Online Actor Critic Learning for Optimal Tracking in Servo Positioning Systems

Hyochan Lee
CCS Graduate School of Mobility
Korea Advanced Institute of Science and Technology
Daejeon, Korea
hyochanlee@kaist.ac.kr

Kyunghwan Choi
CCS Graduate School of Mobility
Korea Advanced Institute of Science and Technology
Daejeon, Korea
kh.choi@kaist.ac.kr

***Abstract*—This paper proposes an online actor critic learning-based optimal tracking control method for output-feedback servo positioning systems under unknown external disturbances. The servo system is reformulated into a control-affine form, where the uncertain dynamics are compactly represented as a lumped unknown function. An online identifying filter is introduced to estimate these dynamics, while an actor-critic neural network structure is used to approximate the value function and optimal control input. The proposed method yields an approximate solution to the Hamilton–Jacobi–Bellman equation, with adaptive update laws ensuring asymptotic convergence of the Bellman residual error. Lyapunov-based analysis guarantees the stability of the closed-loop system. Simulation results confirm the effectiveness of the proposed method in achieving robust tracking under time-varying disturbances.**

***Index Terms*—Learning based control, Intelligent control, Reinforcement learning, Optimal control, Motor control**

I. Introduction

Servo positioning systems are critical in a wide range of industrial applications requiring high-precision motion control, including robotics, semiconductor manufacturing, and automatic assembly lines. These systems often encounter unmodeled dynamics, time-varying disturbances, and parameter uncertainties, which significantly challenge robust and accurate trajectory tracking. Therefore, the development of intelligent and adaptive control strategies capable of ensuring high-performance tracking under uncertain conditions remains an important research focus [1], [2].

Various nonlinear control methods have been proposed to address these challenges, including feedback linearization, backstepping, sliding mode control, and disturbance observer (DOB)-based techniques [3], [4]. Although effective under certain assumptions, many of these model-based approaches depend heavily on accurate knowledge of system dynamics or disturbance characteristics. Their performance may deteriorate in the presence of modeling errors or unknown time-varying disturbances. In recent years, data-driven control and reinforcement learning (RL)-based strategies have emerged as promising alternatives, offering the potential to learn control policies directly from system interactions without explicit modeling [5], [6].

This paper presents an online actor critic learning-based optimal tracking control method for output feedback servo positioning systems. The control design is formulated using a control-affine representation of the system, and an online identifying filter is utilized to estimate unknown nonlinear dynamics in real time. An actor-critic neural network architecture is adopted to approximate both the value function and optimal control policy without requiring full-state information or prior model identification. The convergence of the learning process and closed-loop stability are rigorously established through Lyapunov-based analysis. Simulation results demonstrate the effectiveness of the proposed approach compared to existing learning-based and disturbance rejection methods.

II. Preliminaries

A. Servo Model Dynamics

The servo drive primarily consists of two key components: the stator coil and the rotor. When a stator voltage v_s is applied, it induces a current i_s that produces a rotational torque defined by $T_e(i_s) = k_T i_s$, where $k_T > 0$ is the torque constant. This torque drives the rotor, resulting in angular speed ω and position θ, while acting against the back-electromotive force (back-EMF) generated by the stator coil, modeled as $b_F(\omega) = k_e \omega$, for some $k_e > 0$. The resulting electromechanical dynamics of the system from the stator input voltage v_s to the rotor position θ can be described by the following third-order nonlinear system:

$$\dot{\theta} = \omega, \qquad J\dot{\omega} = -B\omega + T_e(i_s) - T_L, \tag{1}$$

$$L\dot{i}_s = -Ri_s - b_F(\omega) + v_s, \quad \forall t \geq 0, \tag{2}$$

where T_L denotes the mismatched external disturbance torque arising from an uncertain load. In these equations, the positive constants J, B, L, R, k_T, k_e represent the rotor's moment of inertia and viscous friction coefficient, the stator coil's inductance and resistance, and the torque and back-EMF constants, respectively.

These physical parameters are typically affected by variations in voltage, current, temperature, and other environmental

*This work was supported in part by a Korea Research Institute for Defense Technology planning and advancement (KRIT) grant funded by Korea government DAPA (Defense Acquisition Program Administration) (No. KRIT-CT-22-087, Synthetic Battlefield Environment based Integrated Combat Training Platform) and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00554087) *(Corresponding author: Kyunghwan Choi)*

conditions. To account for such variations, each parameter is decomposed into a nominal part and an uncertain deviation component: $J = J_0 + \Delta J, B = B_0 + \Delta B, L = L_0 + \Delta L, R = R_0 + \Delta R, k_T = k_{T,0} + \Delta k_T, k_e = k_{e,0} + \Delta k_e$.

