Robust Attitude Control of a Three-Degree-of-Freedom Satellite via Integration of the Super-Twisting Algorithm and Deep Reinforcement Learning with Hyperparameter Tuning Using Taguchi Design of Experiments

Mostafa Sarjoughian, Hojat Taei*

Department of Aerospace Engineering, Faculty of Engineering, University of Isfahan, Isfahan, Iran

* h.taei@eng.ui.ac.ir

ABSTRACT

This study presents a hybrid control framework for the attitude regulation of a three-degree-of-freedom satellite subject to parametric uncertainties, external disturbances, actuator constraints, and implementation imperfections. The core robust controller is formulated using the Super-Twisting Algorithm, which guarantees finite-time convergence and robustness while effectively suppressing the high-frequency chattering typically associated with conventional sliding mode control. To enhance tracking precision and improve adaptability under nonlinear and uncertain conditions, deep reinforcement learning is incorporated as an adaptive compensator within the control loop. Three representative algorithms, namely Deep Deterministic Policy Gradient, Twin Delayed Deep Deterministic Policy Gradient, and Proximal Policy Optimization, are investigated and comparatively evaluated in terms of stability, convergence behavior, and control efficiency. To systematically tune the learning hyperparameters and reduce the computational burden associated with manual trial-and-error procedures, the Taguchi design of experiments method is employed to perform multi-objective optimization considering both tracking performance and control effort. The performance index is defined as a composite measure that combines time-weighted tracking error and control energy. Numerical simulations together with experimental validation on a satellite attitude simulator demonstrate that the proposed hybrid control architecture reduces settling time and control effort while improving disturbance rejection capability, without compromising stability or steady-state tracking accuracy.

KEYWORDS

Satellite Attitude Control; Super-Twisting Algorithm; Deep Reinforcement Learning; Taguchi Design of Experiments.

1. Introduction

Space mission complexity has driven increasingly stringent demands on satellite attitude control in terms of robustness, precision, and reliability. Sliding Mode Control (SMC) offers strong robustness to matched uncertainties but suffers from chattering; the Super-Twisting Algorithm (STA) effectively mitigates this drawback while preserving robustness [1-3].

Deep Reinforcement Learning (DRL) has emerged as a compelling model-free paradigm for nonlinear uncertain systems, including spacecraft attitude control, offering adaptability without requiring precise dynamic models [4-6]. However, hyperparameter sensitivity and limited robustness guarantees hinder standalone DRL deployment in safety-critical applications, motivating hybrid robust–learning architectures [7]. Efficient hyperparameter tuning is essential for DRL performance; the Taguchi method provides a systematic, low-experiment alternative to exhaustive search [8, 9].

This paper presents a hybrid STA–DRL framework for three-DOF rigid satellite attitude control, in which DDPG, TD3, and PPO agents augment the STA baseline with Taguchi-tuned hyperparameters. Performance is evaluated under nominal conditions, disturbances, and parametric uncertainties using an ITSE-based composite index, validated through comparative simulations.

2. System Modeling

In this study, the attitude dynamics of a rigid satellite are formulated in an affine nonlinear form suitable for robust and learning-based control design. The simulation framework employed is schematically illustrated in Fig. 1.

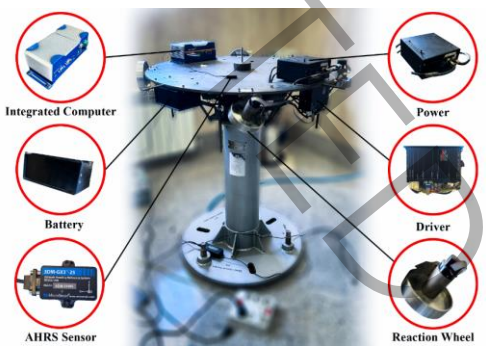


Figure 1. General structure of the Isfahan University satellite attitude simulator [10]

Let $\eta \in \mathbb{R}^3$ denote the Euler angle vector and $\omega \in \mathbb{R}^3$ represent the angular velocity vector expressed in the body-fixed frame. The satellite attitude dynamics can be expressed as the following affine nonlinear system:

$$\begin{cases} \dot{\eta} = f_1(\eta, \omega) \\ \dot{\omega} = f_2(\eta, \omega) + g(\eta)\tau \end{cases} \quad (1)$$

where $\tau \in \mathbb{R}^3$ denotes the control torque vector.

The nonlinear functions $f_1(\cdot)$, $f_2(\cdot)$, and the control distribution matrix $g(\cdot)$ are defined as:

$$f_1(\eta, \omega) = R\omega$$
$$f_2(\eta, \omega) = I^{-1}(-\omega \times (I\omega) + mg(r_s \times K)) \quad (2)$$
$$g(\eta) = I^{-1}$$

This affine representation provides a compact and suitable foundation for the subsequent development of the super-twisting-based robust controller and the deep reinforcement learning framework.

3. Control Structures

Two control structures are investigated to evaluate the contribution of the learning-based component: STA and hybrid STA–DRL. The learning-based controller operates in parallel with the super-twisting algorithm, generating an adaptive torque component that augments the baseline control law,

$$\tau = \tau_{STA} + \tau_{DRL} \quad (3)$$

This structure preserves the inherent robustness of the sliding-mode controller while allowing performance refinement through reinforcement learning.

3. Results

Table 1 and Fig. 3 summarize the experimental evaluation of the four control strategies implemented on the satellite attitude simulator.

Table 1. Performance comparison of each control scenario and maximum tolerable disturbance level

Method	MSE	ISE	ITSE	CE[1]	T_s	TCD[2]	BIT[3]	WNTP[4]
STA	1.719×10^{-8}	2.55	629.9	24.2	26	0.17	3.15	1×10^{-3}
STA-DDPG	3.719×10^{-8}	2.937	971	23.58	23	0.25	3.25	3×10^{-3}
STA-TD3	4.567×10^{-8}	2.757	875	23.1	22	0.255	3	3×10^{-3}
STA-PPO	6.005×10^{-8}	2.824	677.2	26.23	24	0.135	2.8	2×10^{-3}

As shown in Table 1, all controllers achieve comparable tracking accuracy (MSE ~10^{-8}), with STA yielding the lowest error indices. STA–TD3 improves settling time (26s→22s) and reduces control effort, while also providing the highest tolerance to torque command disturbances (0.255 N·m) and white-noise torque (3×10^{-3}). STA–DDPG shows the greatest resilience to direct impulse torques (3.25 N·m). Figure 3 confirms stable closed-loop behavior across all strategies, with hybrid controllers exhibiting smoother control inputs and reduced steady-state chattering, especially STA–TD3.

[1] Control Effort
[2] Torque Command Disturbance
[3] Body Impulse Torque
[4] White-Noise Torque Power

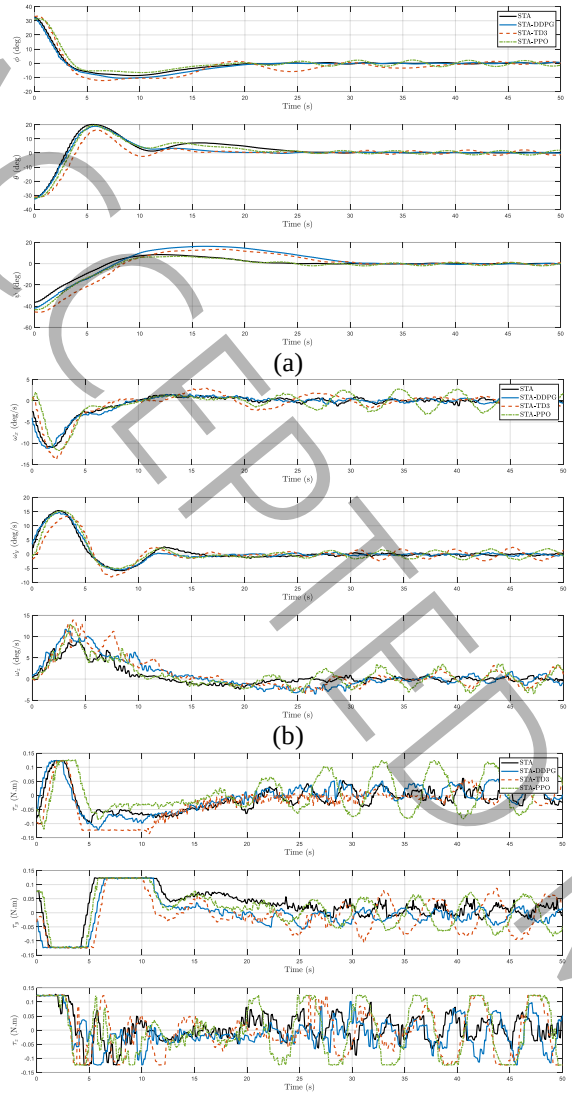


Figure 3. Experimental implementation results of four control strategies in the presence of inherent test-platform disturbances: (a) attitude angles, (b) angular velocities, and (c) control inputs

