A Data-Driven Model-Reference Adaptive Control Approach Based on Reinforcement Learning

Mohammed Abouheaf[*], Wail Gueaieb[†], Davide Spinello[‡] and Salah Al-Sharhan[§]

[*]College of Technology, Architecture & Applied Engineering
Bowling Green State University, Bowling Green, 43402, OH, USA
Email: mabouhe@bgsu.edu
[†]School of Electrical Engineering and Computer Science
[‡]Department of Mechanical Engineering
University of Ottawa, Ottawa, Ontario, Canada K1N 6N5
Email: {wgueaieb, dspinell}@uottawa.ca
[§]Machine Intelligence Research Labs
Auburn, Washington 98071-2259, USA
Email: salah27@ieee.org

Abstract—Model-reference adaptive systems refer to a consortium of techniques that guide plants to track desired reference trajectories. Approaches based on theories like Lyapunov, sliding surfaces, and backstepping are typically employed to advise adaptive control strategies. The resulting solutions are often challenged by the complexity of the reference model and those of the derived control strategies. Additionally, the explicit dependence of the control strategies on the process dynamics and reference dynamical models may contribute in degrading their efficiency in the face of uncertain or unknown dynamics. A model-reference adaptive solution is developed here for autonomous systems where it solves the Hamilton-Jacobi-Bellman equation of an error-based structure. The proposed approach describes the process with an integral temporal difference equation and solves it using an integral reinforcement learning mechanism. This is done in real-time without knowing or employing the dynamics of either the process or reference model in the control strategies. A class of aircraft is adopted to validate the proposed technique.

Index Terms—Model-Reference Control, Integral Bellman Equation, Integral Reinforcement Learning, Adaptive Critics

I. INTRODUCTION

Optimal tracking control problems for linear time-invariant systems are typically solved using the Linear Quadratic Tracking Regulator (LQT) [1]. This is done offline via solving a set of differential equations backward-in-time. The guidance and tracking problems cover a wide range of autonomous and intelligent applications [2]. Some of which require sophisticated modeling to describe the dynamics, that when available allow to develop model-reference adaptive mechanisms that are robust and effective in real-world applications. Model-reference adaptive control employs theories in the class of sliding surfaces, backstepping, and Lyapunov-based methods to develop online adaptive strategies [3]–[5]. These methods, being based on systems' models, inherit their inherent complexity, uncertainties and approximations. Additionally, in networked multi-agent systems, the order of the underlying individual dynamics along with the network's connectivity may further raise the complexity of the problem.

Herein, a model-reference adaptive control (MRAC) approach, also known as model reference adaptive system (MRAS), is designed to guide dynamic systems using an online integral reinforcement learning (IRL) mechanism, where the stability is guaranteed for a broad range of slowly varying dynamics under standard conditions that apply to a large class of systems. The sums of squared error-based optimization approach is employed to design an adaptive control system in [6] for underactuated systems of moderate dimensions (less than 12). An MRAS is developed to stablize the lateral motion of a 5-DOF car-trailer system in [3], where the yaw rates of the car and trailer are assigned by using a Lyapunov stability criterion. An adaptive switch controller is developed to control a hydraulic actuator system in [7]. It coordinates among a PID control unit and MRAS to improve the tracking performance and hence the response time. In [8], a robust MRAS based on augmented error dynamics is developed where solving a set of linear matrix inequalities results in the error feedback matrix. This approach tackled the high oscillations in the actuator response due to the tuning with high gain adaptive rates. A decentralized MRAS is developed in [9] for large-scale interconnected systems, such as robotic manipulators. An adaptive control approach combined with anti-windup structure is developed for autonomous underwater vehicles (AUVs) in [10]. Another augmented MRAS with anti-windup commentator is applied to tackle the limitations caused by the actuator saturation for a class of aircraft in [11]. An adaptive control scheme is developed for uncertain hypersonic flight vehicles in [12]. A barrier function technique is employed to advise a novel handling of the angle-of-attack of this aircraft. An MRAC approach is employed to design an adaptive cruise system for vehicles in [13]. A combined approach that uses L_1 adaptive control and a backstepping method is employed to achieve fast adaptation and dynamic positioning of an underactuated surface vessel in [5]. It uses

a nonlinear observer to estimate the uncertainties and to guarantee a chattering-free performance. A leader-follower tracking problem is solved using distributed estimators under uncertain dynamical environment in [4]. The work used the sliding surface approach to coordinate among the different tasks. An adaptive mechanism is employed to trace the peaks of unknown fields for a mulit-agent system in [14]. Pinning control ideas are employed in [15]–[17] to leader-followers tracking problems with multi-agent systems under a fixed graph topology.

The approximation features of machine learning (ML) can be applied to solve adaptive control problems [18], [19]. Such approaches have been adopted in several applications ranging from the autonomous control of aircraft and underactuated sea vessels to regulating distributed generation sources [20]–[23]. Reinforcement learning (RL) provides the agent with a mechanism to learn the best control strategy-to-follow [19]. This is done according to multiple interactions between the agent and the environment to explore how to transit to a better state. The attempted policies are either penalized or rewarded in order to maximize the sum of a cumulative reward [19], [24]. The RL approaches are solved using two-step techniques like policy iteration (PI) and value iteration (VI) [24]–[26]. The Integral Reinforcement Learning approaches have been used to solve control problems with nonlinear-performance indices [27]. An adaptive critics is usually built as a neural network platform to approximate the strategy-to-follow using an actor neural network and the associated value which is modeled by a critic neural network [24]. These two-step approaches are employed to solve a class of cooperative control problems for multi-agent systems communicating over graphs [15], [28]. An adaptive Fuzzy-RL is adopted to control the flocking motion of a swarm of robots in [29]. The PI solutions are implemented using regression models like least squares and batch least squares [25], [30].

