

ICL-RCTAC hierarchical learning algorithm optimizes trajectory tracking and environmental disturbance adaptation for AUV in the horizontal plane

Nguyen Van Hoa¹, Pham Van Tuan², Nguyen Quang Vinh^{1*},
Nguyen Bich Van³, Nguyen Thi Thu Huong⁴

¹Institute of Missile, Academy of Military Science and Technology, 17 Hoang Sam, Nghia Do, Hanoi, Vietnam;

²Underwater Weapons Department, Naval Academy, 30 Tran Phu, Nha Trang, Khanh Hoa, Vietnam;

³Institute for Artificial Intelligence, VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam;

⁴Institute of Technology, General Department of Defense Industry, 3 Cau Vong, Dong Ngac, Hanoi, Vietnam.

*Corresponding author: vinhquang2808@gmail.com

Received 21 Mar. 2026; Revised 27 Jun. 2026; Accepted 15 Sep. 2026; Published 15 Oct. 2026.

DOI: <https://doi.org/10.54939/1859-1043.j.mst.114.2026.3-11>

ABSTRACT

ICL-RCTAC is a hierarchical reinforcement learning algorithm for adaptive optimal trajectory tracking of an Autonomous Underwater Vehicle (AUV) in the horizontal plane. It addresses two key limitations of existing continuous-time reinforcement learning: the warm-up phase and the persistent excitation (PE) condition, via a two-phase framework. In Phase I, Integral Concurrent Learning (ICL) exploits stored historical data to identify stable Critic weights under a finite-excitation (FE) condition, with convergence formally proven in Proposition 1. Phase II then activates an Actor–Critic loop under relaxed conditions for efficient policy optimization. Lyapunov analysis confirms uniform ultimate boundedness (UUB) of the closed-loop system. Simulations in MATLAB/Simulink 2024a show faster convergence, lower tracking error, and reduced control effort than PID and RCTAC, confirming the algorithm's practical potential for AUV applications.

Keywords: Autonomous underwater vehicle; Reinforcement learning; Integral concurrent learning; Lyapunov stability; Adaptive control; Optimal control.

1. INTRODUCTION

AUVs play a key role in marine research, exploration, and defense missions, operating underwater for long periods unmanned [1] to perform tasks like seabed surveying, water-quality monitoring, pipeline inspection, and maritime reconnaissance [2]. Reliable tracking remains difficult due to AUVs' nonlinear dynamics, time-varying hydrodynamics, and unpredictable disturbances such as ocean currents and turbulence. Model-based methods PID, sliding mode, backstepping, model predictive control perform well with an accurate model, but their effectiveness declines as parameters vary or disturbances arise, as is common in underwater operations [3, 4]. Recent research thus focuses on reinforcement learning and adaptive optimal control, letting controllers refine policies through interaction with the environment and reducing reliance on precise modeling [5, 6]. Within continuous-time reinforcement learning, the Actor–Critic framework is widely used for nonlinear optimal control; however, many methods require a warm-up phase and the PE condition for convergence, hard to maintain in real AUV operations with limited inputs [7], raising energy use and causing transient oscillations. ICL addresses this via convergence under the FE condition using stored data [8], while RCTAC relaxes convergence requirements to speed learning, though their integration remains limited.

This paper proposes ICL-RCTAC, combining an ICL phase for Critic weight stabilization with an RCTAC-based Actor–Critic framework for policy optimization. Contributions: (i) a two-phase architecture removing the warm-up stage and PE requirement via the FE condition; (ii) relaxed

convergence for continuous-time Actor–Critic; (iii) a Lyapunov-based UUB proof.

2. MATHEMATICAL MODEL OF THE AUV

Theoretical foundations

The AUV model uses Earth-fixed and body-fixed frames [9]: state $\boldsymbol{\eta} = [x, y, z, \phi, \theta, \psi]^T$ in the Earth frame and $\mathbf{v} = [u, v, w, p, q, r]^T$ in the body frame, linked by kinematic equations.

$$\mathbf{M}\dot{\mathbf{v}} + \mathbf{C}(\mathbf{v})\mathbf{v} + \mathbf{D}(\mathbf{v})\mathbf{v} + \mathbf{g}(\boldsymbol{\eta}) = \boldsymbol{\tau} + \boldsymbol{\tau}_{env}, \quad \dot{\boldsymbol{\eta}} = \mathbf{J}(\boldsymbol{\eta})\mathbf{v} \quad (1)$$

For an AUV operating at near-constant depth with negligible pitch and roll ($\varphi \approx 0, \theta \approx 0$) (a standard assumption [3, 7]), the 6-DOF system decouples into horizontal and vertical subsystems. The states (v, r, ψ, x, y) in (2) describe trajectory tracking in the OXY plane:

$$\dot{\boldsymbol{\chi}} = \mathbf{M}^{-1}[\mathbf{F}\boldsymbol{\chi} + \mathbf{G} + \mathbf{T}\delta_v] \quad (2)$$

Where: $\mathbf{G} = [0, 0, 0, u_0 c\psi - v s\psi, u_0 s\psi + v c\psi]^T$; $\mathbf{T} = [Y_{\delta_v} u_0^2, N_{\delta_v} u_0^2, 0, 0, 0]^T$;
 $\boldsymbol{\chi} = [v, r, \psi, x, y]^T$;

$$\mathbf{M} = \begin{bmatrix} m + Y_{\dot{v}} & Y_{\dot{r}} & 0 & 0 & 0 \\ N_{\dot{v}} & I_{zz} + N_{\dot{r}} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}; \quad \mathbf{F} = \begin{bmatrix} Y_{uv}u_0 + Y_{v|v}|v| & Y_{ur}u_0 + Y_{r|r}|r| & 0 & 0 & 0 \\ N_{uv}u_0 + N_{v|v}|v| & N_{ur}u_0 + N_{r|r}|r| & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$c(\cdot) = \cos(\cdot), \quad s(\cdot) = \sin(\cdot)$$

From (2), the heading-angle dynamics can be written as

$$\dot{\psi} = r; \quad \dot{r} = (I_{zz} - N_{\dot{r}})^{-1} [N_{\dot{v}}\dot{v} + N_{ur}u_0r] + (I_{zz} - N_{\dot{r}})^{-1} N_{\delta_v}u_0^2\delta_v \quad (3)$$

The model (3) can be rewritten in the affine nonlinear form

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})\mathbf{u} \quad (4)$$

Where: the control input is $u = \delta_v$ with $|\delta_v| \leq \delta_{max} = \pi / 6 \text{ rad}$, the system dynamics are $f(x) = (I_{zz} - N_{\dot{r}})^{-1} [N_{\dot{v}}\dot{v} + N_{ur}u_0r]$ and the input gain is $g(x) = (I_{zz} - N_{\dot{r}})^{-1} N_{\delta_v}u_0^2$.

