ISA Transactions xxx (xxxx) xxx

Contents lists available at ScienceDirect

ISA Transactions

journal homepage: www.elsevier.com/locate/isatrans

Research article

Adaptive optimal trajectory tracking control of AUVs based on reinforcement learning

Zhifu Li *, Ming Wang, Ge Ma

School of Mechanical and Electrical Engineering, Guangzhou University, Guangzhou, 510006, China

ARTICLE INFO

Article history:
Received 27 February 2022
Received in revised form 16 November 2022
Accepted 2 December 2022
Available online xxxx

Keywords:
Reinforcement learning (RL)
Optimal control
Neural networks (NNs)
Autonomous underwater vehicle (AUV)
Input saturation

ABSTRACT

In this paper, an adaptive model-free optimal reinforcement learning (RL) neural network (NN) control scheme based on filter error is proposed for the trajectory tracking control problem of an autonomous underwater vehicle (AUV) with input saturation. Generally, the optimal control is realized by solving the Hamilton–Jacobi–Bellman (HJB) equation. However, due to its inherent nonlinearity and complexity, the HJB equation of AUV dynamics is challenging to solve. To deal with this problem, an RL strategy based on an actor–critic framework is proposed to approximate the solution of the HJB equation, where actor and critic NNs are used to perform control behavior and evaluate control performance, respectively. In addition, for the AUV system with the second-order strict-feedback dynamic model, the optimal controller design method based on filtering errors is proposed for the first time to simplify the controller design and accelerate the response speed of the system. Then, to solve the model-dependent problem, an extended state observer (ESO) is designed to estimate the unknown nonlinear dynamics, and an adaptive law is designed to estimate the unknown model parameters. To deal with the input saturation, an auxiliary variable system is utilized in the control law. The strict Lyapunov analysis guarantees that all signals of the system are semi-global uniformly ultimately bounded (SGUUB). Finally, the superiority of the proposed method is verified by comparative experiments.

1. Introduction

With the development of the marine economy, more and more countries pay attention to marine science and technology. Autonomous Underwater Vehicle (AUV), as a vital equipment of ocean exploration, has been widely applied to underwater tasks such as seabed mapping [1], pipeline maintenance [2], field source search [3] and so on. In these tasks, AUV needs to track the desired trajectory autonomously, so AUV trajectory tracking control has become an active research topic and has attracted wide attention. Many control algorithms have been successfully applied to AUV trajectory tracking tasks, including PID control [4], backstepping control [5], sliding mode control [6], model predictive control [7] and so on.

Due to the enormous energy consumption of underwater navigation, it is necessary to consider the optimal characteristics of AUV trajectory tracking control. Solving the optimal control problem is equivalent to solving the Hamilton–Jacobi–Bellman (HJB) equation [8]. However, the HJB equation of AUV dynamics is challenging to solve due to its inherent nonlinearity and complexity. Therefore, some of the existing optimal control algorithms [9–11] are unsuitable for AUV. As a bridge between optimal control and adaptive control, reinforcement learning (RL) is a goal-oriented learning tool, which can be used to solve optimal control problems to avoid solving HJB equations without analytical form. In this context, using RL neural network (NN) to solve the HJB equation has become a popular method. In [12–14], the actor–critic structure called adaptive dynamic programming was first proposed to create an RL-based optimal control scheme. Subsequently, in [15–19], the optimal trajectory tracking control algorithm based on the adaptive RL method is gradually developed and has become popular in recent years. However, this method is rarely used in AUV trajectory tracking control. In [20], a data-driven optimal RL control scheme with prescribed performance was created for the unmanned surface vehicle control, which simultaneously pursues control optimality and prescribed control accuracy. In [21], an optimal backstepping trajectory tracking control method for surface vessels was proposed, in which the HJB equation is solved by actor–critic approximation at each step of backstepping. In [22], a self-learning-based optimal tracking control scheme is constructed for unmanned surface vehicles with attitude and velocity constraints using adaptive RL and backstepping techniques. However, the above control methods still need partial model information.

* Corresponding author.
E-mail addresses: lizhifu8@163.com, aulzf@gzhu.edu.cn (Z. Li).

https://doi.org/10.1016/j.isatra.2022.12.003
0019-0578/

Please cite this article as: Z. Li, M. Wang and G. Ma, Adaptive optimal trajectory tracking control of AUVs based on reinforcement learning. ISA Transactions (2022), https://doi.org/10.1016/j.isatra.2022.12.003.

In practice, it is almost impossible to establish an accurate dynamic model of the nonlinear system, especially for the complex AUV system. The inertial mass and fluid dynamics are difficult to identify accurately. For this reason, integral RL is proposed as an alternative formulation to the Bellman equation to eliminate the requirement for the dynamics of uncertain systems [23,24]. In [20], the integral RL technique was used to avoid the necessity of uncertain system dynamics, but it still needs the inertia matrix information. Furthermore, actor–critical-identifier (ACI) structures have been widely used in developing optimal RL controllers [25–27], where identifiers are used to estimate the dynamics of uncertain systems. An advantage of using the ACI architecture is that the learning of actors, critics and identifier is continuous and synchronous, there is no need to understand the nonlinear dynamics of the system, and disturbances and state unmeasurability problems can be dealt with simultaneously when necessary [28].

Usually, AUVs are modeled with second-order dynamics in strict-feedback, so two primary schemes are currently used to design optimal trajectory tracking controllers based on adaptive RL. First, the second-order dynamic model is written in linear form when designing the optimal controller, and then solved [15,20]. In this scheme, the position and velocity components are combined into one vector, and then the controller is designed, which will also lead to the doubling of the order of the matrix in the design process. The other scheme is called optimal backstepping, that is, the optimal property [21,22] is considered at each step of the backstepping method. Compared with the linearization method, this method can design the controller without increasing the order of the matrix. However, it requires four NNs (i.e., two actor–critic structures) to complete the solution. In summary, although the optimal trajectory tracking problem based on adaptive RL methods for solving AUVs has been solved, the solution process is more complicated. Therefore, it is necessary to propose a new approach to simplify the design of the controller.

In addition, in practical applications, the AUV actuator has power constraints, so it is meaningful to consider the input saturation problem when designing the control algorithm. In [29,30], the regular operation of the AUV actuator was guaranteed by directly limiting the control input. However, the control truncation between the saturated and unsaturated input usually occurs, which causes potentially unstable control behavior. To deal with this problem, an additional auxiliary term is designed to compensate for control truncation between saturated and unsaturated inputs [31,32].

Inspired by the above literature, in this paper, an adaptive model-free optimal RL NN trajectory tracking control method based on filtering error is designed for the AUV system. An optimal control method based on filtering errors is proposed to simplify the design of the controller and speed up the response of the system. The actor–critic is used to solve the HJB equation, where actor and critic NNs are used to perform control behavior and evaluate control performance, respectively. Then, an extended state observer (ESO) is designed to estimate the unknown nonlinear dynamics, and an adaptive law is designed to estimate the unknown model parameters. In addition, an auxiliary variable system is utilized to deal with the input saturation. The main contributions are as follows:

(1) For the trajectory tracking control problem of the AUV, which is modeled by second-order strict-feedback dynamics, unlike the traditional error-based performance metrics, a new filtered error-based performance metric is proposed for the first time in this paper, considering both the error and the derivative of the error. It not only effectively speed up the response of the system, but also simplifies the controller design by avoiding the technical and mathematical complexity due to the use of backstepping techniques [21,22] or linearization [15,20].

(2) Unlike previous work [20–22], the model information of AUV must be known, the proposed optimal RL control method estimates the unknown nonlinear dynamics by designing an ESO, and designs an adaptive law to estimate the unknown model parameters, so that the controller is completely independent of the model information.

(3) Because of the enormous energy consumption in deep-sea long-distance navigation, this paper considers the optimal characteristics in the control of the AUV and realizes the balance between performance and cost, so it has great practical significance.

This paper is structured as follows: Section 2 covers the AUV model transformations and control objectives. The design and stability analysis of the adaptive optimal RL controller is given in Section 3. Section 4 verifies the proposed control algorithm by simulation experiments. Finally, the conclusion of this paper is summarized in Section 5.

2. Problem description

2.1. AUV systems with input saturation

The motion model of a 3-DOF AUV in the *xoy* plane is as follows [33]:

$$\begin{cases} \dot{\eta}(t) = R(\eta_z)v(t) \\ M\dot{v}(t) + C(v)v(t) + D(v)v(t) + d = \delta \end{cases} \tag{1}$$

where $\eta = [\eta_x, \eta_y, \eta_z]^T \in R^3$ denotes the position coordinate (η_x, η_y) and the yaw angle η_z of the AUV in the inertial coordinate system, respectively; $v = [v_x, v_y, v_z]^T \in R^3$ denotes the surge, sway and yaw velocities of the AUV in the body-fixed coordinate system, respectively; $\delta = [\delta_1, \delta_2, \delta_3] = sat(\delta_c) \in R^3$ is the control input with saturation, it is designed as

$$\delta_i = sat(\delta_{ci}) = \begin{cases} \delta_{imax}, & \text{if } \delta_{ci} \geq \delta_{imax} \\ \delta_{ci}, & \text{if } |\delta_{ci}| < \delta_{imax} \\ -\delta_{imax}, & \text{if } \delta_{ci} \leq -\delta_{imax} \end{cases} \tag{2}$$

where $\delta_c = [\delta_{c1}, \delta_{c2}, \delta_{c3}]^T$ is the actual controller designed later, and $\delta_{max} = [\delta_{1max}, \delta_{2max}, \delta_{3max}]^T$ denotes the upper limit of the control input that can guarantee the normal operation of the AUV. $R(\eta_z)$ is the coordinate transformation matrix satisfying $R^T(\eta_z) = R^{-1}(\eta_z)$ and $\|R(\eta_z)\| = 1$; $M \in R^{3\times3}$ is the AUV inertia matrix; $C(v) \in R^{3\times3}$ is the Coriolis centripetal matrix satisfying $C(v) = -C^T(v)$; $D(v) \in R^{3\times3}$ is the damping matrix; d is the unmodeled dynamics and external disturbances. The M, $C(v)$ and $D(v)$ have the following forms:

$$M = \begin{bmatrix} m_{11} & 0 & 0 \\ 0 & m_{22} & 0 \\ 0 & 0 & m_{33} \end{bmatrix},$$

$$C(v) = \begin{bmatrix} 0 & 0 & -m_{22}v_y \\ 0 & 0 & m_{11}v_x \\ m_{22}v_y & -m_{11}v_x & 0 \end{bmatrix},$$

$$D(v) = \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & d_{23} \\ 00 & d_{32} & d_{33} \end{bmatrix}.$$

Assumption 1. $M, C(v), D(v)$ and d are unknown but bounded.

