J. of Marine Eng. Vol. 21, Iss. 46, 2025. (71-79)

Journal of Marine Engineering

Journal homepage: marine-eng.ir

Depth control of an AUV robot using reinforcement learning (RL)

Ali Hasanvand[1], **Mohammad Saeed Seif**[2*]

[1] *Postdoctoral Researcher, Faculty of Mechanical Engineering, Sharif University of Technology, Ali.hasanvand@sharif.edu*
[2] *Professor, Faculty of Mechanical Engineering, Sharif University of Technology, Seif@sharif.edu*

ARTICLE INFO

Article History:
Received: 13 May 2025
Last modification: 16 Jul 2025
Accepted: 17 Jul 2025
Available online: 17 Jul 2025

Article type:
Research paper

Keywords:
Reinforcement learning
Depth Control
AUV Robot
Underactuated

ABSTRACT

Nowadays, the use of advanced methods for controlling the motion of underwater robots has led to improved efficiency and enhanced operational quality. In this research, a method based on reinforcement learning has been developed for the depth control of AUV robots. This method learns the robot's movement pattern based on a reward criterion and makes the optimal decision for motion and control surface adjustments accordingly. Depth control using reinforcement learning improves the robot's performance and selects the most optimal control signal based on the robot's conditions and rewards. In this study, a linear dynamic model of pitch motion was used to develop the depth control model. For each desired state, the scenario is repeated 500 times to update the Q-matrix during simulation. Subsequently, by assigning rewards to each signal, the optimal value is determined. After completing the scenario, the optimal value from the Q-matrix is selected to determine the control signal for the fin. The results showed that the use of reinforcement learning significantly enhances the quality of the AUV robot's control system, resulting in minimal overshoot and oscillation in performance.

[Downloaded from marine-eng.ir on 2026-11-13]

[DOI: 10.61882/marineeng.21.46.6]

ISSN: 2645-8136

DOI: http://dx.doi.org/10.61882/marineeng.21.46.6

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license [https://creativecommons.org/licenses/by/4.0/]

نشریه مهندسی دریا، جلد ۲۱، شماره ۴۶، ۱۴۰۴، (۷۹-۷۱)

نشریه مهندسی دریا

صفحه اصلی نشریه: marine-eng.ir

کنترل عمق عملیاتی ربات زیرسطحی AUV با روش یادگیری تقویتی (RL)

علی حسنوند[۱]، محمد سعید سیف[۲]*

[۱] پسا دکتری، دانشکده مهندسی مکانیک، دانشگاه صنعتی شریف، Ali.hasanvand@sharif.edu

[۲] استاد، دانشکده مهندسی مکانیک، دانشگاه صنعتی شریف، Seif@sharif.edu

اطلاعات مقاله

تاریخچه مقاله:

تاریخ دریافت مقاله: ۱۴۰۴/۰۲/۲۳

تاریخ اصلاح مقاله: ۱۴۰۴/۰۴/۲۵

تاریخ پذیرش مقاله: ۱۴۰۴/۰۴/۲۶

تاریخ انتشار مقاله: ۱۴۰۴/۰۴/۲۶

نوع مقاله:

مقاله پژوهشی

کلمات کلیدی:

یادگیری تقویتی

کنترل عمق

ربات AUV

Underactuated

چکیده

امروزه استفاده از روش‌های پیشرفته برای کنترل حرکات ربات‌های زیرسطحی سبب بهبود راندمان و افزایش کیفیت عملیات آن ها شده است. در این پژوهش با استفاده از روش یادگیری تقویتی برای حرکت عمقی ربات‌های AUV روشی توسعه داده شده است که براساس معیار پاداش الگوی حرکت ربات را یاد گرفته و براساس آن بهترین تصمیم را برای حرکت و کنترل سطوح کنترلی اتخاذ می‌کند. کنترل حرکت عمقی براساس یادگیری تقویتی سبب بهبود عملکرد ربات میگردد و بهینه‌ترین سیگنال کنترلی را براساس شرایط لحظه‌ای ربات و پاداش‌ها اتخاذ می‌کند. در این پژوهش از مدل دینامیکی خطی حرکت پیچ برای توسعه مدل حرکت عمقی استفاده شده است. برای هر هدف مطلوب ۵۰۰ مرتبه سناریو تکرار می‌شود تا در حین شبیه‌سازی ماتریس Q به روزرسانی شود. در ادامه با ارائه پاداش به هر سیگنال مقدار مطلوب مشخص می‌گردد. پس از پایان سناریو، با انتخاب مقدار بهینه از ماتریس Q، مقدار سیگنال کنترلی برای بالک مشخص می‌گردد. نتایج نشان داد که استفاده از روش یادگیری تقویتی کمک شایانی به کیفیت سیستم کنترل ربات‌های AUV می‌کند تا جایی که مقدار فرارفت و نوسان کمی در عملکرد مشاهده شد.

ISSN: 2645-8136

DOI: http://dx.doi.org/10.61882/marineeng.21.46.6

حق نشر: © ۲۰۲۵ توسط نویسندگان. این اثر برای انتشار با دسترسی آزاد، تحت شرایط و ضوابط مجوز (CC BY) ارسال شده است.

[Downloaded from marine-eng.ir on 2026-11-13]

[DOI: 10.61882/marineeng.21.46.6]

۱ – مقدمه

استفاده از روش‌های یادگیری تقویتی[1] برای کنترل حرکات ربات زیرسطحی AUV به دلیل توانایی این روش‌ها در یادگیری سیاست‌های بهینه در محیط‌های پویا و نامعلوم، مورد توجه قرار گرفته است. پژوهش‌های اخیر نشان داده‌اند که الگوریتم‌های RL مانند Q-Learning و DQN[2] می‌توانند برای کنترل مسیر و عمق AUVها با دقت بالا استفاده شوند. به عنوان مثال، استفاده از DQN می‌تواند خطای تعقیب مسیر را تا ۳۰٪ در مقایسه با روش‌های کنترل کلاسیک کاهش دهد [۱].