4. Conclusion

This study presented and experimentally validated a hybrid attitude control framework combining STA with DRL (DDPG, TD3, and PPO), in which STA provides baseline robustness while the DRL agent acts as an adaptive corrective term. All controllers achieve stable and accurate attitude regulation, with differences primarily in transient performance, control effort, and disturbance robustness. STA–TD3 offers the best overall trade-off, followed by STA–DDPG; STA alone excels in error minimization, while PPO is less competitive in practical metrics. Experimental results confirm stable behavior under sensor noise, communication delays, and inherent platform disturbances, with performance distinctions reflected in control chattering, effort, and reaction wheel momentum regulation. These findings highlight the importance of incorporating actuator constraints, command smoothing, and torque-variation penalties in both control and reward design for practical deployment.

5. References

[1] K. Lu, Y. Xia, Finite-time attitude control for rigid spacecraft-based on adaptive super-twisting algorithm, IET Control Theory & Applications, 8(15) (2014) 1465–1477.
[2] Y. Su, S. Shen, Adaptive predefined-time fault-tolerant attitude tracking control for rigid spacecraft with guaranteed performance, Acta Astronautica, 214 (2024) 677–688.
[3] C. Xiao, Y. Guo, C.-q. Xie, A.-j. Li, C.-q. Wang, Adaptive super-twisting sliding mode attitude coordination control for spacecraft formation flying with actuator saturation, Advances in Space Research, 72(10) (2023) 4244–4255.
[4] M. Tipaldi, R. Iervolino, P.R. Massenio, Reinforcement learning in spacecraft control applications: Advances, prospects, and challenges, Annual Reviews in Control, 54 (2022) 1–23.
[5] W. Retagne, J. Dauer, G. Waxenegger-Wilfing, Adaptive satellite attitude control for varying masses using deep reinforcement learning, Frontiers in Robotics and AI, 11 (2024) 1402846.
[6] S. Oghim, J. Park, H. Bang, H. Leeghim, Deep reinforcement learning-based attitude control for spacecraft using control moment gyros, Advances in Space Research, 75(1) (2025) 1129–1144.
[7] M. Ran, J. Li, L. Xie, Reinforcement-learning-based disturbance rejection control for uncertain nonlinear systems, IEEE Transactions on Cybernetics, 52(9) (2021) 9621–9633.
[8] X. Zhang, X. Chen, L. Yao, C. Ge, M. Dong, Deep neural network hyperparameter optimization with orthogonal array tuning, in: International conference on neural information processing, Springer, 2019, pp. 287–295.
[9] P. Arevalo, A. Cano, O. Fedoseienko, F. Jurado, A data-driven approach to microgrid fault detection and classification using Taguchi-optimized CNNs and wavelet transform, Applied Soft Computing, 170 (2025) 112667.
[10] M. Sarjoughian, M. Malekzadeh, N. Sayyaf, Hybrid Control of Spacecraft: Super-Twisting Algorithm Based on Taguchi-Driven Deep Reinforcement Learning, Results in Engineering, (2026) 110530.

طراحی کنترل مقاوم وضعیت ماهواره سه‌درجه‌آزادی با تلفیق الگوریتم فوق‌پیچشی و یادگیری تقویتی عمیق و تنظیم فراپارامترها با طرح‌ریزی آزمایش تاگوچی

مصطفی سرجوقیان، حجت طائی*

گروه مهندسی هوافضا، دانشکده فنی و مهندسی، دانشگاه اصفهان، اصفهان، ایران

* h.taei@eng.ui.ac.ir

چکیده

در این پژوهش، یک چارچوب کنترلی ترکیبی برای کنترل وضعیت ماهواره سه‌درجه‌آزادی در حضور نامعینی‌های پارامتری، اغتشاش‌های خارجی، محدودیت‌های عملگر و خطاهای اجرایی ارائه می‌شود. هسته‌ی مقاوم سامانه بر پایه الگوریتم فوق‌پیچشی طراحی شده است تا ضمن حفظ پایداری و مقاوم بودن، پدیده لرزش کاهش یابد. به‌منظور بهبود دقت رهگیری و افزایش قابلیت سازگاری در مواجهه با غیرخطی‌بودن و عدم‌قطعیت‌های پیش‌بینی‌ناپذیر، از یادگیری تقویتی عمیق به‌عنوان یک مؤلفه اصلاحی تطبیقی در کنار کنترل‌کننده مقاوم استفاده شده است. در این راستا، سه الگوریتم شاخص شامل گرادیان سیاست تعیینی عمیق، گرادیان سیاست تعیینی عمیق تأخیردار دوقلو و بهینه‌سازی سیاست مجاورتی مورد مطالعه و مقایسه قرار گرفته‌اند. همچنین، برای کاهش هزینه محاسباتی تنظیم فراپارامترها و ایجاد یک رویه نظام‌مند و تکرارپذیر، از روش طرح‌ریزی آزمایش تاگوچی جهت بهینه‌سازی چندهدفه فراپارامترهای یادگیری تقویتی عمیق بهره گرفته شده است. معیار ارزیابی عملکرد شامل ترکیبی از دقت رهگیری زمان‌وزن‌شده و میزان تلاش کنترلی در نظر گرفته شده است. نتایج شبیه‌سازی عددی و پیاده‌سازی آزمایشگاهی روی شبیه‌ساز وضعیت ماهواره نشان می‌دهد ساختار ترکیبی پیشنهادی، در مقایسه با کنترل مقاوم پایه، می‌تواند زمان نشست و تلاش کنترلی را کاهش داده و استحکام در برابر اغتشاشات را بهبود دهد، در حالی که پایداری و دقت رهگیری حفظ می‌شود.

کلمات کلیدی

کنترل وضعیت ماهواره، کنترل مقاوم، الگوریتم فوق‌پیچشی، یادگیری تقویتی عمیق، طرح‌ریزی آزمایش تاگوچی.

۱- مقدمه

گسترش مأموریت‌های مداری و افزایش پیچیدگی سناریوهای عملیاتی، ضرورت دستیابی به کنترل وضعیت دقیق و پایدار را بیش از پیش برجسته کرده است. در این میان، پدیده‌هایی نظیر نامعینی دینامیکی، اغتشاش‌های محیطی، محدودیت‌های گشتاور و بروز خطا و کاهش کارایی عملگرها، می‌توانند کیفیت رهگیری و پایداری نهایی را به‌طور معناداری تضعیف کنند. ازاین‌رو، توسعه راهبردهایی که هم دقت رهگیری را تضمین کند و هم در برابر شرایط نامطلوب عملیاتی مقاوم باشد، همچنان موضوع پژوهش‌های گسترده در ادبیات کنترل سامانه‌های فضایی است.