To derive the control-affine representation, the second equation in (2) is reformulated by incorporating the expressions for electromagnetic torque $T_e(i_s) = k_T i_s$ and back-EMF $b_F(\omega) = k_e\omega$. Substituting these relations into the system dynamics yields the angular acceleration $\dot{\omega}$ as follows:

$$\begin{aligned}\dot{\omega} &= -\frac{B}{J}\omega + \frac{k_T}{J}\left(-\frac{L}{R}\dot{i}_s - \frac{k_e}{R}\omega + \frac{1}{R}v_s\right) - \frac{T_L}{J} \\ &= -\frac{B}{J}\omega - \frac{k_T k_e}{JR}\omega - \frac{k_T L}{JR}\dot{i}_s + \frac{k_T}{JR}v_s - \frac{T_L}{J} \\ \Rightarrow \frac{k_T k_e}{JR}\omega &= -\dot{\omega} - \frac{B}{J}\omega - \frac{k_T L}{JR}\dot{i}_s + \frac{k_T}{JR}v_s - \frac{1}{J}T_L \\ \omega &= -\frac{JR}{k_T k_e}\dot{\omega} - \frac{BR}{k_T k_e}\omega - \frac{L}{k_e}\dot{i}_s + \frac{1}{k_e}v_s - \frac{R}{k_T k_e}T_L\end{aligned}$$

Based on the above derivation, the third-order model can be reduced to a second-order control-affine system of the form:

$$\dot{x} = g_0 u + f, \tag{3}$$

where $x = \theta$, $u = v_s$, and $g_0 = \frac{1}{k_{e,0}}$ is a known constant derived from the nominal back-EMF coefficient. The term f represents all model uncertainties and external disturbances, and its time derivative is assumed to be bounded as $|\dot{f}| \le \bar{f}$, where $\bar{f} > 0$ is a known positive constant. It can be expressed as

$$f = -\frac{JR}{k_T k_e}\dot{\omega} - \frac{BR}{k_T k_e}\omega - \frac{L}{k_e}\dot{i}_s - \frac{\Delta k_e}{k_{e,0}(k_{e,0} + \Delta k_e)}v_s - \frac{R}{k_T k_e}T_L.$$

B. Neural Network Approximation

For a continuous function $\phi(x) : \mathbb{R}^n \to \mathbb{R}$ defined on a compact set $\Omega \subset \mathbb{R}^n$, a neural network (NN) can approximate $f(x)$ using the following model:

$$\phi(x) = W^T \sigma(x), \tag{4}$$

where $W \in \mathbb{R}^{p\times 1}$ is the weight vector, p is the number of neurons, and $\sigma(x) = [\sigma_1(x), \ldots, \sigma_p(x)]^T$ is the basis function vector.

In this work, the radial basis function (RBF) is employed as the activation function:

$$\sigma_i(x) = \exp\left(-\frac{(x - m_i)^T (x - m_i)}{\mu_i^2}\right),$$

where $m_i \in \mathbb{R}^n$ is the center of the i-th receptive field and $\mu_i > 0$ is its width.

The NN approximation with bounded error is expressed as

$$\phi(x) = W^{*T}\sigma(x) + \varepsilon(x), \tag{5}$$

where W^* is the optimal weight that minimizes the supremum of the error over Ω:

$$W^* = \arg\min_{W\in\mathbb{R}^p}\left\{\sup_{x\in\Omega}\|\phi(x) - W^T\sigma(x)\|\right\}.$$

The approximation error $\varepsilon(x)$ can be made arbitrarily small by selecting a sufficiently large number of neurons p in accordance with the universal approximation theorem [7].

III. Proposed Control Methods

This section introduces the optimal tracking control statement based on the control-affine form (3), and proposes a actor critic learning-based solution to solve it.

A. Optimal Tracking Control Statement

Consider the control-affine system in (3), and define the tracking error as $e(t) = x(t) - r(t)$, where $r(t)$ is a desired reference trajectory. The tracking error dynamics can be written as

$$\dot{e} = g_0 u + f - \dot{r}, \tag{6}$$

To evaluate the tracking control performance of a given control input u, the following value function is defined as

$$V(e) = \int_t^\infty r(e(s), u(s))\, ds, \tag{7}$$

where the reward function is given by $\rho(e,u) = e^2 + g_0^2 u^2$.

The optimal value function $V^*(e)$ is obtained by minimizing the value function over admissible inputs:

$$V^*(e) = \min_{u\in\Psi(\Omega)}\left(\int_t^\infty \rho(e(s), u(s))\, ds\right) = \int_t^\infty \rho(e(s), u^*(s))\, ds \tag{8}$$

where $\Omega \subset \mathbb{R}$ is a compact set containing the origin, and $\Psi(\Omega)$ is the set of admissible control functions.

By taking the time derivative of the optimal value function $V^*(e)$ along the system trajectories, the Hamilton–Jacobi–Bellman (HJB) equation is obtained as

$$\begin{aligned}H\left(e, u^*, \frac{dV^*}{de}\right) &= e^2 + g_0^2 u^{*2} \\ &\quad + \frac{dV^*}{de}(g_0 u^* + f - \dot{r}) = 0,\end{aligned} \tag{9}$$

where $H(\cdot)$ denotes the Hamiltonian. The optimal control input u^* is obtained by minimizing the Hamiltonian with respect to u^*, which leads to the following first-order optimality condition:

$$\frac{\partial H}{\partial u^*} = 0. \tag{10}$$

Solving this condition (10) gives

$$u^* = -\frac{1}{2g_0}\frac{dV^*}{de}. \tag{11}$$

The derivative of the optimal value function can be expressed as

$$\frac{dV^*}{de} = 2\gamma_e e + 2f + V_0(e,x), \tag{12}$$

where $V_0(e,x)$ is defined as

$$V_0(e,x) = \frac{dV^*}{de} - 2\gamma_e e - 2f.$$

Substituting this expression into the optimal control law yields

$$u^* = -\frac{1}{g_0}\left(\gamma_e e + f + \frac{1}{2}V_0\right). \tag{13}$$

Since the nonlinear function f and the optimal value function $V^*(e)$ are unknown, obtaining an analytical solution to the HJB equation becomes nearly impossible. To overcome this challenge, a actor critic learning framework combined with online identifying filter techniques is proposed to approximate the solution.