در تحقیقات دیگری، از روش‌های مبتنی بر Actor-Critic مانند PPO[3] برای بهبود پایداری و کارایی کنترل حرکات AUV استفاده شده است. این روش‌ها به دلیل توانایی در مدیریت فضای عمل پیوسته و کاهش نوسانات در فرآیند یادگیری، برای کنترل دقیق‌تر AUVها مناسب هستند. روش PPO در مقایسه با روش‌های مبتنی بر ارزش، عملکرد بهتری در محیط‌های دارای نویز و اغتشاشات دارد [۲]. همچنین، این مطالعه تأکید می‌کند که ترکیب RL با مدل‌های دینامیکی AUV می‌تواند زمان همگرایی الگوریتم را کاهش دهد.

یکی از چالش‌های اصلی در استفاده از RL برای کنترل AUVها، نیاز به حجم زیاد داده‌های آموزشی و زمان محاسباتی طولانی است. برای حل این مشکل، برخی محققان از روش‌های ترکیبی مانند Transfer Learning و Imitation Learning استفاده کرده‌اند. مطالعات نشان می‌دهد که با استفاده از یادگیری انتقالی، می‌توان مدل‌های آموزش‌دیده در شبیه‌ساز را به محیط واقعی تعمیم داد و در نتیجه، هزینه‌های آزمایش‌های فیزیکی را کاهش داد [۳]. این پژوهش همچنین پیشنهاد می‌کند که ترکیب RL با کنترلرهای مدل پیش‌بین[4] می‌تواند عملکرد سیستم را در شرایط عملیاتی پیچیده بهبود بخشد.

اخیراً، استفاده از روش‌های یادگیری تقویتی چندعامله[5] برای کنترل هماهنگ چندین AUV در مأموریت‌های اکتشافی و نظارتی مورد توجه قرار گرفته است. این روش‌ها امکان همکاری بین AUVها را فراهم می‌کنند و می‌توانند منجر به بهبود کارایی در انجام وظایف پیچیده مانند نقشه‌برداری از بستر دریا شوند. الگوریتم‌های مبتنی بر MADDPG[6] می‌توانند هماهنگی بهتری بین‌AUVها ایجاد کنند و در مقایسه با روش‌های تک‌عامله، دقت بیشتری در ردیابی مسیرهای از پیش تعیین‌شده دارند [۴]. همچنین، استفاده از این روش‌ها می‌تواند مصرف انرژی را تا ۲۰٪ کاهش دهد، که برای مأموریت‌های طولانی‌مدت زیرسطحی پیچیده است [۵].

یکی دیگر از زمینه‌های تحقیقاتی نوظهور، ترکیب یادگیری تقویتی با شبکه‌های عصبی اسپایکی[7] برای کنترل AUV ها با مصرف انرژی بهینه است. این روش‌ها به دلیل شباهت بیشتر به سیستم‌های عصبی بیولوژیکی، می‌توانند محاسبات را با کارایی بالاتری انجام دهند و برای کاربردهای ناگهانی مناسب‌تر هستند. SNNهای آموزش‌دیده با RL می‌توانند زمان پاسخگویی AUV ها را در مواجهه با موانع پیش بینی نشده تا ۴۰٪ بهبود ببخشند [۶]. همچنین، پژوهش‌های [۷] و [۸] به بررسی تأثیر معماری‌های مختلف شبکه‌های عصبی در ترکیب با RL پرداختند و نتایج نشان داد که استفاده از LSTM[8] می‌تواند به AUVها در یادگیری رفتارهای پیچیده در محیط‌های پویا کمک کند.

نوآوری پژوهش حاضر در به‌کارگیری یک چارچوب یادگیری تقویتی عمیق برای کنترل هوشمند حرکت عمقی ربات‌های زیرسطحی با استفاده از سطوح کنترلی است. در این روش، با ترکیب مدل دینامیکی خطی‌شده حرکت پیچ و الگوریتم Q-Learningپیشرفته، سیاست بهینه کنترل به‌صورت خودکار و مبتنی بر داده‌های محیطی یادگیری می‌شود. طراحی یک الگوی پاداش تطبیقی که نه‌تنها خطای عمق، بلکه مصرف انرژی و پایداری دینامیکی را نیز به‌صورت همزمان بهینه‌سازی می‌کند میتواند کمک شایانی به سیستم کنترل این دسته ربات‌ها داشته باشد. همچنین، استفاده از تکرار ۵۰۰ مرحله‌ای برای هر سناریو و به‌روزرسانی ماتریس Q به‌صورت همگرا، دقت و کارایی کنترلر را در شرایط غیرخطی زیرسطحی بهبود می‌بخشد. این رویکرد قادر است بدون نیاز به مدل دقیق دینامیکی، بهترین سیگنال کنترلی را برای بالک‌ها نسبت به روش‌های کلاسیک کنترل نشان دهد و همچنین انعطاف‌پذیری و مقاومت بیشتری در برابر اغتشاشات محیطی از خود نشان دهد.