در سال‌های اخیر، روش‌های مقاوم مبتنی بر کنترل مود لغزشی[۱] به دلیل برخورداری از ویژگی‌های مقاومتی مناسب در برابر اغتشاش‌ها و نامعینی‌ها، مورد توجه ویژه قرار گرفته‌اند. با این حال، چالش اصلی کنترل مود لغزشی کلاسیک، ظهور پدیده لرزش[۲] فرمان و تبعات آن بر عملگرها و کیفیت عملکرد است. در پاسخ به این مسئله، استفاده از کنترل لغزشی مرتبه‌بالا و به‌ویژه الگوریتم فوق‌پیچشی[۳] به‌عنوان یک راهبرد مؤثر برای کاهش لرزش[۴] و حفظ ویژگی‌های مقاومتی مطرح شده است [۱]. در حوزه کنترل وضعیت و رهگیری ماهواره/فضاپیما، سو[۵] و همکاران [۲] کنترلگرهای مقاومِ تحمل‌خطا[۶] را با بهره‌گیری از سازوکارهای مبتنی بر رهیافت فوق-پیچشی و مشاهده‌گرها توسعه داده‌اند. همچنین شیائو[۷] و همکاران [۳] در زمینه هماهنگ‌سازی وضعیت در پرواز آرایشی[۸] با تکیه بر سازوکارهای فوق‌پیچشی گزارش‌هایی ارائه کرده‌اند. در امتداد این خط پژوهش، خداوردیان و ملک‌زاده [۴] نیز نشان داده‌اند که ترکیب ایده‌های لغزشی با ساختارهای پیش‌بینانه می‌تواند برای حذف هم‌زمان خطاهای وضعیت و ارتعاشات سازه‌ای در فضاپیماهای انعطاف‌پذیر مفید واقع شود. علاوه بر این، شی[۹] و همکاران [۵] در قالب کنترل‌های پیشرفته برای رهگیری وضعیت، به نقش ساختارهای مقاوم و لحاظ قیود پرداخته‌اند.

با وجود پیشرفت‌های یادشده، تکیه صرف بر طراحی‌های مدل‌محور در سامانه‌هایی با دینامیک‌های غیرخطی، تغییرپذیر و دارای عدم‌قطعیت‌های پیش‌بینی‌ناپذیر، می‌تواند محدودیت ایجاد کند. در چنین شرایطی، رویکردهای داده‌محور و به‌ویژه یادگیری تقویتی عمیق[۱۰] به دلیل قابلیت تولید سیاست کنترلی بدون نیاز به مدل دقیق و امکان سازگاری پس از آموزش، به‌صورت جدی وارد ادبیات کنترل سامانه‌های فضایی شده‌اند. تیپالدی[۱۱] و همکاران [۶] یک مرور جامع از دستاوردها، چشم‌اندازها و چالش‌های یادگیری تقویتی عمیق در کاربردهای کنترل فضاپیما ارائه کرده‌اند. در سطح پیاده‌سازی‌های کاربردی، رتاین[۱۲] و همکاران [۷] نشان داده‌اند که یادگیری تقویتی عمیق می‌تواند برای کنترل وضعیت ماهواره در حضور تغییرات شدید جرم/اینرسی مفید باشد و حتی با تکیه بر مشاهده‌های پشته‌ای[۱۳] به سازگاری بهتر دست یابد. از منظر محدودیت‌های پیاده‌سازی و محاسبات بلادرنگ، لی[۱۴] و همکاران [۸] یک چارچوب یادگیری تقویتی عمیق برای کنترل وضعیت نانوماهواره با عملگرهای مغناطیسی[۱۵] را با تاکید بر اجرا در سخت‌افزار و قیود بلادرنگ گزارش کرده‌اند. همچنین در زمینه سامانه‌های چابک با ژیروسکوپ‌های ممان‌کنترل، اوغیم[۱۶] و همکاران [۹] طراحی و پیاده‌سازی

---

[1] Sliding Mode Control (SMC)
[2] Chattering
[3] Super-Twisting Algorithm (STA)
[4] Chattering
[5] Su
[6] Fault-Tolerant
[7] Xiao
[8] Formation Flying
[9] Shi
[10] Deep Reinforcement Learning (DRL)
[11] Tipaldi
[12] Retagne
[13] Stacked Observations
[14] Lee
[15] Magnetic Attitude Control
[16] Oghim

کنترلگرهای یادگیری تقویتی عمیق را برای رهگیری/مانور وضعیت مطرح کرده‌اند. افزون بر این، وو[1] و همکاران [۱۰] استفاده از یادگیری تقویتی عمیق از نوع سیاست تعیینی[2] را در مسائل کنترل ماهواره‌ای گزارش نموده‌اند و مسلی[3] و همکاران [۱۱] نیز به کارگیری سیاست‌های مبتنی بر گرادیان سیاست قطعی عمیق دوقلو با به‌روزرسانی تأخیری[4] را در سناریوهای تثبیت وضعیت مطالعه کرده‌اند.

با وجود قابلیت‌های یادگیری تقویتی عمیق، چالش‌های کلیدی همچنان پابرجاست: (۱) حساسیت عملکرد به انتخاب فراپارامتر[5] شبکه‌ها و الگوریتم یادگیری، (۲) هزینه محاسباتی زیاد آموزش در محیط‌های شبیه‌سازی دقیق، و (۳) دشواری تضمین کیفیت/پایداری در حضور قیود عملیاتی و اغتشاش‌های سخت. به‌طور نمونه، در مطالعاتی نظیر لی و همکاران [۸] و رتاین و همکاران [۷]، مسئله بار محاسباتی و قیود پیاده‌سازی، به‌عنوان عامل تعیین‌کننده در انتخاب ساختار الگوریتم و شبکه مطرح شده است. از سوی دیگر، پژوهش‌هایی در سامانه‌های مشابهِ کنترلی/پایدارسازی نیز نشان می‌دهند که افزودن ناظرها و سازوکارهای جبران اغتشاش می تواند به بهبود کیفیت سیاست‌های یادگرفته‌شده یا کاهش اثر اغتشاش کمک کند؛ برای نمونه ران[6] و همکاران [۱۲] یک چارچوب مبتنی بر یادگیری تقویتی عمیق همراه با ناظر حالت توسعه‌یافته[7] را برای حذف اغتشاش در یک سامانه پایدارساز گزارش کرده‌اند. بنابراین، ترکیب ساختارهای مقاوم (مانند الگوریتم فوق‌پیچشی) با یادگیری تقویتی عمیق، از منظر مهندسی کنترل، می‌تواند به‌عنوان یک مسیر طبیعی برای هم‌زمان‌سازی «مقاومت ذاتی» و «توان یادگیری/سازگاری» تلقی شود.

نکته بنیادین دیگر، تنظیم نظام‌مند فراپارامترهای یادگیری تقویتی عمیق است. در عمل، جست‌وجوهای متداول مانند جست‌وجوی شبکه‌ای[8] یا تصادفی[9] در مسائل کنترلی با شبیه‌سازی‌های سنگین، هزینه‌بر و کم‌بازده می‌شوند. در این چارچوب، استفاده از طرح‌ریزی آزمایش[10] و به‌طور خاص روش تاگوچی[11] و آرایه‌های متعامد برای کاهش تعداد آزمایش‌ها و استخراج اثر عوامل به‌صورت سیستماتیک، در ادبیات یادگیری ماشین/یادگیری عمیق به‌کار گرفته شده است. ژانگ[12] و همکاران [۱۳] روش تنظیم متعامد[13] را برای تنظیم فراپارامترها توسعه داده‌اند. چندر[14] و همکاران [۱۴] یک چارچوب مبتنی بر تنظیم متعامد برای تنظیم فراپارامترهای مدل‌های یادگیری ماشین/عمیق ارائه کرده‌اند و رانکوویچ[15] و همکاران [۱۵] نسخه توسعه‌یافته‌ای را با رویکرد یکپارچه و کارامدتر معرفی نموده‌اند. همچنین در کاربردهای عملی، آروالو[16] و همکاران [۱۶] از سازوکار تاگوچی برای تنظیم مؤثر فراپارامترهای شبکه‌های کانولوشنی[17] استفاده کرده‌اند و گربچیچ[18] و همکاران [۱۷] نیز در مسئله جست‌وجوی پارامتر/طراحی به مزیت‌های طراحی آزمایش و پیمایش کارامد فضا اشاره کرده‌اند. افزون بر این، لین[19] و همکاران [۱۸] نیز نمونه‌ای از استفاده از تاگوچی را برای تنظیم سازوکارهای یادگیری عمیق گزارش کرده‌اند. مجموعه این نتایج، دلالت می‌کند که بهره‌گیری از طراحی آزمایش مبتنی بر تاگوچی می‌تواند در سناریوهای یادگیری تقویتی عمیق با شبیه‌سازی‌های محاسباتی سنگین، به کاهش هزینه آزمایش‌ها و افزایش تکرارپذیری کمک کند.