3. HIERARCHICAL LEARNING ALGORITHM ICL-RCTAC

3.1. Optimal trajectory-tracking control problem

Consider the affine nonlinear system (4), with infinite-horizon cost function [7]:

$$V(\mathbf{x}_0, \mathbf{u}(\cdot)) = \int_0^\infty (\mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{u}^T \mathbf{R} \mathbf{u}) d\tau \quad (5)$$

where $\mathbf{Q} > 0$ and $\mathbf{R} > 0$. Set $L(\mathbf{x}, \mathbf{u}) = \mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{u}^T \mathbf{R} \mathbf{u}$. The Hamiltonian function is defined as

$$H(\mathbf{x}, \mathbf{u}, \Delta V) = L(\mathbf{x}, \mathbf{u}) + \Delta V^T \dot{\mathbf{x}}, \quad \forall \mathbf{x}, \mathbf{u}. \quad (6)$$

The optimal value function $V^*(\mathbf{x})$ must satisfy the HJB equation [7]: $\min_u [H(\mathbf{x}, \mathbf{u}, \Delta V)] = 0$

The optimal control policy is $\partial H / \partial \mathbf{u} = 2\mathbf{R}\mathbf{u}^* + \mathbf{g}(\mathbf{x})^T \nabla V^*(\mathbf{x}) = 0$ obtained from the stationary condition:

$$\mathbf{u}^*(\mathbf{x}) = -(\mathbf{R}^{-1} \mathbf{g}(\mathbf{x})^T \Delta V^*(\mathbf{x})) / 2 \quad (7)$$

Substituting (7) into the Hamiltonian leads to the optimal HJB equation

$$\mathbf{x}^T \mathbf{Q} \mathbf{x} + (\mathbf{u}^*)^T \mathbf{R} \mathbf{u}^* + \nabla V^*(\mathbf{x})^T \left[\mathbf{f}(\mathbf{x}) - \mathbf{g}(\mathbf{x}) (\mathbf{R}^{-1} (\mathbf{g}(\mathbf{x}))^T \nabla V^*(\mathbf{x})) / 2 \right] = 0 \quad (8)$$

Because (8) is difficult to solve analytically, (8) is approximated using Critic and Actor networks with weight vectors $\mathbf{W}_c \in \mathbb{R}^{n_c}$, $\mathbf{W}_a \in \mathbb{R}^{n_a}$ respectively [4]:

$$V(\mathbf{x}) = \boldsymbol{\phi}^T \mathbf{W}_c \text{ and } \mathbf{u}(\mathbf{x}) = \boldsymbol{\sigma}^T \mathbf{W}_a \quad (9)$$

Where: $\boldsymbol{\phi}$, $\boldsymbol{\sigma}$ are bounded activation-function vectors. The gradient of the approximated $V(\mathbf{x})$ by:

$$\nabla V(\mathbf{x}) = \nabla \boldsymbol{\phi}^T \mathbf{W}_c \quad (10)$$

3.2. Phase 1: Task-oriented system parameter adaptation

Assumption 1: The environmental disturbance $\mathbf{d}(\mathbf{x}, t)$ and ε_c and ε_a are bounded [7], there exists $d_{\max}, \varepsilon_{c, \max}, \varepsilon_{a, \max} > 0 : \|\mathbf{d}(\mathbf{x}, t)\| \leq d_{\max}, \|\varepsilon_c\| \leq \varepsilon_{c, \max}, \|\varepsilon_a\| \leq \varepsilon_{a, \max}$.

Integrating both sides over the interval $[t, t + \Delta t]$, with $\Delta t > 0$, yields

$$\int_t^{t+\Delta t} H(x, u, \Delta V) d\tau = 0 \quad (11)$$

Substituting (10) into (11) results in a model that is linear with respect to the Critic weights:

$$\mathbf{W}_c \boldsymbol{\xi}_i = \psi_i + \delta_{Bell, i} \quad (12)$$

Where: $\boldsymbol{\xi}_i = \int_t^{t+\Delta t} \nabla \boldsymbol{\phi}(\mathbf{x}(\tau)) d\tau$ is the integral regression vector, $\psi_i = \int_t^{t+\Delta t} L(\mathbf{x}(\tau), \mathbf{u}(\tau)) d\tau$ is the accumulated observed cost, and residual error $\delta_{Bell, i}$ arises from the neural-network approximation error ε_c and satisfies $\|\delta_{Bell, i}\| \leq \bar{\Delta}_{Bell}$, where $\bar{\Delta}_{Bell}$ is a known upper bound [8].