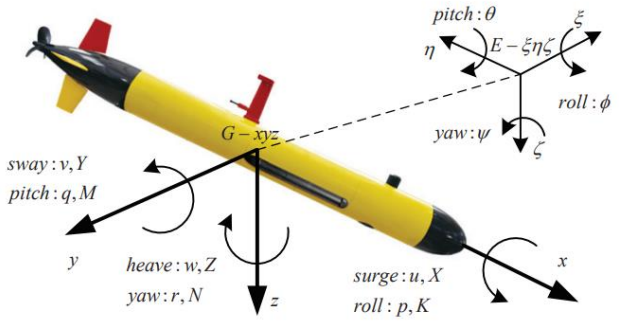
۲ – معادلات حرکت

شبیه‌سازی معادلات حرکت یک ربات زیرسطحی به حل همزمان شش معادله دیفرانسیل غیرخطی کوپل شده احتیاج دارد. برای مدلسازی دینامیکی همواره دو دستگاه مختصات تعریف می‌شود. اولی فیکس شده در قاب مرجع متصل به زمین و بی حرکت است. دومی متصل به بدنه ربات می‌باشد. برای کاهش پارامترها، دستگاه مختصات متصل به بدنه در مرکز بویانسی در نظر گرفته می‌شود

[1] Reinforcement Learning (RL)
[2] Deep Q-Networks
[3] Proximal Policy Optimization
[4] MPC
[5] Multi-Agent RL
[6] Multi-Agent Deep Deterministic Policy Gradient
[7] Spiking Neural Networks (SNN)
[8] Long Short-Term Memory

[Downloaded from marine-eng.ir on 2026-11-13]
[DOI: 10.61882/marineeng.21.46.6]

[۱۱-۹]. شکل (۱) وضعیت نسبی دستگاه مختصات‌های تعریف شده را نشان می‌دهد.



شکل ۱- دستگاه مختصات های تعریف شده و شماتیک حرکت‌های شش درجه آزادی [۱۲]

با استفاده از فرضیات می‌توان حرکت پیچ را که یک حرکت مهم در کنترل عمق ربات‌های underactuated بحساب می‌آید از سایر حرکات جدا کرد. با خطی سازی معادله حرکت پیچ ربات AUV، مدل دینامیکی حرکت پیچ استخراج می‌شود. معادله حاکم بر حرکت پیچ در رابطه (۱) ارائه شده است. معادله (۱) شامل ترم‌های بازگرداننده، ممان هیدرودینامیکی، دینامیک جسم صلب و نیروی بالک‌ها است.

$$\left(I_Y + M_{\dot{q}}\right)\dot{q} + M_q q + M_\theta \theta + M_u u = \tau_{dis} \qquad (۱)$$

که در اینجا I_Y ممان اینرسی جرمی حول محور Y، $M_{\dot{q}}$ جرم افزوده حرکت پیچ، M_q دمپینگ حرکت پیچ، M_θ ضریب بازگرداننده، M_u ضریب عملگر کنترلی، τ_{dis} اغتشاش خارجی، u سیگنال کنترل و $\theta, q, \dot{q}$ به ترتیب شتاب، سرعت و مقدار پیچ هستند. در ادامه بعد از حل مدل دینامیکی در هر گام زمانی، برای محاسبه میزان زاویه پیچ براساس معادلات سینماتیکی عمل شده است. مقدار تغییرات زاویه پیچ به صورت معادله (۲) بیان میگردد.

$$\dot{\theta} = q \qquad (۲)$$

مدل فضای حالت فرمت استاندارد مناسبی است که می‌تواند پل مناسبی بین دینامیک و کنترل سیستم باشد. فرمت عمومی فضای حالت به مانند معادله (۳) است.

$$\begin{aligned} \dot{x} &= Ax + Bu + E \\ y &= Cx \end{aligned} \qquad (۳)$$

که در آن x بردار متغیرهای حالت $_{(n\times 1)}$، $\dot{x}$ نرخ زمانی بردار حالت $_{(n\times 1)}$، u ورودی کنترلر $_{(p\times 1)}$، y بردار خروجی $_{(q\times 1)}$، A ماتریس سیستم $_{(n\times n)}$، B ماتریس ورودی $_{(n\times p)}$ و C ماتریس خروجی $_{(q\times n)}$ است. مدل دینامیکی حرکت پیچ در فرمت ماتریسی به شکل (۴) بیان میگردد.

$$\begin{bmatrix} \dot{\theta} \\ \dot{q} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \dfrac{M_q}{I_Y + M_{\dot{q}}} & \dfrac{M_\theta}{I_Y + M_{\dot{q}}} \end{bmatrix} \begin{bmatrix} \theta \\ q \end{bmatrix} + \begin{bmatrix} 0 \\ \dfrac{M_u}{I_Y + M_{\dot{q}}} \end{bmatrix} u + \begin{bmatrix} \tau_{dis1} \\ \tau_{dis2} \end{bmatrix} \qquad (۴)$$

$$C = \begin{bmatrix} 1 & 0 \end{bmatrix} \qquad (۵)$$

برای کنترل عمق ربات، با استفاده از بالک، مقدار پیچ به شکلی کنترل می‌شود که عمق ربات به سمت مقدار مطلوب حرکت کند. با فرض حرکت ربات در نزدیکی سرعت ۱ متر برثانیه، در نهایت معادله حرکت عمقی ربات با استفاده از روابط سینماتیکی به شکل معادله (۶) بیان می‌شود.

$$\begin{aligned} \dot{z} &= \sin(\theta) \\ z &= \int_0^t \dot{z} dt + z_0 \end{aligned} \qquad (۶)$$

که در اینجا z عمق لحظه‌ای ربات می‌باشد و می‌توان با حل مدل دینامیکی یک درجه آزادی حرکت پیچ، مقدار عمق را در هر گام زمانی تخمین زد. برای مدلسازی از مشخصات دینامیکی یک نمونه ربات زیرسطحی واقعی بهره گرفته شده است. از ضرایب هیدرودینامیکی ربات AUV isimi برای مدلسازی دینامیکی در این مقاله استفاده شده است [۱۳]. در جدول (۱) ضرایب هیدرودینامیکی این ربات ارائه شده است.

