Journal of Systems Engineering and Electronics
Vol. 33, No. 1, February 2022, pp.170 – 179

Reinforcement learning based parameter optimization of active disturbance rejection control for autonomous underwater vehicle

SONG Wanping[1], CHEN Zengqiang[1,2,*], SUN Mingwei[1], and SUN Qinglin[1]

1. College of Artificial Intelligence, Nankai University, Tianjin 300350, China;
2. Key Laboratory of Intelligent Robotics of Tianjin, Nankai University, Tianjin 300350, China

Abstract: This paper proposes a liner active disturbance rejection control (LADRC) method based on the Q-Learning algorithm of reinforcement learning (RL) to control the six-degree-of-freedom motion of an autonomous underwater vehicle (AUV). The number of controllers is increased to realize AUV motion decoupling. At the same time, in order to avoid the oversize of the algorithm, combined with the controlled content, a simplified Q-learning algorithm is constructed to realize the parameter adaptation of the LADRC controller. Finally, through the simulation experiment of the controller with fixed parameters and the controller based on the Q-learning algorithm, the rationality of the simplified algorithm, the effectiveness of parameter adaptation, and the unique advantages of the LADRC controller are verified.

Keywords: autonomous underwater vehicle (AUV), reinforcement learning (RL), Q-learning, linear active disturbance rejection control (LADRC), motion decoupling, parameter optimization.

DOI: 10.23919/JSEE.2022.000017

1. Introduction

Underwater robot plays a very important role in the development of marine resources and protection of marine rights and interests. Autonomous underwater vehicle (AUV) has been widely used in marine research and national security [1−4]. In recent years, AUV has been successfully applied to complex underwater motion such as seabed imaging and seabed mapping [5,6]. The motion of an AUV has six degrees of freedom. Because each operation on the AUV will have different degrees of influence on its various degrees of freedom, the AUV has the characteristics of strong coupling and strong nonlinearity. In addition, the complex dynamics of the underwater environment makes it more difficult to control AUVs [7]. Therefore, it is of practical significance to control the movement of AUVs in accordance with the required performance. Many control methods in classical control theory, modern control theory and intelligent control theory have been applied to motion control of AUVs. For example, proportional-integral-derivative (PID) control, sliding mode control, fuzzy control and adaptive control, and many combination methods of the above methods [8−11]. PID control is a feedback control based on error signals and is currently one of the main AUV control methods. However, when PID control faces a system with strong nonlinearity and strong coupling, the dynamic performance of the system is poor and the overshoot is large. Han proposed the active disturbance rejection control (ADRC) method in the 1990s [12,13]. On the basis of this, Gao proposed a linear active disturbance rejection control (LADRC) method [14], which greatly reduced the number of parameters of the ADRC controller and made the whole system easy to debug and apply. ADRC has been applied to the control problems of fighter aircraft's high angle of attack tracking [15], ship course control [16] and the power system [17], demonstrating its superior control performance. A good controller should have a certain degree of adaptive ability in resisting disturbances while having a good control performance [18,19], so the selection of parameters of the controller has been the focus of many experts and scholars. At present, many algorithms have been used to calculate controller parameters, such as the adaptive controller combined with fuzzy control algorithm can realize parameter self-adjustment [10,20,21]. However, establishment of fuzzy rules depends on professional experience and model, so the application scope of fuzzy control is limited. Reinforcement learning has

Manuscript received November 26, 2020.
*Corresponding author.
This work was supported by the National Natural Science Foundation of China (61973175; 61973172) and Tianjin Natural Science Foundation (19JCZDJC32800).

been widely used in artificial intelligence and machine learning [22,23]. Control algorithms based on reinforcement learning can optimize control strategies by interacting with unknown environments. The temproal-difference (TD) method in reinforcement learning is a model-independent reinforcement learning algorithm. The algorithm updates strategies by updating the value function, and the new status and immediate rewards generated after the execution of the strategy are used to update the value function again. The TD method includes on-policy Sarsa algorithm and off-policy Q-learning algorithm [24]. The effectiveness of Q-learning algorithm has been verified in many fields [16,17,25]. At present, in most control method researches, the AUV model is always decoupled [23,26], so the authenticity of the controlled model is reduced. Therefore, the LADRC controller based on the Q-learning algorithm is used to control the six-degree-of-freedom AUV model, and the related structure of the controller and the Q-learning algorithm are designed. Through Matlab simulation experiments, the control effect of the new controller is compared with that of the PID and LADRC controllers with fixed parameters. The results show that the LADRC controller based on the Q-learning algorithm can achieve the better control effect.

The main contributions of this paper are summarized as follows:

(i) The LADRC controller is adopted to stabilize the AUV system.

(ii) Controllers are added for motion decoupling: two LADRC controllers are used to control AUV motion in yaw and pitch planes.

(iii) The Q-learning algorithm is applied to realize parameter self-adaptation of the LADRC controller.

(iv) The state division and reward design of Q-learning algorithm are constructed for the controlled content. The scale of the algorithm is simplified and the effectiveness of the algorithm is guaranteed.

2. Motion and modeling of AUV

2.1 Coordinate frames and rigid body dynamics equation

Six degrees of freedom motion equations of AUV can be described using the earth-fixed coordinate frame and the body-fixed coordinate frame shown in Fig.1, both of which are right-handed. The origin of the body-fixed coordinate frame is located at the AUV center of buoyancy.

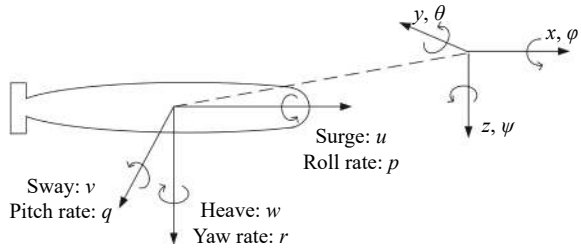


Fig. 1 Coordinate frames and motion parameters

The motion of AUV can be described by these vectors:

$$\boldsymbol{\eta}_1 = [x \quad y \quad z]^{\mathrm{T}}, \boldsymbol{\eta}_2 = [\varphi \quad \theta \quad \psi]^{\mathrm{T}},$$

$$\boldsymbol{\upsilon}_1 = [u \quad v \quad w]^{\mathrm{T}}, \boldsymbol{\upsilon}_2 = [p \quad q \quad r]^{\mathrm{T}},$$

$$\boldsymbol{\tau}_1 = [X \quad Y \quad Z]^{\mathrm{T}}, \boldsymbol{\tau}_2 = [K \quad M \quad N]^{\mathrm{T}},$$

where $\boldsymbol{\eta}$ describes the position and orientation of the AUV in the earth-fixed coordinate frame, $\boldsymbol{\upsilon}$ describes the linear and angular velocities of the AUV, and $\boldsymbol{\tau}$ describes the total forces and moments acting on the AUV in the body-fixed coordinate frame. The meanings of the symbols are summarized in Table 1.

