following error dynamics and optimize process variables simultaneously. Integral reinforcement learning is employed to adapt three key strategies, including observation, closed-loop stabilization, and reference trajectory tracking. Implementation uses an approximate projection estimation method under mild conditions on learning parameters.**

I. Introduction

Model-following techniques based on the optimal control framework have been used for trajectory-tracking control problems [1]. However, these solutions require offline solving of differential equations and full knowledge of process dynamics. Model Reference Adaptive Systems (MRAS) have been employed for real-time tracking control, but they have limitations such as dependence on process dynamics and lack of optimization for dynamic variables [2]. In this work, we propose a new model-free control architecture using an observer approach specifically designed for Linear Time Invariant (LTI) systems. Our approach, based on Integral Reinforcement Learning (IRL), ensures convergence under mild conditions on learning parameters.

The model-following applications involve various systems, including hypersonic aircraft, autonomous vehicles, underactuated systems, and robotic manipulators [3–6]. Several existing solutions exhibit the aforementioned limitations in model-following, such as dual mode predictive control [7], sliding mode surfaces [8], adaptive backstepping and $\mathcal{L}_1$ adaptive control [4], Lyapunov-based MRAS [9], model predictive control [5], and barrier function-based MRAS [3]. Graphical games have been used to address leader-follower control problems for LTI agents interacting through graph topologies [10–12], relying on pinning control to ensure synchronization. Model-based approaches using sum-of-squares polynomials [9], model-predicted control [13], and sliding surface-based observers [8] have been explored. Additionally, observer-based approaches [14] have been employed, but they lack model-free strategies.

This work was partially supported by the National Science Foundation under grant numbers S&AS-1849264, CPS-1851588, and CPS-2038589, and by the National Sciences and Engineering Research Council of Canada (NSERC) under grant number RGPIN-2022-04937.

[1]M. I. Abouheaf and M. A. Mayyas are with the Robotics Engineering, Bowling Green State University, Bowling Green, OH, 43403, USA, email: {mabouhe,mmayyas}@bgsu.edu. [2]K. G. Vamvoudakis is with the Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA, e-mail: kyriakos@gatech.edu. [3]H. A. Hashim is with the Department of Mechanical and Aerospace Engineering, Carleton University, Ottawa, Ontario, K1S-5B6, Canada, e-mail: Hhashim@carleton.ca.

Reinforcement Learning (RL) is a machine learning tool that utilizes temporal difference structures to find optimal strategies in dynamic learning environments [15–17]. This implies seeking rewards or penalties to maximize cumulative rewards. RL solutions are implemented using Value Iteration (VI) and Policy Iteration (PI) techniques [17–19], and parameter estimation approaches such as Recursive Least Squares (RLS) and Batch Least Squares (BLS) are employed to find underlying strategies [19,20]. In the continuous-time domain, the control setup results in the integral Bellman optimality equation [21,22], which is solved using IRL approaches. As the Bellman equation cannot be solved analytically, adaptive critic structures approximating RL solutions are used [23–27]. RL approaches have been applied to various problems such as Linear Quadratic Regulator (LQR) [28], model-following control [29], output-based regulation of multi-agent systems [30], and control of flexible wing aircraft [31].

Contributions: This work proposes a customized control structure to solve the model-following control problem. The structure consists of three model-free strategies, adept at regulating model-following tracking errors, observation errors, and optimizing closed-loop performance. The observer strategy, with its flexible-order error dynamics, offers advantages over low-order schemes and serves as an additional model-following structure to guide internal process dynamics. Moreover, unlike many existing approaches, this method optimizes both model-following error dynamics and closed-loop dynamic performance by solving the underlying LQR problem.

Mathematical notation: In this paper, $\mathbb{R}$ represents the set of real numbers, $\mathbb{Z}0^{+}$ denotes non-negative integers, and $\mathbb{N}$ stands for positive whole numbers. The Kronecker product is denoted by $\otimes$. The gradient of function $\mathcal{M}$ is referred to as $\nabla\mathcal{M}$. Let $\|\varkappa\|_\infty = \sup_{k\in\mathbb{N}}\|\varkappa(k)\|_\infty$ define the $\mathcal{L}_\infty-$ norm of a sequence $\{\varkappa(k)\}_{k=0}^{\infty}$ with $\mathcal{L}_2 \stackrel{\text{def}}{=} \{\varkappa : \|\varkappa\|_2 < \infty\}$ and $\mathcal{L}_\infty \stackrel{\text{def}}{=} \{\varkappa : \|\varkappa\|_\infty < \infty\}$.

Structure: The paper is structured as follows: Section II explains the overall control scheme consisting of three model-free strategies to achieve optimization goals. In Section III, the optimal control setup leading to the integral Bellman optimality equation is discussed, along with stability characteristics. Section IV presents the model-free IRL solution and its actor-critic implementation, introducing an approximate projection technique for stable adaptations of actor-critic weights. Section V validates the IRL solution using an unstable dynamic process and a nonlinear reference-trajectory. Finally, Section VI summarizes the main findings.

II. Problem Formulation

The model-following problem encounters challenges due to complex mathematical manipulations of reference-tracking error dynamics and the need for simultaneous optimization of other process variables. To address this, an observer-based strategy is presented to simplify the control scheme and enhance computational efficiency.

The process dynamics structure is described below.

$$\dot{\mathcal{X}} = \mathbf{A}\,\mathcal{X} + \mathbf{B}\,\mathbf{u} \quad \text{and} \quad Y = C\,\mathcal{X}, \tag{1}$$

where $\mathcal{X} \in \mathbb{R}^n$, $\mathbf{u} \in \mathbb{R}^m$, and $Y \in \mathbb{R}^p$ represent vectors of the states, input control signals, and output signals, respectively. Moreover, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are the dynamic parameters characterizing the process.

The process described by (1) is required to follow another dynamical model given by: $\dot{\hat{\mathcal{X}}} = \hat{\mathbf{A}}\,\hat{\mathcal{X}} + \hat{\mathbf{B}}\,(\mathbf{u}^{\pi_{\mathbf{Ob}}} + \mathbf{u})$, where $\hat{\mathcal{X}} \in \mathbb{R}^n$ is a vector of either the desired or observed states, $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ are parameters of the desired dynamical system or approximated process, and $\mathbf{u}^{\pi_{Ob}} \in \mathbb{R}^m$ is the control signal due to an observer strategy π_{Ob}.

Remark 1: The parameters $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ can represent the desired dynamical performance or an approximation of the process ($\mathbf{A}$ and $\mathbf{B}$). The objective is to regulate the observation errors without requiring partial or full knowledge of the process dynamics.