B. Actor-Critic based Optimal Control Solution with Online Identifying Filter

A primary difficulty in solving the HJB equation is the estimation of the optimal value function $V^*(e)$ or its derivative $\frac{dV^*}{de}$. To address this issue, the term $V_0(e,x)$, which appears in the structure of the value function derivative, is approximated by a neural network (NN) model as

$$V_0(e,x) = W_V^{*T}\sigma_V(e,x) + \varepsilon_V(e,x), \tag{14}$$

where W_V^* is the unknown ideal weight vector, $\sigma_V(e,x)$ is a designed basis function vector, and $\varepsilon_V(e,x)$ represents the NN approximation error.

Assumption 1: The ideal weight vector W_V^* of the neural network approximation is assumed to be bounded as follows:

$$\|W_V^*\| \le \bar{W}_V,$$

where $\bar{W}_V > 0$ is a positive constant. ◊

Substituting this approximation into the expressions for $\frac{dV^*}{de}$ and u^* yields

$$\frac{dV^*}{de} = 2\gamma_e e + 2f + W_V^{*T}\sigma_V + \varepsilon_V, \tag{15}$$

$$u^* = -\frac{1}{g_0}\left(\gamma_e e + f + \frac{1}{2}(W_V^{*T}\sigma_V + \varepsilon_V)\right). \tag{16}$$

However, the ideal weight W_V^* and the nonlinear function f are still unknown. To address this problem, an online actor-critic neural network framework is introduced to approximate the ideal weight W_V^*, and an identifying filter is employed to compensate the unknown dynamics f.

The online identifying filter is constructed to provide an online estimate of the unknown dynamics f, which accounts for the model uncertainties and external disturbances such as load torque. The filter dynamics are described as follows:

$$\dot{\hat{x}} = g_0 u + \hat{f} + \gamma_{f,1}\tilde{x} \tag{17}$$

$$\hat{f} = \zeta + \gamma_{f,2}\tilde{x} \tag{18}$$

$$\dot{\zeta} = -\gamma_{f,2}\zeta - \gamma_{f,2}^2\tilde{x} + \gamma_{f,2}\left(\hat{f} + \gamma_{f,1}\tilde{x}\right) \tag{19}$$

where $\tilde{x} = x - \hat{x}$ is the output state filtering error, and $\gamma_{f,1}, \gamma_{f,2} > 0$ are filter learning rates.

Remark 1: The estimated dynamics $\hat{f}$ generated by the identifying filter is incorporated into the control design as a feedforward term to actively compensate for the unknown disturbances. This approach aligns with the concept of active disturbance rejection control. The convergence and stability properties of the proposed filter-based estimation scheme are analyzed in Section IV. ◊

The critic NN is designed to evaluate the control performance by approximating the derivative of the value function. It is formulated as

$$\frac{d\hat{V}^*}{de} = 2\gamma_e e + \hat{f} + \hat{W}_c^T\sigma_V(e,x), \tag{20}$$

where $\hat{W}_c$ is the weight vector of the critic NN and $\sigma_V(e,x)$ denotes the basis function vector. The weight update law for the critic is given by

$$\dot{\hat{W}}_c = -\gamma_c\sigma_V(e,x)\sigma_V^T(e,x)\hat{W}_c, \tag{21}$$

where $\gamma_c > 0$ is a learning rate of the critic NN.

The actor NN is designed to approximate the $W_V^{*T}\sigma_V$ term, which is used in the actual control input based on the critic NN evaluation. It is defined as

$$\hat{u}^* = -\frac{1}{g_0}\left(\gamma_e e + \hat{f} + \frac{1}{2}\hat{W}_a^T\sigma_V(e,x)\right), \tag{22}$$

where $\hat{W}_a$ is the actor weight vector. The weight update law for the actor is given by

$$\dot{\hat{W}}_a = -\sigma_V(e,x)\sigma_V^T(e,x)\left(\gamma_a\left(\hat{W}_a - \hat{W}_c\right) + \gamma_c\hat{W}_c\right), \tag{23}$$

where $\gamma_a > 0$ is a learning rate of the actor NN. Based on (20) and (22), the approximated HJB equation can be expressed as follows:

$$\begin{aligned}\hat{H}\left(e,\hat{u}^*,\frac{d\hat{V}^*}{de}\right) &= e^2 + g_0^2\hat{u}^{*2} + \frac{d\hat{V}^*}{de}\left(g_0\hat{u}^* + f - \dot{r}\right)\\ &= e^2 + \left(\gamma_e e + \hat{f} + \frac{1}{2}\hat{W}_a^T\sigma_V(e,x)\right)^2\\ &\quad + \left(2\gamma_e e + \hat{f} + \hat{W}_c^T\sigma_V(e,x)\right)\times\\ &\quad \left(f - \gamma_e e - \hat{f} - \frac{1}{2}\hat{W}_a^T\sigma_V(e,x) - \dot{r}\right).\end{aligned}$$