The tracking error is defined as

$$\eta_e(t) = \eta(t) - \eta_d(t) \tag{3}$$

where $\eta_d = [\eta_{xd}, \eta_{yd}, \eta_{zd}]^T$ represents the desired trajectory. The filter tracking error is then defined as

$$s(t) = \Lambda\eta_e(t) + \dot{\eta}_e(t). \tag{4}$$

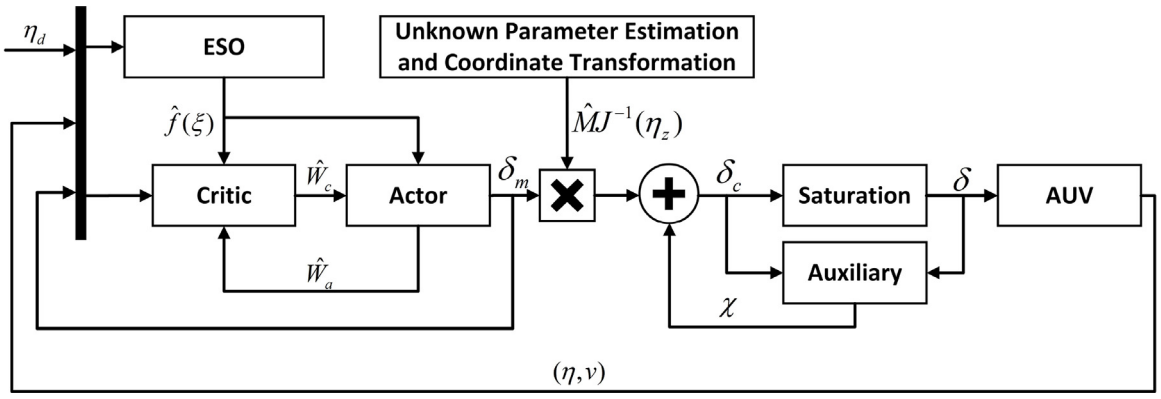

Fig. 1. Control system structure of the proposed scheme.

Let $F(\xi) = \dot{R}(\eta_z)v(t) - R(\eta_z)M^{-1}[C(v)v(t) + D(v)v(t) + d]$, then we have

$$\dot{s}(t) = \Lambda\dot{\eta}_e(t) + F(\xi) + R(\eta_z)M^{-1}\delta - \ddot{\eta}_d(t). \tag{5}$$

To facilitate the solution, let $\omega(t) = R(\eta_z)v(t)$, then (1) can be rewritten as

$$\begin{cases} \dot{\eta}(t) = \omega(t) \\ \dot{\omega}(t) = \breve{C}(v)\omega(t) + \breve{D}(v)\omega(t) + \breve{d} + \delta_m \end{cases} \tag{6}$$

where $\breve{C}(v) = -R(\eta_z)M^{-1}C(v)R^{-1}(\eta_z)$, $\breve{D}(v) = -R(\eta_z)M^{-1}D(v)R^{-1}(\eta_z) + \dot{R}(\eta_z)R^{-1}(\eta_z)$, $\breve{d} = -R(\eta_z)M^{-1}d$, $\delta_m = R(\eta_z)M^{-1}\delta$.

Define $f(\xi) = \breve{C}(v)\omega(t) + \breve{D}(v)\omega(t) + \breve{d}$, then (6) can be further expressed as

$$\begin{cases} \dot{\eta}(t) = \omega(t) \\ \dot{\omega}(t) = f(\xi) + \delta_m. \end{cases} \tag{7}$$

The new filter tracking error is then defined as

$$\breve{s}(t) = \Lambda\eta_e(t) + \dot{\eta}_e(t). \tag{8}$$

In addition

$$\dot{\breve{s}}(t) = \Lambda\dot{\eta}_e(t) + f(\xi) + \delta_m - \ddot{\eta}_d(t). \tag{9}$$

Assumption 2. The unknown dynamics $f(\xi)$ and its derivative $\dot{f}(\xi)$ are bounded, that is, there exist positive constants γ_f and $\gamma_{\dot{f}}$ such that $\|f(\xi)\| \le \gamma_f$ and $\|\dot{f}(\xi)\| \le \gamma_{\dot{f}}$.

Remark 1. $s(t)$ and $\breve{s}(t)$ are numerically identical. Still, their first derivatives have different forms, which are used in the stability analysis of the two models, where the intermediate control law δ_m is related to $\breve{s}(t)$, and the final control law δ is related to $s(t)$.

2.2. Control objective

Based on the filtering error, an optimal RL NN controller completely independent of the model information is designed for the AUV such that (1) the AUV follows the desired trajectory η_d and (2) all error signals are semi-global uniformly ultimately bounded (SGUUB).

3. Optimal reinforcement learning neural network controller design

In this section, an ESO is first designed to estimate the unknown dynamics $f(\xi)$ of the system. Secondly, an optimal intermediate control law is designed based on the transformed model (6), in which the critic and actor NN are used to evaluate the control performance and execute the control behavior, respectively. Then, by inverse transformation and designing adaptive law for unknown model parameters, the control law which acts on the original system completely independent of the model information is obtained. Furthermore, the problem of input constraints is considered, and an auxiliary system is used to deal with the control truncation between saturation and unsaturation. The controller structure is shown in Fig. 1.

3.1. Extended state observer design

In order to estimate the unknown dynamics $f(\xi)$, based on system (7), the following ESO is designed as

$$\begin{cases} \tilde{\eta} = x_1 - \hat{x}_1 \\ \dot{\hat{x}}_1 = \hat{x}_2 + \beta_1\tilde{\eta} \\ \dot{\hat{x}}_2 = \hat{x}_3 + \beta_2\tilde{\eta} + \delta_m \\ \dot{\hat{x}}_3 = \beta_3\tilde{\eta} \end{cases} \tag{10}$$

where $\beta_{i=1,2,3} \in R^{3\times 3}$.

According to the ESO (10), combined with the AUV model (7) we get

$$\hat{\eta} = \hat{x}_1, \quad \hat{\omega} = \hat{x}_2, \quad \hat{f}(\xi) = \hat{x}_3$$

where $\hat{\eta}$ is an estimate of the AUV position state vector η, $\hat{\omega}$ is an estimate of the AUV conversion velocity state vector ω, and $\hat{f}(\xi)$ is an estimate of the AUV system unknown dynamics $f(\xi)$.

Let $\tilde{x}_1 = x_1 - \hat{x}_1$, $\tilde{x}_2 = x_2 - \hat{x}_2$ and $\tilde{x}_3 = x_3 - \hat{x}_3$, according to (7) and (10), the error dynamics are defined as

$$\begin{cases} \dot{\tilde{x}}_1 = \tilde{x}_2 - \beta_1\tilde{x}_1 \\ \dot{\tilde{x}}_2 = \tilde{x}_3 - \beta_2\tilde{x}_1 \\ \dot{\tilde{x}}_3 = \dot{f}(\xi) - \beta_3\tilde{x}_1. \end{cases} \tag{11}$$

Define $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \tilde{x}_3]^T$, and the error dynamics (11) can be rewritten in a compact form that

$$\dot{\tilde{X}} = A(e)\tilde{X} + B\dot{f}(\xi) \tag{12}$$

where

$$A(e) = \begin{bmatrix} -\beta_1 & I & 0 \\ -\beta_2 & 0 & I \\ -\beta_3 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ I \end{bmatrix}.$$

Then the following theorem can be given.

Theorem 1. *Considering the closed-loop system consisting of the AUV system (7) and the proposed ESO (10) with Assumptions 1 and 2, if there is a positive definite matrix Γ satisfying:*

$$A\Gamma + \Gamma A^T = -\lambda I \tag{13}$$

and choosing the positive design parameters λ *and* $\beta_{i=1,2,3}$ *appropriately with satisfying:*

$$\lambda > \gamma_f \tag{14}$$

$$\beta_1\beta_2 > \beta_3 \tag{15}$$

then error signals $\tilde{x}_1$, $\tilde{x}_2$ *and* $\tilde{x}_3$ *are SGUUB.*

Proof. See Appendix A.

3.2. Controller design

For the converted system (6), the long-term performance index is defined as

$$J(\check{s}) = \int_t^{\infty} e^{-\gamma(\varsigma-t)} r\big(\check{s}(\varsigma), \delta_m(\check{s})\big) d\varsigma \tag{16}$$

where γ is the discount factor, and $r(\check{s}, \delta_m) = \check{s}^T(t)Q\check{s}(t) + \delta_m^T(\check{s})\delta_m(\check{s})$ is the cost function at time t, and Q is a constant matrix that can be traded off between control performance and control cost. The optimal control problem is to solve the control strategy $\delta_m \in \Omega$ such that the long-term performance index (16) is minimized [34].

Remark 2. The boundedness of the long-term performance index (16) is guaranteed by introducing a discount factor γ. In general, $\gamma = 1$ can be used only if it is known in advance that the reference trajectory η_d is generated by an asymptotically stable command system. However, if the reference trajectory η_d is a general bounded signal, γ needs to satisfy $0 < \gamma < 1$ to ensure the boundedness of the long-term performance index (16).

Based on the long-term performance index (16), the Hamiltonian function can be obtained as

$$H(\check{s}, \delta_m, J_{\check{s}}) = r(\check{s}, \delta_m) + J_{\check{s}}^T(\check{s})\dot{\check{s}}(t) - \gamma J(\check{s}) \tag{17}$$

where $J_{\check{s}}(\check{s}) = \partial J(\check{s})/\partial \check{s}$ is the gradient of $J(\check{s})$ versus $\check{s}$.

The optimal long-term performance metric is defined as

$$\begin{aligned} J^*(\check{s}) &= \min_{\delta_m \in \Psi(\Omega)} \int_t^{\infty} e^{-\gamma(\varsigma-t)} r\big(\check{s}(\varsigma), \delta_m(\check{s})\big) d\varsigma \\ &= \int_t^{\infty} e^{-\gamma(\varsigma-t)} r\big(\check{s}(\varsigma), \delta_m^*(\check{s})\big) d\varsigma \end{aligned} \tag{18}$$

where δ_m^* is the optimal control law, $\Psi(\Omega)$ is the set of control policies on Ω that satisfy the control performance, and $\Omega \in R^3$ is a compact set.

Based on (9), (17), (18) and ESO (10), we get

$$\begin{aligned} H(\check{s}, \delta_m^*, J_{\check{s}}^*) &= r(\check{s}, \delta_m^*) + J_{\check{s}}^{*T}(\check{s})\dot{\check{s}}(t) - \gamma J^*(\check{s}) \\ &= \check{s}^T(t)Q\check{s}(t) + \delta_m^{*T}\delta_m^* + J_{\check{s}}^{*T}(\check{s}) \big[\Lambda\dot{\eta}_e(t) + \hat{f}(\xi) \\ &\quad + \delta_m^* - \ddot{\eta}_d(t) \big] - \gamma J^*(\check{s}) = 0. \end{aligned} \tag{19}$$

Then, the optimal control law δ_m^* can be implemented by solving $\partial H(\check{s}, \delta_m^*, J_{\check{s}}^*)/\partial \delta_m^* = 0$, thus we have

$$\delta_m^* = -\frac{1}{2} J_{\check{s}}^*(\check{s}). \tag{20}$$

Substituting (20) into (19), we obtain

$$\begin{aligned} H(\check{s}, \delta_m^*, J_{\check{s}}^*) &= \lambda(Q)\|\check{s}\|^2 + J_{\check{s}}^{*T}\big[\Lambda\dot{\eta}_e(t) + \hat{f}(\xi) - \ddot{\eta}_d(t)\big] \\ &\quad - \frac{1}{4} J_{\check{s}}^{*T}(\check{s}) J_{\check{s}}^*(\check{s}) - \gamma J^*(\check{s}) = 0 \end{aligned} \tag{21}$$

where $\lambda(Q)$ represents any eigenvalue of the matrix Q.