The main contribution of this work is the development of a generalized data-driven MRAC solution that is independent of the plant dynamics. It approximates an optimal control Hamiltonian structure to find an integral temporal difference expression. Then, it solves the underlying integral Bellman equation using IRL in real-time. The proposed solution is flexible in terms of the system order, unlike those associated with solving the optimal tracking problems of high-order systems [1], [31]. The algorithm runs online without explicitly solving the underlying optimal tracking problem. The remaining parts of the paper are arranged as follows: Section II formulates the model-reference adaptive problem. The duality between the optimal control and adaptive control ideas is explained in Section III. Section IV introduces the IRL implementation of the adaptive control solution using a Value Iteration approach. Section V shows the simulation results for an autonomous aircraft system. Finally, some concluding remarks are presented in Section VI.

II. PROBLEM FORMULATION

The MRAS decides on the control strategy-to-follow along the trajectory of the reference model dynamics [2]. This is done based on minimizing the error between the output $y(t) \in \mathbb{R}^P$ of the process and that of the reference model $y^{ref}(t) \in \mathbb{R}^P$. Herein, it is assumed that the dynamics of the process is linearizable into a state-space model in the form

$$\dot{X} = A\,X + B\,u \qquad\qquad y = C\,X \qquad (1)$$

where $X \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$ are the states and control input signals of the aircraft, respectively. The drift dynamics, control input, and output measurement matrices are denoted by the unknown matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{p\times n}$, respectively.

The objective of the optimization problem and hence the adaptive controller is to decide on the adaptive strategy π defined by three control gain matrices $\{K_0, K_\Delta, K_{2\Delta}\}$ of proper dimensions, with $u^\pi = K_0\,e(t) + K_\Delta\,e(t-\Delta) + K_{2\Delta}\,e(t-2\Delta)$ along the reference trajectory in order to minimize the error $e(t) = y(t) - y^{ref}(t)$; i.e., $\lim_{t\to\infty} \|e(t)\| = 0$. The symbol Δ refers to a sampling time interval. The adaptive control solution works in a different way when compared to the optimal control one. It is concerned with finding a control law to minimize the tracking errors in real-time without targeting the optimal control gains. In this work, a single output system is considered (i.e., $p = 1$). The solution provided herein combines an optimal control framework with an adaptive control mechanism that is based on IRL to solve the MRAC problem as shown in Fig. 1. It is worth to note that, the proposed adaptive strategy depends on only the reference signal $y^{ref}(t)$ and the process output $y(t)$. It does not rely on the state-space matrices.
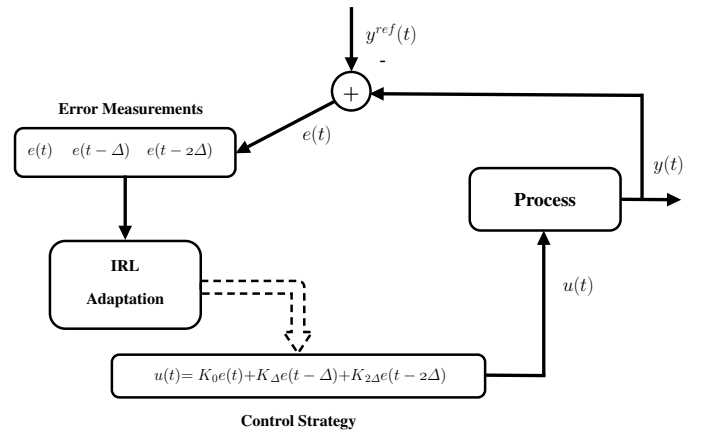

Fig. 1: Model-reference adaptive system

III. OPTIMAL AND ADAPTIVE CONTROL

This section discusses the optimal control framework and hence the underlying adaptive system. The structure of the control strategy depends on the history of error samples, i.e., t, $t - \Delta$, $t - 2\Delta$, etc. The width of the time window to consider is a design parameter which is left to the discretion of

the designer, which gives flexibility to the proposed approach. A wider time window may lead to a higher performance but at the expense of computational complexity. Herein, a time window of 3 samples is chosen, i.e., $t, t - \Delta, t - 2\Delta$, which yields a second-order error dynamic system. This is unlike many classical adaptive structures which are mostly based on zero- or first-order error strategies [1], [2], [11]. The error data are stacked together in a vector $\boldsymbol{E}(t) = [\, e(t) \ e(t - \Delta) \ e(t - 2\Delta)\,]^T \in \mathbb{R}^3$. A performance index $J^\pi(t)$ is used to measure the quality of the applied strategy $\boldsymbol{u}^\pi(t)$ by evaluating a cost or utility function U over the infinite horizon, such that

$$J^\pi(t) = \int_t^\infty U(\boldsymbol{E}(\tau), \boldsymbol{u}^\pi(\tau))\, d\tau,$$

where $U(\boldsymbol{E}(\tau), \boldsymbol{u}^\pi(\tau)) = \frac{1}{2}\left(\boldsymbol{E}^T(\tau)\, \boldsymbol{Q}\, \boldsymbol{E}(\tau) + \boldsymbol{u}^{\pi T}(\tau)\, \boldsymbol{R}\, \boldsymbol{u}^\pi(\tau)\right)$. The weighting matrices are denoted by $0 < \boldsymbol{Q} \in \mathbb{R}^{3\times3}$ (positive definite) and $0 < \boldsymbol{R} \in \mathbb{R}$.