The reference dynamical behavior is determined by a command generator, represented as: $\mathbf{Y}^{ref}(t) = f(t)$, where $\mathbf{Y}^{ref}(t) \in \mathbb{R}^q$. The objective of the optimization problem is to let an output $\mathbf{Y}^s(t) \in \mathbb{R}^q$ of system (1) follow the reference-trajectory $\mathbf{Y}^{ref}(t) \in \mathbb{R}^q$ (i.e., $\lim_{t\to\infty} \|\mathbf{e}^{Mf}(t)\| \to \mathbf{0}$, $\mathbf{e}^{Mf}(t) = \mathbf{Y}^{ref}(t) - \mathbf{Y}^s(t)$). To clarify mathematical notation, any time-dependent function $g(t)$ will be denoted as g_t.

To address the challenges, we divide the overall control strategy into three sub-strategies: (i) $\mathbf{u}_t^{\pi_{Ob}} \in \mathbb{R}^m$ observes the process states, forming an additional model-following loop to compare the process outputs (1) with the desired or approximated outputs. (ii) $\mu_t^{\pi_{Cl}} \in \mathbb{R}^m$ optimizes the closed-loop performance of the dynamic system, and (iii) $\mathbf{u}_t^{\pi_{Mf}} \in \mathbb{R}^m$ reflects the model-following actions. These interactive strategies are implemented in a model-free fashion, resulting in the main control strategy $\mathbf{u}_t = \mu_t^{\pi_{Cl}} + \mathbf{u}_t^{\pi_{Mf}}$. The detailed control scheme will be explained in the following sections.

A. Observing the Desired Dynamic Performance

This strategy aims to find the desired states using an observer-like structure. Hence, the desired or approximated dynamic process is described by:

$$\dot{\hat{\mathcal{X}}} = \hat{\mathbf{A}}\,\hat{\mathcal{X}} + \hat{\mathbf{B}}\,(\mathbf{u}^{\pi_{Ob}} + \mathbf{u}) \;\text{ and }\; \hat{Y} = C\,\hat{\mathcal{X}}, \tag{2}$$

where $\hat{\mathcal{X}} \in \mathbb{R}^n$ and $\hat{Y} \in \mathbb{R}^p$ are vectors of the desired or observed states and output signals, respectively.

The observer control signal $\mathbf{u}^{\pi_{Ob}}$ relies on a flexible-order of tracking-error dynamics defined by a vector $\mathbf{E}^{Ob}$. The size of this vector varies according to the number of error samples $e_t^{Ob} = \mathbf{Y}_t - \hat{\mathbf{Y}}_t$ collected at fixed-time intervals δ, such that $\mathbf{E}_t^{Ob} = \begin{bmatrix} e_t^{Ob} & e_{t+\delta}^{Ob} & e_{t+2\delta}^{Ob} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{3p}$. The observer strategy π_{Ob} is selected using an adaptive learning mechanism, and the resulting control signal is given by $\mathbf{u}_{t+\delta}^{\pi_{Ob}} = \mathbf{u}_t^{\pi_{Ob}} + \mu_t^{\pi_{Ob}}$, where $\mu_t^{\pi_{Ob}} = \pi_{Ob}\,\mathbf{E}_t^{Ob}$, $\mu_t^{\pi_{Ob}} \in \mathbb{R}^m$. The strategy π_{Ob} is selected to minimize the performance index given by $J_t^{\pi_{\mathbf{Ob}}} = \int_t^\infty U_\tau^{Ob}\left(\mathbf{E}_\tau^{Ob}, \mu_\tau^{\pi_{Ob}}\right) d\tau$. The cost function U^{Ob} is used to minimize the observation errors.

Assumption 1: The dynamic system (2) defined by $(\hat{\mathbf{A}}, \mathbf{C})$ is observable and the process (1) is observable as well. □

B. Closed-Loop Strategy

The model-following strategy regulates the trajectory-tracking error dynamics while stabilizing and optimizing the process's performance. To achieve this, a closed-loop feedback strategy π_{Cl} is advised based on real-time observed states, ensuring stability of the closed-loop dynamical system, provided it is stabilizable. The objective function associated with this strategy is given by $J_t^{\pi_{\mathbf{Cl}}} = \int_t^\infty U_\tau^{Cl}\left(\hat{\mathcal{X}}_\tau, \mu_\tau^{\pi_{Cl}}\right) d\tau$, where U_t^{Cl} is a cost function. The resulting strategy takes the form of linear feedback: $\mu_t^{\pi_{Cl}} = \pi_{Cl}\,\hat{\mathcal{X}}_t^{Cl}$, $\mu_t^{\pi_{Cl}} \in \mathbb{R}^m$. This strategy solves the underlying Linear Quadratic Regulation (LQR) problem for the desired or observed system (2).

Assumption 2: There exists a stabilizing control strategy π_{Cl} that can stabilize the closed-loop dynamics of the desired or approximated process $\dot{\hat{\mathcal{X}}} = (\hat{\mathbf{A}} + \hat{\mathbf{B}}\,\pi_{Cl})\hat{\mathcal{X}}$. □

C. Online Model-Following Strategy

The model-following strategy aims to regulate the errors e_t^{Mf} between the selected outputs of the process $\mathbf{Y}_t^s$ and those of the reference system $\mathbf{Y}_t^{ref}$ (i.e., $e_t^{Mf} = \mathbf{Y}^{ref} - \mathbf{Y}_t^s$). Similar to the observer strategy, the model-following error samples are collected at a fixed-time interval δ, resulting in $\mathbf{E}_t^{Mf} = \begin{bmatrix} e_t^{Mf} & e_{t+\delta}^{Mf} & e_{t+2\delta}^{Mf} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{3q}$. Three error samples are considered for both the observer and model-following strategies. The model-following strategy π_{Mf} is determined online using the control law $\mathbf{u}_{t+\delta}^{\pi_{Mf}} = \mathbf{u}_t^{\pi_{Mf}} + \mu_t^{\pi_{Mf}}$, where $\mu_t^{\pi_{Mf}} = \pi_{Mf}\,\mathbf{E}_t^{Mf} \in \mathbb{R}^m$. The performance index to evaluate the quality of π_{Cl} is defined as $J_t^{\pi_{\mathbf{Mf}}} = \int_t^\infty U_\tau^{Mf}\left(\mathbf{E}_\tau^{Mf}, \mu_\tau^{\pi_{Mf}}\right) d\tau$, where U_t^{Mf} represents the model-following cost function.