---

[1] Wu
[2] Deterministic Policy
[3] Mosali
[4] Twin Delayed Deep Deterministic Policy Gradient (TD3)
[5] Hyperparameter
[6] Ran
[7] Extended State Observer (ESO)
[8] Grid Search
[9] Random Search
[10] Design of Experiments (DoE)
[11] Taguchi
[12] Zhang
[13] Orthogonal Array (OA)Tuning
[14] Chandar
[15] Rankovic
[16] Arevalo
[17] Convolutional Neural Network
[18] Grbcic
[19] Lin

با جمع‌بندی مطالعات پیشین می‌توان مشاهده کرد که هر یک از سه مسیر پژوهشی یادشده، بخشی از مسئله کنترل وضعیت مقاوم را پوشش می‌دهند، اما به‌تنهایی پاسخ کاملی برای چالش مورد نظر این پژوهش فراهم نمی‌کنند. روش‌های مقاوم مبتنی بر کنترل مود لغزشی و الگوریتم فوق‌پیچشی از نظر تضمین پایداری و مقابله با اغتشاشات کراندار مزیت دارند، اما در بهبود هم‌زمان پاسخ گذرا، کاهش تلاش کنترلی و سازگاری با تغییرات پیچیده دینامیکی محدودیت‌هایی نشان می‌دهند. در مقابل، روش‌های یادگیری تقویتی عمیق قابلیت یادگیری سیاست کنترلی از تعامل با محیط و سازگاری با شرایط غیرخطی را دارند، اما عملکرد آن‌ها به‌شدت به تنظیم فراپارامترها، کیفیت آموزش و قیود اجرایی وابسته است. از سوی دیگر، روش‌های طرح‌ریزی آزمایش مانند تاگوچی می‌توانند هزینه تنظیم فراپارامترها را کاهش دهند، اما به‌تنهایی یک راهبرد کنترلی محسوب نمی‌شوند. بنابراین، شکاف اصلی موجود در ادبیات، نبود یک چارچوب یکپارچه است که در آن هسته مقاوم الگوریتم فوق‌پیچشی حفظ شود، عامل یادگیری تقویتی عمیق به‌عنوان مؤلفه اصلاحی کراندار به آن افزوده شود و فراپارامترهای یادگیری تقویتی عمیق نیز با یک روند نظام‌مند و قابل تکرار تنظیم شوند. پژوهش حاضر برای پاسخ به همین شکاف طراحی شده است.

با توجه به پیشینه فوق، در این پژوهش یک چارچوب کنترل وضعیت برای ماهواره سه‌درجه‌آزادی ارائه می‌شود که در آن، از الگوریتم فوق‌پیچشی به‌عنوان کنترلگر مقاوم پایه و از سه الگوریتم یادگیری تقویتی عمیق شامل گرادیان سیاست قطعی عمیق[1]، گرادیان سیاست قطعی عمیق دوقلو با به‌روزرسانی تأخیری و بهینه‌سازی سیاست مجاورتی[2] برای تولید سیاست کنترلی بهره گرفته می شود؛ سپس، تنظیم فراپارامترهای یادگیری تقویتی عمیق با اتکا بر طراحی آزمایش مبتنی بر تاگوچی انجام می‌پذیرد تا اثر عوامل کلیدی و انتخاب سطح‌های مناسب به‌صورت نظام‌مند استخراج شود. همچنین، معیار ارزیابی عملکرد بر مبنای یک تابع هدف ترکیبی شامل انتگرال خطای مربعی وزن‌دهی‌شده با زمان[3] و تلاش کنترلی تعریف می‌گردد تا هم کیفیت گذرا و هم هزینه کنترلی به‌طور هم‌زمان لحاظ شود. بدین ترتیب، انتظار می‌رود چارچوب پیشنهادی بتواند مزیت‌های «کاهش لرزش و مقاومت در برابر اغتشاش/نامعینی» را در کنار «توان یادگیری و سازگاری سیاست‌های یادگیری تقویتی عمیق» فراهم آورد.

با وجود آنکه در پژوهش‌های اخیر ترکیب‌هایی از یادگیری تقویتی با کنترل مود لغزشی، کنترل مقاوم یا ناظرهای اغتشاش گزارش شده است، تمایز پژوهش حاضر در نحوه سازمان‌دهی مکمل الگوریتم فوق‌پیچشی و یادگیری تقویتی عمیق است. در ساختار پیشنهادی، عامل یادگیری تقویتی عمیق جایگزین کنترل‌کننده مقاوم نمی‌شود، بلکه الگوریتم فوق‌پیچشی به‌عنوان هسته مقاوم و پایدارکننده حفظ شده و یادگیری تقویتی عمیق به‌صورت یک ترم اصلاحی کراندار در کنار آن قرار می‌گیرد. بنابراین، نقش الگوریتم فوق‌پیچشی تأمین پایداری و مقاومت پایه در برابر اغتشاشات و عدم‌قطعیت‌های کراندار است، در حالی که یادگیری تقویتی عمیق برای بهبود پاسخ گذرا، کاهش تلاش کنترلی و افزایش مقاومت عملی کنترل‌کننده به کار گرفته می‌شود. این موضوع، چارچوب حاضر را از رویکردهایی که در آن‌ها یادگیری تقویتی نقش کنترل‌کننده اصلی، تنظیم‌کننده مستقیم بهره‌ها یا تخمین‌گر اغتشاش را ایفا می‌کند، متمایز می‌سازد.

نتایج عددی نیز نشان می‌دهد که افزودن یادگیری تقویتی عمیق لزوماً به بهبود همه شاخص‌ها در همه الگوریتم‌ها منجر نمی‌شود، اما در ترکیب‌های TD3–STA و DDPG–STA موجب کاهش زمان نشست، کاهش تلاش کنترلی و افزایش توان دمپ اغتشاش نسبت به الگوریتم فوق‌پیچشی تنها شده است. افزون بر این، در این پژوهش فراپارامترهای سه الگوریتم DDPG، TD3 و PPO با روش تاگوچی و آرایه متعامد L_{27} تنظیم شده‌اند؛ ازاین‌رو، نوآوری مقاله در طراحی یک ساختار مکمل STA–DRL، مقایسه سه عامل یادگیری تقویتی عمیق در چارچوبی یکسان، تنظیم نظام‌مند فراپارامترها و اعتبارسنجی عددی و آزمایشگاهی آن برای کنترل وضعیت ماهواره سه‌درجه‌آزادی است.

در ادامه، ابتدا مدل‌سازی سینماتیکی و دینامیکی ماهواره سه‌درجه‌آزادی و تعریف متغیرهای حالت، خطاها، و قیود عملگر ارائه می شود. سپس، طراحی کنترل‌کننده مقاوم مبتنی بر الگوریتم فوق‌پیچشی و صورت‌بندی چارچوب یادگیری تقویتی عمیق شامل تعریف

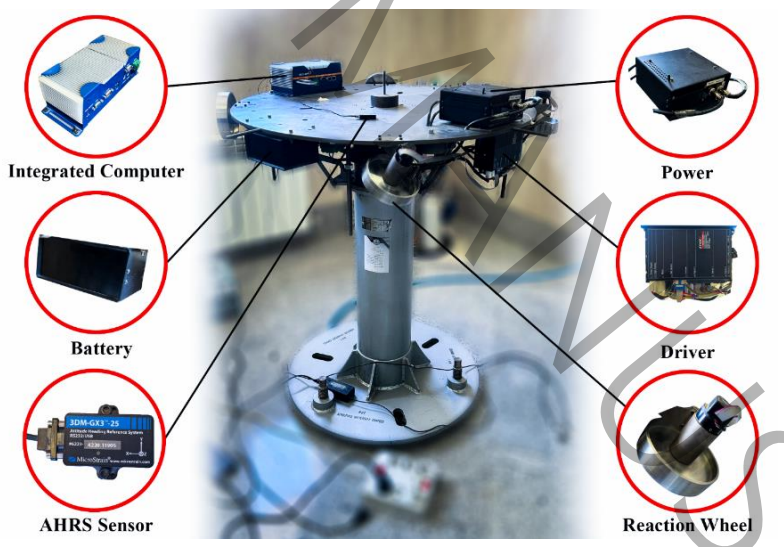
[1] Deep Deterministic Policy Gradient (DDPG)
[2] Proximal Policy Optimization (PPO)
[3] Integral of Time-weighted Squared Error (ITSE)