شکل ۲- ربات *AUV ISIMI* [۱۳]

جدول ۱- ضرایب هیدرودینامیکی ربات *AUV ISIMI* [۱۳]

	parameter	value	unit
1	I_y	7.5	kg.m^2
2	$M_{\dot{q}}$	0.09	kg.m^2
3	M_q	-5	kg.m^{2}s
4	M_θ	-1.4	kg.m^2
5	M_u	-0.8	kg.m^2

[Downloaded from marine-eng.ir on 2026-11-13]

[DOI: 10.61882/marineeng.21.46.6]

۳ – یادگیری تقویتی برای کنترل عمق

روش یادگیری تقویتی یکی از شیوه‌های اصلی یادگیری ماشین است که در آن یک عامل[9] از طریق تعامل با محیط و دریافت بازخورد (پاداش یا جریمه) یاد می‌گیرد تا رفتار بهینه را برای رسیدن به اهداف بلندمدت انتخاب کند. برخلاف یادگیری نظارت‌شده که به داده‌های ازپیش‌برچسب‌خورده نیاز دارد، RL بر پایه‌ی آزمون و خطا و بهینه‌سازی سیاست‌های تصمیم‌گیری استوار است. این روش در کاربردهای متنوعی مانند رباتیک، بازی‌های هوش مصنوعی (مثل AlphaGo)، مدیریت منابع و سیستم‌های توصیه‌گر مورد استفاده قرار گرفته است. الگوریتم‌های کلیدی RL مانند Q-Learning، DQN و Policy Gradient به عامل کمک می‌کنند تا در محیط‌های پیچیده با عدم قطعیت، تصمیم‌های هوشمندانه بگیرد.

همانطور که بیان شد، یادگیری تقویتی یک روش یادگیری ماشین است که در آن یک عامل از طریق تعامل با محیط، اقداماتی را انجام می‌دهد و بازخورد (پاداش یا جریمه) دریافت می‌کند. هدف عامل یادگیری یک سیاست بهینه است که حداکثر پاداش تجمعی را در طول زمان به دست آورد. این فرآیند معمولاً با استفاده از مفاهیم ریاضی مانند معادله بلمن در رابطه (۷) مدلسازی می‌شود.

$$V^{\pi}(s)=\sum_{a}\pi(a|s)\sum_{s',r}p(s',r|s,a)\left[r+\gamma V^{\pi}(s')\right] \qquad (۷)$$

که در آن $V^{\pi}(s)$ مقدار حالت s تحت سیاست π را نشان می‌دهد، γ فاکتور تخفیف است و $p(s',r|s,a)$ تابع انتقال محیط است. در روش‌های پیشرفته‌تر مانند Q-Learning، عامل جدولی از مقادیر $Q(s,a)$ را یاد می‌گیرد که نشان‌دهنده کیفیت انجام عمل a در حالت s است. الگوریتم با به‌روزرسانی مقدار Q بر اساس معادله زیر کار میکند.

$$Q(s,a)\leftarrow \\ Q(s,a)+\alpha(r+\gamma\max Q(s',a')-Q(s,a)) \qquad (۸)$$

که در آن α نرخ یادگیری است. این روش‌ها به عامل اجازه می‌دهند تا در محیط‌های پیچیده، سیاست بهینه را حتی بدون مدل دقیق محیط یاد بگیرند.

در ربات‌های AUV، کنترل عمق یکی از چالش‌های اساسی است که باید با دقت بالا و در شرایط واقعی اقیانوس انجام شود. یادگیری تقویتی با قابلیت یادگیری با تعامل مستقیم با محیط، یک راه‌حل مؤثر برای طراحی کنترلرهای هوشمند است. در این روش، عامل به‌عنوان کنترلر AUV عمل می‌کند و با دریافت حالت‌هایی مانند عمق فعلی، زاویه حمله، سرعت و داده‌های سنسورهای فشار، اقدام مناسب را برای تنظیم زوایای بالک‌ها را انتخاب می‌کند. محیط و معادلات دینامیکی AUV، پاداش را محاسبه می‌کند که مبتنی بر خطای عمق (تفاوت با عمق مطلوب) و مصرف انرژی است. با استفاده از الگوریتم‌های RL عامل به‌تدریج یک سیاست بهینه یاد می‌گیرد که نه‌تنها خطای عمق را به حداقل می‌رساند، بلکه نوسانات و تغییرات ناگهانی جریان آب را نیز جبران می‌کند. این روش به‌ویژه برای ربات های زیرسطحی که سطوح کنترلی دارند، می‌توانند منجر به کنترل دقیق‌تر و مقاوم‌تر در برابر اغتشاشات شوند.

۴ – آنالیز الگوریتم و تحلیل عملکرد

برای بررسی عملکرد الگوریتم و روش شناسی، عملکرد الگوریتم تحت مقادیر متفاوت نرخ یادگیری و اپیزود بررسی و تحلیل شد. مقادیر فرارفت، زمان نشست و پاداش‌های انباشته شده برای تمامی حالت‌ها بررسی شدند. در جدول زیر مقادیر نامبرده شده برای نرخ یادگیری و اپیزودهای مختلف ارائه شده است.