Table 1 Symbols and their meanings

Motion	Position and angle	Linear and angular velocity (Force and moment)
Surge	x	$u(X)$
Sway	y	$v(Y)$
Heave	z	$w(Z)$
Roll	φ	$p(K)$
Pitch	θ	$q(M)$
Yaw	ψ	$r(N)$

The coordinate transformation of the translational velocity between earth-fixed and body-fixed coordinate frames can be expressed as

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \boldsymbol{J}_1 \begin{bmatrix} u \\ v \\ w \end{bmatrix} \tag{1}$$

where

$$\boldsymbol{J}_1 = \begin{bmatrix} \cos\psi\cos\theta & \cos\psi\sin\theta\sin\varphi - \sin\psi\cos\varphi & \cos\psi\sin\theta\cos\varphi + \sin\psi\sin\varphi \\ \sin\psi\cos\theta & \sin\psi\sin\theta\sin\varphi + \cos\psi\cos\varphi & -\cos\psi\sin\varphi + \sin\psi\sin\theta\cos\varphi \\ -\sin\theta & \cos\theta\sin\varphi & \cos\theta\cos\varphi \end{bmatrix}.$$

The coordinate transformation of the rotational velocity between two coordinate systems can be expressed as

$$\begin{bmatrix}\dot{\varphi}\\ \dot{\theta}\\ \dot{\psi}\end{bmatrix}=\boldsymbol{J}_2\begin{bmatrix}p\\ q\\ r\end{bmatrix} \tag{2}$$

where

$$\boldsymbol{J}_2=\begin{bmatrix}1 & \sin\varphi\tan\theta & \cos\varphi\tan\theta\\ 0 & \cos\varphi & -\sin\varphi\\ 0 & \sin\varphi/\cos\theta & \cos\varphi/\cos\theta\end{bmatrix}.$$

The positions of the AUV centers of gravity and buoyancy are defined in the body-fixed coordinate frame as follows:

$$\boldsymbol{r}_G=[x_g \quad y_g \quad z_g]^{\mathrm{T}},\boldsymbol{r}_B=[x_b \quad y_b \quad z_b]^{\mathrm{T}}.$$

According to the theory of rigid body dynamics, the motion equations of a six degrees of freedom rigid body defined by body-fixed coordinates are as follows:

$$\begin{aligned}
&m\left[(\dot{u}-vr+wq)-x_g\left(q^2+r^2\right)+y_g(pq-\dot{r})+\right.\\
&\qquad\left.z_g(pr+\dot{q})\right]=X,\\
&m\left[(\dot{v}-wp+ur)-y_g\left(r^2+p^2\right)+z_g(qr-\dot{p})+\right.\\
&\qquad\left.x_g(qp+\dot{r})\right]=Y,\\
&m\left[(\dot{w}-uq+vp)-z_g\left(q^2+p^2\right)+y_g(rq+\dot{p})+\right.\\
&\qquad\left.x_g(rp-\dot{q})\right]=Z,\\
&I_x\dot{p}+(I_z-I_y)qr+m\left[y_g(\dot{w}+pv-qu)-\right.\\
&\qquad\left.z_g(\dot{v}+ru-pw)\right]=K,\\
&I_y\dot{q}+(I_x-I_z)rp+m\left[z_g(\dot{u}+wq-vr)-\right.\\
&\qquad\left.x_g(\dot{w}+pv-uq)\right]=M\\
&I_z\dot{r}+(I_y-I_x)pq+m\left[x_g(\dot{v}+ur-pw)-\right.\\
&\qquad\left.y_g(\dot{u}+qw-vr)\right]=N
\end{aligned} \tag{3}$$

where m is AUV's weight and I_x, I_y, I_z are the moments of inertia of mass m of AUV to three coordinate axes.

2.2 Force and motion equation

At present, there are many submarine motion equations used in the world, but each equation differs only in mathematical description and mathematical processing method. The AUV model referred in this paper is a remote environmental monitoring units (REMUS) autonomous underwater vehicle [27]. The total forces and moments acting on AUV can be expressed as follows:

$$\begin{aligned}
X&=X_{\mathrm{HS}}+X_{u|u|}u|u|+X_{\dot{u}}\dot{u}+X_{wq}wq+\\
&\quad X_{qq}qq+X_{vr}vr+X_{rr}rr+X_{\mathrm{prop}}\\
Y&=Y_{\mathrm{HS}}+Y_{v|v|}v|v|+Y_{r|r|}r|r|+Y_{\dot{v}}\dot{v}+Y_{\dot{r}}\dot{r}+Y_{ur}ur+\\
&\quad Y_{wp}wp+Y_{pq}pq+Y_{uv}uv+Y_{uu\delta_r}u^2\delta_r\\
Z&=Z_{\mathrm{HS}}+Z_{w|w|}w|w|+Z_{q|q|}q|q|+Z_{\dot{w}}\dot{w}+Z_{\dot{q}}\dot{q}+\\
&\quad Z_{uq}uq+Z_{vp}vp+Z_{rp}rp+Z_{uw}uw+Z_{uu\delta_s}u^2\delta_s\\
K&=K_{\mathrm{HS}}+K_{p|p|}p|p|+K_{\dot{p}}\dot{p}+K_{\mathrm{prop}}\\
M&=M_{\mathrm{HS}}+M_{w|w|}w|w|+M_{q|q|}q|q|+M_{\dot{w}}\dot{w}+\\
&\quad M_{\dot{q}}\dot{q}+M_{uq}uq+M_{vp}vp+M_{rp}rp+\\
&\quad M_{uw}uw+M_{uu\delta_s}u^2\delta_s\\
N&=N_{\mathrm{HS}}+N_{v|v|}v|v|+N_{r|r|}r|r|+N_{\dot{v}}\dot{v}+N_{\dot{r}}\dot{r}+\\
&\quad N_{ur}ur+N_{wp}wp+N_{pq}pq+N_{uv}uv+N_{uu\delta_r}u^2\delta_r
\end{aligned} \tag{4}$$