The mathematical development of the adaptive control solution combines together ideas from optimal control, adaptive control, and IRL, as is detailed next.

Theorem 1. *Let $V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t))$ be a nonnegative multi-input single-output quadratic convex function which is nil only at the origin. Then,*

1) $V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t))$ satisfies a continuous-time Hamilton-Jacobi-Bellman (CTHJB) equation $H(\boldsymbol{E}(t), \nabla V(\boldsymbol{E}(t), \boldsymbol{u}^{\pi*}(t)), \boldsymbol{u}^{\pi*}(t)) = 0$.
2) $V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t))$ is a Lyapunov function.

   Proof:

1) The continuous-time Hamiltonian function for the optimal control problem is given by

$$H(\boldsymbol{E}(t), \lambda(t), \boldsymbol{u}^\pi(t)) = \lambda^T(t)\, \dot{\boldsymbol{Z}}^\pi(t) + U(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)),$$

where λ is a Lagrange multiplier associated with the constraint $\dot{\boldsymbol{Z}}^\pi(t) = 0$ with $\boldsymbol{Z}^\pi(t) = [\, \boldsymbol{E}^T(t) \ \boldsymbol{u}^\pi(t)^T \,]^T$. Let the value function V be quadratic and convex in $\boldsymbol{Z}^\pi(t)$, such that

$$V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)) = \frac{1}{2}\boldsymbol{Z}^{\pi T}(t)\, \boldsymbol{H}\, \boldsymbol{Z}^\pi(t), \qquad (2)$$

where $\boldsymbol{H} \equiv \begin{bmatrix} \boldsymbol{H_{EE}} & \boldsymbol{H_{Eu}} \\ \boldsymbol{H_{uE}} & \boldsymbol{H_{uu}} \end{bmatrix} \in \mathbb{R}^{4\times4} > \boldsymbol{0}$, $\boldsymbol{H_{uu}} \in \mathbb{R}$, and $\boldsymbol{H_{uE}} \in \mathbb{R}^{1\times3}$. The Hamilton-Jacobi theory provides a relation between Lagrange multiplier λ and the value function $V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t))$ [1], where $\lambda = \nabla V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)) = \partial V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t))/\partial \boldsymbol{Z}^\pi(t)$. Therefore, this value function satisfies the continuous-time Bellman equation given by

$$\nabla V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t))^T\, \dot{\boldsymbol{Z}}^\pi(t) + U(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)) = 0. \qquad (3)$$

This expression is an infinitesimal equivalent of $J^\pi(t) \stackrel{\text{def}}{=} V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)) = \int_t^\infty U(\boldsymbol{E}(\tau), \boldsymbol{u}^\pi(\tau))\, d\tau$. Hence, (3) is equivalent to

$$H(\boldsymbol{E}(t), \nabla V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)), \boldsymbol{u}^\pi(t)) =$$
$$\dot{V}(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)) + U(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)) = 0. \qquad (4)$$

Applying the Bellman stationarity condition yields the optimal strategy $u^{\pi*}(t)$ which is calculated as $u^{\pi*}(t) = \arg\min_{u^\pi(t)} H(\boldsymbol{E}(t), \nabla V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)), \boldsymbol{u}^\pi(t)))$. This yields the CTHJB

$$H(\boldsymbol{E}(t), \nabla V(\boldsymbol{E}(t), \boldsymbol{u}^{\pi*}(t)), \boldsymbol{u}^{\pi*}(t)) =$$
$$\dot{V}^*(\boldsymbol{E}(t), \boldsymbol{u}^{\pi*}(t)) + U(\boldsymbol{E}(t), \boldsymbol{u}^{\pi*}(t)) = 0, \qquad (5)$$

where V^* is the optimal solution of the CTHJB equation.
2) The value function V is a Lyapunov candidate function. Its derivative according to (4) is $\dot{V}(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)) = -U(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)) \leq 0$. Thus, $V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t))$ is a Lyapunov function.
   ∎

The following result explains how to apply Theorem 1 to come up with a temporal difference form (the integral Bellman equation) in order to adapt the control gains $\{K_0, K_\Delta, K_{2\Delta}\}$ in real-time.

Theorem 2. *The value function $V^*(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t))$ satisfies the following IRL-Bellman equation*

$$V^*(\boldsymbol{E}(t), \boldsymbol{u}^{\pi*}(t)) = \int_t^{t+\Delta} U(\boldsymbol{E}(\tau), \boldsymbol{u}^{\pi*}(\tau))\, d\tau$$
$$+ V^*(\boldsymbol{E}(t + \Delta), \boldsymbol{u}^{\pi*}(t + \Delta)) \qquad (6)$$

where the optimal control gains $\{K_0^*, K_\Delta^*, K_{2\Delta}^*\}$ are decided using Bellman optimality principles.