The optimal control law δ_m^* can be obtained by combining Eqs. (20) and (21). However, the high nonlinearity and complexity of the AUV system make this equation difficult to solve. Therefore, the classical framework of RL, actor–critic NN, is used for online learning to obtain δ_m^*.

To achieve the desired control performance, the optimal value function (18) is rewritten as

$$\begin{aligned} J^*(\check{s}) &= K_{\check{s}}\|\check{s}(t)\|^2 - K_{\check{s}}\|\check{s}(t)\|^2 + J^*(\check{s}) \\ &= K_{\check{s}}\|\check{s}(t)\|^2 + J^o(\check{s}) \end{aligned} \tag{22}$$

where $K_{\check{s}}$ is a positive design constant and $J^o(\check{s}) = J^*(\check{s}) - K_{\check{s}}\|\check{s}\|^2$.

Remark 3. The Eq. (22) decomposes the term $K_{\check{s}}\|\check{s}(t)\|^2$ to better realize the tracking control of the system. Although there are many optimal control methods, such as [17,35,36], which can guarantee state boundedness and system stability, there are few research results to solve the trajectory tracking control problem. In the design, the desired tracking performance can be obtained by decomposing the term $K_{\check{s}}\|\check{s}(t)\|^2$ from the optimal value function and selecting an appropriate parameter $K_{\check{s}}$. The method can also be easily extended to higher dimensional systems by replacing the $K_{\check{s}}\|\check{s}(t)\|^2$ term with a norm expression.

Then (20) can be rewritten as

$$\delta_m^* = -K_{\check{s}}\check{s}(t) - \frac{1}{2} J_{\check{s}}^o(\check{s}) \tag{23}$$

where $J_{\check{s}}^o(\check{s}) = \partial J^o(\check{s})/\partial \check{s}$ is the gradient of $J^o(\check{s})$ versus $\check{s}$.

NNs can approximate unknown nonlinear functions with arbitrary accuracy on compact set Ω_z [37]. Here, the function $J^o(\check{s})$ is approximated by NN and defined as

$$J^o(\check{s}) = W^{*T}\varphi(\check{s}) + \epsilon(\check{s}) \tag{24}$$

where $W^* \in R^N$ is the ideal weight vector and N is the number of neurons; $\varphi(\check{s})$ is the basis function vector; $\epsilon(\check{s})$ is an approximation error. According to [37], there exist two unknown positive constants d_W and d_ϵ such that $\|W^*\| \le d_W$ and $\|\epsilon(\check{s})\| \le d_\epsilon$.

Based on (24), $J^*(\check{s})$ and δ_m^* can be rewritten as

$$J^*(\check{s}) = K_{\check{s}}\|\check{s}(t)\|^2 + W^{*T}\varphi(\check{s}) + \epsilon(\check{s}) \tag{25}$$

$$\delta_m^* = -K_{\check{s}}\check{s}(t) - \frac{1}{2}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}} W^* - \frac{1}{2}\frac{\partial\epsilon(\check{s})}{\partial\check{s}} \tag{26}$$

where $\partial^T\varphi(\check{s})/\partial\check{s} \in R^{3\times N}$ and $\partial\epsilon(\check{s})/\partial\check{s} \in R^3$ are the gradients of $\varphi(\check{s})$ and $\epsilon(\check{s})$ to $\check{s}$, respectively.

Substituting (25) and (26) into (19), we have

$$\begin{aligned} H(\check{s}, \delta_m^*, W^*) = &- \big(K_{\check{s}}^2 - \lambda(Q) + \gamma K_{\check{s}}\big)\|\check{s}(t)\|^2 \\ &- 2K_{\check{s}}\check{s}^T(t)\big[-\Lambda\dot{\eta}_e(t) - \hat{f}(\xi) + \ddot{\eta}_d(t)\big] \\ &- W^{*T}\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\big[K_{\check{s}}\check{s}(t) - \Lambda\dot{\eta}_e(t) - \hat{f}(\xi) + \ddot{\eta}_d(t)\big] \\ &- \frac{1}{4}\left\|\frac{\partial^T\varphi(\check{s})}{\partial\check{s}} W^*\right\|^2 \\ &- \gamma W^{*T}\varphi(\check{s}) + \rho(t) \end{aligned} \tag{27}$$

where $\rho(t) = \big(\partial\epsilon(\check{s})/\partial\check{s}^T\big)\delta_m^* + (1/4)\|\partial\epsilon(\check{s})/\partial\check{s}\|^2 + \big(\partial\epsilon(\check{s})/\partial\check{s}^T\big)\big[\Lambda\dot{\eta}_e(t) + \hat{f}(\xi) - \ddot{\eta}_d(t)\big] - \gamma\epsilon(\check{s})$.

Since the optimal weight matrix W^* is unknown, $\hat{W}$ is used to estimate W^*. And perform that optimal RL NN algorithm by constructing the following critic and actor NN:

$$\hat{J}^*(\check{s}) = K_{\check{s}}\|\check{s}(t)\|^2 + W_c^T\varphi(\check{s}) \tag{28}$$

$$\delta_m = -K_{\check{s}}\check{s}(t) - \frac{1}{2}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a(t) \tag{29}$$

where $\hat{J}^*(\check{s})$ represents an estimate of $J^*(\check{s})$ and δ_m represents an estimate of δ_m^*; $\hat{W}_c$ and $\hat{W}_a$ represent the weight vectors of the

critic and actor NN, respectively. The weight error is defined as

$$\tilde{W}_c = \hat{W}_c - W^*, \tag{30}$$

$$\tilde{W}_a = \hat{W}_a - W^*. \tag{31}$$

Substituting (28), (29) into (19) to obtain the approximate HJB equation

$$\begin{aligned} &H(\check{s}, \delta_m, \hat{W}) \\ &= [\lambda(Q) - \gamma K_{\check{s}}]\|\check{s}(t)\|^2 + \left\| K_{\check{s}}\check{s}(t) + \frac{1}{2}\frac{\partial^T \varphi(\check{s})}{\partial \check{s}}\hat{W}_a(t) \right\|^2 \\ &\quad - \left[2K_{\check{s}}\check{s}(t) + \frac{\partial^T \varphi(\check{s})}{\partial \check{s}}\hat{W}_c(t) \right]^T \\ &\quad \times \left[K_{\check{s}}\check{s}(t) + \frac{1}{2}\frac{\partial^T \varphi(\check{s})}{\partial \check{s}}\hat{W}_a(t) - \Lambda\dot{\eta}_e(t) - \hat{f}(\xi) + \ddot{\eta}_d(t) \right] \\ &\quad - \gamma \hat{W}_c(t)\varphi(\check{s}). \end{aligned} \tag{32}$$

Bellman error is defined as

$$e_c(t) = H(\check{s}, \delta_m, \hat{W}) - H(\check{s}, \delta_m^*, J_{\check{s}}^*) = H(\check{s}, \delta_m, \hat{W}). \tag{33}$$

Let $E_c(t) = (1/2)e_c^2(t)$, then the critic NN updating law that leads to the minimum Bellman error is designed as

$$\begin{aligned} &\dot{\hat{W}}_c(t) \\ &= -\frac{l_c}{1 + \|\phi(t)\|^2} e_c(t) \frac{\partial e_c(t)}{\partial \hat{W}_c(t)} \\ &= -\frac{l_c}{1 + \|\phi(t)\|^2}\phi(t)\left\{ \phi^T(t)\hat{W}_c(t) - \left[K_{\check{s}}^2 - \lambda(Q) + \gamma K_{\check{s}}\right]\|\check{s}(t)\|^2 \right. \\ &\quad \left. - 2K_{\check{s}}\check{s}^T(t)\left[-\Lambda\dot{\eta}_e(t) - \hat{f}(\xi) + \ddot{\eta}_d(t)\right] + \frac{1}{4}\left\| \frac{\partial^T \varphi(\check{s})}{\partial \check{s}}\hat{W}_a(t) \right\|^2 \right\} \end{aligned} \tag{34}$$

where $l_c > 0$ is the learning rate of the critic NN; $\phi(t) = -(\partial\varphi(\check{s})/\partial\check{s}^T) \times \left[K_{\check{s}}\check{s}(t) + (1/2)(\partial^T\varphi(\check{s})/\partial\check{s})\hat{W}_a(t) - \Lambda\dot{\eta}_e(t) - \hat{f}(\xi) + \ddot{\eta}_d(t)\right] - \gamma\varphi$.

And the updating law of actor NN is designed as

$$\begin{aligned} \dot{\hat{W}}_a(t) =& \frac{1}{2}\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\check{s}(t) - l_a\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a(t) \\ &+ \frac{l_c}{4\left(1 + \|\phi(t)\|^2\right)}\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a(t)\phi^T(t)\hat{W}_c(t) \end{aligned} \tag{35}$$

where $l_a > 0$ is the learning rate of the actor NN.

Assumption 3 (*[38] Persistent Excitation (PE)*)**.** There are constants $T > 0$, $\overline{\phi} > 0$, $\underline{\phi} > 0$ in the interval $[t, t+T]$, for all t, if the following inequality holds

$$\underline{\phi} I_3 \le \int_t^{t+T} \phi(\vartheta)\phi^T(\vartheta)d\vartheta \le \overline{\phi} I_3 \tag{36}$$

then the signal $\phi\phi^T$ is said to be continuously excited in the interval $[t, t+T]$, where $I_3 \in R^{3\times 3}$ is identity matrix.

Remark 4. The PE condition in Assumption 3 is to ensure stable control performance. In the proof of Theorem 2, there is a Lyapunov function (B.1) containing the term $(1/2)\tilde{W}_c^T(t)\tilde{W}_c(t)$. By calculating the time derivative along the critic update law (34), the term $\phi(t)\phi^T(t)$ is obtained to occur until (B.5). Then, based on the PE assumption, the inequality (B.6) is obtained, so that Lemma 1 can be applied. Finally, the boundedness of the optimal control is proved. An exploratory signal consisting of sine waves of different frequencies can be added to the control input to ensure PE qualitatively.

Remark 5. The HJB should satisfy $H(\check{s}, \delta_m, \hat{W}) \to H(\check{s}, \delta_m^*, W^*) \to 0$ when the control system is optimal, i.e. $\delta_m \to \delta_m^*$. Therefore, to ensure that the system can achieve the optimal, the Bellman residual error of the critic NN is defined as $e_c(t) = H(\check{s}, \delta_m, \hat{W})$, and the critic NN update law (34) is derived by calculating the negative gradient $\dot{\hat{W}}_c(t) = -\frac{l_c}{1+\|\phi(t)\|^2}\frac{\partial E_c(t)}{\partial \hat{W}_c(t)}$, to ensure that $H(\check{s}, \delta_m, \hat{W}) \to 0$, that is, the system achieves the optimal. The updated law of the actor NN (35) is then derived based on the stability analysis.