where X_{HS}, Y_{HS}, Z_{HS}, K_{HS}, M_{HS}, N_{HS} are hydrostatics; $X_{u|u|}$, $Y_{v|v|}$, $Y_{r|r|}$, $Z_{w|w|}$, $Z_{q|q|}$, $K_{p|p|}$, $M_{w|w|}$, $M_{q|q|}$, $N_{v|v|}$, $N_{r|r|}$ are hydrodynamic damping coefficients. Y_{uv}, $Y_{uu\delta_r}$, Z_{uw}, $Z_{uu\delta_s}$, M_{uw}, $M_{uu\delta_s}$, N_{uv}, $N_{uu\delta_r}$ are lift coefficients and lift moment coefficients of body and control fin. X_{prop}, K_{prop} are propeller thrust and torque. δ_s, δ_r are the AUV's pitch fin angle and rudder angle. The remaining coefficients are additional mass coefficients.

Substitute (4) into the right end of (3). Organize the formula so that all the left end of the formula are acceleration terms. The nonlinear equations of motion can be obtained after sorting out

$$\begin{bmatrix}\dot{u}\\ \dot{v}\\ \dot{w}\\ \dot{p}\\ \dot{q}\\ \dot{r}\end{bmatrix}=\begin{bmatrix}m-X_{\dot{u}} & 0 & 0 & 0 & mz_g & -my_g\\ 0 & m-Y_{\dot{v}} & 0 & -mz_g & 0 & mx_g-Y_{\dot{r}}\\ 0 & 0 & m-Z_{\dot{w}} & my_g & -mx_g-Z_{\dot{q}} & 0\\ 0 & -mz_g & my_g & I_{xx}-K_{\dot{p}} & 0 & 0\\ mz_g & 0 & -mx_g-M_{\dot{w}} & 0 & I_{yy}-M_{\dot{q}} & 0\\ -my_g & mx_g-N_{\dot{v}} & 0 & 0 & 0 & I_{zz}-N_{\dot{r}}\end{bmatrix}^{-1}\begin{bmatrix}\sum X\\ \sum Y\\ \sum Z\\ \sum K\\ \sum M\\ \sum N\end{bmatrix} \tag{5}$$

where $\sum X,\cdots,\sum N$ is the other terms except the term containing acceleration. Six degrees of freedom nonlinear motion equations of AUV can be obtained by combining (5) with (1) and (2).

2.3 AUV system model

The attitude of the REMUS vehicle is controlled by horizontal fins and vertical fins. The horizontal fins of the REMUS vehicle can control the pitching fin angle δ_s, so

that vehicle can carry out pitching motion. The vertical fins can control the rudder angle δ_r, to control the heading motion of the vehicle. In addition, this paper assumes that the propeller speed is constant at 1 500 rpm, and the REMUS vehicle maintains a speed of 1.51 m/s [27].

As can be seen from Fig. 2, taking depth control as an example, the depth set value $\hat{z}$ is used as the controller input to get the appropriate control quantity. The input of AUV motion control is fin angle δ_s, rudder angle δ_r, and propeller thrust X_{prop}.

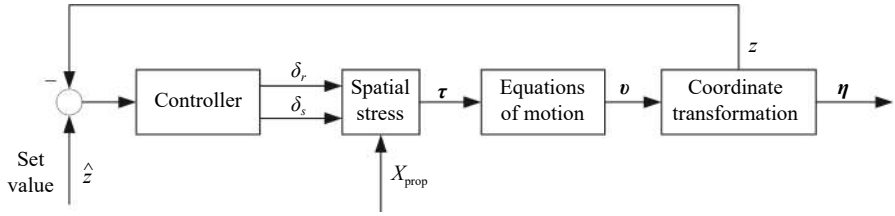


Fig. 2 AUV system work flowchart

3. LADRC controller

LADRC does not rely on the accurate mathematical model, and treats various uncertain factors in the controlled object as the total disturbance, uses linear extended state observer (LESO) to estimate the total disturbance and eliminate it, so as to suppress the influence of the disturbance [13]. The LADRC controller for an n-order system is shown in Fig.3.

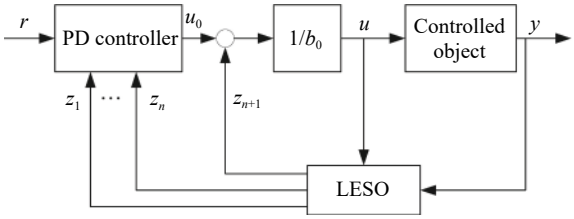


Fig. 3 LADRC basic control structure

The following content takes AUV depth control as the control target. According to (5), the AUV model can be regarded as a first-order system

$$\dot{y} = f + b_0 u \tag{6}$$

where f is the total disturbance. Set the state variable $x_1 = y$, $x_2 = f$, then $\boldsymbol{x} = \begin{bmatrix} y & f \end{bmatrix}^{\mathrm{T}}$ is the extended state including disturbance. Equation (6) is transformed into the description of the extended state space

$$\begin{cases} \dot{x}_1 = x_2 + b_0 u \\ \dot{x}_2 = \dot{f} \\ y = x_1 \end{cases}. \tag{7}$$

Construct an extended state observer for (7) to estimate the extended state x_2 [28] as

$$\begin{cases} \dot{z}_1 = z_2 + b_0 u + \beta_1 (y - z_1) \\ \dot{z}_2 = \beta_2 (y - z_1) \end{cases} \tag{8}$$

where $\boldsymbol{Z} = \begin{bmatrix} z_1 & z_2 \end{bmatrix}^{\mathrm{T}}$, $\boldsymbol{Z} \to \boldsymbol{x}$ is the state vector of the observer. For an n-order system, the observer gain coefficient [14] can be taken as: $[\beta_1 \beta_2 \cdots \beta_{n+1}] = [\omega_0 \alpha_1 \ \omega_0^2 \alpha_2 \cdots \omega_0^{n+1} \alpha_{n+1}]$, and $\alpha_i = \dfrac{(n+1)!}{i!(n+1-i)!}$, so the observer's gains in (8) are $\beta_1 = 2\omega_0$, $\beta_1 = {\omega_0}^2$, where ω_o is the observer bandwidth.