Assumption 3: The strategies π_{Cl} and π_{Mf} stabilize the process around the desired reference-trajectory $\mathbf{Y}^{ref}$. □

D. Overall Control Solution

The control mechanism implies the existence of kernel solution structures that realize the interactive optimization goals of the sub-control problems (i.e., $\operatorname{argmin}_{\pi_{Ob}} J_t^{\pi_{\mathbf{Ob}}}$, $\operatorname{argmin}_{\pi_{Cl}} J_t^{\pi_{\mathbf{Cl}}}$, and $\operatorname{argmin}_{\pi_{Mf}} J_t^{\pi_{\mathbf{Mf}}}$). Since the process is an LTI system, the kernel solutions can take quadratic forms in the observer errors, observed states, and model-following errors. Assumptions 1, 2, and 3 are made to ensure the availability of such strategies that can stabilize the closed-loop dynamics and follow the desired reference-trajectory. Moreover, this solution form can also be attempted for nonlinear systems, given the data-driven structure of its

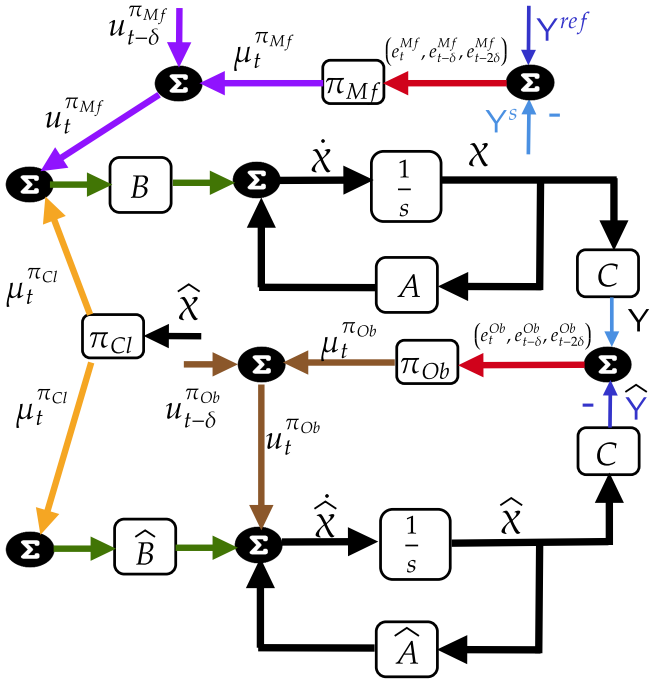


Fig. 1: The overall adaptive control scheme

proposed strategies. The overall control scheme of the model-following solution is depicted in Fig. 1.

III. Optimal Control Foundation

The aim of each sub-control problem is to minimize its respective cost function, given by $U_t^i\left(\mathcal{F}_t^i, \mu_t^{\pi_i}\right) = \frac{1}{2}\left(\mathcal{F}_t^{i\,T}\,\mathbf{Q}^i\,\mathcal{F}_t^i + \mu_t^{\pi_i\,T}\,\mathbf{R}^i\,\mu_t^{\pi_i}\right)$, where $i \in \{Ob, Cl, Mf\}$ refers to each sub-control problem, $\mathbf{Q}^i \in \mathbb{R}^{n\times n}$ and $\mathbf{R}^i \in \mathbb{R}^{m\times m}$ are weighting matrices. Each optimization problem is solved using a Hamiltonian structure, which is given by

$$H^i(\mathcal{F}_t^i, \lambda_t^{\pi_i}, \mu_t^{\pi_i}) = \lambda_t^{\pi_i\,T}\,\dot{\mathbf{Z}}_t^{\pi_i} + U_t^i\left(\mathcal{F}_t^i, \mu_t^{\pi_i}\right), \tag{3}$$

where $H^i, i \in \{Ob, Cl, Mf\}$ is a Hamiltonian function for each sub-control problem i, $\mathcal{F}_t^i \in \left\{\mathbf{E}_t^{Ob}, \hat{\mathcal{X}}_t^{Cl}, \mathbf{E}_t^{Mf}\right\}$, and $\lambda_t^{\pi_i} \in \mathbb{R}^{(n+m)}$ is a Lagrange multiplier associated with constraint $\dot{\mathbf{Z}}_t^{\pi_i}: \quad \mathbf{Z}^{\pi_i} = [\,\mathcal{F}_t^{i\,T}\ \mu_t^{\pi_i\,T}\,]^T \in \mathbb{R}^{(n+m)}$.

The following result demonstrates how the kernel solution forms can be selected.

Lemma 1: Let $V_t^i(\mathbf{Z}_t^{\pi_i}) > 0, V_t^i(0) = 0$ be a solving value function satisfying the Hamiltonian (3). Then, $V_t^i(\mathbf{Z}_t^{\pi_i})$ represents a Lyapunov Function.

Proof: The function V_t^i takes advantage of the LTI dynamic properties of the underlying sub-control problems. Therefore, its structure can be chosen as follows: $J_t^{\pi_i} \overset{\text{def}}{=} V_t^i(\mathbf{Z}_t^{\pi_i}) = \frac{1}{2}\mathbf{Z}_t^{\pi_i}\,\mathcal{S}^i\,\mathbf{Z}_t^{\pi_i\,T}$, where $\mathbf{0} < \mathcal{S}^i \equiv \begin{bmatrix}\mathcal{S}_{\mathcal{F}\mathcal{F}}^i & \mathcal{S}_{\mathcal{F}\mu^\pi}^i \\ \mathcal{S}_{\mu^\pi\mathcal{F}}^i & \mathcal{S}_{\mu^\pi\mu^\pi}^i\end{bmatrix} \in \mathbb{R}^{4\times 4}$, $\mathcal{S}_{\mathcal{F}\mathcal{F}}^i \in \mathbb{R}^{3\times 3}$, and $\mathcal{S}_{\mu^\pi\mu^\pi}^i \in \mathbb{R}$. This form represents a candidate Lyapunov function under the given assumptions. Furthermore, the Hamilton-Jacobi (HJ) theory establishes the relation between the value function V_t^i and the Lagrange multiplier $\lambda_t^{\pi_i}$ as follows: $\lambda_t^{\pi_i} = \nabla V_t^i = \partial V_t^i / \partial \mathbf{Z}_t^{\pi_i}$. Moreover, the solution of each underlying optimal sub-control problem yields a solution for its corresponding Bellman equation (i.e., $H^i(\mathcal{F}_t^i, \nabla V_t^i, \mu_t^{\pi_i}) = 0$) which implies that $\frac{\partial V_t^i}{\partial \mathbf{Z}_t^{\pi_i}}^T \dot{\mathbf{Z}}_t^{\pi_i} + U_t^i\left(\mathcal{F}_t^i, \mu_t^{\pi_i}\right) = 0$. This is an infinitesimal form of the Hamilton-Jacobi-Bellman (HJB) equation, given by: $\dot{V}_t^i + U_t^i\left(\mathcal{F}_t^i, \mu_t^{\pi_i}\right) = 0$. Since $\dot{V}_t^i \le 0$, then V_t^i is a Lyapunov function. ■

The realization of a model-free control strategy involves constructing a temporal difference structure, which can be adapted using various approximate dynamic programming forms, as explained in the following result.