جدول ۲- تحلیل عملکرد الگوریتم

no. Episode	Learning rate	Cumulative reward (-)	Overshoot (%)	Settling Time (s)
500	0.05	13521	2.3	7.46
500	0.1	15030	1.1	3.82
500	0.15	13733	2	5.16
500	0.2	-8620	27	12.22
1	0.1		-	-
100	0.1		3.4	5.26
200	0.1		5.2	3.34
300	0.1		4.2	3.82
400	0.1		1	3.8
500	0.1		1.1	3.82
600	0.1		3.3	2.86

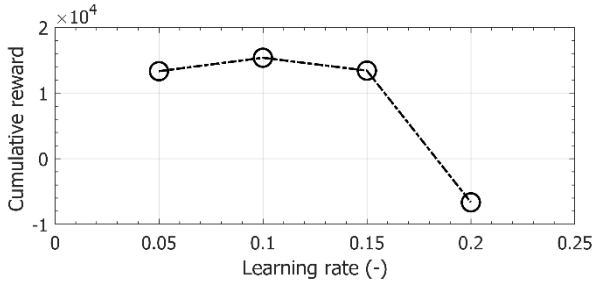
مقادیر جدول (۲) نشان می‌دهد که افزایش مقدار اپیزود مقدار فرارفت را افزایش داده اما زمان نشست را کاهش می‌دهد و با کاهش مقدار اپیزود این رویکرد برعکس می‌شود. بهترین حالت برای انتخاب اپیزود ۵۰۰ انتخاب شده است به گونه‌ای که پارامترهای مورد نظر در محدوده مناسبی قراردارند. برای تحلیل نرخ یادگیری نتایج برای مقادیر بالا سبب ناپایداری می‌شود در صورتی مقدار کم این پارامتر نرخ همگرایی را کاهش می‌دهد. پاداش‌های انباشته شده در شکل (۳) ارائه شده است که مقدار ۰/۱، مقدار بهینه می‌باشد.

[9] agent

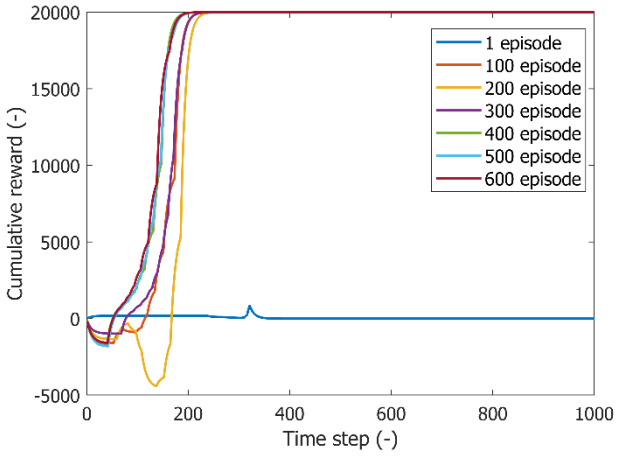
[Downloaded from marine-eng.ir on 2026-11-13]

[DOI: 10.61882/marineeng.21.46.6]

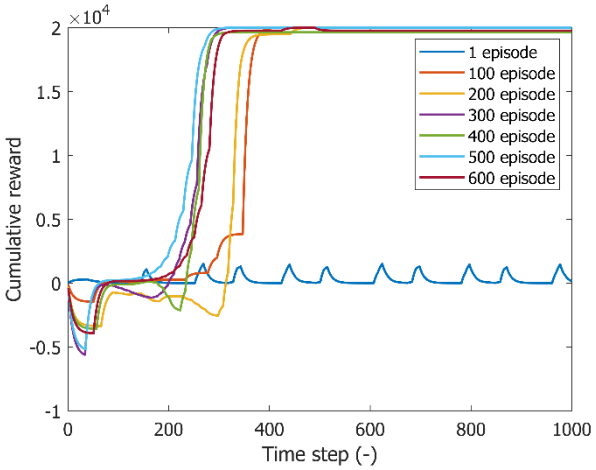
شکل‌های (۴) و (۵) نیز به صورت نموداری مقدار پاداش انباشته شده براساس تعداد اپیزود را نشان می‌دهد. شکل نشان می‌دهد با افزایش مقدار اپیزود، پاداش انباشته شده زیادتر می‌شود. با افزایش مقدار اپیزود، الگوریتم زمان بیشتری را برای آموزش جدول Q سپری می‌کند و بهبود عملکرد امری طبیعی است اما با افزایش مقدار اپیزود میزان محاسبات افزایش یافته و بهبود عملکرد تا مقدار مشخصی قابل دسترسی است. برای همین منظور مقدار ۵۰۰ اپیزود برای آموزش انتخاب شد.