With a well-tuned LESO, we can get the estimate of the second state in (7), if the controller adopt the following form:

$$u = \frac{u_0 - z_2}{b_0}, \tag{9}$$

then (6) will be simplified as an integrator without dynamic uncertainty

$$\dot{y} = u_0. \tag{10}$$

Then a simple P control can be employed as

$$u_0 = \omega_c (r - y) \tag{11}$$

where ω_c is the controller bandwidth, and r is the given value of the system.

Finally, (8), (9), and (11) are combined into a LADRC controller for first-order systems.

4. LADRC controller based on reinforcement learning

In recent years, reinforcement learning has attracted extensive attention. For a sequential decision making process with Markov property, through the interaction between agent and environment, the strategy is constantly updated and optimized to finally realize value maximization.

4.1 Q-learning

Given the five elements of reinforcement learning [24]: action set A, state set S, reward R, attenuation factor γ, exploration rate ε, solve the optimal action value function q_* and the optimal strategy π_*. The Q-learning algorithm has two strategies:

(i) Greedy strategy

Q-learning uses the greedy strategy to update the value function as follows:

$$\pi_*(a|s)=\begin{cases}1, & \text{if } a=\underset{a\in A}{\operatorname{argmax}}\, q_*(s,a)\\ 0, & \text{else}\end{cases}.$$

(ii) $\varepsilon-$greedy strategy

The $\varepsilon-$greedy strategy is adopted to select new actions. By setting a value ε, the action that currently has the greatest action value is greedily accessed with the probability of $1-\varepsilon$, while the action is randomly selected from all m optional actions with the probability of ε.

$$\pi(a|s)=\begin{cases}\dfrac{\varepsilon}{m+1-\varepsilon}, & \text{if } a=\underset{a\in A}{\operatorname{argmax}}\, Q_*(s,a)\\ \dfrac{\varepsilon}{m}, & \text{else}\end{cases}$$

Q-learning uses this strategy to encourage exploration in action selection, so that as many actions as possible can be accessed. The steps of Q-learning algorithm are as follows:

Step 1 Algorithm initialization: state set S, action set A, learning rate α, attenuation factor γ, exploration rate ε.

Step 2 Initialize state $s\in S$.

Step 3 Use the $\varepsilon-$greedy strategy to select action a in the current state.

Step 4 Perform action a in current state s to get new state s' and reward R.

Step 5 Update value function

$$Q(s,a)=Q(s,a)+\alpha\left(R+\gamma\max_a Q(s',a)-Q(s,a)\right),\ s=s'.$$

Step 6 Learning ends when the termination condition is reached; otherwise, return to Step 3.

The complete algorithm flowchart is shown in Fig.4.

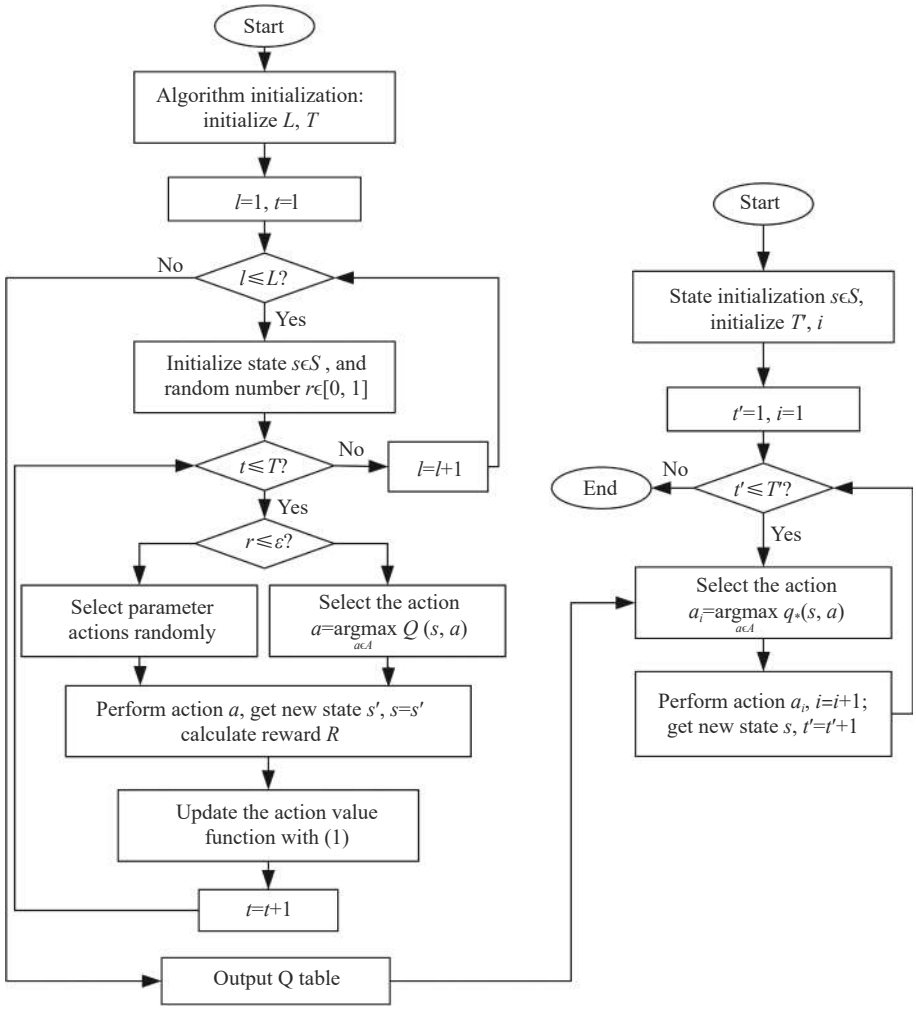


Fig. 4 Q-Learning algorithm flowchart

4.2 Q-learning algorithm design

In this subsection, based on the established model of six degrees of freedom AUV model, Q-learning algorithm is combined to realize the design of the adaptive LADRC

controller.