Historical data are stored in the ICL stack $\mathcal{S} = \{\boldsymbol{\xi}_k, \psi_k\}_{k=1}^M$ which is used to minimize the least-squares cost function [8], set $\boldsymbol{\Xi} = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_M]^T \in \mathbb{R}^{M \times n_c}$ and $\boldsymbol{\Psi} = [\psi_1, \psi_2, \dots, \psi_M]^T \in \mathbb{R}^M$.

$$\min_{\mathbf{W}_c} J_{ICL}(\mathbf{W}_c) = (1/2M) \sum_{i=1}^M (\boldsymbol{\xi}_i^T \mathbf{W}_c - \psi_i)^2 = (1/2M) \|\boldsymbol{\Xi} \mathbf{W}_c - \boldsymbol{\Psi}\|^2 \quad (13)$$

Assumption 2 (FE condition) [8]: $\mathbf{W}_c$ converges to a neighborhood of $\mathbf{W}_c^*$ if $\mathbf{S}$ satisfies $\text{rank}(\boldsymbol{\Xi}) = n_c$, $\det(\boldsymbol{\Xi}^T \boldsymbol{\Xi}) \neq 0$, so $\boldsymbol{\Xi}^T \boldsymbol{\Xi}$ is invertible. The $\mathbf{W}_c$ update law in Phase 1, with learning rate $\alpha_{ICL} > 0$, is given by $\dot{\mathbf{W}}_c = -\alpha_{ICL} \nabla J_{ICL} = -(\alpha_{ICL} / M) \boldsymbol{\Xi}^T [\boldsymbol{\Xi} \mathbf{W}_c - \boldsymbol{\Psi}]$. When the convergence condition $\|\Delta \mathbf{W}_c(t)\| < \varepsilon_{ICL}$, where ε_{ICL} is a small convergence threshold, the resulting stabilized weight vector $\mathbf{W}_{c_0} = \mathbf{W}_{ident}$ is used as the initial Critic weight for Phase 2.

Proposition 1: Under Assumption 2, the Critic weight error $\tilde{\mathbf{W}}_c = \mathbf{W}_c - \mathbf{W}_c^*$ under update law (Assumption 2) converges to the compact set $\Omega_1 = \{\|\tilde{\mathbf{W}}_c\| \leq r_{ICL}\}$.

Proof: From Assumption 2 have $\dot{\tilde{\mathbf{W}}}_c = -\alpha_{ICL} \boldsymbol{\Xi}^T (\boldsymbol{\Xi} \tilde{\mathbf{W}}_c + \boldsymbol{\varepsilon}_{ICL}) / M$. Define $V_1 = (1/2) \|\tilde{\mathbf{W}}_c\|^2$

$\Rightarrow \dot{V}_1 = \tilde{\mathbf{W}}_c^T \dot{\tilde{\mathbf{W}}}_c$. Thus: $\dot{V}_1 = -\alpha_{ICL} \tilde{\mathbf{W}}_c^T \Xi^T (\Xi \tilde{\mathbf{W}}_c + \boldsymbol{\varepsilon}_{ICL}) / M$ and $\lambda_{\min}(\Xi^T \Xi) > 0$.

Applying Young's inequality: $\dot{V}_1 \leq -(\alpha_{ICL} / M) \lambda_{\min}(\Xi^T \Xi) \|\tilde{\mathbf{W}}_c\|^2 + (\alpha_{ICL} / M) \|\Xi\| \varepsilon_{ICL \max} \|\tilde{\mathbf{W}}_c\|$. Setting $r_{ICL} = \varepsilon_{ICL \max} \|\Xi\| / (\lambda_{\min}(\Xi^T \Xi))$, one obtains $\dot{V}_1 < 0$ whenever $\|\tilde{\mathbf{W}}_c\| > r_{ICL}$. Therefore $\tilde{\mathbf{W}}_c$ converges to Ω_1 .

3.3. Phase 2: Actor-critic policy optimization

The stabilized Critic weight $\mathbf{W}_c = \mathbf{W}_{c_0}$ is used to evaluate the control policy. The instantaneous Bellman error and the historical ICL error are combined to form the Critic objective function $\mathbf{J}_c$.

$$\mathbf{J}_c = (1/2) [L(\mathbf{x}, \mathbf{u}) + \nabla \phi^T \mathbf{W}_c \dot{\mathbf{x}}]^2 + (\alpha_{ICL} / 2M) \|\Xi \mathbf{W}_c - \Psi\|^2 \quad (14)$$

Here $\nabla \mathbf{J}_{Bell} = \nabla \phi [L(\mathbf{x}, \mathbf{u}) + \nabla \phi^T \mathbf{W}_c \dot{\mathbf{x}}]$. The gradient of the Critic objective is given by

$$\nabla \mathbf{J}_c = \nabla \mathbf{J}_{Bell} + (\alpha_{ICL} / M) \Delta \Xi^T (\Delta \Xi \mathbf{W}_c - \Psi) \quad (15)$$

With $\alpha_c > 0$ denotes the learning rate of the Critic network, the Critic weight-update law becomes

$$\dot{\mathbf{W}}_c = -\alpha_c (\nabla \mathbf{J}_{Bell} + (\alpha_{ICL} / M) \Delta \Xi^T (\Delta \Xi \mathbf{W}_c - \Psi)) \quad (16)$$

When $\|\nabla \mathbf{J}_c\| < \varepsilon_r$, a sufficiently small Actor network is allowed to begin updating [5]. $\mathbf{W}_a$ are updated to minimize the approximate Hamiltonian. The Actor objective function is defined as.

$$\mathbf{J}_a = (1/2) \|\boldsymbol{\sigma}^T \mathbf{W}_a + (1/2) \mathbf{R}^{-1} \mathbf{g}(\mathbf{x})^T \nabla \phi \mathbf{W}_c\|^2 \quad (17)$$

The gradient of the Actor objective is

$$\nabla \mathbf{J}_a = \boldsymbol{\sigma} [\boldsymbol{\sigma}^T \mathbf{W}_a + (1/2) \mathbf{R}^{-1} \mathbf{g}(\mathbf{x})^T \nabla \phi \mathbf{W}_c] \quad (18)$$

Therefore, the Actor weight-update law $\dot{\mathbf{W}}_a = -\alpha_a \nabla \mathbf{J}_a$ becomes

$$\dot{\mathbf{W}}_a = -\alpha_a \boldsymbol{\sigma} [\boldsymbol{\sigma}^T \mathbf{W}_a + (1/2) \mathbf{R}^{-1} \mathbf{g}(\mathbf{x})^T \nabla \phi \mathbf{W}_c] \quad (19)$$

The process is repeated until both the Actor and Critic networks converge to the neighborhood of the optimal solution.