Considering Assumption 1, through the inverse transformation and the design of an adaptive law for the unknown parameters M, the desired control law δ_d, which acts on the system (1) completely independent of the model information, is obtained as

$$\delta_d = \hat{M}R^{-1}(\eta_z)\delta_m \tag{37}$$

where $\hat{M} = diag(\hat{m}_{11}, \hat{m}_{22}, \hat{m}_{33})$ represents the estimation of the unknown parameter M, the adaptive law of $\hat{m}_{ii}$ is designed as

$$\dot{\hat{m}}_{ii} = -\Psi_i(\delta_{mi}s_i + \Omega_i\hat{m}_{ii}), \quad i = 1, 2, 3 \tag{38}$$

where Ψ_i and Ω_i are design constants. Then the estimation error $\tilde{M} = diag(\tilde{m}_{11}, \tilde{m}_{22}, \tilde{m}_{33})$ of M is defined as

$$\tilde{M} = \hat{M} - M. \tag{39}$$

Considering the input saturation (2), we define

$$\Delta\delta = \delta - \delta_c. \tag{40}$$

Then, the auxiliary system is defined as

$$\dot{\chi} = \begin{cases} -K_\chi\chi - \frac{2\left\|s^T\Delta\delta\right\| + 3\Delta\delta^T\Delta\delta}{\|\chi\|^2}\chi + 2\Delta\delta, & \text{if } \|\chi\| \ge \kappa \\ 0, & \text{if } \|\chi\| < \kappa \end{cases} \tag{41}$$

where $K_\chi > 0$ is a constant matrix and $\kappa > 0$ is a small constant.

Therefore, the final control law acting on the AUV system (1) is given as

$$\begin{aligned} \delta_c &= \delta_d + \chi \\ &= -\hat{M}R^{-1}(\eta_z)\left[K_{\check{s}}\check{s}(t) + \frac{1}{2}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a(t)\right] + \chi. \end{aligned} \tag{42}$$

3.3. Main results

Theorem 2. *Consider the AUV system (1) under Assumptions 1–3, the critic NN (28), and the actor NN (29), the adaptive optimal RL NN controller (42) has updating laws (10), (34), (35), (38) and (41), with the bounded initial conditions, the system error signal is SGUUB. The condition is that there is a positive definite matrix Γ which satisfies (13), and the parameters λ and $\beta_{i=1,2,3}$ are properly designed according to (15). And error signals η_e, s, $\tilde{W}_a$, $\tilde{W}_c$ and $\tilde{M}$ converge to compact sets Ω_1, Ω_2, Ω_3, Ω_4 and Ω_5, which are defined as*

$$\Omega_1 = \left\{ \eta_e \in R^n \middle| \|\eta_e\| \le \max(\sqrt{\Pi}, \sqrt{\Xi}) \right\} \tag{43}$$

$$\Omega_2 = \left\{ s \in R^n \middle| \|s\| \le \sqrt{\Pi} \right\} \tag{44}$$

$$\Omega_3 = \left\{ \tilde{m}_{ii} \in R^n \middle| \sum_{i=1}^{3} m_{ii}^{-1}\Psi_i^{-1}\tilde{m}_{ii}^2 \le \Pi \right\} \tag{45}$$

$$\Omega_4 = \left\{ \tilde{W}_a \in R^n \middle| \|\tilde{W}_a\| \le \sqrt{\Xi} \right\} \tag{46}$$

$$\Omega_5 = \left\{ \tilde{W}_c \in R^n \middle| \|\tilde{W}_c\| \le \sqrt{\Xi} \right\} \tag{47}$$

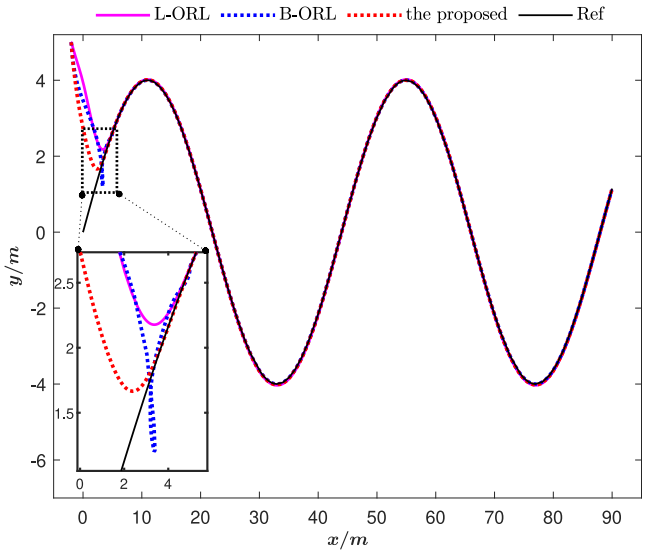

Fig. 2. Desired and actual trajectories of the AUV compared to L-ORL and B-ORL.

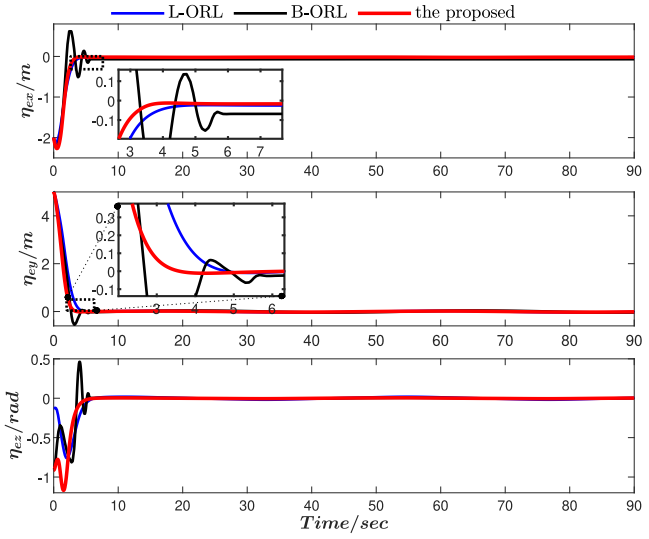

Fig. 3. Tracking error of the AUV compared to L-ORL and B-ORL.

where $\Pi = 2(\zeta + V_2(0))$, $\Xi = 2(\kappa + V_1(0))$. *In addition, the selection of the parameters* Λ, $K_{\check{s}}$, K_χ, l_a *and* l_c *should satisfy*

$$\Lambda > I, \quad K_{\check{s}} > \frac{7 + 2\varrho^2 + \|\Lambda\|_F^2}{2}, \quad K_\chi > \frac{5}{2} I,$$

$$l_a > l_c^2 + \frac{\bar{\phi}}{16} W^{*T} W^*, \quad l_c > \frac{1}{16} \sup_{t \geq 0} \left\{ W^{*T} \frac{\partial \varphi(\check{s})}{\partial \check{s}^T} \frac{\partial^T \varphi(\check{s})}{\partial \check{s}} W^* \right\}.$$

Proof. See Appendix B.

4. Numerical simulation

The specific parameters of the model are as follows: $m_{11} = 25.8$, $m_{22} = 24.6612$, $m_{33} = 2.76$, $d_{11} = 0.7225 + 1.3274|v_x| + 5.8664v_y^2$, $d_{22} = 0.8612 + 36.2823|v_y| + 8.05|v_z|$, $d_{23} = -0.1079 + 0.845|v_y| + 3.45|v_z|$, $d_{32} = -0.1052 - 5.0437|v_y| - 0.13|v_z|$, $d_{33} = 1.9 - 0.08|v_y| + 0.75|v_z|$, and $d = [0, 0.01v_x^2 + 0.5, -0.1v_z^3 + \sin(v_y)]^T$.

The desired tracking trajectory signal is $\eta_d(t) = [t, 4\sin(\frac{t}{7}), arctan(\frac{4}{7}\cos(\frac{t}{7}))]^T$. The initial position of the AUV is set to be $\eta_0 = [-2, 5, -\pi/8]^T$. The discount factor $\gamma = 0.6$. The ESO gain $\beta_1 = diag(6, 6, 6)$, $\beta_2 = diag(8, 8, 8)$, $\beta_3 = diag(16, 16, 16)$. The upper limit of the control input is represented as $\delta_{max} = [80, 80, 5]^T$. The number of hidden neurons in the NNs is selected as 12, and $\hat{W}_c(0) = [0.1, \ldots, 0.1]^T$, $\hat{W}_a(0) = [0.1, \ldots, 0.1]^T$. The control parameters are properly selected as: $\Lambda = diag(1.2, 1.2, 1.2)$, $K_\chi = diag(30, 30, 30)$, $\Psi_i = 0.1$, $\Omega_i = 0.01$, $l_c = 0.1$, $l_a = 0.3$, $K_{\check{s}} = 6$. In addition, we ensure continuous excitation by adding the exploration signal $n(t) = [0.3\sin(8t)^2\cos(2t) + 0.3\sin(20t)^2\cos(7t), 0.2\sin(6t)^2\cos(4t) + 0.3\sin(12t)^2\cos(5t), 0.2\sin(8t)^2\cos(6t) + 0.1\sin(8t)^2\cos(3t)]^T$ to the control input.

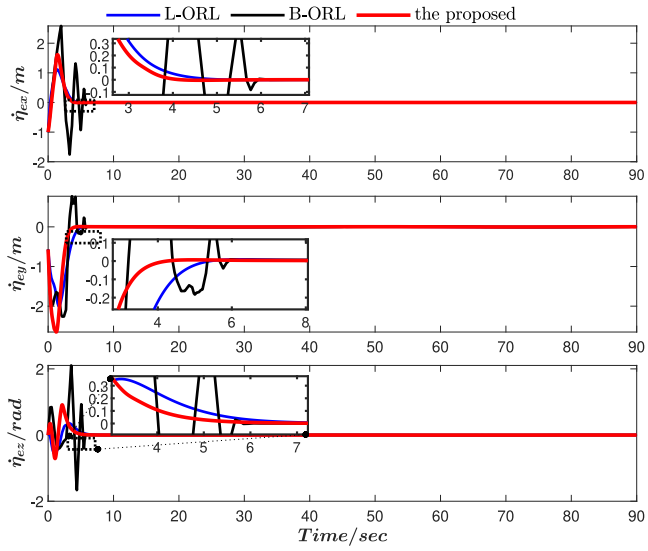

Fig. 4. Tracking error derivative of the AUV compared to L-ORL and B-ORL.