فضای مشاهده، فضای کنش و تابع پاداش بیان می‌گردد. در گام بعد، روش طرح‌ریزی آزمایش تاگوچی برای تنظیم فراپارامترهای یادگیری تقویتی عمیق و نحوه اجرای آزمایش‌ها تشریح می‌شود. پس از آن، نتایج شبیه‌سازی و مقایسه عملکرد روش‌های DDPG، TD3 و PPO در کنار حالت پایه الگوریتم فوق‌پیچشی گزارش و تحلیل می‌گردد. در نهایت، جمع‌بندی دستاوردها و نتیجه‌گیری به همراه پیشنهاد مسیرهای ادامه پژوهش ارائه خواهد شد.

۲- مدل‌سازی دینامیکی شبیه‌ساز وضعیت ماهواره

شکل ۱ نمایی از شبیه‌ساز ماهواره مستقر در آزمایشگاه فضایی دانشگاه اصفهان را ارائه می‌دهد. این سامانه به‌منظور بازآفرینی شرایط دینامیکی حرکت دورانی یک ماهواره در محیط آزمایشگاهی طراحی شده و از سه بخش اصلی تشکیل شده است. بخش نخست، سازه نگهدارنده ثابت است که وظیفه تثبیت کل مجموعه را بر عهده دارد. بخش دوم، سامانه یاتاقان هوایی است که با ایجاد تعلیق تقریباً بدون اصطکاک، امکان شبیه‌سازی دقیق حرکت آزاد دورانی را فراهم کرده و اثرات مزاحم ناشی از اصطکاک مکانیکی را به حداقل می‌رساند. بخش سوم، سکوی اصلی است که تمامی عملگرها، حسگرها و زیرسامانه‌های کنترلی بر روی آن نصب شده‌اند و به‌عنوان بدنه معادل ماهواره در آزمایش‌های کنترلی عمل می‌کند.

سامانه تولید گشتاور در این شبیه‌ساز مبتنی بر سه چرخ واکنشی[1] است که در راستای محورهای اصلی اینرسی بدنه آرایش یافته‌اند. این پیکربندی امکان اعمال گشتاور کنترلی مستقل حول هر یک از محورهای اصلی را فراهم می‌کند. افزون بر این، به‌منظور افزایش قابلیت اطمینان و تحمل‌پذیری خطا[2]، یک چرخ واکنشی افزونه و با زاویه ۵۴/۷ درجه با محورهای دیگر چرخ‌ها در ساختار سامانه پیش‌بینی شده است که در شرایط عادی غیرفعال بوده و تنها در صورت بروز نقص یا ازکارافتادگی یکی از چرخ‌های اصلی وارد مدار کنترلی می‌شود. این رویکرد، تداوم عملکرد سامانه و حفظ قابلیت کنترل ماهواره را در شرایط خرابی تضمین می‌کند.



شکل ۱ ساختار کلی شبیه ساز دینامیک وضعیت ماهواره - دانشگاه اصفهان [۱۹]

Fig. 1. General structure of the satellite attitude dynamics simulator — University of Isfahan [19]

مدل دینامیکی سکوی شبیه‌ساز با استفاده از زاویه‌های اویلر $\boldsymbol{\eta} = [\varphi, \theta, \psi]^T$ توصیف می‌شود. در این رابطه، ω (سرعت زاویه‌ای) به مشتق زمانی زاویه‌های اویلر به‌صورت زیر مرتبط است [۲۰]:

$$\dot{\eta} = R(\eta)\omega \quad (۱)$$

[1] Reaction Wheels (RW)
[2] Fault Tolerance

ماتریس چرخش $R(\eta)$ به‌صورت زیر تعریف می‌شود:

$$\boldsymbol{R}^{-1}(\boldsymbol{\eta}) = \begin{bmatrix} 1 & 0 & -\sin\theta \\ 0 & \cos\phi & \sin\phi\cos\theta \\ 0 & -\sin\phi & \cos\phi\cos\theta \end{bmatrix} \quad (۲)$$

تکانه زاویه‌ای کلی شبیه‌ساز ماهواره به‌عنوان مجموع سهم سکو و چرخ‌های واکنشی تعریف می‌شود.

$$h_t = I\omega + h_w \quad (۳)$$

در اینجا، I ماتریس اینرسی سکوی شبیه‌ساز را نشان می‌دهد. تکانه زاویه‌ای چرخ‌های واکنشی به‌صورت $h_w = I_w\omega_w$ بیان می‌شود، که در آن I_w ماتریس اینرسی چرخ‌ها و ω_w بردار سرعت زاویه‌ای آن‌ها است که مستقیماً از انکودرهای موتور اندازه‌گیری می‌شود. با اعمال معادله دینامیکی اویلر در چارچوب بدنه، گشتاور اختلال کلی وارد بر شبیه‌ساز به‌صورت زیر تعیین می‌شود:

$$d = \dot{h}_t + \omega \times h_t \quad (۴)$$

با نادیده گرفتن اثرات یاتاقان هوایی و اختلالات آیرودینامیکی که در حدود $10^{-4}\,N.m$ هستند، گشتاور اختلال d به عدم تعادل سکو نسبت داده شده و می‌توان آن را به‌صورت زیر بیان کرد:

$$\boldsymbol{d}(\boldsymbol{t}) = \left(mg\boldsymbol{r}_s\right) \times \boldsymbol{K} = mg \begin{bmatrix} r_x \\ r_y \\ r_z \end{bmatrix} \times \begin{bmatrix} -\sin\theta \\ \sin\phi\cos\theta \\ \cos\phi\cos\theta \end{bmatrix} = mg \begin{bmatrix} r_y\cos\phi\cos\theta - r_z\sin\phi\cos\theta \\ -r_x\cos\phi\cos\theta - r_z\sin\theta \\ r_x\sin\phi\cos\theta + r_y\sin\theta \end{bmatrix} \quad (۵)$$

که در آن K بردار یکه در راستای نیروی گرانش است و $\boldsymbol{r}_s = \left[r_x, r_y, r_z\right]^T$ بردار مکان از مرکز هندسی تا مرکز جرم سکو را نشان می‌دهد.

با جایگذاری معادله (۳) در معادله (۴)، معادله دینامیکی شبیه‌ساز به‌صورت زیر به‌دست می‌آید:

$$d = I\dot{\omega} + \dot{h}_w + \omega \times \left(I\omega + h_w\right) \quad (۶)$$

زمانی که صفحه و چرخ‌ها به‌عنوان یک سیستم واحد در نظر گرفته شوند، هیچ گشتاور خارجی بر آن وارد نمی‌شود. بنابراین، با توجه به قانون بقای تکانه زاویه‌ای، گشتاوری که بر روی سکو وارد می‌شود برابر با مشتق زمانی تکانه زاویه‌ای چرخ‌ها است:

$$\boldsymbol{\tau} = -\dot{\boldsymbol{h}}_w - \boldsymbol{\omega} \times \boldsymbol{h}_w \quad (۷)$$