Lemma 2: Let $V_t^{*i}(\mathbf{Z}_t^{\pi_{*i}}) > 0, V_t^{*i}(0) = 0$ represent the optimal solution of (3) following the optimal strategy π_{*i}. Thus, $V_t^{*i}(\mathbf{Z}_t^{\pi_{*i}})$ corresponds to the optimal solution of the integral Bellman optimality expression provided by:

$$V_t^{*i}(\mathbf{Z}_t^{\pi_{*i}}) = \int_t^{t+\delta} U_\tau^{*i}\left(\mathcal{F}_\tau^i, \mu_\tau^{\pi_{*i}}\right) d\tau + V_t^{*i}(\mathbf{Z}_{t+\delta}^{\pi_{*i}}). \tag{4}$$

Proof: The Hamiltonian $H^i(\mathcal{F}_t^i, \nabla V_t^i, \mu_t^{\pi_i}) = 0$ can be reformulated using Euler approximation as follows:

$$V_t^i(\mathbf{Z}_t^{\pi_i}) = \int_t^{t+\delta} U_\tau^i\left(\mathcal{F}_\tau^i, \mu_\tau^{\pi_i}\right) d\tau + V_t^i(\mathbf{Z}_{t+\delta}^{\pi_i}). \tag{5}$$

The optimal strategy takes on a linear form, which can be expressed as follows:

$$\mu_t^{\pi_{*i}} = -\mathcal{S}_{\mu^\pi\mu^\pi}^{i^{-1}}\,\mathcal{S}_{\mu^\pi\mathcal{F}}^i\,\mathcal{F}_t^i. \tag{6}$$

This strategy leads to an optimal function $V_t^{*i}(\mathbf{Z}_t^{\pi_{*i}})$, which solves (4) and the HJB equation $H^i(\mathcal{F}_t^i, \nabla V_t^{*i}, \mu_t^{\pi_{*i}}) = 0$. ■

The subsequent result indicates that the observer-based model-following strategy achieves asymptotic stabilization of both the observer and model-following errors. These tracking errors are denoted as $\mathbf{e}_t^i, i \in \{Ob, Mf\}$ for clarity.

Theorem 1: Let the initial values of functions $V_0^i(\mathbf{Z}_0^{\pi_i}), \forall i$ be bounded by upper values $\Upsilon^i, \forall i \in \{Ob, Mf\}$. Then, the trajectory-tracking dynamical error systems are asymptotically stable (i.e., $\lim_{t\to\infty}\|\mathbf{e}_t^i\| \to \mathbf{0}$).

Proof: The integral Bellman equation (5) yields a Lyapunov function, as per Lemma 1. Hence, $V_t^i(\mathbf{Z}_t^{\pi_i}) \le V_0^i(\mathbf{Z}_0^{\pi_i}) \le \Upsilon^i$ and $V_t^i(\mathbf{Z}_t^{\pi_i}) \in \mathcal{L}_\infty, \forall i \in \{Ob, Mf\}$. This, Assumption 1, and Assumption 3 reveal that the trajectory-tracking errors $\left\{e_t^i, e_{t+\delta}^i, e_{t+2\delta}^i\right\} \in \mathcal{L}_\infty$ and hence the stabilizing strategy, derived using the kernel solution $\mathcal{S}^i$, is $\pi_i \in \mathcal{L}_\infty$. The HJB equation $H^i(\mathcal{F}_t^i, \nabla V_t^i, \mu_t^{\pi_i}) = 0$ signifies that $\dot{V}_t^i = -U_t^i\left(\mathcal{F}_t^i, \mu_t^{\pi_i}\right) \le 0$. Therefore, $\dot{V}_t^i \in \mathcal{L}_\infty$ and $\dot{\mathbf{Z}}_t^{\pi_i} \in \mathcal{L}_\infty$. This HJB equation yields $\lim_{t\to\infty}\|V_t^i\| \to \mathbf{0}$ with $\int_0^t \frac{1}{2}\left(\mathcal{F}_\tau^{i\,T}\,\mathbf{Q}^i\,\mathcal{F}_\tau^i + \mu_\tau^{\pi_{*i}\,T}\,\mathbf{R}^i\,\mu_\tau^{\pi_{*i}}\right) d\tau \le V_0^i(\mathbf{Z}_0^{\pi_i})$. Then, $\int_0^t \frac{1}{2}\mathcal{F}_\tau^{i\,T}\left(\mathbf{Q}^i + \pi_{*i}^T\,\mathbf{R}^i\,\pi_{*i}\right)\mathcal{F}_\tau^i\,d\tau \le V_0^i(\mathbf{Z}_0^{\pi_i})$. This reveals that $\mathcal{F}_t^i \in \mathcal{L}_2$ and $\dot{V}_t^i \in \mathcal{L}_2$. Therefore, according to Barbalat's Lemma $\lim_{t\to\infty}\dot{V}_t^i \to \mathbf{0}$, implying asymptotic stabilization of the model-following and observation errors. ■

IV. IRL Solution Algorithm

The analytical solution of the coupled integral Bellman optimality equation (4) is challenging, requiring the use of

approximate learning mechanisms like RL. Consequently, a model-free IRL solution is developed to identify the optimal strategies to follow.

A. Integral Reinforcement Learning Algorithm

Algorithm 1 presents an online IRL solution, which solves the integral temporal difference equations (4) with the optimal strategies (6). This model-free approach utilizes error measurements e_t^{Mf} and e_t^{Ob} alongside observed states $\hat{\mathcal{X}}$. The algorithm employs an adaptive critic to approximate the underlying optimal strategy using the actor structure, while the critic assesses the quality of the attempted strategy.

Algorithm 1 Integral Reinforcement Learning Algorithm

1: Initialize the states $\mathcal{F}_0^i, \forall i$ and strategies $\pi_i, \forall i$.
2: Compute $\mathcal{S}^{i(r+1)}, \forall i$ by solving the equation:

$$V_t^{i(r+1)}(\mathbf{Z}_t^{\pi_{i(r)}}) - V_t^{i(r+1)}(\mathbf{Z}_{t+\delta}^{\pi_{i(r)}}) = \int_t^{t+\delta} U_\tau^{i(r)}\left(\mathcal{F}_\tau^i, \mu_\tau^{\pi_{i(r)}}\right) d\tau, \tag{7}$$

where r is an iterative index.
3: Update the strategy and determine the control signal

$$\mu_{t+\delta}^{\pi_i^0 0(r+1)} = -\mathcal{S}_{\mu^\pi\mu^\pi}^{i(r+1)^{-1}} \mathcal{S}_{\mu^\pi\mathcal{F}}^{i(r+1)} \mathcal{F}_{t+\delta}^i. \tag{8}$$

4: Terminate upon convergence of $\|\mathcal{S}^{i(r+1)} - \mathcal{S}^{i(r)}\|, \forall i$.