شکل ۳- پاداش‌های انباشته شده براساس نرخ یادگیری



شکل ۴- پاداش‌های انباشته شده براساس تعداد اپیزود تعقیب عمق ۱ متر

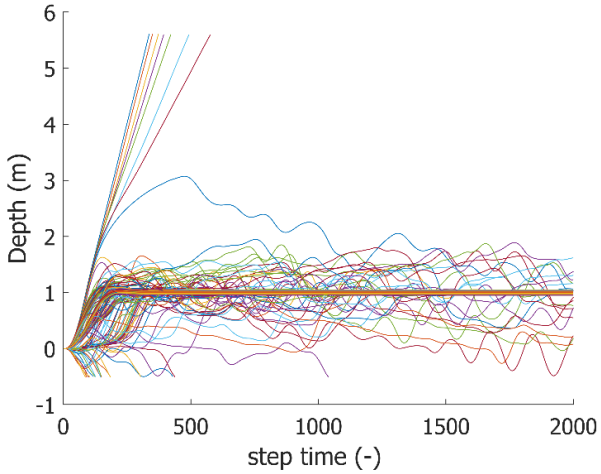


شکل ۵- پاداش‌های انباشته شده براساس تعداد اپیزود تعقیب عمق ۲ متر

۵ – نتایج و بحث

در بخش‌های قبلی به مرور مدل دینامیکی، روش و الگوریتم یادگیری تقویتی پرداخته شد. پس از پیاده سازی الگوریتم‌ها در بستر متلب ۲۰۲۳، شبیه سازی‌های مورد نظر صورت گرفته است. در شبیه سازی‌های این پژوهش نرخ یادگیری ۰/۱ و نرخ تخفیف ۰/۵ در نظر گرفته شده است. به ازای هر نقطه هدف ۵۰۰ مرتبه در سیکل داخلی با استفاده از مدل دینامیکی شبیه سازی‌ها انجام شد. پس از شبیه سازی وضعیت در ماتریس Q مقدار دهی شده است تا بتواند بهترین عمل را در شرایط مشابه داشته باشد. ماتریس Q یک ماتریس دو بعدی بوده است و به شکلی تعریف شده است که سطر آن سیگنال کنترل و ستون آن مقدار عمق لحظه‌ای ربات بوده است. در هر سناریو ربات با استفاده از الگوریتم تقویتی سعی دارد تا نقطه مطلوب را تعقیب کند. به ازای هر عمل یک پاداش برای الگوریتم در نظر گرفته شده است و در ماتریس Q جایگذاری گردیده است. در نهایت الگوریتم آموزش دیده براساس انباشتگی پاداش‌ها در فاز آموزش، برای بخش تصمیم گیری و تولید سیگنال کنترلی استفاده می‌شود. ربات براساس عمق لحظه‌ای مقدار سیگنال کنترلی را از ماتریس Q انتخاب می‌کند و به این شیوه سیگنال کنترلی تعیین گردیده است. در ادامه نتایج شبیه سازی‌ها ارائه شده است.

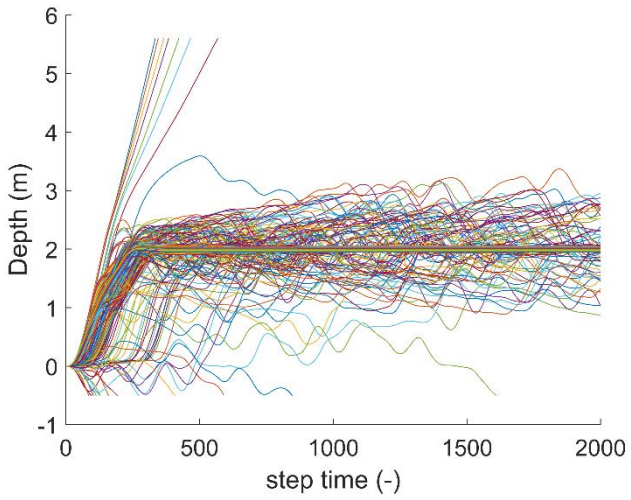
در شکل‌های (۶-۸) نتیجه شبیه سازی ۵۰۰ سیکل داخلی برای آموزش ماتریس Q برای عمق‌های ۱ تا ۳ متر ارائه شده است. همانطور که در قسمت‌های قبلی نیز ارائه شد، برای آموزش الگوریتم ۵۰۰ مرتبه شبیه سازی تکرار شده است. در این شبیه سازی‌ها ۲۰ درصد تصمیمات به صورت تصادفی و ۸۰ درصد به صورت مقدار بهینه انتخاب شده اند تا الگوریتم بتواند شرایط جدیدتری را تجربه کند و آموزش بیشتری را داشته باشد.



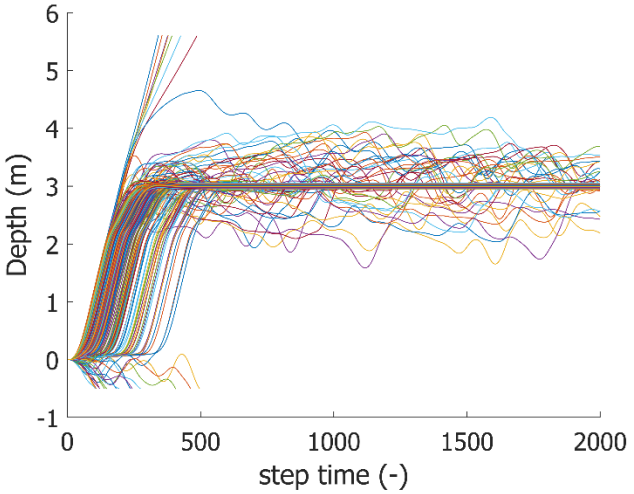
شکل ۶- نتیجه آموزش ۵۰۰ سیکل تعقیب عمق ۱ متر

[Downloaded from marine-eng.ir on 2026-11-13]

[DOI: 10.61882/marineeng.21.46.6]



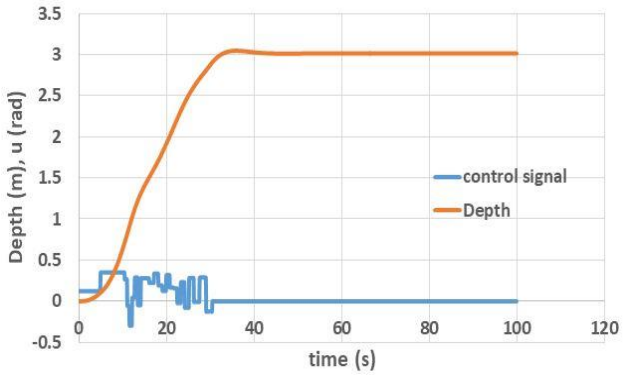
شکل ۷- نتیجه آموزش ۵۰۰ سیکل تعقیب عمق ۲ متر