And the Bellman residual error for the approximated Hamiltonian is defined as

$$\delta_B = \hat{H}\left(e,\hat{u}^*,\frac{d\hat{V}^*}{de}\right) - H\left(e,u^*,\frac{dV^*}{de}\right) = \hat{H}\left(e,\hat{u}^*,\frac{d\hat{V}^*}{de}\right), \tag{24}$$

The solution $\hat{u}^*$ minimizes the approximated Hamiltonian $\hat{H}\left(e,\hat{u}^*,\frac{d\hat{V}^*}{de}\right)$. In particular, if the condition $\hat{H}\left(e,\hat{u}^*,\frac{d\hat{V}^*}{de}\right) = 0$ is satisfied, then $\hat{u}^*$ is the unique solution that also satisfies the true Hamiltonian equation $H\left(e,u^*,\frac{dV^*}{de}\right) = 0$. Fig. 1 illustrates the overall structure of the proposed control system.

Remark 2: The NN update laws given in (21) and (23) are designed to asymptotically drive the approximated Hamiltonian $\hat{H}\left(e,\hat{u}^*,\frac{d\hat{V}^*}{de}\right)$ toward zero. This facilitates convergence of the actor and critic networks to the optimal control policy. The closed-loop stability and convergence properties of this learning mechanism are analyzed in Section IV. ◊

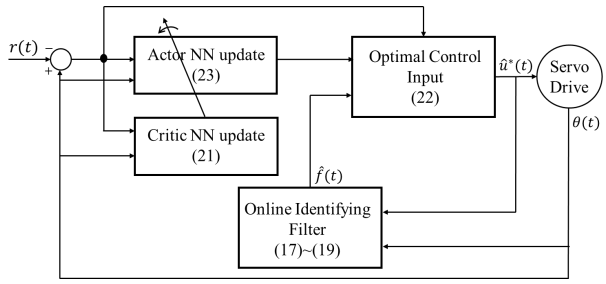

Fig. 1: Proposed online actor critic learning based optimal control solution

IV. Closed-loop Analysis

This section presents the stability and convergence analysis of the proposed control framework presented in Section III. A Lyapunov-based approach is employed to verify the boundedness of the closed-loop system and the asymptotic behavior of the learning dynamics under the actor-critic structure with the identifying filter.

Lemma 1: The actor and critic NN update laws defined in (21) and (23) guarantee that the Bellman residual error δ_B converges to 0 asymptotically as $t \to \infty$. ♢

Proof: According to the optimality condition, the optimized solution $\hat{u}^*$ satisfies the Bellman residual error condition $\delta_B = 0$. When this condition holds and the solution is unique, it is equivalent to the following first-order condition:

$$\frac{\partial \hat{H}(e,\hat{u}^*,\frac{d\hat{V}^*}{de})}{\partial \hat{W}_a} = \frac{1}{2}\sigma_V(e,x)\sigma_V^T(e,x)\left(\hat{W}_a - \hat{W}_c\right) = 0. \quad (25)$$

To analyze the satisfaction of this optimality condition, consider the following positive definite function:

$$P = \left(\hat{W}_a - \hat{W}_c\right)^T \left(\hat{W}_a - \hat{W}_c\right), \quad (26)$$

which plays the role of a Lyapunov-like candidate. If $P = 0$, then the optimality condition is satisfied, implying convergence of the actor and critic neural network weights.

Based on the neural network update laws in (21) and (23), the time derivative of P is computed as

$$\frac{dP}{dt} = \frac{dP}{d\hat{W}_c}\dot{\hat{W}}_c + \frac{dP}{d\hat{W}_a}\dot{\hat{W}}_a. \quad (27)$$

Using the identity $\frac{dP}{d\hat{W}_a} = -\frac{dP}{d\hat{W}_c} = 2(\hat{W}_a - \hat{W}_c)$ and substituting the update laws yields

$$\begin{aligned}\frac{dP}{dt} &= -2\gamma_c(\hat{W}_a - \hat{W}_c)^T\sigma(x,e)\sigma^T(x,e)\hat{W}_c \\ &\quad - 2(\hat{W}_a - \hat{W}_c)^T\sigma(x,e)\sigma^T(x,e)\left(\gamma_a(\hat{W}_a - \hat{W}_c) + \gamma_c\hat{W}_c\right) \\ &= -2(\hat{W}_a - \hat{W}_c)^T\sigma(x,e)\sigma^T(x,e)\left(\gamma_a(\hat{W}_a - \hat{W}_c)\right).\end{aligned}$$

This expression can be rearranged as

$$\frac{dP}{dt} = -2\gamma_a(\hat{W}_a - \hat{W}_c)^T\sigma(x,e)\sigma^T(x,e)(\hat{W}_a - \hat{W}_c) \le 0. \quad (28)$$