Algorithm: ICL-RCTAC — Hierarchical Learning Controller

Initialization:

Initialize $\mathbf{W}_c(0)$, $\mathbf{W}_a(0)$; $\mathcal{S} \leftarrow \emptyset$; set phase $\leftarrow 1$; Select $T, \alpha_{ICL}, \alpha_c, \alpha_a > 0$; $\varepsilon_{ICL}, \varepsilon_a > 0$

Main loop:

1. **for** each sampling instant t **do** Measure state $x(t)$; apply control $u(t)$ to plant
2. **if** phase=1 **then** $\triangleright$ Phase 1: ICL-based Critic Initialization
3. Compute integral regression pair: $\{\xi_i, \psi_i\}$, Store into memory stack:
 $\mathcal{S} \leftarrow \mathcal{S} \cup \{\xi_i, \psi_i\}$
4. **if** FE condition satisfied: $\text{rank}(\Xi) = n_c$ **then** Update Critic:
 $\dot{\mathbf{W}}_c = -(\alpha_{ICL} / M) \Xi^T [\Xi \mathbf{W}_c - \Psi]$ **end if**
5. **if** $\|\dot{\mathbf{W}}_c\| < \varepsilon_{ICL}$ **then** $\triangleright$ Phase 1 converged $\rightarrow$ switch to Phase 2. Set $\mathbf{W}_{c_0} \leftarrow \mathbf{W}_c$; Phase

$\leftarrow 2$ end if
 6. **else** $\triangleright$ Phase 2: Actor–Critic Policy Optimization
 7. Compute: $\mathbf{u} = \boldsymbol{\sigma} \mathbf{W}_a^T$; $\nabla V = \nabla \boldsymbol{\phi}^T \mathbf{W}_c$; $\delta = L(\mathbf{x}, \mathbf{u}) + \nabla V^T (\mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})\mathbf{u})$
 8. Update Critic $\dot{\mathbf{W}}_c = -\alpha_c (\nabla \mathbf{J}_{Bell} + (\alpha_{ICL} / M) \Delta \Xi^T (\Delta \Xi \mathbf{W}_c - \Psi))$
 9. **if** $\|\nabla J_c\| < \varepsilon_a$ **then** $\triangleright$ Critic sufficiently converged: Update Actor: $\dot{\mathbf{W}}_a = -\alpha_a \nabla \mathbf{J}_a$;
 end if
 10. **end if**
 11. **end for**
Output: Converged weights $\mathbf{W}_c^*$, $\mathbf{W}_a^*$; control policy $\mathbf{u}^* = \boldsymbol{\sigma} \mathbf{W}_a^{*T}$

4. LYAPUNOV STABILITY ANALYSIS AND UUB CONVERGENCE

4.1. Assumptions and lemmas

Assumption 3. The activation functions $\boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\sigma}(\mathbf{x})$ satisfy [7]: Lipschitz continuity: There exist positive constants L_ϕ, L_σ such that: $\|\boldsymbol{\sigma}(\mathbf{x}) - \boldsymbol{\sigma}(\mathbf{y})\| \leq L_\sigma \|\mathbf{x} - \mathbf{y}\|$ and $\|\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{y})\| \leq L_\phi \|\mathbf{x} - \mathbf{y}\|$. Boundedness: $\boldsymbol{\phi}_{\max}, \boldsymbol{\sigma}_{\max} \in \Omega$, $\exists \boldsymbol{\phi}_{\max}, \boldsymbol{\sigma}_{\max} > 0$ such that: $\|\boldsymbol{\phi}\| \leq \boldsymbol{\phi}_{\max}$ and $\|\boldsymbol{\sigma}\| \leq \boldsymbol{\sigma}_{\max}$

Lemma (Boundedness of Critic–Actor gradients and coupling term) [7, 8, 10, 11]: Under Assumption 1, the gradients of the Critic and Actor objective functions are bounded as $\|\mathbf{J}_c \delta(\mathbf{x}, \mathbf{W}_c, \mathbf{W}_a)\| \leq L_\delta (1 + \|\tilde{\mathbf{W}}_c\|)$ (dependent on $\|\nabla \phi\|$) and $\|\nabla \mathbf{J}_a H\| \leq K_a (1 + \|\tilde{\mathbf{W}}_c\| + \|\tilde{\mathbf{W}}_a\|)$ (depending on $L_\sigma, \sigma_{\max} \|\mathbf{R}\|$). Moreover, the coupling term between the two networks satisfies $\|\tilde{\mathbf{W}}_c^T \nabla \boldsymbol{\phi}^T \mathbf{g}(\mathbf{x}) \tilde{\mathbf{u}}\| \leq L_{ca} \|\tilde{\mathbf{W}}_c\| \|\tilde{\mathbf{W}}_a\| + c_\varepsilon$, where $L_{ca} = L_\phi \sup_{\mathbf{x} \in \Omega} \|\mathbf{g}(\mathbf{x})\| L_\sigma \sigma_{\max}$, and c_ε is a constant dependent on the approximation-error bounds $\varepsilon_{c,\max}, \varepsilon_{a,\max}$.

4.2. Lyapunov stability analysis

Theorem: Consider the AUV system (3) with the ICL-RCTAC controller described in Algorithm: ICL-RCTAC. If Assumptions 1–3 and the Lemma are satisfied, then the state error, the Critic weight error, and the Actor weight error of the closed-loop system are UUB.