B. Actor-Critic Implementation

Real-time parameter estimation involves two steps. First, we use structures $\hat{\mathcal{S}}^i, \forall i$ to approximate matrices $\mathcal{S}^i, \forall i$ by solving the underlying integral Bellman optimality equations. Second, we approximate the optimal strategies $\pi_{*i}, \forall i$ in the form of $\hat{\pi}_i, \forall i$. The vector-indices of the process are defined as follows: $n = 3$ and $m = p = s = 1$. To approximate each function $V_t^i(\mathbf{Z}_t^{\pi_i})$, we use the following critic structure:

$$\hat{V}_t^i(\mathbf{Z}_t^{\hat{\pi}_i}) = \frac{1}{2} \mathbf{Z}_t^{\hat{\pi}_i T} \hat{\mathcal{S}}^i \mathbf{Z}_t^{\hat{\pi}_i}, \forall i \tag{9}$$

where $\hat{\pi}_i^T \in \mathbb{R}^3$ and $\mathbf{0} < \hat{\mathcal{S}}^{i\,T} \in \mathbb{R}^{4\times4}$ are the weights of the actor and critic structures, respectively. The Bellman optimality equation is written as $\tilde{V}_{t,t+\delta}^i(\mathbf{Z}_{t,t+\delta}^{\hat{\pi}_i}) = \int_t^{t+\delta} U_\tau^{*i}\left(\mathcal{F}_\tau^i, \mu_\tau^{\pi_{*i}}\right) d\tau, \forall i$ with $\tilde{V}_{t,t+\delta}^i(\mathbf{Z}_{t,t+\delta}^{\hat{\pi}_i}) = \hat{V}_t^i(\mathbf{Z}_t^{\hat{\pi}_i}) - \hat{V}_{t+\delta}^i(\mathbf{Z}_{t+\delta}^{\hat{\pi}_i}), \forall i$. This equation can be reformulated as follows:

$$\mathbf{\Theta}^i \tilde{\mathbf{Z}}_t^{\hat{\pi}_i} = \mathbf{\Phi}_t^i, \tag{10}$$

where $\tilde{\mathbf{Z}}_t^{\hat{\pi}_i} = \left\{\left(\mathbf{Z}_t^{\zeta_i} \bigotimes \mathbf{Z}_t^{\eta_i}\right), i \in \{Ob, Cl, Mf\}, \ \zeta_{Ob} = 1, \ldots, (\ell \times p + m), \ \zeta_{Cl} = 1, \ldots, (n \times p + m), \ \zeta_{Ob} = 1, \ldots, (\nu \times s + m), \ \eta_{Ob} = \zeta_{Ob}, \ldots, (\ell \times p + m), \ \eta_{Cl} = \zeta_{Cl}, \ldots, (n \times p + m), \ \eta_{Ob} = \zeta_{Ob}, \ldots, (\nu \times s + m)\}$, $\mathbf{\Theta}^i$ is a vector that is calculated by reshaping matrix $\frac{1}{2}\hat{\mathcal{S}}^i$ to associate its entries with $\tilde{\mathbf{Z}}_t^{\hat{\pi}_i}$, and $\mathbf{\Phi}_t^i = \int_t^{t+\delta} U_\tau^{*i}\left(\mathcal{F}_\tau^i, \mu_\tau^{\pi_{*i}}\right) d\tau$.

Similarly, the best strategy-to-follow (8) is approximated by an actor structure $\hat{\pi}_i$ such that

$$\hat{\pi}_i \mathcal{F}_t^i = \boldsymbol{\phi}_t^i \quad \text{and} \quad \boldsymbol{\phi}_t^i = -\hat{\mathcal{S}}_{\mu^{\hat{\pi}}\mu^{\hat{\pi}}}^{i^{-1}} \hat{\mathcal{S}}_{\mu^{\hat{\pi}}\mathcal{F}^i}^i \mathcal{F}_t^i. \tag{11}$$

The actor-critic weights will be tuned using projection-based parameter estimation approach. The following result outlines the convergence characteristics of the employed projection adaptation approach.

Theorem 2: Let the actor weights $\hat{\pi}_i$ and critic weights $\hat{\mathcal{S}}^i$ be calculated using Algorithm 1. Then,

a. The actor and critic weights converge to a set of weights $\hat{\pi}_i^*$ and $\hat{\mathcal{S}}^{*i}$, respectively.
b. The actor and critic weights' deviations from the optimal solution (i.e., $\hat{\pi}_i^*$ and $\mathbf{\Theta}^{i*}$) remain bounded, given mild conditions on the learning rates.

Proof: *a.* The tuning errors in the adapted critic and actor weights are optimized using the Hamiltonian functions $H_{\mathbf{\Theta}}^i$ and $H_{\hat{\pi}}^i$, respectively, as follows. $H_{\mathbf{\Theta}}^i(\mathbf{\Theta}^i, \lambda_{\mathbf{\Theta}}^i, f_{\mathbf{\Theta}}^i) = \frac{1}{2}\left(\mathbf{\Theta}^{i(r+1)} - \mathbf{\Theta}^{i(r)}\right)\left(\mathbf{\Theta}^{i(r+1)} - \mathbf{\Theta}^{i(r)}\right)^{\mathrm{T}} + \lambda_{\mathbf{\Theta}}^i f_{\mathbf{\Theta}}^i$ and $H_{\hat{\pi}}^i(\hat{\pi}_i, \lambda_{\hat{\pi}}^i, f_{\hat{\pi}}^i) = \frac{1}{2}\left(\hat{\pi}_i^{r+1} - \hat{\pi}_i^r\right)\left(\hat{\pi}_i^{r+1} - \hat{\pi}_i^r\right)^{\mathrm{T}} + \lambda_{\hat{\pi}}^i f_{\hat{\pi}}^i$, where $\lambda_{\mathbf{\Theta}}^i$ and $\lambda_{\hat{\pi}}^i$ are Lagrange multipliers associated with the optimization constraints $f_{\mathbf{\Theta}}^i = \mathbf{\Theta}^{i(r+1)}\tilde{\mathbf{Z}}_r^{\hat{\pi}_i} - \mathbf{\Phi}_r^i$ and $f_{\hat{\pi}}^i = \hat{\pi}_i^{r+1}\mathcal{F}_r^i - \boldsymbol{\phi}_r^i$, respectively. To determine the critic and actor adaptation laws, the Hamiltonian optimization conditions are applied as follows: $\frac{\partial H_{\mathbf{\Theta}}^i}{\partial \mathbf{\Theta}^{i(r+1)}} = 0$, $\frac{\partial H_{\mathbf{\Theta}}^i}{\lambda_{\mathbf{\Theta}}^i} = 0$, $\frac{\partial H_{\hat{\pi}}^i}{\hat{\pi}_i^{r+1}} = 0$, and $\frac{\partial H_{\hat{\pi}}^i}{\lambda_{\hat{\pi}}^i} = 0$. This yields $\left(\mathbf{\Theta}^{i(r+1)} - \mathbf{\Theta}^{i(r)}\right)^{\mathrm{T}} + \lambda_{\mathbf{\Theta}}^i \mathbf{Z}_t^{\hat{\pi}_i} = 0$, $f_{\mathbf{\Theta}}^i = 0$, $\left(\hat{\pi}_i^{r+1} - \hat{\pi}_i^r\right)^{\mathrm{T}} + \lambda_{\hat{\pi}}^i \mathcal{F}_t^i = 0$, and $f_{\hat{\pi}}^i = 0$, respectively. Further manipulation results in the following critic and actor update laws: $\mathbf{\Theta}^{i(r+1)} = \mathbf{\Theta}^{i(r)} - \frac{\tilde{\mathbf{Z}}_r^{\hat{\pi}_i\,\mathrm{T}}}{\tilde{\mathbf{Z}}_r^{\hat{\pi}_i\,\mathrm{T}}\tilde{\mathbf{Z}}_r^{\hat{\pi}_i}}\left(\mathbf{\Theta}^{i(r)}\tilde{\mathbf{Z}}_r^{\hat{\pi}_i} - \mathbf{\Phi}_r^i\right)$ and $\hat{\pi}_i^{r+1} = \hat{\pi}_i^r - \frac{\mathcal{F}_t^{i\,\mathrm{T}}}{\mathcal{F}_t^{i\,\mathrm{T}}\mathcal{F}_t^i}\left(\hat{\pi}_i^r\mathcal{F}_r^i - \boldsymbol{\phi}_r^i\right)$. The actor-critic adaptation forms can be modified while maintaining the overall optimization objectives by controlling the adaptation paces and addressing possible singularity issues, ensuring non-divergent behavior. The modifications can be incorporated as follows