Since $P \ge 0$ and $\dot{P} \le 0$, it follows that $P \to 0$ as $t \to \infty$, ensuring that the optimality condition is asymptotically satisfied. ■

Theorem 1: The closed-loop system under the proposed actor-critic update laws and identifying filter is uniformly ultimately bounded. ♢

Proof: To analyze the stability of the closed-loop system, consider the following Lyapunov function candidate:

$$L = \frac{1}{2}e^2 + \frac{1}{2}\tilde{f}^2 + \frac{1}{2}\tilde{W}_c^T\tilde{W}_c + \frac{1}{2}\tilde{W}_a^T\tilde{W}_a, \quad (29)$$

where $\tilde{W}_c = \hat{W}_c - W_V^*$, $\tilde{W}_a = \hat{W}_a - W_V^*$ and $\tilde{f} = f - \hat{f}$. The time derivative of L is derived as follows:

$$\begin{aligned}\dot{L} &= e(g_0\hat{u}^* + f - \dot{r}) + \tilde{f}\dot{\tilde{f}} + \tilde{W}_c\dot{\tilde{W}}_c + \tilde{W}_a\dot{\tilde{W}}_a \\ &= -\gamma_e e^2 - \frac{1}{2}e\hat{W}_a^T\sigma_V + e\tilde{f} - e\dot{r} + \tilde{f}(-\gamma_{f,2}\tilde{f} + \dot{f}) \\ &\quad - \gamma_c\tilde{W}_c^T\sigma_V\sigma_V^T\hat{W}_c - \tilde{W}_a^T\sigma_V\sigma_V^T(\gamma_a(\hat{W}_a - \hat{W}_c) + \gamma_c\hat{W}_c)\end{aligned}$$

Applying Young's inequality: $e\tilde{f} \le \frac{1}{2}e^2 + \frac{1}{2}\tilde{f}^2$, $-e\dot{r} \le \frac{1}{2}e^2 + \frac{1}{2}\dot{r}^2$, and $-\frac{1}{2}e\hat{W}_a^T\sigma_a \le \frac{1}{4}e^2 + \frac{1}{4}\hat{W}_a^T\sigma_a\sigma_a^T\hat{W}_a$, the Lyapunov function derivative satisfies the following inequality:

$$\begin{aligned}\dot{L} &\le -\gamma_e e^2 + \frac{1}{2}e^2 + \frac{1}{2}\dot{r}^2 + \frac{1}{2}e^2 + \frac{1}{2}\tilde{f}^2 + \frac{1}{4}e^2 \\ &\quad + \frac{1}{4}\hat{W}_a^T\sigma_V\sigma_V^T\hat{W}_a + \tilde{f}(-\gamma_{f,2}\tilde{f} + \dot{f}) \\ &\quad - \gamma_c\tilde{W}_c^T\sigma_V\sigma_V^T\hat{W}_c - \tilde{W}_a^T\sigma_V\sigma_V^T\left(\gamma_a(\hat{W}_a - \hat{W}_c) + \gamma_c\hat{W}_c\right) \\ &= -\left(\gamma_e - \frac{5}{4}\right)e^2 - \gamma_c\tilde{W}_c^T\sigma_V\sigma_V^T\hat{W}_c - \gamma_a\tilde{W}_a^T\sigma_V\sigma_V^T\hat{W}_a \\ &\quad + (\gamma_a - \gamma_c)\tilde{W}_a^T\sigma_V\sigma_V^T\hat{W}_c + \frac{1}{4}\hat{W}_a^T\sigma_V\sigma_V^T\hat{W}_a \\ &\quad + \tilde{f}\left(-(\gamma_{f,2} - 1)\tilde{f} + \dot{f}\right) + \frac{1}{2}\dot{r}^2\end{aligned}$$

Using the relationships $\tilde{W}_a = \hat{W}_a - W_V^*, \tilde{W}_c = \hat{W}_c - W_V^*$ it follows that: $\tilde{W}_c^T\sigma_V\sigma_V^T\hat{W}_c = \frac{1}{2}\tilde{W}_c^T\sigma_V\sigma_V^T\tilde{W}_c + \frac{1}{2}\hat{W}_c\sigma_V\sigma_V^T\hat{W}_c - \frac{1}{2}\left(W_V^{*T}\sigma_V\right)^2$, and $\hat{W}_a^T\sigma_V\sigma_V^T\hat{W}_a = \frac{1}{2}\tilde{W}_a^T\sigma_V\sigma_V^T\tilde{W}_a + \frac{1}{2}\hat{W}_a\sigma_V\sigma_V^T\hat{W}_a - \frac{1}{2}\left(W_V^{*T}\sigma_V\right)^2$. Based on these expressions, the Lyapunov derivative can be rewritten as follows:

$$\begin{aligned}\dot{L} &\le -\left(\gamma_e - \frac{5}{4}\right)e^2 - \frac{\gamma_c}{2}\tilde{W}_c^T\sigma_V\sigma_V^T\tilde{W}_c \\ &\quad - \frac{\gamma_a}{2}\tilde{W}_a^T\sigma_V\sigma_V^T\tilde{W}_a + (\gamma_a - \gamma_c)\tilde{W}_a^T\sigma_V\sigma_V^T\hat{W}_c \\ &\quad - \frac{\gamma_c}{2}\hat{W}_c\sigma_V\sigma_V^T\hat{W}_c - \left(\frac{\gamma_a}{2} - \frac{1}{4}\right)\hat{W}_a\sigma_V\sigma_V^T\hat{W}_a \\ &\quad + \left(\frac{\gamma_c}{2} + \frac{\gamma_a}{2}\right)(W_V^{*T}\sigma_V)^2 + \tilde{f}\left(-(\gamma_{f,2} - 1)\tilde{f} + \dot{f}\right) + \frac{1}{2}\dot{r}^2\end{aligned}$$