Proof: Consider the composite Lyapunov function V_L , which is positive definite on $\Omega \times \mathbb{R}^{N_c} \times \mathbb{R}^{N_a}$.

$$V_L = V^*(\mathbf{x}) + (1/2\alpha_c) \tilde{\mathbf{W}}_c^T \tilde{\mathbf{W}}_c + (1/2\alpha_a) \tilde{\mathbf{W}}_a^T \tilde{\mathbf{W}}_a \quad (20)$$

Where: $\alpha_c > 0$ and $\alpha_a > FE_{th} / [\lambda_{\min}(\mathbf{R}) \sigma_{\min}^2]$ (with FE_{th} - FE threshold) the weight estimation errors are defined as $\tilde{\mathbf{W}}_c = \mathbf{W}_c - \mathbf{W}_c^*$, $\tilde{\mathbf{W}}_a = \mathbf{W}_a - \mathbf{W}_a^*$. The time derivative of the Lyapunov function;

$$\dot{V}_L = \dot{V}^*(\mathbf{x}) + (1/\alpha_c) \tilde{\mathbf{W}}_c^T \dot{\mathbf{W}}_c + (1/\alpha_a) \tilde{\mathbf{W}}_a^T \dot{\mathbf{W}}_a \quad (21)$$

Using the HJB equation (7) and the control error $\tilde{\mathbf{u}} = \mathbf{u} - \mathbf{u}^*$, we obtain

$$\dot{V}^*(\mathbf{x}) = -\mathbf{x}^T \mathbf{Q} \mathbf{x} - (\mathbf{u}^*)^T \mathbf{R} \mathbf{u}^* + (\nabla V^*)^T \mathbf{g}(\mathbf{x}) \tilde{\mathbf{u}} + (\nabla V^*)^T \mathbf{d} \quad (22)$$

Since $\exists \lambda_{\min}(\mathbf{Q}) > 0: \mathbf{x}^T \mathbf{Q} \mathbf{x} \geq \lambda_{\min}(\mathbf{Q}) \|\mathbf{x}\|^2$, there exists $-L(\mathbf{x}, \mathbf{u}^*) \leq -\lambda_{\min}(\mathbf{Q}) \|\mathbf{x}\|^2 - (\mathbf{u}^*)^T \mathbf{R} \mathbf{u}^*$ such that.

coupling term becomes $(\nabla V^*)^T \mathbf{g}(\mathbf{x}) \tilde{\mathbf{u}} = -2\mathbf{u}^T \mathbf{R} \boldsymbol{\sigma}^T \tilde{\mathbf{W}}_a - 2\mathbf{u}^T \mathbf{R} \boldsymbol{\varepsilon}_a$. Applying Young's inequality

with $\kappa_1 > 0$, $0 < \kappa_2 < \lambda_{\min}(\mathbf{Q})$. $\|(\nabla V)^T \mathbf{d}\| \leq \|\nabla V\| d_{\max} \leq L_V \|\mathbf{x}\| d_{\max} \leq (\kappa_2 / 2) \|\mathbf{x}\|^2 + (1 / 2\kappa_2) L_V^2 d_{\max}^2$.

$$-2\mathbf{u}^T \mathbf{R} \boldsymbol{\sigma}^T \tilde{\mathbf{W}}_a \leq \kappa_1 \|\tilde{\mathbf{W}}_a\|^2 + (1 / \kappa_1) \|\mathbf{u}^T \mathbf{R} \boldsymbol{\sigma}\|^2 \leq \kappa_1 \|\tilde{\mathbf{W}}_a\|^2 + (1 / \kappa_1) \|\mathbf{R}\|^2 \boldsymbol{\sigma}_{\max}^2 \|\mathbf{u}^*\|^2 \quad (23)$$

Consequently, the derivative of the optimal value function satisfies

$$\dot{V}^*(\mathbf{x}) \leq -(\lambda_{\min}(\mathbf{Q}) - \kappa_2) \|\mathbf{x}\|^2 + \kappa_1 \|\tilde{\mathbf{W}}_a\|^2 + C_1 \quad (24)$$

$C_1 = (1 / \kappa_1) \|\mathbf{R}\|^2 \boldsymbol{\sigma}_{\max}^2 \|\mathbf{u}^*\|^2 + 2 \|\mathbf{R}\| \boldsymbol{\sigma}_{\max} \|\mathbf{u}^*\| \varepsilon_{a,\max} + (L_V^2 d_{\max}^2 / 2\kappa_2)$, and L_V : the Lipschitz constant of $V^*(\mathbf{x})$.

$$\dot{V}_{Critic} = (1 / \alpha_c) \tilde{\mathbf{W}}_c^T \dot{\tilde{\mathbf{W}}}_c = (1 / \alpha_c) \tilde{\mathbf{W}}_c^T \dot{\mathbf{W}}_c, \quad (\text{By } \dot{\mathbf{W}}_c^* = 0) \quad (25)$$

From the Critic update law (16), the derivative of the Critic term becomes:

$$\dot{V}_{Critic} = -\tilde{\mathbf{W}}_c^T \nabla J_{Bell} - (\alpha_{ICL} / M) \tilde{\mathbf{W}}_c^T \Xi^T \Xi \tilde{\mathbf{W}}_c + (\alpha_{ICL} / M) \tilde{\mathbf{W}}_c^T \Xi^T \varepsilon_{Critic} \quad (26)$$

Applying Young's inequality with $c_1, c_2 > 0$, $(1 / 2c_1) + (1 / 2c_2) < (\alpha_{ICL} / M) \sigma_{\min}$, and $\nabla \mathbf{J}_{Bell} \leq \Delta_{Bell}$, Equation (26) is transformed into:

$$\dot{V}_{Critic} \leq -\beta_1 \|\tilde{\mathbf{W}}_c\|^2 + C_2 \quad (27)$$

Let: $\beta_1 = [(\alpha_{ICL} / M) \sigma_{\min} - (1 / 2c_1) - (1 / 2c_2)]$ and $C_2 = c_1 \Delta_{Bell}^2 / 2 + c_2 \|(\alpha_{ICL} / M) \Xi^T \varepsilon_{Critic}\|^2 / 2$

$$\dot{V}_{Actor} = (1 / \alpha_a) \tilde{\mathbf{W}}_a^T \dot{\tilde{\mathbf{W}}}_a = (1 / \alpha_a) \tilde{\mathbf{W}}_a^T \dot{\mathbf{W}}_a, \quad (\text{By } \dot{\mathbf{W}}_a^* = 0) \quad (28)$$

From (19), equation (28) becomes: $\dot{V}_{Actor} = -\tilde{\mathbf{W}}_a^T \boldsymbol{\sigma} \boldsymbol{\sigma}^T \tilde{\mathbf{W}}_a - (1 / 2) \mathbf{R}^{-1} \tilde{\mathbf{W}}_a^T \boldsymbol{\sigma} \mathbf{g}^T \nabla \phi \tilde{\mathbf{W}}_c$

By Lemma $\|\nabla \mathbf{J}_a\| \leq K_a (1 + \|\tilde{\mathbf{W}}_c\| + \|\tilde{\mathbf{W}}_a\|)$. Using $\boldsymbol{\sigma} \boldsymbol{\sigma}^T \geq \lambda_{\min}(\mathbf{R}) \boldsymbol{\sigma}_{\min}^2 \mathbf{I}$ and $\|\boldsymbol{\sigma} \boldsymbol{\sigma}^T\| \leq \boldsymbol{\sigma}_{\max}^2 \mathbf{I}$.