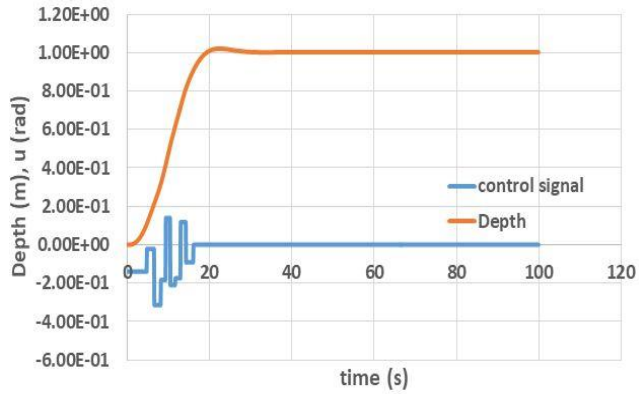
شکل ۸- نتیجه آموزش ۵۰۰ سیکل تعقیب عمق ۳ متر

نمودارهای ارائه‌شده در شکل‌های (۶–۸) نشان‌دهنده‌ی وجود برخی نوسانات و پراکندگی‌ها در روند یادگیری ماتریس Q هستند که عمدتاً ناشی از رویکرد تصادفی‌سازی ۲۰ درصدی اقدامات در فرآیند آموزش است. این استراتژی که بر اساس اصل اکتشاف-بهره‌برداری طراحی شده، به‌صورت عمدی امکان انتخاب اقدامات غیربهینه را در طول یادگیری فراهم می‌کند تا ربات بتواند فضای حالت‌ها و اقدامات را به‌طور کامل کشف کرده و از گیرکردن در بهینه‌های محلی جلوگیری شود. در نتیجه، در برخی سیکل‌های شبیه‌سازی (به‌ویژه در مراحل اولیه آموزش)، اقدامات تصادفی منجر به اعمال سیگنال‌های کنترلی نادرست می‌شوند که این امر به‌صورت خطاهای مقطعی در عمق مطلوب (۱ تا ۳ متر) نمود پیدا می‌کند. با این حال، همانطور که روند نمودارها نشان می‌دهد، با افزایش تعداد تکرارها، از میزان این پراکندگی‌ها کاسته شده و همگرایی به سمت مقادیر بهینه افزایش می‌یابد. این رفتار نشان می‌دهد که الگوریتم به‌تدریج از فاز اکتشاف به فاز بهره‌برداری انتقال یافته و با به‌روزرسانی ماتریس Q، دقت کنترلر بهبود می‌یابد. بنابراین، پراکندگی‌های مشاهده‌شده نه تنها ضعف سیستم نیستند، بلکه بخشی ضروری از فرآیند یادگیری برای دستیابی به یک سیاست کنترلی مقاوم و تطبیقی محسوب می‌شوند.

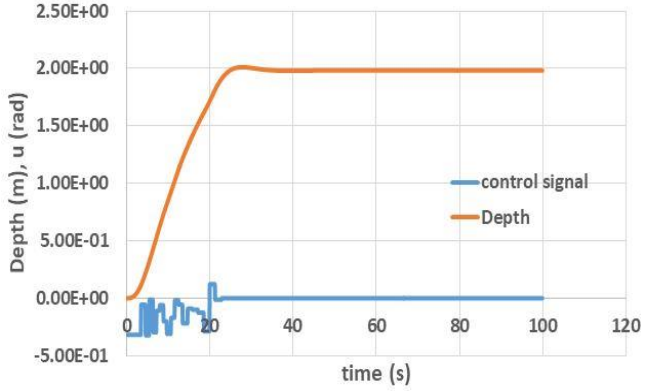
پس از آموزش الگوریتم در ادامه نتیجه کنترل عمق برای عمق های ۱ تا ۳ متر در شکل های (۹-۱۲) ارائه شده است.



شکل ۹- سیگنال کنترل و عمق لحظه ای ربات با الگوی تقویت شده برای تعقیب عمق ۳ متر



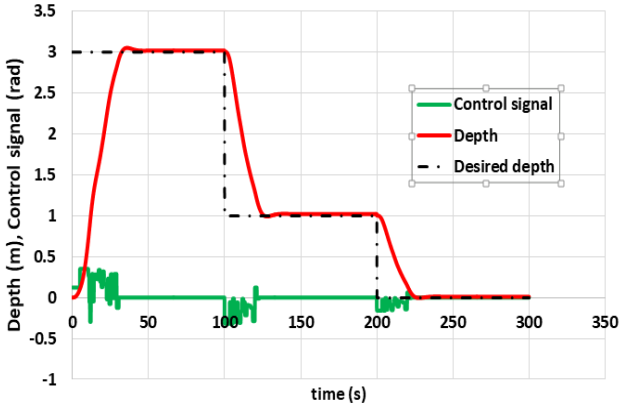
شکل ۱۰- سیگنال کنترل و عمق لحظه ای ربات با الگوی تقویت شده برای تعقیب عمق ۱ متر



شکل ۱۱- سیگنال کنترل و عمق لحظه ای ربات با الگوی تقویت شده برای تعقیب عمق ۲ متر

[Downloaded from marine-eng.ir on 2026-11-13]

[DOI: 10.61882/marineeng.21.46.6]



شکل ۱۲- کنترل عمق ربات با استفاده از کنترل یادگیری تقویتی

نتایج شبیه‌سازی‌ها نشان می‌دهد که سیستم کنترل مبتنی بر یادگیری تقویتی با دقت بالا توانسته است حرکت عمقی AUV را در عمق‌های مختلف (۱ تا ۳ متر) به‌خوبی مدیریت کند. با توجه به نمودارهای ارائه‌شده، خطای عمق پس از تکمیل فرآیند یادگیری به مقدار ناچیزی کاهش یافته و ربات قادر به حفظ پایداری مطلوب حتی در شرایط دینامیکی متغیر است. یکی از دستاوردهای کلیدی این پژوهش، بهینه‌سازی مصرف انرژی و کاهش استفاده غیرضروری از بالک است. الگوریتم طراحی‌شده با یادگیری سیاست‌های بهینه، تنها در مواقع لازم اقدام به تنظیم زاویه بالک‌ها می‌کند و از حرکات تند و پرش‌های ناگهانی که منجر به اتلاف انرژی می‌شود، اجتناب می‌نماید. این رویکرد نه‌تنها کارایی سیستم را افزایش داده، بلکه باعث کاهش مصرف انرژی و افزایش دامنه عملیاتی ربات شده است. نتایج نشان می‌دهد که ترکیب یادگیری تقویتی با مدل دینامیکی خطی‌شده حرکت پیچ، یک راه‌حل مؤثر برای کنترل دقیق و کم‌مصرف ربات‌های زیرسطحی محسوب می‌شود.