Applying Young's inequality: $(\gamma_a - \gamma_c)\tilde{W}_a^T\sigma_V\sigma_V^T\hat{W}_c \le \frac{(\gamma_a - \gamma_c)}{2}\tilde{W}_a^T\sigma_V\sigma_V^T\tilde{W}_a + \frac{(\gamma_a - \gamma_c)}{2}\hat{W}_c\sigma_V\sigma_V^T\hat{W}_c$, the Lyapunov

function derivative satisfies the following inequality:

$$\begin{aligned}\dot{L} \leq & -\left(\gamma_e-\frac{5}{4}\right)e^2-\frac{\gamma_c}{2}\tilde{W}_c^T\sigma_V\sigma_V^T\tilde{W}_c-\frac{\gamma_a}{2}\tilde{W}_a^T\sigma_V\sigma_V^T\tilde{W}_a \\ & +\frac{(\gamma_a-\gamma_c)}{2}\tilde{W}_a^T\sigma_V\sigma_V^T\tilde{W}_a+\frac{(\gamma_a-\gamma_c)}{2}\hat{W}_c^T\sigma_V\sigma_V^T\hat{W}_c \\ & -\frac{\gamma_c}{2}\hat{W}_c^T\sigma_V\sigma_V^T\hat{W}_c-\left(\frac{\gamma_a}{2}-\frac{1}{4}\right)\hat{W}_a^T\sigma_V\sigma_V^T\hat{W}_a \\ & +\left(\frac{\gamma_c}{2}+\frac{\gamma_a}{2}\right)\left(W_V^{*T}\sigma_V\right)^2+\tilde{f}\left(-(\gamma_{f,2}-1)\tilde{f}+\dot{f}\right)+\frac{1}{2}\dot{r}^2 \\ = & \left(\gamma_e-\frac{5}{4}\right)e^2-\frac{\gamma_c}{2}\tilde{W}_c^T\sigma_V\sigma_V^T\tilde{W}_c-\frac{\gamma_c}{2}\tilde{W}_a^T\sigma_V\sigma_V^T\tilde{W}_a \\ & -\left(\gamma_c-\frac{\gamma_a}{2}\right)\hat{W}_c^T\sigma_V\sigma_V^T\hat{W}_c-\left(\frac{\gamma_a}{2}-\frac{1}{4}\right)\hat{W}_a^T\sigma_V\sigma_V^T\hat{W}_a \\ & +\left(\frac{\gamma_c}{2}+\frac{\gamma_a}{2}\right)\left(W_V^{*T}\sigma_V\right)^2+\tilde{f}\left(-(\gamma_{f,2}-1)\tilde{f}+\dot{f}\right)+\frac{1}{2}\dot{r}^2\end{aligned}$$

Under the boundedness assumptions $|\dot{f}| \leq \bar{f}$ and $\|W_V^*\| \leq \bar{W}_V$, the following inequality is obtained:

$$\begin{aligned}\dot{L} \leq & -\left(\gamma_e-\frac{5}{4}\right)e^2-\frac{\gamma_c}{2}\tilde{W}_c^T\sigma_V\sigma_V^T\tilde{W}_c-\frac{\gamma_c}{2}\tilde{W}_a^T\sigma_V\sigma_V^T\tilde{W}_a \\ & -\left(\gamma_c-\frac{\gamma_a}{2}\right)\hat{W}_c^T\sigma_V\sigma_V^T\hat{W}_c-\left(\frac{\gamma_a}{2}-\frac{1}{4}\right)\hat{W}_a^T\sigma_V\sigma_V^T\hat{W}_a \\ & +\left(\frac{\gamma_c}{2}+\frac{\gamma_a}{2}\right)(W_V^{*T}\sigma_V)^2-\frac{(\gamma_{f,2}-1)}{2}\tilde{f}^2 \\ & -\left(\frac{\gamma_{f,2}-1}{2}\right)\tilde{f}\left(\tilde{f}-\frac{2\bar{f}}{\gamma_{f,2}-1}\right)+\frac{1}{2}\dot{r}^2\end{aligned}$$