$-\boldsymbol{\sigma}^T \tilde{\mathbf{W}}_a \boldsymbol{\sigma} \tilde{\mathbf{W}}_a^T \leq -\lambda_{\min}(\mathbf{R}) \boldsymbol{\sigma}_{\min}^2 \|\tilde{\mathbf{W}}_a\|^2$. Using the Cauchy-Schwarz inequality [12]:

$$-(1 / 2) \mathbf{R}^{-1} \tilde{\mathbf{W}}_a^T \boldsymbol{\sigma} \mathbf{g}^T \nabla \phi \tilde{\mathbf{W}}_c \leq (1 / 2) \mathbf{R}^{-1} \|\tilde{\mathbf{W}}_a\| \|\boldsymbol{\sigma}\| \|\mathbf{g}^T \nabla \phi\| \|\tilde{\mathbf{W}}_c\| \quad (29)$$

Applying Young's inequality with $c_3 > 0$, $\|\boldsymbol{\sigma}\| \leq \boldsymbol{\sigma}_{\max}$ and $\mathbf{G}_g = \|\mathbf{g}^T \nabla \phi\|$

$$(1 / 2) \mathbf{R}^{-1} \|\tilde{\mathbf{W}}_a\| \|\boldsymbol{\sigma}\| \mathbf{G}_g \|\tilde{\mathbf{W}}_c\| \leq (1 / 2) \mathbf{R}^{-1} \boldsymbol{\sigma}_{\max} \mathbf{G}_g \left[(1 / 2c_3) \|\tilde{\mathbf{W}}_a\|^2 + (c_3 / 2) \|\tilde{\mathbf{W}}_c\|^2 \right] \quad (30)$$

Let: $\kappa_3 = [\lambda_{\min}(\mathbf{R}) \boldsymbol{\sigma}_{\min}^2 - \mathbf{R}^{-1} \boldsymbol{\sigma}_{\max} \mathbf{G}_g / 4c_3]$ and $\beta_2 = [c_3 \mathbf{R}^{-1} \boldsymbol{\sigma}_{\max} \mathbf{G}_g / 4]$. Combining (29) and (30):

$$\dot{V}_{Actor} \leq -\kappa_3 \|\tilde{\mathbf{W}}_a\|^2 + \beta_2 \|\tilde{\mathbf{W}}_c\|^2 \quad (31)$$

Combining inequalities (24), (27), and (31) into (21), we obtain:

$$\dot{V}_L \leq -(\lambda_{\min}(\mathbf{Q}) - \kappa_2) \|\mathbf{x}\|^2 - (\beta_1 - \beta_2) \|\tilde{\mathbf{W}}_c\|^2 - (\kappa_3 - \kappa_1) \|\tilde{\mathbf{W}}_a\|^2 + (C_1 + C_2) \quad (32)$$

Where: ensures state error convergence $\Gamma_1 = \lambda_{\min}(\mathbf{Q}) - \kappa_2 > 0$, ensures Critic convergence $\Gamma_2 = \beta_1 - \beta_2 > 0$, ensures Actor convergence $\Gamma_3 = \kappa_3 - \kappa_1 > 0$ and $C = C_1 + C_2$.

$$\dot{V}_L \leq -\Gamma_1 \|\mathbf{x}\|^2 - \Gamma_2 \|\tilde{\mathbf{W}}_c\|^2 - \Gamma_3 \|\tilde{\mathbf{W}}_a\|^2 + C \quad (33)$$

The state vector is defined as: $\xi \triangleq [\mathbf{x}, \tilde{\mathbf{W}}_c, \tilde{\mathbf{W}}_a]^T$ and $\Gamma \triangleq \min\{\Gamma_1, \Gamma_2, \Gamma_3\} > 0$. Do $(u^*)^T R u^* \geq 0$:

$$\dot{V}_L \leq -\Gamma \|\xi\|^2 + C \quad (34)$$

Equation (34) is the standard Lyapunov inequality proving UUB [12], with attraction radius $r_{ub} = \sqrt{C/\Gamma}$: $\|\xi\|^2 > r_{ub}^2$: $\dot{V}_L < 0$, and the system converges toward the origin; when $\|\xi\|^2 \leq r_{ub}^2$, $\dot{V}_L$ may become positive and the trajectory oscillates within the bounded set $\Omega \triangleq \{\xi \in \mathbb{R}^n : \|\xi\| \leq r_{ub}\}$. Convergence rate: from (34), using the comparison lemma [12]:

$$V_L(t) \leq V_L(0)e^{-(2\Gamma/(\lambda_{\max}(P))t)} + r_{ub}^2(1 - e^{-(2\Gamma/(\lambda_{\max}(P))t)}) \quad (35)$$

Where P is related to the Lyapunov structure and $\lambda_{\max}(P)$ denotes the largest eigenvalue.

Therefore, all errors are bounded, satisfying the UUB property. $\square$

5. RESULTS AND DISCUSSION

5.1. Simulation setup

To evaluate the effectiveness of the ICL-RCTAC controller for an AUV model with mass 295 kg, length 2.4 m, diameter 0.386 m, and inertia matrix $I = \text{diag}[160.203, 160.203, 4.03]$. The AUV dynamic parameters are listed in Table 1.