In order to implement parameter self-adaptation using the Q-learning algorithm, the dynamic parameter adjustment process is considered to be equivalent to the action selection process in the Q-learning algorithm. Therefore, reasonable state division and the design of reward function R have become important contents. There are two main considerations in controller design:

(i) Coupling between AUV heave motion and yaw motion.

(ii) With the increase of the types of states to be divided and the control parameters, the dimension of the state set S increases and the Q table becomes larger, which will lead to an increase in the amount of calculation in the learning process.

Aiming at the first problem, the AUV yaw controller is considered to be added in this paper, so that the AUV can maintain course stability during the sinking process.

There are two main solutions to the second problem. The first one is to reduce the number of control parameters. According to (8), (9) and (11), the parameters needed to be adjusted by the LADRC controller are ω_c, ω_o and b_0. It is worth mentioning that in the simulation experiment, the parameter b_0 can be approximated by the model calculation. For AUV system without time delay, b_0 can take the approximate value of the actual value of the system, while the LESO can still work normally [29,30]. Therefore, the parameters to be adjusted in the adaptive LADRC controller are simplified to ω_c and ω_o, and b_0 is fixed according to model calculation and experience. Finally, the structure of LADRC controller based on the Q-learning algorithm is shown in Fig.5, where ω_c, ω_o and $\tilde{\omega}_c$, $\tilde{\omega}_o$ are parameters of AUV depth and yaw controller respectively.

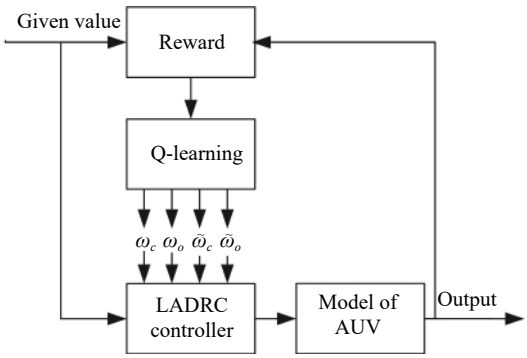


Fig. 5 AUV control system based on Q-Learning

The second one is to design the state division method. In order to avoid doubling the dimension of the controlled state set S caused by the dual controllers, this paper constructs a Q-learning state division method based on the main controlled state and constructs a reward design method that is not limited to the error of the divided state. Taking AUV sinking depth and attitude angle θ as the main controlled states, the division of the states is shown in Table 2, with a total of 25 states. e is defined as $e = \text{depth} - \text{state}$, where "depth" is set at 10 m and "state" is the real-time depth of AUV. The main function of this division is reflected in the "Initialize state s" and "get new state s'" processes on the left side of Fig. 4. The yaw motion of AUV is taken as the secondary controlled state. Its state error and the main controlled state participate in the reward design in the value function. The process is embodied in the "calculate reward R" on the left of Fig. 4.

Table 2 Division of the states

θ/rad	e/m				
	(−11,0.3]	(−0.3,0.1]	(−0.1,0.1)	[0.1,0.3)	[0.3,11)
(−1,−0.3]	1	6	11	16	21
(−0.3,−0.1]	2	7	12	17	22
(−0.1,0.1)	3	8	13	18	23
[0.1,0.3)	4	9	14	19	24
[0.3,1)	5	10	15	20	25

Then, the four-dimensional parameter space which can be selected by ω_c, ω_o, $\tilde{\omega}_c$, and $\tilde{\omega}_o$ is established. The parameter selection range here is

$$\omega_c \in [0.05 : 0.1 : 0.65],$$
$$\omega_o \in [2.4 : 0.05 : 2.7],$$
$$\tilde{\omega}_c \in [1.7 : 0.05 : 2],$$
$$\tilde{\omega}_o \in [2.7 : 0.05 : 3],$$

in total 2 401 parameter combinations are available.

After the above state and parameter division, the Q-table size of LADRC controller based on the Q-learning algorithm (Q-LADRC) is 2 401×25. Similarly, since there are six parameters to be adjusted for the dual PID controller, the scale of the Q-table of PID controller based on the Q-learning algorithm (Q-PID) is 117 649×25.

5. Simulation results analysis

As one of the reinforcement learning methods, Q-learning algorithm is most widely used. Therefore, Q-learning algorithm is compared with another reinforcement learning algorithm in this section, and it proves the advantages of Q-learning algorithm in some aspects. In order to verify the view that the controller parameter

changes caused by AUV state changes will improve the AUV control performance in the process of AUV sinking and resisting external disturbance, this section simulates AUV's sinking motion and adds the disturbance of set values to verify the controller's disturbance rejection performance based on the Q-learning algorithm. In addition, the LADRC method is compared with PID method to verify the superiority of LADRC method in some aspects.

5.1 Comparison between Q-learning algorithm and Sarsa algorithm

In addition to the off-policy Q-learning algorithm, the temporal-difference method in reinforcement learning also has the on-policy Sarsa algorithm. Sarsa algorithm adopts the $\varepsilon-$greedy strategy in both value function update and action selection. In order to conduct comparative experiments, the structural design and simplification of Sarsa algorithm are the same as the Q-learning algorithm in Subsection 4.2, which will not be repeated here.

Because Sarsa algorithm is relatively conservative in updating the value function, the convergence speed of the algorithm itself will be slower. Fig. 6 shows the length of each episodes between the Q-LADRC controller and the Sarsa-based LADRC controller (S-LADRC) in 1 500 episodes.