Based on the gain conditions $\gamma_e > \frac{5}{4}$, $\gamma_c > \frac{\gamma_a}{2}$, $\gamma_{f,2} > 1$, and $\frac{2\bar{f}}{\gamma_{f,2}-1} \approx 0$, the Lyapunov derivative is bounded by

$$\begin{aligned}\dot{L} \leq & -\left(\gamma_e-\frac{5}{4}\right)e^2-\frac{\gamma_c}{2}\tilde{W}_c^T\sigma_V\sigma_V^T\tilde{W}_c \\ & -\frac{\gamma_c}{2}\tilde{W}_a^T\sigma_V\sigma_V^T\tilde{W}_a-\frac{(\gamma_{f,2}-1)}{2}\tilde{f}^2+c,\end{aligned} \tag{30}$$

where $C(t) = \left(\frac{\gamma_c}{2}+\frac{\gamma_a}{2}\right)(W_V^{*T}\sigma_V)^2+\frac{1}{2}\dot{r}^2$ is bounded such that $C(t) \leq c$. Let $\lambda_v^{\min}$ denote the minimum eigenvalue of $\sigma_V\sigma_V^T$. Then, the derivative can be further bounded as

$$\dot{L} \leq -\alpha L + c, \tag{31}$$

where $\alpha = \min\left\{2\left(\gamma_e-\frac{5}{4}\right),\ \gamma_c\lambda_v^{\min},\ (\gamma_{f,2}-1)\right\}$.

■

V. Simulation Validation

A. Simulation Setup

This subsection provides detailed information about the simulation environment and the design parameters used for the proposed control strategy. The proposed controller was implemented in the MATLAB/Simulink 2024b environment using the servo motor model described by equations (1) and (2). The motor parameters were configured as follows: $J = 4.3\times10^{-5}$, $B = 0.1J$, $k_T = k_e = 0.068$, $R = 0.078$, $L = 1.3\times10^{-4}$.

The actor and critic NNs were each implemented with 16 neurons. The basis function for the i-th neuron ($i = 1,\ldots,16$) was defined as:

$$\sigma_{V,i}(e,x) = \exp\left(-\frac{\left([e,x]^T-\frac{1}{2}[i,i]^T\right)^T\left([e,x]^T-\frac{1}{2}[i,i]^T\right)}{3^2}\right).$$

The design gains were selected as:

$$\gamma_e = 20, \quad \gamma_{f,1} = 100, \quad \gamma_{f,2} = 227.5, \quad \gamma_c = 42, \quad \gamma_a = 23.$$

The initial weights of the NNs were initialized as:

$$\hat{W}_a(0) = [6,\ldots,6]^T \in \mathbb{R}^{16\times1}, \quad \hat{W}_c(0) = [4,\ldots,4]^T \in \mathbb{R}^{16\times1}.$$

For comparative analysis, the proposed method was evaluated alongside two representative online learning-based control approaches.

- Case 1: Identifier-based actor critic learning approach with full state feedback [8],
- Case 2: Extended state observer (ESO)-based actor critic learning controller with output feedback [9],

These two methods were selected based on their relevance to the problem setting. Case 1 represents an ideal learning-based control scheme that requires full state information and involves high computational cost. In contrast, Case 2 adopts an observer-based approach to estimate unmeasured states, making it more suitable for practical output-feedback scenarios while maintaining the online learning framework.

B. Simulation Results

To evaluate the tracking performance, the desired reference trajectory and external load disturbance were defined as:

$$r = 8\sin(2\pi\cdot0.3t)+3\cos(2\pi\cdot0.2t)\cos(2\pi\cdot0.4t)\left(1-e^{-0.2t}\right),$$

$$T_L = 0.4\theta+\sin^2(2\pi\cdot0.2t).$$

The initial condition was set to $\theta(0) = 20$ rad. Fig. 2 illustrates the controlled output responses for three different control strategies, while Fig. 3 presents the corresponding tracking errors. Table I summarizes the computation times and RMSE values for each case. Additionally, Fig. 4 provides RMSE maps for Case 2 and the proposed method with respect to variations in the control gain.

As shown in Figs. 2–3 and Table I, Case 1 exhibited the most stable transient behavior and achieved the best RMSE performance, although it includes slight noise compared to the proposed method. However, this method requires full state feedback and incurs the highest computational cost. In contrast, Fig. 4 visualizes the RMSE performance variation with respect to control gain parameters for Case 2 and the proposed method. The results indicate that the proposed method exhibits more stable performance over a wider range of gain values, whereas Case 2 shows significant degradations in RMSE even with slight gain variations. This implies that Case 2 is more sensitive to gain selection, which makes the tuning process more challenging. This highlights the practical advantage of the proposed method in terms of gain tuning compared to higher-order ESO-based approaches.

TABLE I: Computation Time and RMSE Comparison

Method	Computation Time (s)	RMSE
Case 1	8.182×10^{-6}	3.467
Case 2	6.823×10^{-6}	4.915
Proposed	2.165×10^{-6}	3.747

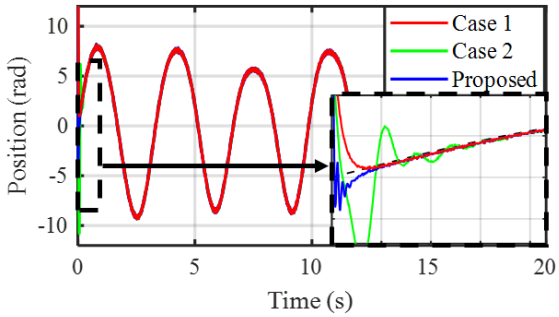

Fig. 2: Controlled output dynamics comparison results.