   Proof: Bellman equation (3) may be discretized using Euler approach, such that

$$\frac{V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)) - V(\boldsymbol{E}(t + \Delta), \boldsymbol{u}^\pi(t + \Delta))}{\Delta} = U(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t))$$

which can be rearranged to

$$V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)) = \int_t^{t+\Delta} U(\boldsymbol{E}(\tau), \boldsymbol{u}^\pi(\tau))\, d\tau$$
$$+ V(\boldsymbol{E}(t + \Delta), \boldsymbol{u}^\pi(t + \Delta)). \qquad (7)$$

This equation is referred to as the Integral Bellman equation, which is equivalent to (3). Applying Bellman optimality principle yields the optimal strategy $\boldsymbol{u}^{\pi*}(t) = \arg\min_{u^\pi(t)} V(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t))$. Using (2) yields

$$\boldsymbol{u}^{\pi*}(t) = -\boldsymbol{H_{uu}^{-1}}\boldsymbol{H_{uE}}\, \boldsymbol{E}(t). \qquad (8)$$

It is noted that this control strategy is model-free because it does not rely on the process dynamics. Hence, the optimal strategy $\pi*$ is defined by $\pi* = -\boldsymbol{H_{uu}^{-1}}\boldsymbol{H_{uE}}$ and the optimal control gains $\boldsymbol{K}^* = [\, K_0^* \ K_\Delta^* \ K_{2\Delta}^* \,] = -\boldsymbol{H_{uu}^{-1}}\boldsymbol{H_{uE}}$. Applying this optimal policy into the integral Bellman equation (7) yields the Integral Bellman optimality form (6). ∎

The next Lemma lays down the stability conditions of the suggested control law.

Lemma 1. *Let $V(\boldsymbol{E}(0), \boldsymbol{u}^\pi(0)) \leq \mathcal{M}$ for an upper bound $\mathcal{M}$. Then, $\lim_{t\to\infty} \dot{V}(\boldsymbol{E}(t), \boldsymbol{u}^\pi(t)) = 0$ and the error system is asymptotically stable (i.e., $\lim_{t\to\infty} e(t) = 0$) for a bounded reference signal $\boldsymbol{y}^{ref}(t)$.*

Proof: $V\left(E(t), u^{\pi}(t)\right)$ is a Lyapunov function. Therefore, $V\left(E(t), u^{\pi}(t)\right) \leq V\left(E(0), u^{\pi}(0)\right) \leq \mathcal{M}$. This implies that, $V\left(E(t), u^{\pi}(t)\right) \in L_{\infty}$. So, $e(t), e(t-\Delta), e(t-2\Delta) \in L_{\infty}$ and H (or equivalently K^*) $\in L_{\infty}$. Taking into consideration that $\int_0^t \dot{V}\left(E(\vartheta), u^{\pi}(\vartheta)\right) d\vartheta = V\left(E(t), u^{\pi}(t)\right) - V\left(E(0), u^{\pi}(0)\right)$ and that V is a Lyapunov function, we get $-\int_0^t \dot{V}\left(E(\vartheta), u^{\pi}(\vartheta)\right) d\vartheta \leq V\left(E(0), u^{\pi}(0)\right)$. This yields $\dot{V}\left(E(t), u^{\pi}(t)\right) \in L_{\infty}$, where the error dynamics $\dot{e}(t), \dot{e}(t-\Delta), \dot{e}(t-2\Delta) \in L_{\infty}$. Using the CTHJB equation (5) and the strategy (8) we infer that $\int_0^t E^T(\vartheta) \left(Q + K^{*T} R K^*\right) E(\vartheta) d\vartheta \leq V\left(E(0), u^{\pi}(0)\right)$, which reveals that $e(t), e(t-\Delta), e(t-2\Delta) \in L_2$ and $\dot{V}\left(E(t), u^{\pi}(t)\right) \in L_2$. Following Barbalat's lemma [2], $\lim_{t \to \infty} \dot{V}\left(E(t), u^{\pi}(t)\right) = 0$ and $\lim_{t \to \infty} e(t) = 0$, which proves the asymptotic stability of the system. ∎

This shows that the choice of the value function (2) and the model-free strategy (8) lead to an optimal solution for the Integral Bellman optimality equation (6).

IV. INTEGRAL REINFORCEMENT LEARNING SOLUTION

The control gains of the MRAC are selected using an online IRL approach that employs the error vector $E(t)$ measured in real-time. This is done by solving IRL-Bellman optimality equation (6) with the model-free optimal policy (8) simultaneously using a Value-Iteration method in real-time. The adaptive control solution requires a real-time adaptation mechanism to solve this system of equations. Therefore, means of adaptive critics are used to implement the online solution of the adaptive control problem. The next Theorem highlights the convergence results of the IRL-Value Iteration process.

Theorem 3. *Consider the value function* (2) *and the control policy* (8). *Then,*
1) *The value function update law* (6) *yields a non-decreasing positive definite sequence* $0 \leq V^0 \leq V^1 \leq V^2 \leq \cdots \leq V^*$, *which asymptotically converges to the solution of the IRL Bellman optimality equation* (6).
2) *The policies generated by* (8) *are stabilizing policies.*

Proof:
1) According to Lemma 1, the initial value function is bounded so that $0 < V^0\left(E(0), u^{\pi}(0)\right) \leq \mathcal{M}$ and the IRL Bellman expression (7) employing a strategy π yields $V^{r+1}\left(E(t), u^{\pi}(t)\right) = \int_t^{t+\Delta} U\left(E(\tau), u^{\pi}(\tau)\right) d\tau + V^r\left(E(t+\Delta), u^{\pi}(t+\Delta)\right)$, where r is a value iteration update index. Hence, $\forall r$, $V^{r+1}\left(E(t), u^{\pi}(t)\right) = \sum_{i=0}^r V^1\left(E(t+i\Delta), u^{\pi}(t+i\Delta)\right) - \sum_{i=1}^r V^0\left(E(t+i\Delta), u^{\pi}(t+i\Delta)\right)$. This leads to $V^{r+1}\left(E(t), u^{\pi}(t)\right) = \int_t^{t+r\Delta} U\left(E(\tau), u^{\pi}(\tau)\right) d\tau + V^0\left(E(t+(r+1)\Delta), u^{\pi}(t+(r+1)\Delta)\right)$. Therefore, the value sequence is non-decreasing, i.e., $0 \leq V^0 \leq V^1 \leq \cdots \leq V^r \leq V^{r+1}$, $\forall r$. According to Lemma 1, the error system is asymptotically stable and hence the cost is bounded so that $0 < \int_0^{\infty} U\left(E(\tau), u^{\pi}(\tau)\right) d\tau \leq \mathcal{U}$. The former sequence of value functions is upper bounded by $\mathcal{M} + \mathcal{U}$. Then, the value function sequence V^r satisfies