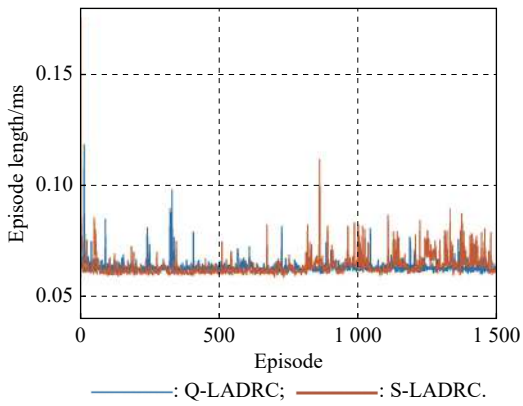


Fig. 6 Length of each episode

Due to the randomness of the AUV state at the beginning of each training and the complexity of the AUV movement, it may take a long time for individual episodes. Excluding the above influencing factors, it can be seen from the Fig. 6 that the Sarsa algorithm adopts the random value update strategy, which makes the most of episodes longer in the later stage of convergence.

Increase the number of episodes, that is, increase the number of training, which enables S-LADRC to achieve similar control effect as Q-LADRC. As shown in Fig. 7, when the number of training of S-LADRC controller reaches 4 500, it has similar depth control effect with the Q-LADRC controller after 1 500 times of training.

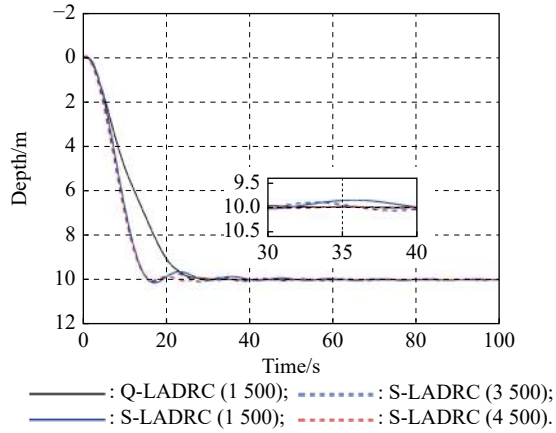


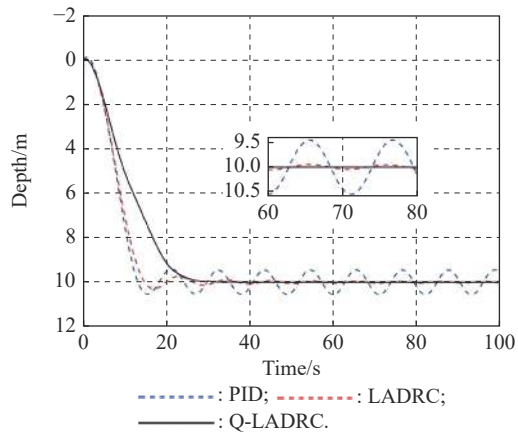
Fig. 7 Depth control effects of two controllers with different training times

The Q-learning algorithm tends to maximize the Q value, while the Sarsa algorithm can avoid errors to a certain extent. Sarsa has a slow convergence speed, but it can improve the training effect by increasing the number of training times.

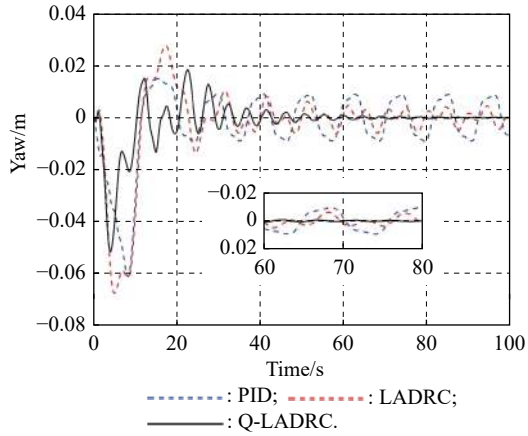
In the simulation experiment of AUV, it is found that the rapidity and smoothness of AUV sinking motion cannot be satisfied at the same time. Therefore, a Q-LADRC controller which can make the AUV motion smoother after Q-Learning training is adopted in the subsequent simulation experiment.

5.2 Parameters fixed controller and controller based on Q-learning algorithm

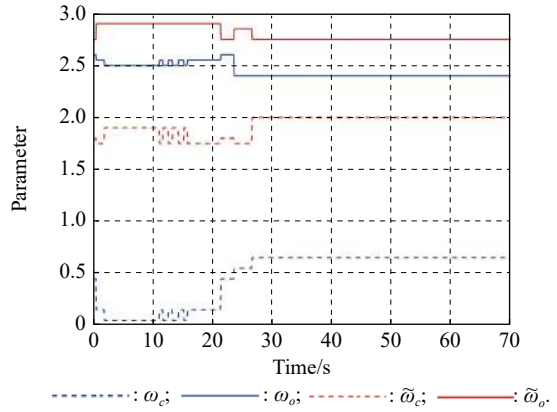
Fig. 8(a) and Fig. 8(b) show the control effect comparison between PID, LADRC controllers with fixed parameters and Q-LADRC controller. Fig. 8(c) shows the changes of parameters caused by the state changes of the AUV when using the Q-LADRC controller.



(a) Time-depth image



(b) Time-yaw image



(c) Parameter variations of the Q-LADRC controller

Fig. 8 Parameters fixed PID, LADRC controller and Q-LADRC controller

It can be seen from the Fig. 8 that the LADRC method is effective in AUV motion control. The PID controller can quickly generate control quantity to meet the requirements of the system, but when the PID controller meets the speed, its control effect in the final stable state of the system is deficient. Compared with PID control method, LADRC gives AUV higher motion stability.

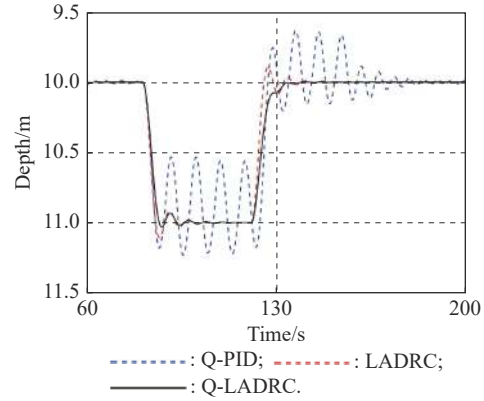
In addition, the data shows that the final depth and yaw error using the Q-LADRC controller are both less than 10^{-3} m. It can be seen from the data and figures that adjusting parameters according to the state in real time has a positive impact on the control effect. At the same time, compared with the controller with fixed parameters, the Q-LADRC controller gives the AUV smaller yaw movement in the process of sinking. Therefore, although the yaw motion state of AUV is not divided in the learning process of the Q table, Q-learning algorithm can update the action value function according to the return value with yaw error, so as to successfully find the parameters of the yaw Q-LADRC controller. This proves that the state division method and reward design of the constructed Q-learning algorithm are reasonable and effective.