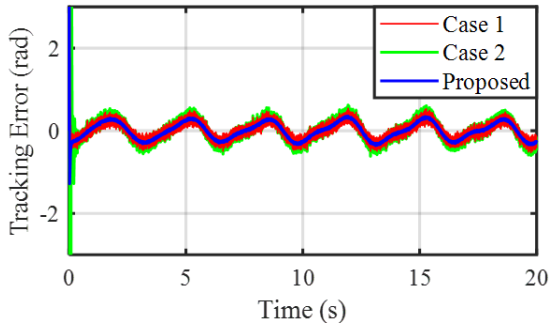

Fig. 3: Tracking error comparison results.

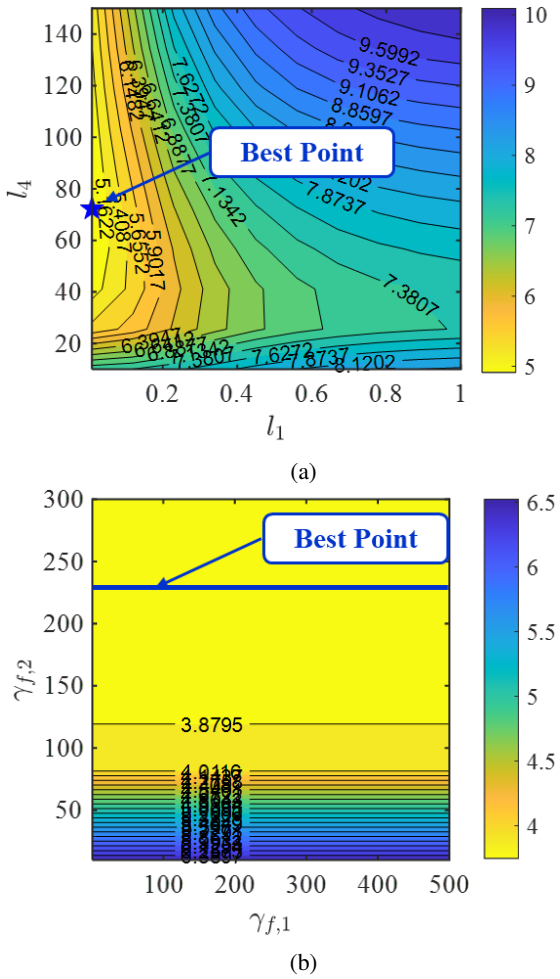

Fig. 4: Changes in RMSE performance of the tuning parameters: (a) Case 2 method and (b) proposed method

VI. Conclusions

This paper presented an online actor critic learning-based optimal tracking control method for output-feedback servo positioning systems with unknown dynamics and external disturbances. A control-affine formulation was used to represent system uncertainties as a lumped unknown function, which was estimated online using an identifying filter. An actor-critic neural network structure was employed to approximate the value function and optimal control policy. Adaptive update laws were designed to minimize the Bellman residual, and a Lyapunov-based analysis established closed-loop stability. Simulation results demonstrated accurate tracking and robust performance under time-varying disturbances. Future work will extend the framework to handle constraints on states and control inputs, and generalize it to complex nonlinear multi-input multi-output (MIMO) systems.

References

[1] C. Liu, G. Luo, Z. Xue, Z. Zhou, and Z. Chen, "A PMSM speed servo system based on internal model control with extended state observer," in *Proc. IECON*, Oct. 2017, pp. 1729–1734.

[2] Y.-J. Lin, P.-H. Chou, and S.-C. Yang, "Parameter identification and controller design for limited-angle servo motor drives using acceleration estimation technique," in *Proc. IECON*, Jan. 2024, pp. 1–6.

[3] K. Yu, Z. Wang, C. Zhao, H. Wang, X. Zhu, and C. H. T. Lee, "Improved universal control scheme with voltage disturbance observer for dual three-phase PMSM drives under single open-phase fault," in *Proc. IECON*, Jan. 2024, pp. 1–6.

[4] Y. Zuo, J. Mei, C. Jiang, X. Yuan, S. Xie, and C. H. T. Lee, "Linear active disturbance rejection controllers for PMSM speed regulation system considering the speed filter," *IEEE Trans. Power Electron.*, vol. 36, no. 12, pp. 14579–14592, Feb. 2021.

[5] B. Luo, Y. Yang, and D. Liu, "Adaptive Q-learning for data-based optimal output regulation with experience replay," *IEEE Trans. Cybern.*, vol. 48, no. 12, pp. 3337–3348, Feb. 2018.

[6] B. Luo, D. Liu, and H.-N. Wu, "Adaptive constrained optimal control design for data-based nonlinear discrete-time systems with critic-only structure," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2099–2111, Jun. 2018.

[7] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Networks*, vol. 22, no. 2, pp. 237–246, Mar. 2009.

[8] G. Wen, C. L. P. Chen, and S. S. Ge, "Simplified optimized backstepping control for a class of nonlinear strict-feedback systems with unknown dynamic functions," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4567–4580, Sep. 2021.

[9] J. Lee, S. You, W. Kim, and J. Moon, "Extended state observer-actor–critic architecture based output-feedback optimized backstepping control for permanent magnet synchronous motors," *Expert Syst. Appl.*, vol. 270, p. 126542, Apr. 2025.