با جایگذاری معادله (۷) در معادله (۶)، مدل دینامیکی شبیه‌ساز ماهواره به‌صورت زیر به‌دست می‌آید:

$$I\dot{\omega} = -\omega \times (I\omega) + \tau + d(t) \quad (۸)$$

با در نظر گرفتن دینامیک یک جسم صلب که تحت تأثیر گشتاورهای کنترلی و اختلالات خارجی قرار دارد، وضعیت‌های سیستم به‌عنوان سرعت‌های زاویه‌ای و متغیرهای وضعیت تعریف می‌شوند. نمایش فضایی وضعیت را می‌توان به‌صورت زیر بیان کرد:

$$\begin{cases} \dot{\eta} = f_1(\eta, \omega) \\ \dot{\omega} = f_2(\eta, \omega) + g(\eta)\tau \end{cases} \quad (۹)$$

که:

$$\begin{aligned} f_1(\eta, \omega) &= R\omega \\ f_2(\eta, \omega) &= I^{-1}(-\omega \times (I\omega) + mg(r_s \times K)) \\ g(\eta) &= I^{-1} \end{aligned} \quad (۱۰)$$

۳- راهبردهای کنترلی پیشنهادی

در این بخش، ابتدا طراحی کنترل‌کننده مقاوم مبتنی بر کنترل لغزشی با الگوریتم فوق‌پیچشی ارائه می‌شود و سپس چارچوب یادگیری تقویتی عمیق شامل سه روش DDPG، TD3 و PPO معرفی می‌گردد. در نهایت، ترکیب STA–DRL و نیز روال طراحی آزمایش‌ها به روش تاگوچی برای تنظیم فراپارامترهای یادگیری تقویتی عمیق مطابق معیار هدف تعریف می‌شود.

۳-۱- کنترل لغزشی فوق‌پیچشی

هدف کنترل‌کننده تضمین آن است که حالت‌های سیستم با کمترین انحراف، مسیرهای مرجع مطلوب را دنبال کنند. بر این اساس، خطای رهگیری به‌صورت زیر تعریف می‌شود:

$$e = \eta - \eta_d$$
$$\dot{e} = R\omega - \dot{\eta}_d \quad (۱۱)$$

که در آن $\boldsymbol{\eta}$ نشان‌دهنده بردار وضعیت واقعی و $\boldsymbol{\eta}_d$ نمایانگر مسیر مرجع مطلوب است. برای تضمین عملکرد دقیق رهگیری، فرضیات زیر اتخاذ می‌شوند:

- تمامی وضعیت‌های سیستم قابل اندازه‌گیری یا قابل برآورد از طریق یک مشاهده‌گر هستند.
- مسیر مرجع و مشتقات اول و دوم آن محدود هستند.
- اختلالات خارجی و مشتقات زمانی آن‌ها محدود در نظر گرفته می‌شوند، به‌گونه‌ای که $|d|, |\dot{d}| \leq W$، که در آن W نشان‌دهنده حداکثر سطح مجاز اختلال است.

ایده اصلی کنترل لغزشی طراحی یک قانون کنترل سوییچینگ است که باعث می‌شود وضعیت‌های سیستم به یک سطح لغزشی برسند و روی آن باقی بمانند. سطح لغزشی به‌صورت زیر تعریف می‌شود:

$$s = \dot{e} + \Lambda e \quad (۱۲)$$

که در آن $\Lambda \in R^{3\times3}$ یک ماتریس افزایش قطری مثبت‌تعریف است. قانون فوق‌پیچشی به‌صورت زیر تعریف می‌شود:

$$\tau_{STA} = g(\eta)^{-1}\left(\ddot{\eta}_d - \Lambda\dot{e} - K_1 |s|^{\frac{1}{2}} sign(s) - K_2 \int_0^t sign(s(\tau))d\tau - f_2 \right) \quad (۱۳)$$

که در آن K_1 و K_2 ماتریس‌های کنترل مثبت‌تعریف هستند. پارامترهای به‌کاررفته برای الگوریتم فوق‌پیچشی، که بر اساس مقادیر بهینه و مقاوم گزارش‌شده در پژوهش سرجوقیان و همکاران [۱۹] انتخاب شده‌اند، در جدول ۱ خلاصه شده‌اند.

جدول ۱. ضرایب بهره کنترل کننده الگوریتم فوق‌پیچشی

Table 1. STA controller gains

پارامتر ضرایب بهره	مقدار
Λ	$0.6\boldsymbol{I}_3$
K_1	$0.1\boldsymbol{I}_3$
K_2	$10^{-2}\boldsymbol{I}_3$

اگرچه الگوریتم فوق‌پیچشی بهبودهای قابل‌توجهی نسبت به کنترل لغزشی معمولی با کاهش لرزش و افزایش پایداری ارائه می‌دهد، اما عملکرد آن ممکن است در محیط‌های با عدم‌اطمینان بالا همچنان محدود باشد.

۳-۲- کنترل‌کننده مبتنی بر یادگیری تقویتی عمیق

برخلاف رویکردهای کلاسیک، در این کار ترکیب الگوریتم فوق‌پیچشی با یادگیری تقویتی عمیق به‌منظور یادگیری مستقیم سیاست‌های کنترلی از طریق تعامل با محیط در نظر گرفته شده است. یادگیری تقویتی عمیق چارچوبی است که در آن عامل از طریق تعامل مستقیم با محیط، راهبردهای تصمیم‌گیری بهینه را فرا می‌گیرد [۲۱]. در چارچوب پیشنهادی، دینامیک شبیه‌ساز ماهواره شامل سینماتیک و حرکت دورانی تحت تأثیر اغتشاشات خارجی به‌عنوان محیط آموزش مدل‌سازی شده است. بردار مشاهده با خطاهای رهگیری وضعیت سه‌محوره و خطاهای سرعت زاویه‌ای تعریف می‌شود که اطلاعات کافی برای تصمیم‌گیری را فراهم می‌کند.

فضای عمل متناظر با بردار پیوسته گشتاور کنترلی اعمالی به عملگرها است که با توجه به محدودیت‌های فیزیکی در بازه $\pm 0.125\ N.m$ کران‌بندی شده است. برای هدایت فرآیند یادگیری، تابع پاداش به‌صورت زیر طراحی می‌شود:

$$\text{Reward} = 1 \Big/ \left(\alpha \frac{ITSE}{100} + \beta \|E\| \right) \quad (۱۴)$$

در این پژوهش، تابع پاداش به‌صورت معکوسِ ترکیبی از دو شاخص عملکردی تعریف شده است تا عامل یادگیرنده به‌سمت کاهش هم‌زمان خطای رهگیری و تلاش کنترلی هدایت شود. در این رابطه، $ITSE$ یا انتگرال خطای مربعی وزن‌دهی‌شده با زمان، معیار تجمعی خطای رهگیری را بیان می‌کند و به خطاهایی که در زمان‌های طولانی‌تر باقی می‌مانند وزن بیشتری می‌دهد. همچنین، جمله $\|E\|$ برای لحاظ‌کردن میزان تلاش کنترلی/کارایی کنترلی در تابع پاداش استفاده شده است. ضرایب وزنی α و β موازنه میان دقت رهگیری و تلاش کنترلی را تعیین می‌کنند که در این پژوهش مقادیر آن‌ها به‌ترتیب $\alpha = 0.6$ و $\beta = 0.4$ انتخاب شده است.

به‌منظور حفظ امکان مقایسه منصفانه میان الگوریتم‌های DDPG، TD3 و PPO، ساختار تابع پاداش و ضرایب آن در تمام آزمایش‌ها ثابت نگه داشته شده‌اند. در غیر این صورت، تنظیم جداگانه ضرایب پاداش برای هر الگوریتم می‌توانست به تابع هدف متفاوتی برای هر روش منجر شود و مقایسه نهایی عملکرد الگوریتم‌ها را تحت تأثیر قرار دهد. بنابراین، در این پژوهش تابع پاداش به‌عنوان معیار مشترک آموزش و ارزیابی در نظر گرفته شده و تنظیم سیستماتیک با روش تاگوچی به فراپارامترهای یادگیری محدود شده است. همچنین، به‌دلیل فرم معکوس تابع پاداش، مقدار پاداش کراندار باقی می‌ماند و از ایجاد مقادیر بسیار بزرگ در فرآیند آموزش جلوگیری می‌شود؛ این موضوع به پایداری عددی آموزش و همگرایی مؤثرتر عامل یادگیرنده کمک می‌کند.