To illustrate the effectiveness of the proposed optimal RL NN controller, we compared the proposed method with the linearization-based optimal RL control method [15] (L-ORL) and the backstepping-based optimal RL control method [21] (B-ORL) for the same control input limits. The comparison tracking results are shown in Fig. 2. The tracking errors for the three controllers are shown in Fig. 3. The time response of the error derivative is shown in Fig. 4. As can be seen in Figs. 3 and 4, since the proposed control method considers both the error and the derivative of the error, it reduces the error and the derivative of the error at the same time, further reducing the fluctuation time of the error. Therefore, it can be seen from Fig. 2 that the proposed control method has a fast response time. Fig. 5 shows the integrated tracking error $z = \|\eta_{ex}, \eta_{ey}, \eta_{ez}\|$, and it can be intuitively seen that the proposed optimal RL NN method has the advantage of high tracking accuracy. Furthermore, to quantitatively evaluate the tracking performance of different control methods, the steady state Integral Absolute Error (IAE) criterion [39] is defined as follows:

$$\text{IAE} = \int_{t_0}^{t} |e(\varsigma)| d\varsigma \tag{48}$$

where t_0 represents the adjustment times, $T = 90$ s is the simulation times and $t_0 \leq t \leq T$. The IAE values of the three controllers are given in Table 1. It can be seen that the performance of the proposed optimal RL NN controller is significantly improved compared to L-ORL and B-ORL, with IAE reductions of approximately 44.6% and 255.4% for error η_{ex}. Similarly, for that error η_{ey}, the IAE is reduced by about 53.6% and 79.5%; About 241.8% and 17.4% reduction in IAE for error η_{ez}.

Fig. 6 shows the curve of the speed state and its desired value over time, and it can be seen that the actual state can well

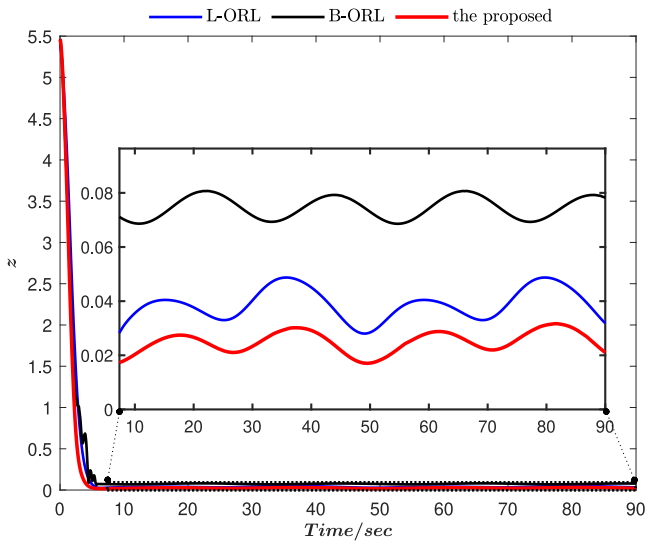

Fig. 5. Comprehensive tracking error z.

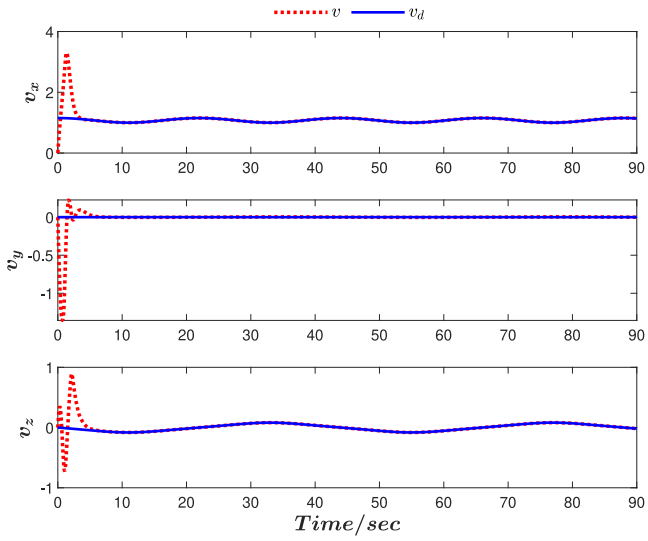

Fig. 6. Desired and actual state in v_x, v_y and v_z.

Table 1
IAE values of the proposed scheme and other two controllers.

Method	IAE		
	η_x [m]	η_y [m]	η_z [rad]
The proposed scheme	1.547	1.169	0.2477
L-ORL	2.237	1.795	0.8466
B-ORL	5.498	2.098	0.2907

Table 2
IAE values of the proposed scheme and other two controllers with sensor noise.

Method	IAE		
	η_x [m]	η_y [m]	η_z [rad]
The proposed scheme	1.724	1.382	0.3429
L-ORL	2.389	1.991	0.8964
B-ORL	5.908	2.411	0.3757

track the desired state. Fig. 7 shows the control input curve of the proposed scheme. It can be seen that the proposed method ensures the regular operation of the AUV when the control input does not violate the constraint value in the initial stage when the error is large. Fig. 8 shows the curves of the weight norms of the actor and critic NNs varying with time, which shows that they are bounded. Fig. 9 shows the cost function of the proposed scheme, the cost function $\hat{J}^*(\check{s})$ can converge to a small value in about 3 seconds. The curve of the HJB equation $H(\check{s}, \delta_m, \hat{W})$ versus time is given in Fig. 10, and it can be seen that it converges to zero, indicating that the system is optimal.

Remark 6. To verify that if the cost function contains the error derivative, the tracking performance will degrade when there is significant noise in the sensor channel, we add noise to the state η when $50 < t < 55$ to do further simulation experiments. The integrated error time response curves and IAE values of the three control algorithms in the presence of sensor noise are given in Fig. 11 and Table 2, respectively. It can be seen that the proposed algorithm can still maintain better tracking performance than the other two algorithms in the presence of noise.

5. Conclusion

An adaptive model-free optimal RL NN control scheme based on filtering error is proposed for the trajectory tracking control

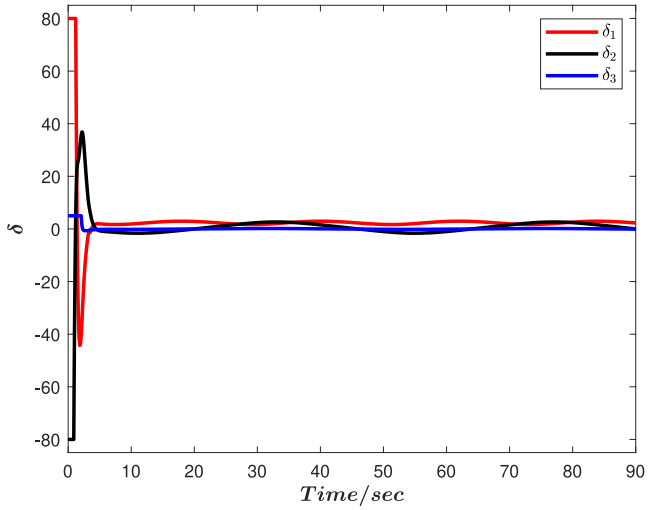

Fig. 7. The control input δ of the proposed method.

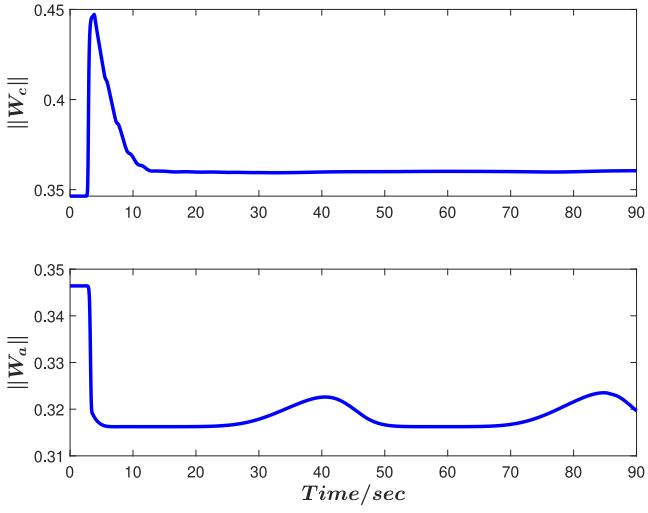

Fig. 8. Norms of actor and critic NNs weights.

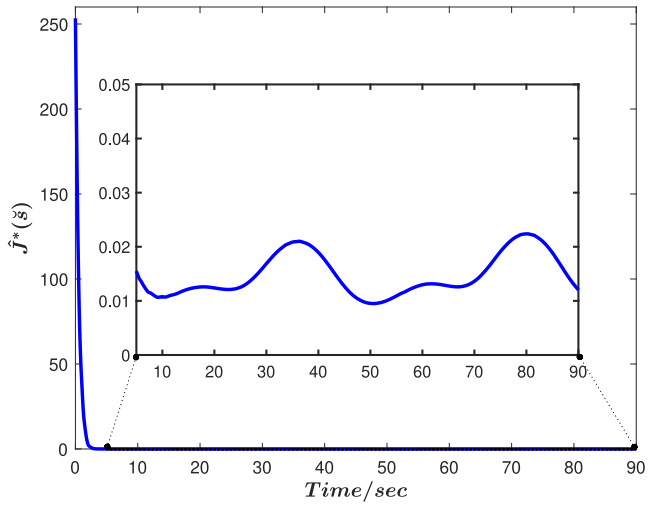

Fig. 9. Approximation cost function $\hat{J}^*(\check{s})$.

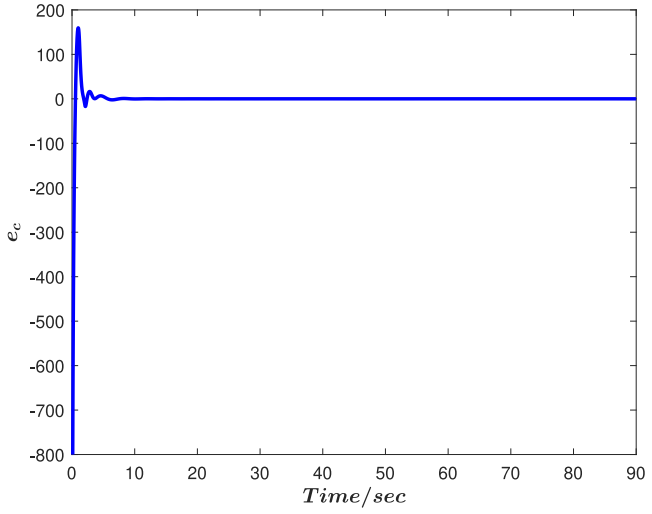

Fig. 10. Bellman error e_c or HJB equation $H(\check{s}, \delta_m, \hat{W})$.