۶ –نتیجه گیری

این پژوهش به توسعه یک چارچوب نوین مبتنی بر یادگیری تقویتی برای کنترل بهینه حرکت عمقی ربات‌های AUV پرداخته است. با استفاده از الگوریتم Q-Learning و طراحی یک سیستم پاداش هوشمند که معیارهای مختلفی از جمله دقت کنترل عمق، مصرف انرژی و پایداری دینامیکی را در نظر می‌گیرد، یک کنترلر تطبیقی و کارآمد طراحی شده است. نتایج شبیه‌سازی‌ها نشان می‌دهد که این روش نه‌تنها قادر به کاهش خطای عمق به میزان قابل‌توجهی است، بلکه در مقایسه با روش‌های کنترل کلاسیک مانند PID، عملکرد پایدارتری در شرایط پیچیده و شرایط غیرخطی از خود نشان می‌دهد. به‌روزرسانی ماتریس Q طی ۵۰۰ تکرار برای هر سناریو، امکان یادگیری عمیق و همگرایی به سیاست بهینه را فراهم کرده است. این رویکرد به‌ویژه در کاربردهای عملیاتی که نیازمند پاسخگویی سریع و دقیق به تغییرات محیطی هستند، می‌تواند عملکرد مناسبی ارائه کند.

این کنترلر هوشمند با توانایی یادگیری از تعامل مستقیم با محیط، می‌تواند در شرایط واقعی اقیانوس که دارای عدم قطعیت‌های فراوانی است، به‌خوبی عمل کند. علاوه بر این، استفاده از بالک های کنترلی در ترکیب با این الگوریتم، امکان دستیابی به حرکات پایدار را حتی در حضور جریان‌های آبی قوی فراهم می‌سازد. در آینده، می‌توان این چارچوب را به حوزه‌های دیگری مانند کنترل چندبعدی AUV (ترکیب عمق و جهت) یا هدایت گروهی چند ربات تعمیم داد. همچنین، ادغام این روش با دیگر تکنیک‌های یادگیری عمیق مانند DDPG یا PPO می‌تواند منجر به توسعه سیستم‌های کنترلی حتی قدرتمندتر و انعطاف‌پذیرتر شود. این پژوهش گامی در جهت بهبود عملکرد کنترلر ربات‌های زیرسطحی و افزایش قابلیت‌های آن‌ها در مأموریت‌های اکتشافی و تحقیقاتی محسوب می‌شود.

۷ –مراجع

1- Zhang, Y., et al. "Deep Reinforcement Learning for Autonomous Underwater Vehicle Path Planning and Control." IEEE Transactions on Robotics, 2022.

2- Li, H., & Wang, J. "PPO-based Control of AUVs in Dynamic Underwater Environments." Ocean Engineering, 2021.

3- Chen, X., et al. "Transfer Learning in Reinforcement Learning for AUV Motion Control." Journal of Marine Science and Technology, 2023.

4- Liu, R., et al. "Multi-Agent Reinforcement Learning for Cooperative AUV Navigation." Autonomous Robots, 2023.

5- Wang, L., & Zhang, K. "Energy-Efficient Multi-AUV Control Using MADDPG." IEEE Journal of Oceanic Engineering, 2022.

6- Patel, S., et al. "Spiking Neural Networks for Real-Time AUV Control." Neural Networks, 2023.

7- Kim, H., & Park, S. "LSTM-Based Reinforcement Learning for Dynamic AUV Motion Planning." Ocean Engineering, 2023.

8- Zhao, Y., et al. "Hybrid RL-SNN Architectures for Autonomous Underwater Vehicles." Journal of Intelligent & Robotic Systems, 2024.

[Downloaded from marine-eng.ir on 2026-11-13]

[DOI: 10.61882/marineeng.21.46.6]

9- Hasanvand, A. and Seif, M.S., 2024. Investigation the effect of length-to-diameter ratio on six-DOF helical and zigzag maneuvers of the SUT glider with internal actuators. Ocean Engineering, 295, p.116819.

10- Hasanvand, A. and Seif, M.S., 2024. Development of state space models for underwater gliders equipped with embedded actuators in vertical and horizontal planes. Ocean Engineering, 309, p.118344.

11- Hasanvand, A. and Seif, M.S., 2024. Adaptive path-following control for high-underactuated underwater glider under hydrodynamic coefficient uncertainties. Journal of Marine Science and Technology, 29(4), pp.1000-1017.

12- Wang, D., Wan, J., Shen, Y., Qin, P. and He, B., 2022. Hyperparameter Optimization for the LSTM Method of AUV Model Identification Based on Q-Learning. Journal of Marine Science and Engineering, 10(8), p.1002.

13- Jun, B.H., Park, J.Y., Lee, F.Y., Lee, P.M., Lee, C.M., Kim, K., Lim, Y.K. and Oh, J.H., 2009. Development of the AUV 'ISiMI'and a free running test in an Ocean Engineering Basin. Ocean engineering, 36(1), pp.2-14.

[Downloaded from marine-eng.ir on 2026-11-13]

[DOI: 10.61882/marineeng.21.46.6]