$$\mathbf{\Theta}^{i(r+1)} = \mathbf{\Theta}^{i(r)} - \frac{\sigma_c^i \tilde{\mathbf{Z}}_r^{\hat{\pi}_i\mathrm{T}}}{\alpha_c^i + \tilde{\mathbf{Z}}_r^{\hat{\pi}_i\mathrm{T}}\tilde{\mathbf{Z}}_r^{\hat{\pi}_i}}\left(\mathbf{\Theta}^{i(r)}\tilde{\mathbf{Z}}_r^{\hat{\pi}_i} - \mathbf{\Phi}_r^i\right) \tag{12}$$

$$\hat{\pi}_i^{r+1} = \hat{\pi}_i^r - \frac{\sigma_c^i \mathcal{F}_t^{i\,\mathrm{T}}}{\alpha_a^i + \mathcal{F}_t^{i\,\mathrm{T}}\mathcal{F}_t^i}\left(\hat{\pi}_i^r\mathcal{F}_r^i - \boldsymbol{\phi}_r^i\right), \tag{13}$$

where $\sigma_c^i, \alpha_c^i, \sigma_a^i$, and $\alpha_a^i, \forall i \in \mathbb{R}$ are positive parameters.

Algorithm 1 and the stability results from Lemma 2 and Theorem 1 show that $\lim_{t\to\infty}\|\mathcal{F}_t^i\| \to \mathbf{0}$ and $\lim_{t\to\infty}\|\tilde{\mathbf{Z}}_t^{\hat{\pi}_i}\| \to \mathbf{0}$. Consequently, the weights $\mathbf{\Theta}^{i(r)}$ and $\hat{\pi}_i^r$ will converge to a solution comprising $\mathbf{\Theta}^{i*}$ and $\hat{\pi}_i^*$, respectively.

b. Let the adaptation errors in the adapted critic and actor weights be denoted as follows: $\mathbf{\Theta}_e^{i(r)} = \mathbf{\Theta}^{i*} - \mathbf{\Theta}^{i(r)}$ and $\hat{\pi}_{i_e}^r = \hat{\pi}_i^r - \hat{\pi}_i^*$, respectively. Then, (12) yields $\mathbf{\Theta}_e^{i(r+1)\,\mathrm{T}} = \mathbf{\Theta}_e^{i(r)\,\mathrm{T}} - \frac{\sigma_c^i \tilde{\mathbf{Z}}_r^{\hat{\pi}_i}\left(\mathbf{\Phi}_r^i - \left(-\mathbf{\Theta}_e^{i(r)} + \mathbf{\Theta}^{i*}\right)\tilde{\mathbf{Z}}_r^{\hat{\pi}_i}\right)^{\mathrm{T}}}{\alpha_c^i + \tilde{\mathbf{Z}}_r^{\hat{\pi}_i\,\mathrm{T}}\tilde{\mathbf{Z}}_r^{\hat{\pi}_i}}$ with $\mathbf{\Phi}_r^i - \mathbf{\Theta}^{i*}\tilde{\mathbf{Z}}_r^{\hat{\pi}_i} = 0$. Further, $\mathbf{\Theta}_e^{i(r+1)\,\mathrm{T}} = \mathbf{\Theta}_e^{i(r)\,\mathrm{T}} - \frac{\sigma_c^i \tilde{\mathbf{Z}}_r^{\hat{\pi}_i}\tilde{\mathbf{Z}}_r^{\hat{\pi}_i\,\mathrm{T}}}{\alpha_c^i + \tilde{\mathbf{Z}}_r^{\hat{\pi}_i\,\mathrm{T}}\tilde{\mathbf{Z}}_r^{\hat{\pi}_i}}\mathbf{\Theta}_e^{i(r)\,\mathrm{T}}$ or simply $\mathbf{\Theta}_e^{i(r+1)\,\mathrm{T}} = \boldsymbol{A}_c^i\,\mathbf{\Theta}_e^{i(r)\,\mathrm{T}}$, where $\boldsymbol{A}_c^i = \left(\mathrm{I}_c^i - \frac{\sigma_c^i \tilde{\mathbf{Z}}_r^{\hat{\pi}_i}\tilde{\mathbf{Z}}_r^{\hat{\pi}_i\,\mathrm{T}}}{\alpha_c^i + \tilde{\mathbf{Z}}_r^{\hat{\pi}_i\,\mathrm{T}}\tilde{\mathbf{Z}}_r^{\hat{\pi}_i}}\right)$ and

I_c^i is an identity matrix. To ensure bounded tuning of the critic weights, the parameters σ_c^i and α_c^i must be chosen such that $0 < \sigma_c^i < 2$ and $0 < \alpha_c^i$. Similarly, (13) yields $\hat{\pi}_{i\,e}^{r+1} = \boldsymbol{A}_a^i \hat{\pi}_{i\,e}^r$ with $\boldsymbol{A}_a^i = \left(I_a^i - \dfrac{\sigma_a^i \mathcal{F}_r^i \mathcal{F}_r^{i\,\mathrm{T}}}{\alpha_a^i + \mathcal{F}_r^{i\,\mathrm{T}} \mathcal{F}_r^i} \right)$, where I_a^i is an identity matrix. Hence, the conditions $0 < \sigma_a^i < 2$ and $0 < \alpha_a^i$ ensure convergence to solution $\hat{\pi}_i^*$. ■