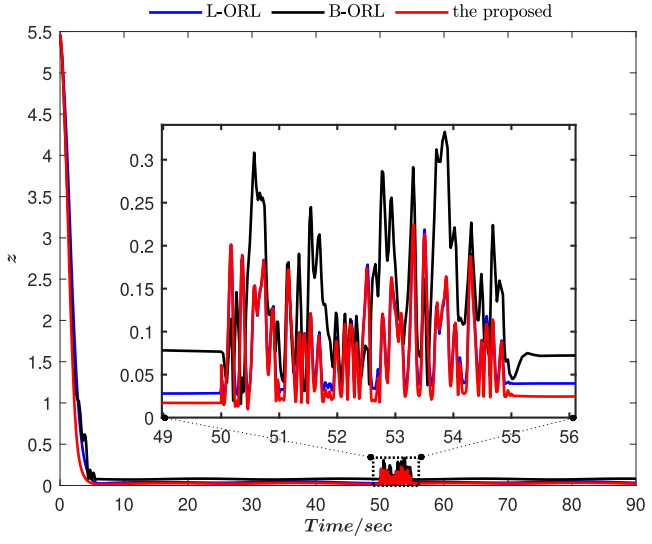

Fig. 11. Comprehensive tracking error z with sensor noise.

problem of AUV with input saturation and completely unknown model information. The proposed performance index based on the filtering error considers both the error and the error derivative, which not only simplifies the design of the controller, but also speeds up the response of the system. In the controller design, the AUV model is transformed first, and then the optimal RL NN control law is designed based on the transformed model using actor–critic structure, in which actor NN and critic NN is used for approximate control strategy and long-term performance index, respectively. Then, to solve the model-dependent problem, an ESO is designed to estimate the unknown nonlinear dynamics, and an adaptive law is designed to estimate the unknown model parameters. Furthermore, the input saturation problem is considered, and an auxiliary variable is designed to embed the control law to deal with the control truncation, to ensure the regular operation of the AUV. It is proved that the system error signal is SGUUB. Finally, a simulation example is given to verify the proposed control method.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 62173102.

Appendix A. Proof of Theorem 1

To facilitate the proof, the following Lemma is provided.

Lemma 1 ([40]). *For a continuous function $V(t) \geq 0$ ($\forall t \in R^+$), if $V(0)$ is bounded and $\dot{V}(t) \leq -z_1 V(t) + z_2$, where z_1, z_2 are constants, then have:*

$$V(t) \leq e^{-z_1 t} V(0) + \frac{z_2}{z_1}(1 - e^{-z_1 t}). \tag{A.1}$$

The Lyapunov function is chosen as follows:

$$V_o(\tilde{X}) = \frac{1}{2}\tilde{X}^T \Gamma \tilde{X}. \tag{A.2}$$

Based on (12), the time derivative of (A.2) can be expressed as

$$\dot{V}_o(\tilde{X}) = \frac{1}{2}\tilde{X}^T(A\Gamma + \Gamma A^T)\tilde{X} + \tilde{X}^T \Gamma B \dot{f}(\xi). \tag{A.3}$$

Then, based on (A.3) can be further written as

$$\begin{aligned}\dot{V}_o(\tilde{X}) &\leq -\frac{\lambda}{2}\|\tilde{X}\|^2 + \|\tilde{X}^T \Gamma B\|\gamma_{\dot{f}} \leq -\frac{\lambda}{2}\|\tilde{X}\|^2 + \frac{\gamma_{\dot{f}}}{2}\|\tilde{X}\|^2 + \frac{\gamma_{\dot{f}}}{2}\|\Gamma\|_F^2 \\ &= -\frac{1}{2}(\lambda - \gamma_{\dot{f}})\|\tilde{X}\|^2 + \frac{\gamma_{\dot{f}}}{2}\|\Gamma\|_F^2 = -\hbar V_o(\tilde{X}) + \upsilon\end{aligned}$$

where

$$\hbar \triangleq \frac{\lambda - \gamma_{\dot{f}}}{\lambda_{min}(\Gamma^{-1})}, \quad \upsilon \triangleq \frac{\gamma_{\dot{f}}}{2}\|\Gamma\|_F^2. \tag{A.4}$$

Considering about (14), $\hbar > 0$. Based on Lemma 1, multiply both sides of (A.4) by $e^{\hbar t}$ to get

$$\frac{d}{dt}(V_o(\tilde{X})e^{\hbar t}) \leq \upsilon e^{\hbar t}. \tag{A.5}$$

Define $\ell = (\upsilon/\hbar)$ and integrate (A.5) over $[0, t]$, we have

$$\begin{aligned}V_1(\tilde{X}) &\leq \ell + [V_o(\tilde{X}(0)) - \ell]e^{-\hbar t} \leq \ell + V_o(\tilde{X}(0))e^{-\hbar t} \\ &\leq \ell + V_o(\tilde{X}(0)).\end{aligned} \tag{A.6}$$

From (A.2) and (A.6), and let $\Phi = 2[V_o(\tilde{X}(0)) + \ell]$, we get

$$\|\tilde{X}\| \le \sqrt{\frac{\Phi}{\lambda_{max}(\Gamma)}} \tag{A.7}$$

then error signals $\tilde{x}_1$, $\tilde{x}_2$ and $\tilde{x}_3$ are SGUUB.

Appendix B. Proof of Theorem 2

Remark 7. For the convenience of stability analysis, the constant matrix $Q = I$ is set, where I is the identity matrix.

Firstly, based on the transformed system (6), the boundedness of the signals $\eta_e(t)$, $\check{s}(t)$, $\tilde{W}_a(t)$ and $\tilde{W}_c(t)$ is proved. The Lyapunov candidate function is selected as

$$V_1(t) = \frac{1}{2}\eta_e^T(t)\eta_e(t) + \frac{1}{2}\check{s}^T(t)\check{s}(t) + \frac{1}{2}\tilde{W}_a^T(t)\tilde{W}_a(t) + \frac{1}{2}\tilde{W}_c^T(t)\tilde{W}_c(t). \tag{B.1}$$

The time derivative of V_1 is

$$\begin{aligned}\dot{V}_1 =& \eta_e^T(\check{s} - \Lambda\eta_e) + \check{s}^T\big(\check{s} - \Lambda\eta_e + \hat{f}(\xi) + \delta_m - \ddot{\eta}_d\big)\\ &+ \tilde{W}_a^T\Big[\frac{1}{2}\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\check{s} - l_a\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a\\ &+ \frac{l_c}{4(1+\|\phi\|^2)}\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a\phi^T\hat{W}_c\Big]\\ &- \tilde{W}_c^T\Big\{\frac{l_c}{1+\|\phi\|^2}\phi\Big[\phi^T\hat{W}_c - (K_{\check{s}}^2 - 1 + \gamma K_{\check{s}})\|\check{s}\|^2\\ &- 2K_{\check{s}}\check{s}^T(-\Lambda\dot{\eta}_e - \hat{f}(\xi) + \ddot{\eta}_d) + \frac{1}{4}\Big\|\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a\Big\|^2\Big]\Big\}.\end{aligned} \tag{B.2}$$

Substituting (29) into (B.2) yields

$$\begin{aligned}\dot{V}_1 =& -\eta_e^T\Lambda\eta_e + \eta_e^T\check{s} - \check{s}^T\Lambda\eta_e - (K_{\check{s}} - 1)\|\check{s}\|^2 + \check{s}^T\hat{f}(\xi) - \check{s}^T\ddot{\eta}_d\\ &- \frac{1}{2}\check{s}^T\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a\\ &+ \frac{1}{2}\tilde{W}_a^T\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\check{s} - l_a\tilde{W}_a^T\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a\\ &+ \frac{l_c}{4(1+\|\phi\|^2)}\tilde{W}_a^T\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a\phi^T\hat{W}_c\\ &- \frac{l_c}{1+\|\phi\|^2}\tilde{W}_c^T\phi\Big[\phi^T\hat{W}_c - (K_{\check{s}}^2 - 1 + \gamma K_{\check{s}})\|\check{s}\|^2\\ &- 2K_{\check{s}}\check{s}^T(-\Lambda\dot{\eta}_e - \hat{f}(\xi) + \ddot{\eta}_d)\\ &+ \frac{1}{4}\Big\|\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a\Big\|^2\Big]\\ =& -\eta_e^T\Lambda\eta_e + \eta_e^T\check{s} - \check{s}^T\Lambda\eta_e - (K_{\check{s}} - 1)\|\check{s}\|^2 + \check{s}^T\hat{f}(\xi) - \check{s}^T\ddot{\eta}_d\\ &- \frac{1}{2}\check{s}^T\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}W^*\\ &- \frac{l_a}{2}\hat{W}_a^T\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a - \frac{l_a}{2}\tilde{W}_a^T\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\tilde{W}_a\\ &+ \frac{l_a}{2}W^{*T}\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}W^*\\ &+ \frac{l_c}{4(1+\|\phi\|^2)}\tilde{W}_a^T\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a\phi^T\hat{W}_c - \frac{l_c}{1+\|\phi\|^2}\tilde{W}_c^T\phi\\ &\times\Big[\phi^T\hat{W}_c - (K_{\check{s}}^2 - 1 + \gamma K_{\check{s}})\|\check{s}\|^2\\ &- 2K_{\check{s}}\check{s}^T(-\Lambda\dot{\eta}_e - \hat{f}(\xi) + \ddot{\eta}_d) + \frac{1}{4}\Big\|\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a\Big\|^2\Big].\end{aligned} \tag{B.3}$$

Based on (27), the following relationship can be obtained

$$\begin{aligned}&-(K_{\check{s}}^2 - 1 + \gamma K_{\check{s}})\|\check{s}\|^2 - 2K_{\check{s}}\check{s}^T(-\Lambda\dot{\eta}_e - \hat{f}(\xi) + \ddot{\eta}_d)\\ =& -\frac{1}{2}\hat{W}_a^T\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}W^* - \phi^TW^* + \frac{1}{4}W^{*T}\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}W^* - \rho.\end{aligned} \tag{B.4}$$

Substituting (B.4) into (B.3), we obtain

$$\begin{aligned}\dot{V}_1 \le& -\lambda_{min}(\Lambda - I)\|\eta_e\|^2 - (K_{\check{s}} - 3 - \frac{1 + \|\Lambda\|_F^2}{2})\|\check{s}\|^2\\ &+ \frac{1}{2}\|\hat{f}(\xi)\|^2 + \frac{1}{2}\|\ddot{\eta}_d\|^2\\ &- \Big(\frac{l_a}{2} - \frac{l_c^2}{2} - \frac{1}{32}W^{*T}\phi\phi^TW^*\Big)\tilde{W}_a^T\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\tilde{W}_a\\ &- \frac{l_c}{1+\|\phi\|^2}\Big(\frac{l_c}{2} - \frac{1}{32}W^{*T}\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}W^*\Big)\tilde{W}_c^T\phi\phi^T\tilde{W}_c\\ &- \Big(\frac{l_a}{2} - \frac{l_c^2}{2}\Big)\hat{W}_a^T\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a\\ &+ (1 + \frac{l_a}{2})W^{*T}\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}W^* + \frac{l_c}{2(1+\|\phi\|^2)}\rho^2.\end{aligned} \tag{B.5}$$