$0 \leq V^0 \leq V^1 \leq V^2 \leq \cdots \leq V^*$ and asymptotically converges to the optimal solution V^* of (6).
2) The optimal policies u^r, $\forall r, \Delta$, satisfy the following IRL-Bellman optimality equation $V^r\left(E(t), u^r(t)\right) - V^r\left(E(t+\Delta), u^r(t+\Delta)\right) = \int_t^{t+\Delta} U\left(E(\tau), u^r(\tau)\right) d\tau$. Using the stability result of Lemma 1 and employing the value function $V^r(\dots)$, or equivalently the policy u^r, over the infinite horizon, we get $V^r\left(E(t), u^r(t)\right) \geq V^r\left(E(t+\Delta), u^r(t+\Delta)\right) \geq V^r\left(E(t+2\Delta), u^r(t+2\Delta)\right) \geq \cdots \geq \lim_{\tau \to \infty} V^r\left(E(\tau), u^r(\tau)\right) = 0$. Thus, the strategies u^r, $\forall r$, are stabilizing and hence admissible. Therefore, $V^r\left(E(t+r\Delta), u^r(t+r\Delta)\right) \geq \cdots \geq V^2\left(E(t+2\Delta), u^2(t+2\Delta)\right) \geq V^1\left(E(t+\Delta), u^1(t+\Delta)\right) \geq V^0\left(E(t), u^0(t)\right)$. ∎

The adaptive critics approximates the value of some strategy π using a critic network and the associated optimal strategy using an actor network. The critic network is motivated by the structure of the value function V, so that

$$\hat{V}\left(E(t), \hat{u}^{\pi}(t)\right) = \frac{1}{2} I^{\pi T}(t) \Theta_c I^{\pi}(t),$$

where $I^{\pi}(t) = [\, E(t) \;\; \hat{u}^{\pi}(t) \,]^T$ and $\hat{u}^{\pi}(t)$ is the control signal using the approximated strategy. The adaptive critic adaptation weights are set as $\Theta_c = H > 0$.

Similarly, the actor network approximation is given by

$$\hat{u}^{\pi}(t) = \Theta_a E(t),$$

where $\Theta_a = [\, K_0 \;\; K_{\Delta} \;\; K_{2\Delta} \,]$ represents the actor adaptation weights.

In order to adapt the MRAC weights (i.e., the actor and critic weights), a gradient descent approach is considered. Thus, the critic and actor adaptation errors are calculated with the help of (6) and (8) which represent the optimal control solution. The critic adaptation error is given by $\varepsilon^{Critic} = \frac{1}{2}\left(\hat{V}\left(E(t), \hat{u}^{\pi}(t)\right) - \tilde{V}(t)\right)^2$ where $\tilde{V}(t) = U\left(E(t), \hat{u}^{\pi}(t)\right) + \hat{V}\left(E(t), \hat{u}^{\pi}(t)\right)$. Thus, the critic adaptation law follows

$$\Theta_c^{(r+1)} = \Theta_c^{(r)} - \zeta_c \, \varepsilon^{Critic(r)} I^{\pi} I^{\pi T}, \qquad (9)$$

where $0 < \zeta_c < 1$ is a critic-adaptation rate.

In a similar fashion, the actor adaptation rule follows a gradient descent approach where the actor adaptation error is given by $\varepsilon^{Actor} = \frac{1}{2}\left(\hat{u}^{\pi} - \tilde{u}\right)^2$ with $\tilde{u} = -\Theta_{c\hat{u}\hat{u}}^{-1} \Theta_{c\hat{u}E}$ which is motivated by the solution structure (8). Hence, the resulting update law for the actor weights is given by

$$\Theta_a^{(r+1)} = \Theta_a^{(r)} - \zeta_a \, \varepsilon^{Actor(r)} E^T \qquad (10)$$

where $0 < \zeta_a < 1$ is an actor-adaptation rate. The synthesis of the adaptive control solution is detailed in Algorithm 1.

It is worth to note that, this actor-critic adaptation mechanism provides an adaptive control solution framework that is based on an optimal control architecture. Further, it gives a flexibility to select the actor form to approximate the optimal strategy. It is different from the classical control structures in the way Lyapunov value functions are used. Herein, it is used

Algorithm 1 Model-Reference Adaptive Control Algorithm

Input:

    Number of adaption steps $\mathcal{N}$

    Adaptation step Δ

    Actor and critic adaptation rates ζ_a and ζ_c

    Weighting matrices Q and R

    Convergence threshold δ and a time window width $\mathcal{L}$

Output:

    Adapted gains $\Theta_a^{(t+k\Delta)}$ and $\Theta_c^{(t+k\Delta)}$, $k = 0, 1, \ldots, \mathcal{N}$