V. Model-Following Validation Results

The model-free IRL solution is validated using a third-order dynamical process with the following parameters: $\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -5 & 10 \\ 0 & -1 & -5 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, and $\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T$. The dynamic parameters of the desired process are as follows: $\hat{\mathbf{A}} = \begin{bmatrix} 0.0132 & 1.0085 & -0.0055 \\ 0.0132 & -5.0286 & 9.9132 \\ -0.0526 & -1.0155 & -4.9374 \end{bmatrix}$ and $\hat{\mathbf{B}} = \begin{bmatrix} -0.0072 & -0.0547 & 1.0527 \end{bmatrix}^T$. The selected output of the process is represented as $\mathbf{Y}^s = \mathcal{X}(2)$. Additionally, the nonlinear reference trajectory Y_t^{ref} is expressed as:

$$Y_t^{ref} = \begin{cases} 1 + \exp^{(-0.01\,t)} \cos\left(\dfrac{1.5\,t}{20}\right), & \text{for } t \le 10 \\ 0.5\ (1 + \exp(-0.01\,(t-10))), & \text{for } 10 < t \le 20 \end{cases}$$

The model-following goal is to regulate the errors $Y_t^{ref} - \mathcal{X}(2)$ such that $\lim_{t\to\infty} \| Y_t^{ref} - \mathcal{X}(2) \| \to 0$. The remaining learning parameters are outlined in Table I. The simulation is performed using MATLAB software for a duration of 20 sec.

TABLE I: Learning and Adaptation Parameters

Parameter	Value	Parameter	Value
$\mathbf{Q}^i$	$0.05\,I_3$	$\mathbf{R}^i$	0.01
δ	0.01 sec	α_c^i	1.8
σ_c^i	0.5	α_a^i	1.8
σ_a^i	0.5		

Discussion: The desired response features an unstable process with a nonlinear reference trajectory, challenging the IRL solution's performance. Simulation results in Figure 2 illustrate closed-loop, observer, and model-following loops with solid, dotted, and dotted-dashed lines, respectively. Critic and actor weights are computed using (12) and (13), and adaptations of the actor weights are depicted in Fig. 2(a). After an exploration phase, the actor weights converge to solutions for the sub-control problems. Control signals $\mu_t^{\pi_{Ob}}$ and $\mu_t^{\pi_{Mf}}$ demonstrate the IRL solution's ability to regulate observation and model-following errors (Figs. 2(b) and 2(d)). The closed-loop control signal $\mu_t^{\pi_{Cl}}$ enables the process to follow the nonlinear reference trajectory and optimize closed-loop performance (Fig. 2(b)). The closed-loop strategy converges to $\pi_{Cl} = [-15.9517, -4.0410, -4.9822]$. Despite the unstable process and nonlinear reference trajectory, the observer and closed-loop strategies stabilize the dynamic process while following the desired reference-trajectory. The IRL solution successfully achieves the control goals simultaneously in a model-free manner, as highlighted in Fig. 2(c).

VI. Conclusion

An innovative model-free integral reinforcement learning approach is proposed to solve the model-following control problem. The control structure comprises three interactive strategies: one for regulating errors between actual and observed states, another for optimizing and stabilizing the closed-loop system, and a third for enabling the process to track a nonlinear reference trajectory. Our solution effectively optimizes the process's dynamic performance while precisely regulating model-following errors. Implementation involves using an approximate projection method to adapt the actor-critic weights for the sub-strategies, with careful management of learning parameters to ensure convergence.

References

[1] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*. John Wiley & Sons, 2012.

[2] K. J. Åström and B. Wittenmark, *Adaptive Control*. Courier Corporation, 2013.

[3] J. Liu, H. An, Y. Gao, C. Wang, and L. Wu, "Adaptive control of hypersonic flight vehicles with limited angle-of-attack," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 2, pp. 883–894, 2018.

[4] H. Chen, Y. Peng, D. Zhang, S. Xie, and H. Yan, "Dynamic positioning for underactuated surface vessel via l1 adaptive backstepping control," *Transactions of the Institute of Measurement and Control*, vol. 43, no. 2, pp. 355–370, 2021.

[5] M. Allenspach and G. J. J. Ducard, "Nonlinear model predictive control and guidance for a propeller-tilting hybrid unmanned air vehicle," *Automatica*, vol. 132, p. 109790, 2021.

[6] M. Abouheaf, D. Boase, W. Gueaieb, D. Spinello, and S. Al-Sharhan, "Real-time measurement-driven reinforcement learning control approach for uncertain nonlinear systems," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106029, 2023.

[7] M. Bagherzadeh, S. Savehshemshaki, and W. Lucia, "Guaranteed collision-free reference tracking in constrained multi unmanned vehicle systems," *IEEE Transactions on Automatic Control*, pp. 1–1, 2021.

[8] C. Wu, A. van der Schaft, and J. Chen, "Robust trajectory tracking for incrementally passive nonlinear systems," *Automatica*, vol. 107, pp. 595–599, 2019.

[9] J. Moore and R. Tedrake, "Adaptive control design for underactuated systems using sums-of-squares optimization," in *2014 American Control Conference*, 2014, pp. 721–728.

[10] M. Abouheaf, F. Lewis, M. Mahmoud, and D. Mikulski, "Discrete-time dynamic graphical games: Model-free reinforcement learning solution," *Control Theory and Technology*, vol. 13, no. 1, pp. 55–69, 2015.