در میان الگوریتم‌های مختلف یادگیری تقویتی عمیق، در این پژوهش سه روش شاخص شامل DDPG، PPO و TD3 مورد بررسی قرار گرفته‌اند. الگوریتم DDPG مبتنی بر روش گرادیان سیاست قطعی است [۹, ۲۲] و به‌طور گسترده در مسائل کنترل پیوسته در یادگیری تقویتی به‌کار می‌رود؛ به‌گونه‌ای که با ترکیب DPG و Q-learning، تابع ارزش–کنش به‌روزرسانی می‌شود [۲۳]. با وجود کارایی مناسب، DDPG نسبت به فراپارامترهایی نظیر نرخ یادگیری و نویز اکتشافی حساس بوده و معمولاً از ناکارآمدی نمونه‌ها رنج می‌برد و به تعاملات گسترده با محیط نیاز دارد. برای غلبه بر این محدودیت‌ها، الگوریتم TD3 معرفی شده است که با کاهش اثر خطاهای ناشی از تقریب تابع، پایداری و عملکرد کلی را بهبود می‌دهد [۲۲].

برخلاف روش‌های مبتنی بر Q-learning که بر برآورد تابع ارزش–کنش بهینه تمرکز دارند، بهینه‌سازی سیاست به‌طور مستقیم سیاستی را می‌آموزد که بازده مورد انتظار را بیشینه می‌کند [۹]. الگوریتم PPO یک روش بدون‌مدل از نوع بازیگر–منتقد است که بر چارچوب گرادیان سیاست بنا شده است [۲۴]. در حالی که DDPG و نسخه توسعه‌یافته آن TD3 از روش‌های گرادیان سیاست قطعی بوده و با بهره‌گیری از تقریب تابع، فضاهای کنش پیوسته را مدیریت می‌کنند، PPO از راهبرد بهینه‌سازی سیاست تصادفی استفاده می‌کند که بر پایداری فرآیند آموزش تأکید دارد.

۳-۳- بهینه‌سازی چندهدفه فراپارامترهای یادگیری تقویتی عمیق با روش تاگوچی

تنظیم فراپارامترهای الگوریتم‌های یادگیری تقویتی عمیق یک مسئله چندبعدی و غیرخطی است که به‌شدت بر پایداری فرآیند آموزش، سرعت همگرایی و کیفیت عملکرد کنترلی نهایی اثر می‌گذارد. از آنجا که برهم‌کنش میان فراپارامترهایی نظیر ضریب تنزیل[1]، نرخ یادگیری[2]، اندازه مینی‌بچ[3] و ساختار شبکه‌های بازیگر–منتقد[4] پیچیده و وابسته به دینامیک سیستم است، استفاده از رویکردهای آزمون‌وخطای متداول نه‌تنها هزینه محاسباتی بالایی دارد، بلکه تضمینی برای دستیابی به ترکیب بهینه فراهم نمی‌کند.

به‌منظور غلبه بر چالش تنظیم تجربی و پرهزینه فراپارامترها، در این پژوهش از روش طراحی آزمایش تاگوچی استفاده شده است. روش تاگوچی یکی از روش‌های نظام‌مند طرح‌ریزی آزمایش است که برای بررسی اثر چندین عامل طراحی بر یک پاسخ مشخص، بدون

[1] Discount factor
[2] Learning rate
[3] Mini-batch size
[4] Actor-Critic Networks

نیاز به آزمودن همه ترکیب‌های ممکن، به کار می‌رود. ایده اصلی این روش آن است که به‌جای انجام جست‌وجوی کامل در فضای پارامترها، مجموعه‌ای محدود، متوازن و ساختاریافته از آزمایش‌ها با استفاده از آرایه‌های متعامد انتخاب شود. در آرایه متعامد، ترکیب سطوح عوامل به‌گونه‌ای تنظیم می‌شود که اثر اصلی هر عامل بر پاسخ نهایی قابل بررسی باشد و در عین حال تعداد آزمایش‌ها به‌صورت قابل توجهی کاهش یابد. بنابراین، روش تاگوچی میان دو نیاز مهم توازن ایجاد می‌کند: از یک سو کاهش هزینه زمانی و محاسباتی آزمایش‌ها، و از سوی دیگر حفظ امکان تحلیل اثر عوامل اصلی بر عملکرد سیستم.

در این روش، ابتدا پارامترهای مؤثر مسئله به‌عنوان «عوامل» انتخاب می‌شوند و برای هر عامل، چند مقدار گسسته به‌عنوان «سطح» تعریف می‌شود. سپس، بر اساس تعداد عوامل و تعداد سطوح، یک آرایه متعامد مناسب انتخاب می‌گردد. هر سطر از این آرایه بیانگر یک آزمایش است؛ یعنی یک ترکیب مشخص از سطوح عوامل که باید اجرا و ارزیابی شود. پس از انجام آزمایش‌ها، مقدار پاسخ یا شاخص عملکرد برای هر سطر محاسبه می‌شود. در مرحله بعد، میانگین پاسخ مربوط به هر سطح از هر عامل بررسی شده و سطحی انتخاب می‌شود که مقدار مطلوب‌تری از شاخص عملکرد را ایجاد کند. به این ترتیب، ترکیب نهایی پارامترها نه بر اساس آزمون‌وخطای پراکنده، بلکه بر اساس یک روند ساختاریافته، قابل تکرار و کم‌هزینه‌تر نسبت به جست‌وجوی کامل تعیین می‌شود.

در مسئله حاضر، عوامل همان فراپارامترهای کلیدی الگوریتم‌های یادگیری تقویتی عمیق هستند؛ از جمله ضریب تنزیل، اندازه مینی‌بچ، نرخ یادگیری، نوع بهینه‌ساز و مشخصات شبکه‌های بازیگر و منتقد. سطوح نیز مقادیر گسسته‌ای هستند که برای هر فراپارامتر در جدول ۲ تعریف شده‌اند. از آنجا که در این پژوهش ده عامل و برای هر عامل سه سطح در نظر گرفته شده است، جست‌وجوی کامل مستلزم انجام $۳^{۱۰}$=۵۹۰۴۹ آزمایش برای هر الگوریتم خواهد بود. چنین تعداد آزمایشی، با توجه به زمان‌بر بودن آموزش عامل‌های یادگیری تقویتی عمیق و اجرای مکرر شبیه‌سازی دینامیک غیرخطی ماهواره، از نظر محاسباتی بسیار پرهزینه است. به همین دلیل، در این پژوهش از آرایه متعامد سه‌سطحی L_{27} استفاده شده است. با این آرایه، به‌جای بررسی ۵۹۰۴۹ ترکیب، تنها ۲۷ ترکیب آزمایشی برای هر الگوریتم ارزیابی می‌شود. بنابراین، برای سه الگوریتم DDPG، TD3 و PPO، مجموعاً ۸۱ آزمایش تاگوچی انجام شده است.

در این پژوهش، شاخص عملکرد J به‌عنوان پاسخ هر آزمایش در نظر گرفته شده است. این شاخص ترکیبی از خطای رهگیری و تلاش کنترلی است؛ بنابراین کاهش مقدار آن به معنای بهبود همزمان دقت رهگیری و کاهش هزینه کنترلی است. پس از اجرای آزمایش‌های متناظر با آرایه L_{27}، مقدار J برای هر ترکیب محاسبه شده و اثر سطوح مختلف هر فراپارامتر بر مقدار این شاخص بررسی می‌شود. در نهایت، برای هر فراپارامتر سطحی انتخاب می‌شود که کمترین میانگین شاخص عملکرد را ایجاد کرده باشد. بدین ترتیب، روش تاگوچی در این پژوهش به‌عنوان یک ابزار تنظیم نظام‌مند فراپارامترها به کار رفته است؛ ابزاری که ضمن کاهش قابل‌توجه تعداد آزمایش‌ها، امکان استخراج ترکیب مناسب فراپارامترهای یادگیری تقویتی عمیق را بر پایه یک روند ساختاریافته، تکرارپذیر و مبتنی بر شاخص عملکرد فراهم می‌سازد.