1: Initialize $E(0)$, $\Theta_a^{(0)}$ and $\Theta_c^{(0)}$ ▷ to zero, for example
2: Get the initial reference signal $y^{ref}(0)$
3: $k \leftarrow 0$
4: Weights_converged $\leftarrow$ false
5: **while** $k < \mathcal{N}$ **and** Weights_converged = false **do**
6:      Calculate the control signal $\hat{u}(k\Delta)$, then apply it to the process to get $y((k+1)\Delta)$
7:      Use previous step and $y^{ref}((k+1)\Delta)$ to calculate $e(t+(k+1)\Delta)$
8:      Find $V(E(k\Delta), \hat{u}(k\Delta))$ and $U(E(k\Delta), \hat{u}(k\Delta))$
9:      Get $E((k+1)\Delta)$ and $\hat{u}((k+1)\Delta)$, and deduce $V(E((k+1)\Delta, \hat{u}((k+1)\Delta))$
10:      Compute $\tilde{V}(k\Delta)$ and $\tilde{u}(k\Delta)$
11:      Update the critic weights $\Theta_c^{(k+1)}$ ▷ using (9)
12:      Update the actor weights $\Theta_a^{(k+1)}$ ▷ using (10)
13:      **if** $k > \mathcal{L}$ **and** $\left\| \Theta_c^{(k+1-l)} - \Theta_c^{(k-l)} \right\| \leq \delta, \; \forall l \in \{0, 1, \ldots, \mathcal{L}\}$, **then**
14:          $\Theta_c^{(*)} \leftarrow \Theta_c^{(k+1)}$
15:          Weights_converged $\leftarrow$ true
16:      **end if**
17:      $k \leftarrow k + 1$
18: **end while**
     return $\Theta_a^{(t+k\Delta)}$ and $\Theta_c^{(t+k\Delta)}$, for $k = 0, 1, \ldots, \mathcal{N}$

---

to find an integral temporal difference structure that can be solved using IRL. This results in a straightforward equation that is easy to manipulate and solve. On the other hand, unlike similar algorithms in the literature, the actor-critic adaption structures, represented by (9) and (10), do not depend on the process model. As a result, the proposed development provide easy-to-implement features of the IRL-based adaptive critics mechanism.

V. RESULTS

The adaptive model-free controller in Algorithm 1 is applied for the control of the longitudinal motion of an aircraft. The underlying longitudinal states of the aircraft (making the vector X in (1)) are the aircraft's angle of attack α and the pitch rate q. The control signal u is the elevator deflection, while the output y is the vertical acceleration. The control scheme does not require the a priori knowledge of the process dynamics. Nevertheless, to simulate the aircraft behavior, the following approximate linearized model is adopted [11]:

$$A = \begin{bmatrix} -8.76 & 0.954 \\ -177 & -9.92 \end{bmatrix} \quad B = \begin{bmatrix} -0.697 \\ -168 \end{bmatrix} \quad C = \begin{bmatrix} -0.8 & -0.04 \end{bmatrix}$$

The simulation and learning parameters of the MRAS are taken as $\Delta = 0.1\,\text{s}$, $\zeta_c = 0.5$, and $\zeta_a = 0.5$. The weighting matrices of the cost function are selected, such that $Q = I_{3\times3}$ (identity matrix) and $R = 1$. The remaining parameters are set to $\mathcal{N} = 180$, $\mathcal{L} = 10$, and $\delta = 10^{-8}$.

Two simulation cases are considered. The first employs a reference-model of a filtered square wave signal with an amplitude of $30\,\text{m/s}^2$. The outcome of this first test case is revealed in Fig. 2. As expected, the tracking error is asymptotically annihilated and the control gains are driven to converge (see Figs. 2(b) and 2(e)).

The second test case challenges the performance of the adaptive learning controller by introducing a dynamic disturbance to the aircraft, such that $\dot{X} = A X + B (\rho u + \Xi X)$, where

$$(\rho, \Xi) = \begin{cases} (0.8052, [\,-0.2760 \;\; -0.0858\,]), \text{ for } 1 \leq k < 60 \\ (0.5693, [\,-0.1959 \;\; -0.0028\,]), \text{ for } 60 \leq k < 120 \\ (0.4187, [\,-0.5324 \;\; -0.0002\,]), \text{ for } k \geq 120 \end{cases}$$

Furthermore, the controller is set to track a distorted reference signal compared to that used earlier. The results are demonstrated in Fig. 3. It is worth noticing how the error and the control gains converge in about 40 iterations (4 s) in both test cases.

VI. CONCLUSION

The paper introduced an online adaptive learning control approach to solve a class of model-reference adaptive problems. The adaptive mechanism solves an integral Bellman equation using a reinforcement learning process. The solution does not rely on any information about the aircraft or the reference dynamics while using only the available measurements. The relation between the optimal control solution and the underlying adaptive control system is explained. A Lyapunov approach is employed to show the stability features of the proposed adaptive learning control mechanism. The convergence properties of the reinforcement learning solution based on a Value Iteration process are highlighted and proven. The salient features of the proposed approach are demonstrated in the control of the longitudinal motion of an aircraft with unknown dynamics.