Table 1. Dynamic parameters of the AUV in the horizontal plane.

Params	Value	Params	Value	Params	Value	Params	Value	Params	Value
$Y_{\dot{v}}$	-272.847	Y_{uv}	-37.7687	$Y_{v v }$	-696.795	N_{ur}	6511.6	$N_{\dot{r}}$	41.7501
$Y_{\dot{r}}$	654.8328	Y_{ur}	-783.5168	$Y_{r r }$	-1633.239	N_{δ_v}	30.2832	N_{uv}	2708.8
$N_{\dot{v}}$	654.8328	Y_{δ_v}	24.9861	$N_{r r }$	6520.0	$N_{v v }$	-1520.74		

The MLP neural network with Gaussian activation functions is used for both Actor and Critic, with learning parameters $\alpha_a = 5 \times 10^{-3}$, $\alpha_{ICL} = 2 \times 10^{-3}$, $\varepsilon_{ICL} = 10^{-4}$. ICL-RCTAC is compared with PID and RCTAC using the indices $ISE = \int_0^T \|e(t)\|^2 dt$, $IAE = \int_0^T \|e(t)\| dt$, and $E_c = \int_0^T \|\tau(t)\|^2 dt$. Environmental disturbances are modeled by $d(\mathbf{x}, t) = d_{\max} \sin(\omega t + \varphi)$, $\omega = \pi/10 \text{ rad/s}$, $d_{\max} = 0.5 \text{ N}$, introduced at $T_1 = 55 \text{ s}$ and $T_2 = 150 \text{ s}$ to evaluate the system's noise immunity.

5.2. Simulation results

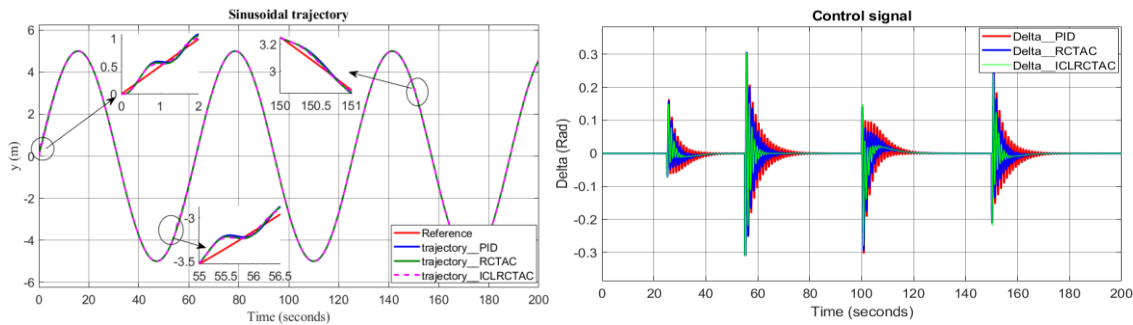


Figure 1. Tracking process and control signals over time.

Figure 1 shows ICL-RCTAC tracking the reference more closely than PID and RCTAC,

especially at $T_1 = 55s$ and $T_2 \approx 150s$, where it has the smallest deviation and fastest recovery, while PID shows the largest transient excursion. Before $t < 55s$, PID already shows pronounced spikes and RCTAC moderate oscillations, whereas ICL-RCTAC maintains smoother, smaller-peak control with the smallest error.

Figures 2 and 3 confirm that both $\mathbf{W}_a$ and $\mathbf{W}_c$ converge rapidly and remain bounded, validating the fast learning and UUB stability of ICL-RCTAC. Phase 2 benefits from the well-initialized Critic weight $\mathbf{W}_{c_0}$ certified by $\|\dot{\mathbf{W}}_c\| < \varepsilon_{ICL}$, which keeps exploiting historical data, eliminating the need for manual warm-up tuning. Quantitatively, the Critic settles at $t \approx 8s$ versus $t \approx 18s$ for RCTAC, and the Actor at $t \approx 8s$ versus $t \approx 15s$, confirming superior initial convergence.

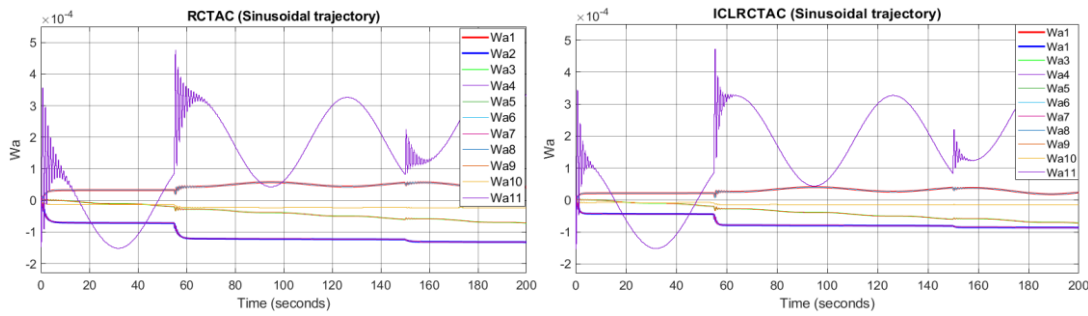


Figure 2. Weighted convergence rate of the Actor network.

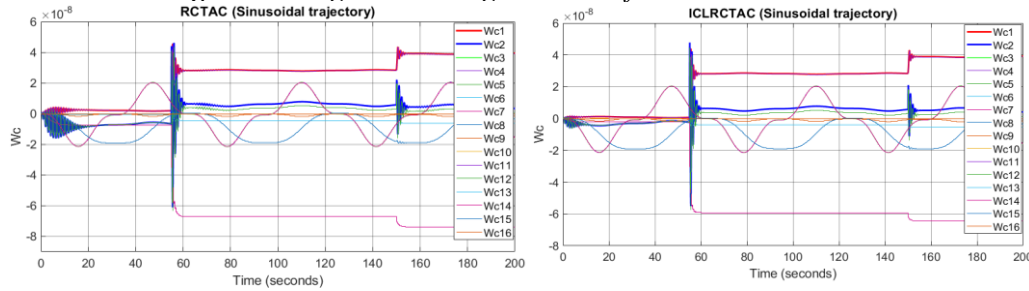


Figure 3. Weighted convergence rate of the Critic network.