Next, the inequality (B.5) can be rewritten as

$$\dot{V}_1 \le -\varpi^TA\varpi + B - \Big(\frac{l_a}{2} - \frac{l_c^2}{2}\Big)\hat{W}_a^T\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a \tag{B.6}$$

where

$$\varpi = [\eta_e^T(t);\ \check{s}^T(t);\ \tilde{W}_a^T(t);\ \tilde{W}_c^T(t)]$$

$$A = diag([a_{11}, a_{22}, a_{33}, a_{44}])$$

$$\begin{aligned}C =& \frac{1}{2}\|\hat{f}(\xi)\|^2 + \frac{1}{2}\|\ddot{\eta}_d\|^2 + \frac{l_c}{2(1+\|\phi\|^2)}\rho^2\\ &+ (1 + \frac{l_a}{2})W^{*T}\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}W^*\end{aligned}$$

and the parameters of matrix A are given as follows: $a_{11} = \lambda_{min}(\Lambda - I)$, $a_{22} = [K_{\check{s}} - 3 - (1 + \|\Lambda\|_F^2)/2]$, $a_{33} = \big[l_a/2 - l_c^2/2 - (1/32)W^{*T}\phi\phi^TW^*\big] \times \big[\partial\varphi(\check{s})/\partial\check{s}^T\big]\big[\partial^T\varphi(\check{s})/\partial\check{s}\big]$, $a_{44} = \big[l_c/\big(1 + \|\phi\|^2\big)\big]\Big\{l_c/2 - (1/32)W^{*T}\big[\partial\varphi(\check{s})/\partial\check{s}^T\big]\big[\partial^T\varphi(\check{s})/\partial\check{s}\big]W^*\Big\}\phi\phi^T$. Select the neural network learning rate to satisfy $l_a \ge l_c^2$, then we have

$$\dot{V}_1 \le -\varpi^TA\varpi + B. \tag{B.7}$$

Based on Assumption 3 of PE, the matrix A is guaranteed to be positive definite by designing parameters Λ, l_c, l_a and $K_{\check{s}}$ as follows:

$$\Lambda > I, \quad K_{\check{s}} > \frac{7 + \|\Lambda\|_F^2}{2}, \quad l_a > l_c^2 + \frac{\bar{\phi}}{16}W^{*T}W^*,$$

$$l_c > \frac{1}{16}\sup_{t\ge0}\Big\{W^{*T}\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}W^*\Big\}.$$

Then (B.6) can be written as

$$\dot{V}_1 \le -a\|\varpi\|^2 + b \tag{B.8}$$

where $a = inf_{t\ge0}\{\lambda_{min}\{A\}\}$, $b = sup_{t\ge0}\{B\}$.

According to Lemma 1, we have

$$\begin{aligned}V_1(t) \le& e^{-at}V_1(0) + \frac{b}{a}(1 - e^{-at})\\ \le& V_1(0) + \varepsilon\end{aligned} \tag{B.9}$$

where $\varepsilon = b/a$, $V_1(0) = (1/2)\big[\eta_e^T(0)\eta_e(0) + \check{s}^T(0)\check{s}(0) + \tilde{W}_a^T(0)\tilde{W}_a(0) + \tilde{W}_c^T(0)\tilde{W}_c(0)\big]$. Let $\Xi = 2(V_1(0) + \varepsilon)$, and then we have

$$\|\eta_e\| \le \sqrt{\Xi}, \quad \|\check{s}\| \le \sqrt{\Xi}, \quad \|\tilde{W}_a\| \le \sqrt{\Xi}, \quad \|\tilde{W}_c\| \le \sqrt{\Xi}$$

thus, the inequality (B.9) guarantees that the signals η_e, $\check{s}$, $\tilde{W}_a$, $\tilde{W}_c$ are SGUUB.

Finally, based on the original system (1), the boundedness of the error signals $\eta_e(t)$, $s(t)$, χ and $\hat{M}$ can be proved. The Lyapunov candidate function is selected as

$$V_2(t) = \frac{1}{2}\eta_e^T(t)\eta_e(t) + \frac{1}{2}s^T(t)s(t) + \frac{1}{2}\chi^T\chi + \frac{1}{2}\sum_{i=1}^{3} m_{ii}^{-1}\Psi_i^{-1}\tilde{m}_{ii}^2. \tag{B.10}$$

The time derivative of V_2 is

$$\begin{aligned}
\dot{V}_2 &= \eta_e^T(s - \Lambda\eta_e) \\
&\quad + s^T\left[s - \Lambda\eta_e + F(\xi) + R(\eta_z)M^{-1}(\delta_c + \Delta\delta) - \ddot{\eta}_d\right] \\
&\quad - \chi^T K_\chi \chi - 2\|s^T\Delta\delta\| \\
&\quad - 3\Delta\delta^T\Delta\delta + 2\chi^T\Delta\delta + \sum_{i=1}^{3} m_{ii}^{-1}\Psi_i^{-1}\tilde{m}_{ii}\dot{\hat{m}}_{ii} \\
&= -\eta_e^T\Lambda\eta_e + s^Ts + \eta_e^Ts - s^T\Lambda\eta_e + s^TF(\xi) \\
&\quad + s^TR(\eta_z)M^{-1}\hat{M}R^{-1}(\eta_z)\delta_m + s^TR(\eta_z)M^{-1}\Delta\delta \\
&\quad + s^TR(\eta_z)M^{-1}\chi - s^T\ddot{\eta}_d - \chi^TK_\chi\chi - 2\|s^T\Delta\delta\| \\
&\quad - 3\Delta\delta^T\Delta\delta + 2\chi^T\Delta\delta \\
&\quad - \sum_{i=1}^{3} m_{ii}^{-1}\tilde{m}_{ii}\delta_{mi}s_i - \sum_{i=1}^{3} m_{ii}^{-1}\Omega_i\tilde{m}_{ii}\hat{m}_{ii} \\
&= -\eta_e^T\Lambda\eta_e + s^Ts + \eta_e^Ts - s^T\Lambda\eta_e + s^TF(\xi) \\
&\quad + s^T(-K_{\check{s}}\check{s} - \frac{1}{2}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a) + s^TR(\eta_z)M^{-1}\Delta\delta \\
&\quad + s^TR(\eta_z)M^{-1}\chi - s^T\ddot{\eta}_d - \chi^TK_\chi\chi - 2\|s^T\Delta\delta\| \\
&\quad - 3\Delta\delta^T\Delta\delta + 2\chi^T\Delta\delta \\
&\quad - \frac{1}{2}\sum_{i=1}^{3} m_{ii}^{-1}\Omega_i(\tilde{m}_{ii}^2 + \hat{m}_{ii}^2 - m_{ii}^2).
\end{aligned} \tag{B.11}$$

Note that

$$\begin{aligned}
&-\frac{1}{2}\sum_{i=1}^{3} m_{ii}^{-1}\Omega_i(\tilde{m}_{ii}^2 + \hat{m}_{ii}^2 - m_{ii}^2) \\
\le &-\frac{1}{2}\sum_{i=1}^{3} m_{ii}^{-1}\Omega_i\tilde{m}_{ii}^2 + \frac{1}{2}\sum_{i=1}^{3} m_{ii}^{-1}\Omega_i m_{ii}^2.
\end{aligned} \tag{B.12}$$

Base on that properties of AUV model, we can define $\|R(\psi)M^{-1}\|_F \triangleq \varrho$, $\varrho > 0$. Then, we have

$$\begin{aligned}
\dot{V}_2 \le &-\eta_e^T(\Lambda - I)\eta_e + \|s\|^2 + \frac{1 + \|\Lambda\|_F^2}{2}\|s\|^2 + \frac{1}{2}\|s\|^2 \\
&+ \|F(\xi)\|^2 - K_{\check{s}}\|s\|^2 + \frac{1}{2}\|s\|^2 \\
&+ \varrho^2\|s\|^2 + \hat{W}_a^T\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a \\
&+ \frac{1}{2}\|\Delta\delta\|^2 + \frac{1}{2}\|\chi\|^2 + \frac{1}{2}\|s\|^2 + \|\ddot{\eta}_d\|^2 - \chi^TK_\chi\chi \\
&+ \frac{1}{2}\|s\|^2 + 2\|\Delta\delta\|^2 - 3\|\Delta\delta\|^2 + 2\|\chi\|^2 + \frac{1}{2}\|\Delta\delta\|^2 \\
&- \frac{1}{2}\sum_{i=1}^{3} m_{ii}^{-1}\Omega_i\tilde{m}_{ii}^2 + \frac{1}{2}\sum_{i=1}^{3} m_{ii}^{-1}\Omega_i m_{ii}^2 \\
\le &-\lambda_{min}(\Lambda - I)\|\eta_e\|^2 - (K_{\check{s}} - \frac{7 + 2\varrho^2 + \|\Lambda\|_F^2}{2})\|s\|^2 \\
&- \lambda_{min}(K_\chi - \frac{5}{2}I)\|\chi\|^2 \\
&+ \hat{W}_a^T\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}}\hat{W}_a + \|F(\xi)\|^2 + \|\ddot{\eta}_d\|^2 \\
&- \frac{1}{2}\sum_{i=1}^{3} m_{ii}^{-1}\Omega_i\tilde{m}_{ii}^2 + \frac{1}{2}\sum_{i=1}^{3} m_{ii}^{-1}\Omega_i m_{ii}^2.
\end{aligned} \tag{B.13}$$

According to the proof conclusion of the transformed system, the inequality $\|\tilde{W}_a\| < \sqrt{\Xi}$ is established. So there is a constant ι such that $\|\hat{W}_a\| < \iota$. So (B.13) can be further written as

$$\begin{aligned}
\dot{V}_2 \le &-\lambda_{min}\{\Lambda - I\}\|\eta_e\|^2 - (K_{\check{s}} - \frac{7 + 2\varrho^2 + \|\Lambda\|_F^2}{2})\|s\|^2 \\
&- \lambda_{min}(K_\chi - \frac{5}{2}I)\|\chi\|^2 \\
&- \frac{1}{2}\sum_{i=1}^{3} m_{ii}^{-1}\Omega_i\tilde{m}_{ii}^2 + \iota^2\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}} + \|F(\xi)\|^2 + \|\ddot{\eta}_d\|^2 \\
&+ \frac{1}{2}\sum_{i=1}^{3} m_{ii}^{-1}\Omega_i m_{ii}^2
\end{aligned} \tag{B.14}$$

We rewrite (B.14) as follows

$$\dot{V}_2 \le -\mu^TC\mu + D \tag{B.15}$$

where

$$\mu = \left[\eta_e^T(t); s^T(t); \chi^T; \sqrt{\sum_{i=1}^{3} m_{ii}^{-1}\Omega_i\tilde{m}_{ii}^2}\right]$$

$$C = diag([c_{11}, c_{22}, c_{33}, c_{44}])$$

$$D = \iota^2\frac{\partial\varphi(\check{s})}{\partial\check{s}^T}\frac{\partial^T\varphi(\check{s})}{\partial\check{s}} + \|F(\xi)\|^2 + \|\ddot{\eta}_d\|^2 + \frac{1}{2}\sum_{i=1}^{3} m_{ii}^{-1}\Omega_i m_{ii}^2$$

and the parameters of matrix C are given as follows: $c_{11} = \lambda_{min}(\Lambda - I)$, $c_{22} = \lambda_{min}\{K_{\check{s}} - [(8 + 2\varrho^2 + \|\Lambda\|_F^2)/2]\}$, $c_{33} = \lambda_{min}[K_\chi - (5/2)I]$, $c_{44} = (1/2)\min\{\Omega_i\}\max\{\Psi_i\}$. Based on Assumption 2 of PE, the matrix C is guaranteed to be positive definite by designing parameters Λ, $K_{\check{s}}$ and K_χ as follows:

$$\Lambda > I, \quad K_{\check{s}} > \frac{7 + 2\varrho^2 + \|\Lambda\|_F^2}{2}, \quad K_\chi > \frac{5}{2}I.$$

Then (B.15) can be written as

$$\dot{V}_2 \le -c\|\mu\|^2 + d \tag{B.16}$$

where $c = inf_{t\ge0}\{\lambda_{min}\{C\}\}$, $d = sup_{t\ge0}\{D\}$.