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] S. Vempaty, E. Lee, and Y. He, "Model-reference based adaptive control for enhancing lateral stability of car-trailer systems," in *ASME International Mechanical Engineering Congress and Exposition*, vol. 50664. American Society of Mechanical Engineers, 2016, p. V012T16A021.
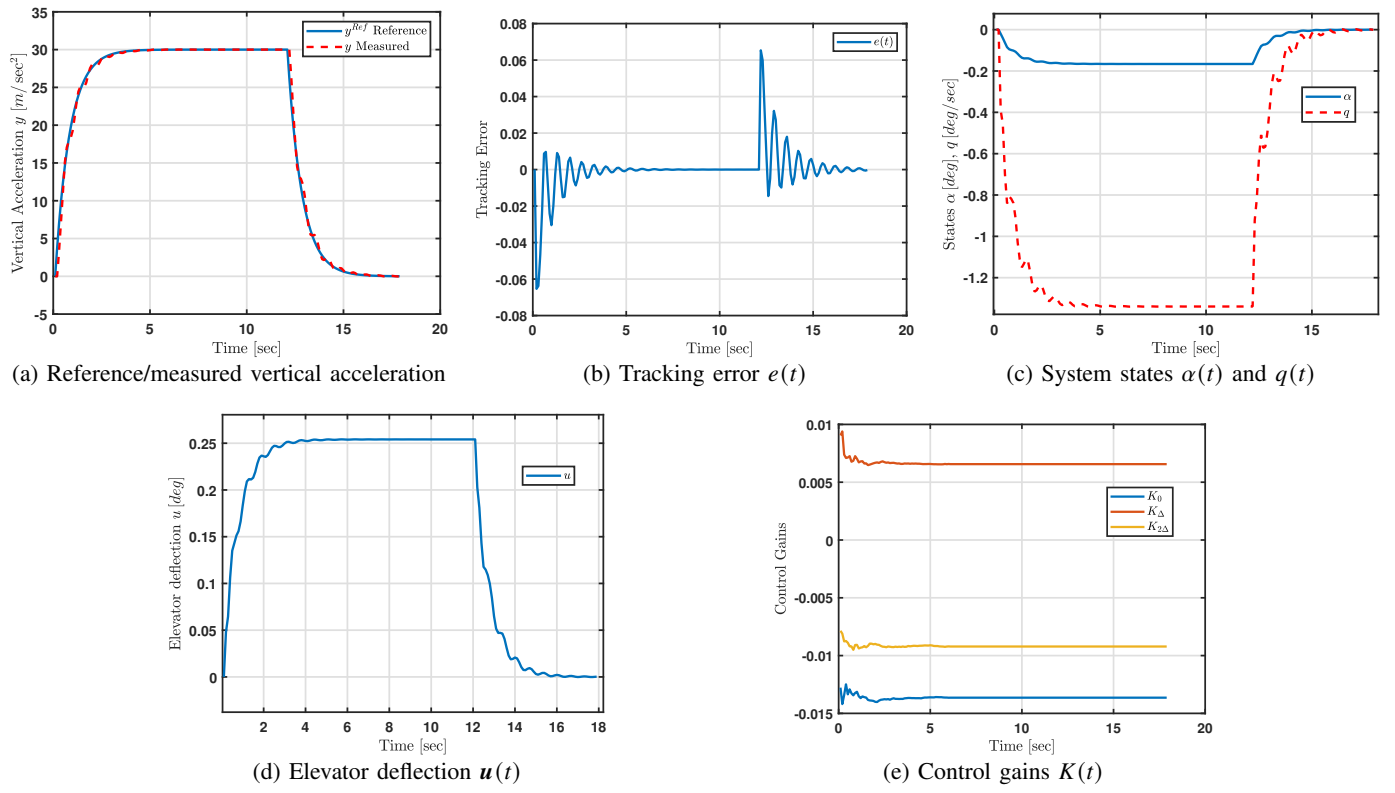
(a) Reference/measured vertical acceleration   (b) Tracking error $e(t)$ (c) System states $\alpha(t)$ and $q(t)$

(d) Elevator deflection $\boldsymbol{u}(t)$ (e) Control gains $K(t)$

Fig. 2: Simulation results of test case 1

(a) Reference/measured vertical acceleration   (b) Tracking error $e(t)$ (c) System states $\alpha(t)$ and $q(t)$

(d) Elevator deflection $\boldsymbol{u}(t)$ (e) Control gains $K(t)$
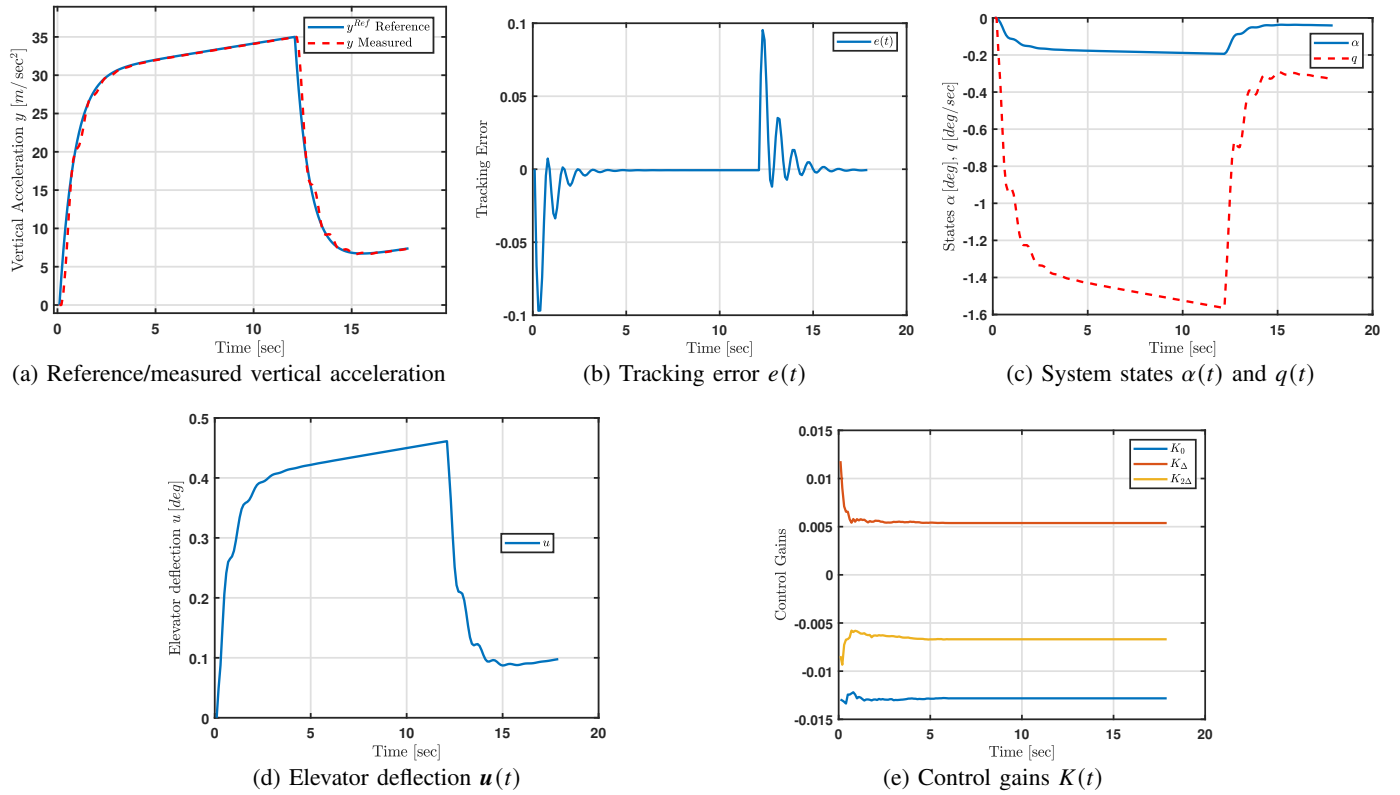
Fig. 3: Simulation results of test case 2