The Q-PID controller can achieve a similar control effect with the Q-LADRC controller on AUV sinking motion control. It will not be repeated here. However, the contradiction between its rapidity and stability still exists when it comes to disturbance rejection.

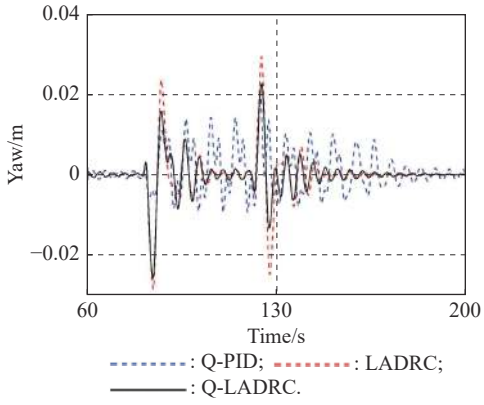
5.3 Change of set value

In order to compare the control effect of the controller in the face of abrupt state change, set value change and other uncertain factors, based on the AUV sinking control in the previous section, change the 80 s to 120 s depth setting from 10 m to 11 m. Simulation studies the control effect of parameters fixed LADRC, Q-LADRC, and Q-PID controllers.

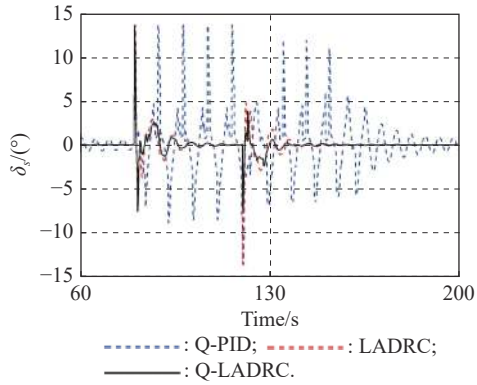
When AUV resisting the disturbance of set value, the overshoot and oscillation of AUV can be reduced by proper parameter adjustment, and the control quantity is still kept in a reasonable range, as shown in Fig. 9(a), Fig. 9(b), and Fig. 9(c). The fin angle of AUV using Q-PID controller changes quickly and responds quickly to the system. However, the disadvantages of PID controller are not changed, PID controller will cause system oscillation and severe overshoot due to excessive initial control force, it takes a long time for AUV to stabilize around the set value of the system. Fig. 9(d) shows the parameter changes of Q-LADRC controller during AUV following the set value.



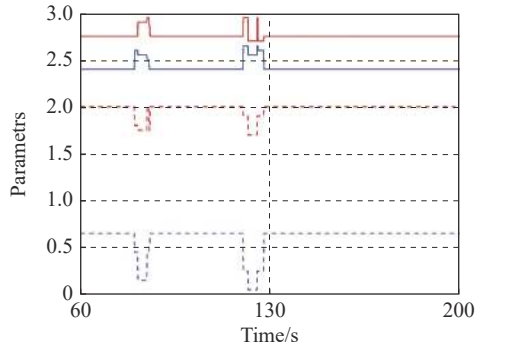
(a) Time-depth image



(b) Time-yaw image



(c) Time-fin angle image



(d) Parameter variations of Q-LADRC controller

Fig. 9 Parameters fixed LADRC, Q-PID and Q-LADRC controller

6. Conclusions

It is an important research topic to design a control method to make AUV have excellent motion performance. While using the LADRC controller to decouple the AUV movement, it also realizes the adaptive adjustment of the controller parameters combined with the reinforcement learning algorithm. Simplifying a part of the structure in the reinforcement learning algorithm avoids the "curse of dimensionality" to a certain extent. In a system with continuously changing state, constant adjustment of controller parameters is beneficial to the final stability of the system. At the same time, simulation experiments verify the effectiveness of the constructed Q-LADRC controller in AUV motion control. Although the value function update of Q-Learning algorithm is relatively risky, the algorithm has faster convergence speed and less time cost. Compared with the controller with fixed parameters, the AUV using Q-LADRC controller has lower overshoot and better motion performance in disturbance rejection. By comparing the control effects of PID and LADRC controllers, it is found that for slowly changing control objects such as AUV, when the control accuracy and stability of the controlled object have higher requirements, the LADRC method has higher applicability than the PID method.

References

[1] ROBERT B. Underwater robots: a review of technologies and applications. Industrial Robot: An International Journal, 2015, 42(3): 186–191.

[2] ZHANG F M, MARANI G, SMITH R N, et al. Future trends in marine robotics [TC Spotlight]. IEEE Robotics & Automation Magazine, 2015, 22(1): 14–122.

[3] RYOSUKE K, SATOSHI O. Development of hovering control system for an underwater vehicle to perform core internal inspections. Journal of Nuclear Science and Technology, 2016, 53(4): 566–573.

[4] MANECIUS S J, ASOKAN T. Station keeping control of underwater robots using disturbance force measurements. Journal of Marine Science and Technology, 2016, 21(1): 70−85.

[5] SATO Y, MAKI T, KUME A, et al. Path replanning method for an AUV in natural hydrothermal vent fields: toward 3D imaging of a hydrothermal chimney. Marine Technology Society Journal, 2014, 48(3): 104–114.

[6] RIBAS D, PALOMERAS N, RIDAO P, et al. Girona 500 AUV: from survey to intervention. IEEE/ASME Trans. on Mechatronics, 2012, 17(1): 46–53.

[7] ANTONELLI G. Underwater robots: motion and force control of vehicle-manipulator systems. Switzerland: Springer, 2010.