Table 2 shows ICL-RCTAC achieves significantly lower ISE and IAE than both PID and RCTAC, with E_c also improved. Simulation results confirm the UUB stability of the proposed algorithm.

Table 2. Evaluation criteria.

Controller	ISE	ISE improve	IAE	IAE improve	E_c	E_c improve
PID	0.042981	0	1.537	0	0.33986	0
RCTAC	0.029903	30.428%	1.1103	27.762%	0.29801	12.313%
ICLRCTAC	0.020137	53.149%	0.86161	43.943%	0.20738	38.98%

6. CONCLUSIONS

The proposed ICL-RCTAC is a hierarchical control framework for uncertain nonlinear systems, integrating ICL with an adaptive Actor–Critic structure. Replacing the restrictive PE requirement with the FE condition removes the need for manual warm-up tuning. Simulation results show smaller tracking errors and lower control effort than PID and RCTAC. UUB analysis confirms closed-loop boundedness, with learning-rate conditions ensuring positive convergence constants. Future work will extend this framework to AUV 6DOF motion with actuator saturation constraints and experimental validation.

REFERENCES

- [1]. Dong Ma, Ye Li, et al., “The state of the art in key technologies for autonomous underwater vehicles: a review.” *Ocean Engineering*, (2025). DOI: <https://doi.org/10.1016/j.eng.2025.08.002>
- [2]. Wibisono, A., Piran, M.J., Song, H.-K., Lee, B.M., “A Survey on Unmanned Underwater Vehicles: Challenges, Enabling Technologies, and Future Research Directions,” *Sensors*, 23, 7321, (2023).
- [3]. Daoliang Li, and Ling Du, “AUV Trajectory Tracking Models and Control Strategies: A Review,” *Journal of Marine Science and Engineering*, 9, 1020, (2021). DOI: <https://doi.org/10.3390/jmse9091020>
- [4]. Bach Nhat Hoang, Doan Trung Thanh, Phan Huy Anh, and Dang Tien Sy, “Research on solutions to enhance trajectory quality in underwater autonomous navigation of unmanned underwater vehicles using IEKF SLAM algorithm for SONAR data”, *J. Mil. Sci. Technol., IITE*, 27–34, (2025). DOI: <https://doi.org/10.54939/1859-1043.j.mst.IITE.2025.27-34>
- [5]. Kairong Duan, Simon Fong, C.L. Philip Chen, “Reinforcement learning-based model-free optimized trajectory tracking strategy design for an AUV”, *Neurocomputing*, 469, 289-297, (2022).
- [6]. Nguyen Van Duc and Sai Van Cuong, “Reinforcement learning algorithm for adaptive intelligent control”, *J. Mil. Sci. Technol., CAPITI*, 62–68, (2024). DOI: <https://doi.org/10.54939/1859-1043.j.mst.CAPITI.2024.62-68>
- [7]. Duan, J., Li, S.E., Guan, Y., Sun, Q., and Cheng, B, “Relaxed actor-critic with convergence guarantees for continuous-time optimal control of nonlinear systems,” *IEEE Transactions on Intelligent Vehicles*, 8, 5, 3299–3311. DOI: <https://doi.org/10.1109/TIV.2023.3255264>
- [8]. Parikh, A., Kamalapurkar, R., and Dixon, W.E, “Integral concurrent learning: Adaptive control with parameter convergence using finite excitation,” *International Journal of Adaptive Control and Signal Processing*, 33, 12, 1775–1787. DOI: <https://doi.org/10.1002/acs.2945>
- [9]. T. I. Fossen, “*Handbook of Marine Craft Hydrodynamics and Motion Control*,” John Wiley & Sons, 2nd Edition, (2021).
- [10]. Chunbin Qin, et al., “Adaptive Dynamic Programming without Persistence Excitation for Tracking in A Special Class of Linear Systems with Input Constraints,” 33rd Chinese Control and Decision Conference, (2021). DOI: <https://doi.org/10.1109/CCDC52312.2021.9602844>
- [11]. S. Bhasin, et al., “A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems,” *Automatica*, 49, 1, 82-92, (2013).
- [12]. H. K. Khalil, “*Nonlinear Systems*,” Prentice-Hall, 3rd ed., (2002).

TÓM TẮT

**Thuật toán học phân cấp ICL-RCTAC điều khiển tối ưu bám quỹ đạo
và thích nghi với nhiễu môi trường cho AUV trong mặt phẳng ngang**

ICL-RCTAC là thuật toán học tăng cường phân cấp cho bài toán bám quỹ đạo tối ưu thích nghi của AUV trên mặt phẳng ngang, giải quyết hai hạn chế của học tăng cường thời gian liên tục — giai đoạn khởi động và điều kiện kích thích liên tục (PE) bằng khung điều khiển hai pha. Pha I sử dụng cơ chế học đồng thời tích phân (ICL) để khai thác dữ liệu lịch sử, nhằm xác định trọng số Critic ổn định dưới điều kiện kích thích hữu hạn (FE), với hội tụ được chứng minh trong Mệnh đề 1. Pha II kích hoạt vòng lặp Actor-Critic với điều kiện nói lỏng để tối ưu hóa chính sách một cách hiệu quả. Phân tích Lyapunov xác nhận rằng hệ vòng kín hội tụ đều và bị chặn (UUB). Mô phỏng MATLAB/Simulink 2024a cho thấy hội tụ nhanh hơn, sai số bám và nỗ lực điều khiển thấp hơn so với PID và RCTAC, qua đó khẳng định tiềm năng ứng dụng thực tiễn của AUV.

Từ khóa: Phương tiện ngầm tự hành; Học tăng cường; Học đồng thời tích phân; Ổn định Lyapunov; Điều khiển thích nghi; Điều khiển tối ưu.