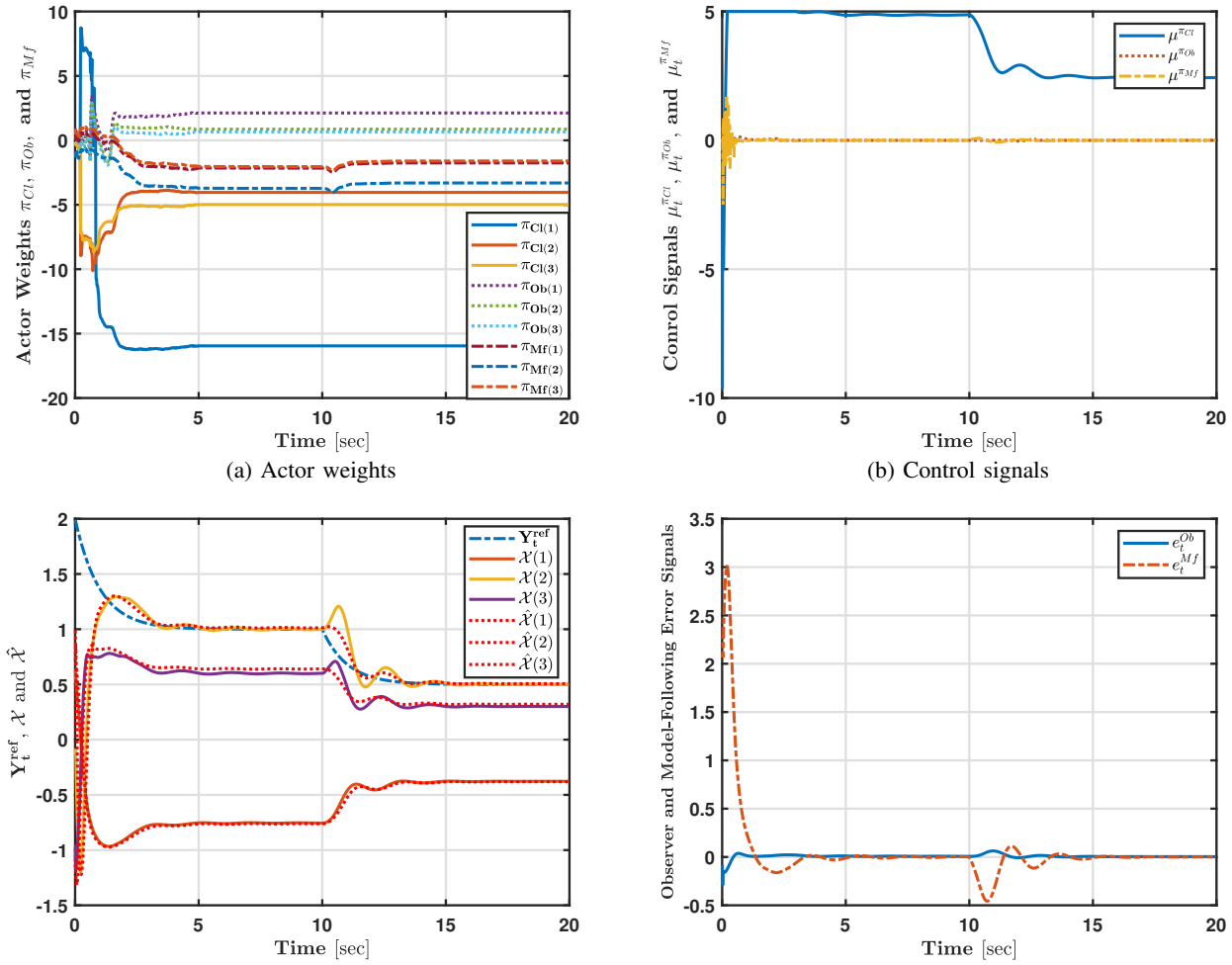
[11] M. Abouheaf and W. Gueaieb, "Multi-agent synchronization using online model-free action dependent dual heuristic dynamic programming approach," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 2195–2201.

[12] M. I. Abouheaf, F. L. Lewis, K. G. Vamvoudakis, S. Haesaert, and R. Babuska, "Multi-agent discrete-time graphical games and reinforcement learning solutions," *Automatica*, vol. 50, no. 12, pp. 3038–3053, 2014.

[13] X. Liu, L. Ma, X. Kong, and K. Y. Lee, "Robust model predictive iterative learning control for iteration-varying-reference batch processes," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 7, pp. 4238–4250, 2021.

[14] W. Li, X. Chu, C. Ma, and Y. Kong, "Finite-time model reference adaptive grasping control with fuzzy state observer for maglev grasping robot system," *IEEE/ASME Transactions on Mechatronics*, pp. 1–12, 2023.

[15] R. S. Sutton, A. G. Barto, and R. J. Williams, "Reinforcement learning is direct adaptive optimal control," *IEEE Control Systems Magazine*, vol. 12, no. 2, pp. 19–22, 1992.



(a) Actor weights

(b) Control signals

(c) Model-following performance

(d) Observation and model-following errors

Fig. 2: Performance of the adaptive control scheme

[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., ser. Second. Massachusetts: MIT Press, 1998.

[17] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*, 1st ed. Massachusetts: Athena Scientific, 1996.

[18] M. I. Abouheaf, M. S. Mahmoud, and F. L. Lewis, "Policy iteration solution for differential games with constrained control policies," in *2019 American Control Conference (ACC)*, 2019, pp. 4301–4306.

[19] L. Buşoniu, D. Ernst, B. De Schutter, and R. Babuška, "Online least-squares policy iteration for reinforcement learning control," in *Proceedings of the 2010 American Control Conference*, 2010, pp. 486–491.

[20] R. Srivastava, R. Lima, K. Das, and A. Maity, "Least square policy iteration for ibvs based dynamic target tracking," in *2019 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2019, pp. 1089–1098.

[21] K. G. Vamvoudakis, D. Vrabie, and F. L. Lewis, "Online adaptive algorithm for optimal control with integral reinforcement learning," *International Journal of Robust and Nonlinear Control*, vol. 24, no. 17, pp. 2686–2710, 2014.

[22] M. Abouheaf, W. Gueaieb, D. Spinello, and S. Al-Sharhan, "A data-driven model-reference adaptive control approach based on reinforcement learning," in *2021 IEEE International Symposium on Robotic and Sensors Environments (ROSE)*, 2021, pp. 1–7.

[23] S. Bhatnagar, R. S. Sutton, M. Ghavamzadeh, and M. Lee, "Natural actor–critic algorithms," *Automatica*, vol. 45, no. 11, pp. 2471–2482, 2009.

[24] B. Kiumarsi and F. L. Lewis, "Actor–critic-based optimal tracking for partially unknown nonlinear discrete-time systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 1, pp. 140–151, 2015.

[25] R. Song, F. Lewis, Q. Wei, H.-G. Zhang, Z.-P. Jiang, and D. Levine, "Multiple actor-critic structures for continuous-time optimal control using input-output data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 4, pp. 851–865, 2015.

[26] X. Zhao, S. Han, B. Tao, Z.-P. Yin, and H. Ding, "Model-based actor-critic learning of robotic impedance control in complex interactive environment," *IEEE Transactions on Industrial Electronics*, pp. 1–1, 2021.

[27] M. I. Abouheaf, H. A. Hashim, M. A. Mayyas, and K. G. Vamvoudakis, "An online model-following projection mechanism using reinforcement learning," *IEEE Transactions on Automatic Control*, pp. 1–8, 2023.

[28] B. Kiumarsi, F. L. Lewis, and Z.-P. Jiang, "h_∞ control of linear discrete-time systems: Off-policy reinforcement learning," *Automatica*, vol. 78, pp. 144–152, 2017.

[29] J. Xu, N. Lin, and R. Chi, "Improved high-order model free adaptive control," in *2021 IEEE 10th Data Driven Control and Learning Systems Conference (DDCLS)*, 2021, pp. 704–708.

[30] Y. Jiang, J. Fan, W. Gao, T. Chai, and F. L. Lewis, "Cooperative adaptive optimal output regulation of nonlinear discrete-time multi-agent systems," *Automatica*, vol. 121, p. 109149, 2020.

[31] M. Abouheaf, N. Q. Mailhot, W. Gueaieb, and D. Spinello, "Guidance mechanism for flexible-wing aircraft using measurement-interfaced machine-learning platform," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 7, pp. 4637–4648, 2020.