[4] J. Hu and G. Feng, "Distributed tracking control of leader–follower multi-agent systems under noisy measurement," *Automatica*, vol. 46, no. 8, pp. 1382–1387, 2010.

[5] 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. [Online]. Available: https://doi.org/10.1177/0142331220952960

[6] 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.

[7] X. Zuo, J.-w. Liu, X. Wang, and H.-q. Liang, "Adaptive pid and model reference adaptive control switch controller for nonlinear hydraulic actuator," *Mathematical Problems in Engineering*, vol. 2017, 2017.

[8] Z. Shi and L. Zhao, "Robust model reference adaptive control based on linear matrix inequality," *Aerospace Science and Technology*, vol. 66, pp. 152–159, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1270963816309683

[9] R. Ben Amor and S. Elloumi, "Decentralized model reference adaptive control for interconnected robotic systems," in *2017 18th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)*, 2017, pp. 235–240.

[10] P. Sarhadi, A. R. Noei, and A. Khosravi, "Model reference adaptive pid control with anti-windup compensator for an autonomous underwater vehicle," *Robotics and Autonomous Systems*, vol. 83, pp. 87–93, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889015303213

[11] J. Chen, J. Wang, and W. Wang, "Model reference adaptive control for a class of aircraft with actuator saturation," in *2018 37th Chinese Control Conference (CCC)*, 2018, pp. 2705–2710.

[12] 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.

[13] A. Abdullahi and S. Akkaya, "Adaptive cruise control: A model reference adaptive control approach," in *2020 24th International Conference on System Theory, Control and Computing (ICSTCC)*, 2020, pp. 904–908.

[14] M. Jadaliha, J. Lee, and J. Choi, "Adaptive control of multi-agent systems for finding peaks of unknown fields," in *Dynamic Systems and Control Conference*, vol. 44182, 2010, pp. 623–630.

[15] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0005109814004282

[16] M. I. Abouheaf and F. L. Lewis, "Multi-agent differential graphical games: Nash online adaptive learning solutions," in *52nd IEEE Conference on Decision and Control*, 2013, pp. 5803–5809.

[17] ——, "Approximate dynamic programming solutions of multi-agent graphical games using actor-critic network structures," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8.

[18] 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.

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

[20] 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.

[21] M. Abouheaf, W. Gueaieb, M. S. Miah, and D. Spinello, "Trajectory tracking of underactuated sea vessels with uncertain dynamics: An integral reinforcement learning approach," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 1866–1871.

[22] M. Abouheaf, S. Qu, W. Gueaieb, R. Abielmona, and M. Harb, "Responding to illegal activities along the canadian coastlines using reinforcement learning," *IEEE Instrumentation Measurement Magazine*, vol. 24, no. 2, pp. 118–126, 2021.

[23] M. Abouheaf, W. Gueaieb, and A. Sharaf, "Load frequency regulation for multi-area power system using integral reinforcement learning," *IET Generation, Transmission & Distribution*, vol. 13, no. 19, pp. 4311–4323, 2019. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-gtd.2019.0218

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

[25] 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.

[26] M. Abouheaf and M. Mahmoud, "Policy iteration and coupled riccati solutions for dynamic graphical games," *International Journal of Digital Signals and Smart Systems*, vol. 1, no. 2, pp. 143–162, 2017.

[27] M. Abouheaf and W. Gueaieb, "Model-free adaptive control approach using integral reinforcement learning," in *2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE)*, 2019, pp. 1–7.

[28] 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.

[29] S. Qu, M. Abouheaf, W. Gueaieb, and D. Spinello, "An adaptive fuzzy reinforcement learning cooperative approach for the autonomous control of flock systems," in *2021 International Conference on Robotics and Automation (ICRA)*, 2021.

[30] 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.

[31] B. Kiumarsi, F. L. Lewis, H. Modares, A. Karimpour, and M.-B. Naghibi-Sistani, "Reinforcement q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, vol. 50, no. 4, pp. 1167–1175, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0005109814000533