[8] PRZEMYSŁAW H. Decoupled PD set-point controller for underwater vehicles. Ocean Engineering, 2009, 36(6): 529–534.

[9] TAHA E, MOHAMED Z, KAMAL Y T. Control for dynamic positioning and way-point tracking of underactuated autonomous underwater vehicles using sliding mode control. Journal of Intelligent & Robotic Systems, 2019, 95(3/4): 1113–1132.

[10] MOHAMMAD H K, SAEED B. Modeling and control of autonomous underwater vehicle (AUV) in heading and depth attitude via self-adaptive fuzzy PID controller. Journal of Marine Science and Technology, 2015, 20(3): 559–578.

[11] XUE Q. Adaptive coordinated tracking control of multiple autonomous underwater vehicles. Ocean Engineering, 2014, 91: 84–90.

[12] HAN J Q. Auto-disturbance-rejection controller and its applications. Control and Decision, 1998, 13(1): 19 –23. (in Chinese)

[13] HAN J Q. From PID to active disturbance rejection control. IEEE Trans. on Industrial Electronics, 2009, 56(3): 900–906.

[14] GAO Z Q. Scaling and bandwidth-parameterization based controller tuning. Proc. of the American Control Conference, 2006: 4989−4996.

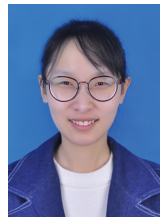
[15] LIU J J, SUN M W, CHEN Z Q, et al. High AOA decoupling control for aircraft based on ADRC. Journal of Systems Engineering and Electronics, 2020, 31(2): 393–402.

[16] CHEN Z Q, QIN B B, SUN M W, et al. Q-learning-based parameters adaptive algorithm for active disturbance rejection control and its application to ship course control. Neurocomputing, 2020, 408: 51–63.

[17] ZHENG Y M, CHEN Z Q, HUANG Z Y, et al. Active disturbance rejection controller for multi-area interconnected power system based on reinforcement learning. Neurocomputing, 2021, 425: 149–159.

[18] LI J H, LEE P M. Design of an adaptive nonlinear controller for depth control of an autonomous underwater vehicle. Ocean Engineering, 2005, 32(17/18): 2165–2181.
[19] LIU S Y, WANG D W, POH E. Non-linear output feedback tracking control for AUVs in shallow wave disturbance condition. International Journal of Control, 2008, 81(11): 1806–1823.
[20] XIANG X B, YU C Y, ZHANG Q. Robust fuzzy 3D path following for autonomous underwater vehicle subject to uncertainties. Computers and Operations Research, 2016, 84: 165–177.
[21] LIANG X, QU X R, WANG N, et al. Three-dimensional trajectory tracking of an underactuated AUV based on fuzzy dynamic surface control. IET Intelligent Transport Systems, 2020, 14(5): 364−370.
[22] LI Y, QIU X H, LIU X D, et al. Deep reinforcement learning and its application in autonomous fitting optimization for attack areas of UCAVs. Journal of Systems Engineering and Electronics, 2020, 31(4): 734–742.
[23] SHEN Y X, SHAO K Y, REN W J, et al. Diving control of autonomous underwater vehicle based on improved active disturbance rejection control approach. Neurocomputing, 2016, 173(3): 1377–1385.
[24] SUTTON R, BARTO A. Reinforcement learning: an introduction. Massachusetts: MIT Press, 1998.
[25] LOW E S, ONG P, CHEAH K C. Solving the optimal path planning of a mobile robot using improved Q-learning. Robotics and Autonomous Systems, 2019, 115: 143–161.
[26] XIANG X B, YU C Y, ZHANG Q, et al. Path-following control of an AUV: fully actuated versus under-actuated configuration. Marine Technology Society Journal, 2016, 50(1): 34–47.
[27] PRESTERO T. Verification of a six-degree of freedom simulation model for the REMUS autonomous underwater vehicle. Massachusetts Institute of Technology, 2001. DOI: 10.1575/1912/3040.
[28] YANG R, SUN M W, CHEN Z Q. Active disturbance rejection control on first-order plant. Journal of Systems Engineering and Electronics, 2011, 22(1): 95–102.
[29] TANG D, GAO Z Q, ZHANG X H. Design of predictive active disturbance rejection controller for turbidity. Control Theory and Applications, 2017, 34(1): 101–108.
[30] XUE W C, HUANG Y. Performance analysis of active disturbance rejection tracking control for a class of uncertain LTI systems. ISA Transactions, 2015, 58: 133–154.

Biographies



SONG Wanping was born in 1998. She received her B.S. degree from Nanjing Agricultural University, Nanjing, China, in 2019. She is currently a graduate student of Nankai University, Tianjin, China. Her current research interests include active disturbance rejection control and reinforcement learning.
E-mail: 1422501596@qq.com

CHEN Zengqiang was born in 1964. He received his B.S., M.E., and Ph.D. degrees from Nankai University, in 1987, 1990, and 1997, respectively. He is currently a professor of control theory and engineering of Nankai University, and deputy director of Institute of Robotics and Information Automation. His current research interests include intelligent predictive control, chaotic systems and complex dynamic network, and multi-agent system control.
E-mail: chenzq@nankai.edu.cn

SUN Mingwei was born in 1972. He received his Ph.D. degree from the Department of Computer and Systems Science, Nankai University, Tianjin, China, in 2000. From 2000 to 2008, he was a Flight Control Engineer with Beijing Electromechanical Engineering Research Institute, Beijing, China. Since 2009, he has been with Nankai University as a professor. His research interests include flight control, guidance, model predictive control, active disturbance rejection control, and nonlinear optimization.
E-mail: smw_sunmingwei@163.com

SUN Qinglin received his B.E. and M.E. degrees in control theory and control engineering from Tianjin University, Tianjin, China, in 1985 and 1990, respectively, and his Ph.D. degree in control science and engineering from Nankai University, Tianjin, China, in 2003. He is currently a professor at the Intelligence Predictive Adaptive Control Laboratory of Nankai University and associate dean of College of Artificial Intelligence. His research interests include self-adaptive control, modeling and control of flexible spacecraft, and embedded control systems.
E-mail: sunql@nankai.edu.cn