According to Lemma 1, we get

$$\begin{aligned}
V_2(t) &\le e^{-ct}V_2(0) + \frac{d}{c}(1 - e^{-ct}) \\
&\le V_2(0) + \zeta
\end{aligned} \tag{B.17}$$

where $\zeta = d/c$, $V_2(0) = (1/2)\left[\eta_e^T(0)\eta_e(0) + s^T(0)s(0) + \chi(0)^T\chi(0) + \sum_{i=1}^{3} m_{ii}^{-1}\Psi_i^{-1}\tilde{m}_{ii}(0)^2\right]$. Let $\Pi = 2(V_2(0) + \zeta)$, and then we have

$$\|\eta_e\| \le \sqrt{\Pi}, \ \|s\| \le \sqrt{\Pi}, \ \|\chi\| \le \sqrt{\Pi}, \ \sum_{i=1}^{3} m_{ii}^{-1}\Psi_i^{-1}\tilde{m}_{ii}^2 \le \Pi$$

thus, the inequality (B.17) guarantees that the error signals η_e, s, and $\tilde{m}_{ii}$ are SGUUB.

References

[1] Ribas D, Palomeras N, Ridao P, Carreras M, Mallios A. Girona 500 AUV: From survey to intervention. IEEE/ASME Trans Mechatronics 2011;17(1):46–53.
[2] Xiang X, Jouvencel B, Parodi O. Coordinated formation control of multiple autonomous underwater vehicles for pipeline inspection. Int J Adv Robot Syst 2010;7(1):75–84.
[3] Li Z, You K, Song S. AUV based source seeking with estimated gradients. J Syst Sci Complex 2018;31(1):262–75.

[4] Rout R, Subudhi B. Inverse optimal self-tuning PID control design for an autonomous underwater vehicle. Int J Syst Sci 2017;48(2):367–75.
[5] Zhu D, Zhao Y, Yan M. A bio-inspired neurodynamics-based backstepping path-following control of an AUV with ocean current. Int J Robot Autom 2012;27(3):298–307.
[6] Shen Z, Zhang X, Zhang N, Guo G. Recursive sliding mode dynamic surface output feedback control for ship trajectory tracking based on neural network observer. Control Theory Appl 2018;35(8):1092–100.
[7] Shen C, Shi Y, Buckham B. Trajectory tracking control of an autonomous underwater vehicle using Lyapunov-based model predictive control. IEEE Trans Ind Electron 2017;65(7):5796–805.
[8] Lewis FL, Vrabie D, Syrmos VL. Optimal control. John Wiley & Sons; 2012.
[9] Tong S, Li Y, Sui S. Adaptive fuzzy tracking control design for SISO uncertain nonstrict feedback nonlinear systems. IEEE Trans Fuzzy Syst 2016;24(6):1441–54.
[10] Lin W-S. Optimality and convergence of adaptive optimal control by reinforcement synthesis. Automatica 2011;47(5):1047–52.
[11] Liu D, Wang D, Li H. Decentralized stabilization for a class of continuous-time nonlinear interconnected systems using online learning optimal control approach. IEEE Trans Neural Netw Learn Syst 2013;25(2):418–28.
[12] Werbos PJ. Neural networks for control and system identification. In: Proc. IEEE conference on decision and control. IEEE; 1989, p. 260–5.
[13] Werbos PJ, Miller WT, Sutton RS. A menu of designs for reinforcement learning over time. In: Neural networks for control. MIT press Cambridge, MA; 1990, p. 67–95.
[14] Werbos PJ. Approximate dynamic programming for realtime control and neural modelling. In: Handbook of intelligent control: Neural, fuzzy and adaptive approaches. Van Nostrand; 1992, p. 493–525.
[15] Wen G, Chen CP, Ge SS, Yang H, Liu X. Optimized adaptive nonlinear tracking control using actor-critic reinforcement learning strategy. IEEE Trans Ind Inf 2019;15(9):4969–77.
[16] Modares H, Lewis FL, Naghibi-Sistani M-B. Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems. Automatica 2014;50(1):193–202.
[17] Modares H, Lewis FL, Naghibi-Sistani M-B. Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks. IEEE Trans Neural Netw Learn Syst 2013;24(10):1513–25.
[18] Dong L, Zhong X, Sun C, He H. Event-triggered adaptive dynamic programming for continuous-time systems with control constraints. IEEE Trans Neural Netw Learn Syst 2016;28(8):1941–52.
[19] Zhao B, Liu D, Luo C. Reinforcement learning-based optimal stabilization for unknown nonlinear systems subject to inputs with uncertain constraints. IEEE Trans Neural Netw Learn Syst 2019;31(10):4330–40.
[20] Wang N, Gao Y, Zhang X. Data-driven performance-prescribed reinforcement learning control of an unmanned surface vehicle. IEEE Trans Neural Netw Learn Syst 2021;32(12):5456–67.
[21] Wen G, Ge SS, Chen CLP, Tu F, Wang S. Adaptive tracking control of surface vessel using optimized backstepping technique. IEEE Trans Cybern. 2018;49(9):3420–31.
[22] Wang N, Gao Y, Liu Y, Li K. Self-learning-based optimal tracking control of an unmanned surface vehicle with pose and velocity constraints. Internat J Robust Nonlinear Control 2022;32(5):2950–68.
[23] Mishra A, Ghosh S. Simultaneous identification and optimal tracking control of unknown continuous-time systems with actuator constraints. Internat J Control 2022;95(8):2005–23.
[24] Modares H, Lewis FL. Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning. Automatica 2014;50(7):1780–92.
[25] Bhasin S, Kamalapurkar R, Johnson M, Vamvoudakis KG, Lewis FL, Dixon WE. A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems. Automatica 2013;49(1):82–92.
[26] Huo X, Karimi HR, Zhao X, Wang B, Zong G. Adaptive-critic design for decentralized event-triggered control of constrained nonlinear interconnected systems within an identifier-critic framework. IEEE Trans Cybern. 2022;52(8):7478–91.
[27] Wang N, Gao Y, Zhao H, Ahn CK. Reinforcement learning-based optimal tracking control of an unknown unmanned surface vehicle. IEEE Trans Neural Netw Learn Syst 2020;32(7):3034–45.
[28] Li X, Ren C, Ma S, Zhu X. Compensated model-free adaptive tracking control scheme for autonomous underwater vehicles via extended state observer. Ocean Eng 2020;217:107976.
[29] Cui R, Yang C, Li Y, Sharma S. Adaptive neural network control of AUVs with control input nonlinearities using reinforcement learning. IEEE Trans Syst Man Cybern Syst 2017;47(6):1019–29.
[30] Peng Z, Wang J, Han Q-L. Path-following control of autonomous underwater vehicles subject to velocity and input constraints via neurodynamic optimization. IEEE Trans Ind Electron 2018;66(11):8724–32.
[31] Yu C, Xiang X, Wilson PA, Zhang Q. Guidance-error-based robust fuzzy adaptive control for bottom following of a flight-style AUV with saturated actuator dynamics. IEEE Trans Cybern. 2019;50(5):1887–99.
[32] Yu H, Guo C, Han Y, Shen Z. Bottom-following control of underactuated unmanned undersea vehicles with input saturation. IEEE Access 2020;8:120489–500.
[33] Guo X, Yan W, Cui R. Integral reinforcement learning-based adaptive NN control for continuous-time nonlinear MIMO systems with unknown control directions. IEEE Trans Syst Man Cybern Syst 2019;50(11):4068–77.
[34] Wen G, Ge SS, Tu F. Optimized backstepping for tracking control of strict-feedback systems. IEEE Trans Neural Netw Learn Syst 2018;29(8):3850–62.
[35] Lv Y, Na J, Zhao X, Huang Y, Ren X. Multi-H_∞ controls for unknown input-interference nonlinear system with reinforcement learning. IEEE Trans Neural Netw Learn Syst 2021.
[36] Wang D, Liu D, Zhang Q, Zhao D. Data-based adaptive critic designs for nonlinear robust optimal control with uncertain dynamics. IEEE Trans Syst Man Cybern Syst 2015;46(11):1544–55.
[37] Sanner RM, Slotine J-JE. Gaussian networks for direct adaptive control. In: 1991 American control conference. IEEE; 1991, p. 2153–9.
[38] Vamvoudakis KG, Lewis FL. Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. Automatica 2010;46(5):878–88.
[39] Begum G, Radhakrishnan TK. Performance assessment of control loops involving unstable systems for set point tracking and disturbance rejection. J Taiwan Inst Chem Eng 2018;85:1–17.
[40] Ge SS, Wang C. Adaptive neural control of uncertain MIMO nonlinear systems. IEEE Trans Neural Netw 2004;15(3):674–92.

Zhifu Li received the B.Sc. And M.Sc. degrees in control theory and control engineering from Central South University, Changsha,Hunan, China, in 2003 and 2006, respectively, and the Ph.D. degree in control theory and control engineering from the South China University of Technology, Guangzhou, Guangdong, China, in 2012. From 2012 to 2015, he was a Postdoctoral Fellow in mechatronics with the South China University of Technology. He has been with the School of Mechanical and Electrical Engineering, Guangzhou University, since 2015, where he is currently an Associate Professor. He has published over 30 research articles. His research interests include nonlinear control and its application, swarm intelligence optimization, applied fractional calculus, robust adaptive control, and machine vision.

Ming Wang received the B.Sc. Degree in machine design, manufacturing and automation from the Xi'an Technological University, Xi'an, Shaanxi, China, in 2020. He is currently pursuing the master's degree in mechanical engineering with the School of Mechanical and Electrical Engineering, Guangzhou University, Guangzhou. His research interests include intelligent control, reinforcement learning, and system identification theory.

Ge Ma received the B.Sc. degree in mathematics from Shandong Jianzhu University, Jinan, Shandong, China, in 2010, and the Ph.D. degree in control theory and control engineering from the South China University of Technology, Guangzhou, Guangdong, China, in 2016. She has been with the School of Mechanical and Electrical Engineering, Guangzhou University, since 2016, where she is currently a Lecturer. Her research interests include control theory, image processing, machine vision, and intelligence optimization.