در این چارچوب، ده فراپارامتر کلیدی الگوریتم‌های یادگیری تقویتی عمیق به‌عنوان عوامل قابل تنظیم انتخاب شده‌اند و برای هر یک سه سطح عملیاتی در نظر گرفته شده است. سطوح تعریف‌شده برای این عوامل در جدول ۲ ارائه شده‌اند. این پارامترها شامل تنظیمات مربوط به فرآیند یادگیری (نظیر ضریب تنزیل، نرخ یادگیری، اندازه مینی‌بچ و نوع بهینه‌ساز) و نیز مشخصات معماری شبکه‌های عصبی بازیگر و منتقد هستند که مستقیماً بر قابلیت تقریب تابع و کیفیت سیاست کنترلی اثرگذارند.

ارزیابی هر ترکیب از فراپارامترها بر اساس یک شاخص عملکرد چندهدفه انجام شده است که به‌طور هم‌زمان دقت رهگیری و کارایی کنترلی را در نظر می‌گیرد. این شاخص به‌صورت زیر تعریف می‌شود:

$$J = \alpha \frac{ITSE}{100} + \beta \|E\| \qquad (۱)$$

در این رابطه، $ITSE$ انتگرال خطای مربعی وزن‌دهی‌شده با زمان بوده و معیار تجمعی کیفیت رهگیری در طول بازه آموزش است؛ وزن‌دهی زمانی موجب حساسیت بیشتر شاخص نسبت به خطاهای پایدار و دیرهنگام می‌شود. تقسیم این کمیت بر ۱۰۰ به‌منظور هم‌مقیاس‌سازی مؤلفه‌ها و جلوگیری از غلبه عددی آن انجام شده است. ترم $\|E\|$ بیانگر اندازه تلاش کنترلی اعمال‌شده به عملگرها بوده و نقش محدودکن33نده در برابر اعمال گشتاورهای بیش‌ازحد ایفا می‌کند.

ضرایب وزنی $\alpha = 0.6$ و $\beta = 0.4$ به‌گونه‌ای انتخاب شده‌اند که دقت رهگیری در اولویت قرار گیرد، در حالی که کارایی کنترلی و محدودیت‌های فیزیکی سامانه نیز به‌طور هم‌زمان لحاظ شوند. در نهایت، ترکیب بهینه فراپارامترهای یادگیری تقویتی عمیق از میان سطوح سه‌گانه معرفی‌شده در جدول فوق و بر مبنای کمینه‌سازی شاخص J استخراج شده و در پیاده‌سازی نهایی ساختار کنترلی مورد استفاده قرار گرفته است.

جدول ۲ . سطوح تاگوچی برای بهینه‌سازی یادگیری تقویتی عمیق

Table 2. Taguchi levels for DRL optimization

پارامتر	فراپارامترها	سطح		
		اول	دوم	سوم
A1	ضریب تنزیل (γ)	۰/۹۵	۰/۹۸	۰/۹۹
A2	اندازه مینی‌بچ	۶۴	۱۲۸	۲۵۶
A3	بهینه‌ساز	Adam	RMSprop	SGDM
A4	نرخ یادگیری	۰/۰۰۰۱	۰/۰۰۰۵	۰/۰۰۱
A5	بازیگر — تعداد لایه‌های پنهان	۱	۲	۳
A6	بازیگر — تعداد نورون در هر لایه	۳۲	۶۴	۱۲۸
A7	بازیگر — تابع فعال‌سازی	Relu	Tanh	Sigmoid
A8	منتقد — تعداد لایه‌های پنهان	۱	۲	۳
A9	منتقد — تعداد نورون در هر لایه	۳۲	۶۴	۱۲۸
A10	منتقد — تابع فعال‌سازی	Relu	Tanh	Sigmoid

۳-۴- سناریوهای کنترلی

شکل ۲ بلوک دیاگرام کنترل وضعیت ماهواره را نشان می‌دهد که در آن دو سناریوی کنترلی تعریف و به‌صورت نظام‌مند با یکدیگر مقایسه شده‌اند.

در سناریوی اول، تنها الگوریتم فوق‌پیچشی به‌کار گرفته می‌شود تا کنترل مقاوم در برابر اغتشاشات خارجی و عدم‌قطعیت‌های مدل فراهم گردد. این سناریو به‌عنوان مبنای مقایسه در نظر گرفته شده و عملکرد کنترل مقاوم کلاسیک را ارزیابی می‌کند.

در سناریوی دوم، ساختار کنترلی به‌صورت ترکیبی STA–DRL پیاده‌سازی شده است. در این چارچوب، علاوه بر حفظ خاصیت مقاوم‌بودن الگوریتم فوق‌پیچشی، یک کنترل‌کننده مبتنی بر یادگیری تقویتی عمیق نیز به سیستم افزوده می‌شود تا از طریق تعامل با محیط، سیاست کنترلی بهینه را بیاموزد و به‌صورت تطبیقی با تغییر شرایط دینامیکی و اغتشاشات سازگار شود. این ترکیب، بهره‌گیری هم‌زمان از پایداری تضمین‌شده روش لغزشی و قابلیت یادگیری و انطباق یادگیری تقویتی عمیق را ممکن می‌سازد.

قانون کنترل پیشنهادی برای ساختار ترکیبی به‌صورت زیر بیان می‌شود:

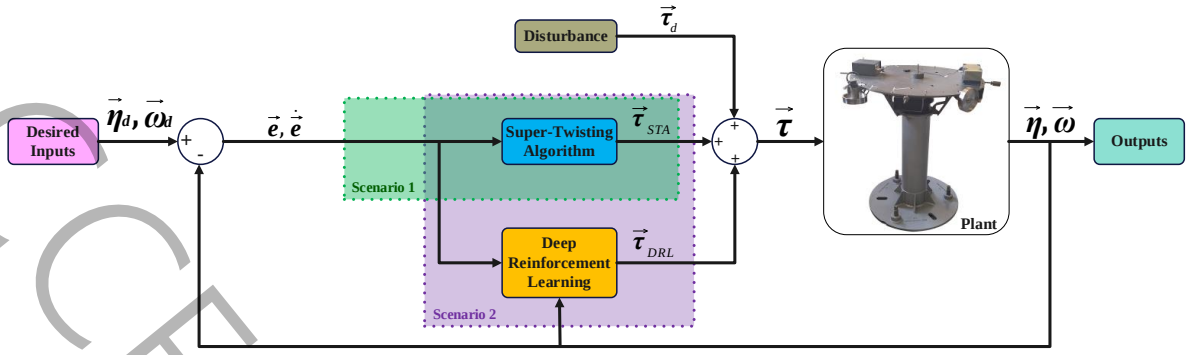
$$\tau = \tau_{STA} + \tau_{DRL} \quad (۲)$$

که در آن τ_{STA} مؤلفه گشتاور مقاوم تولیدشده توسط الگوریتم فوق‌پیچشی و τ_{DRL} گشتاور تطبیقی حاصل از عامل یادگیری تقویتی عمیق است.

۳-۵- تحلیل پایداری ساختار ترکیبی STA–DRL با درنظرگرفتن محدودیت گشتاور

برای بررسی اثر افزودن مؤلفه یادگیری تقویتی عمیق بر پایداری حلقه‌بسته، تحلیل حاضر بر پایه فرم آفاین دینامیک سامانه در روابط (۹) و (۱۰) و سطح لغزش تعریف‌شده در رابطه (۱۲) انجام می‌شود. با مشتق‌گیری از سطح لغزش و استفاده از فرم آفاین دینامیک سامانه، دینامیک سطح لغزش، پس از جداسازی بخش اسمی و جمله‌های نامعین، به‌صورت زیر در نظر گرفته می‌شود:

$$\dot{s} = F_s(x,t) + G_s(x,t)\tau_a + d_s(t) \quad (۳)$$



شکل ۲ . چارچوب کنترلی برای کنترل وضعیت ماهواره

Fig. 2. Control framework for satellite attitude control

در این رابطه، x بردار حالت، τ_a گشتاور واقعی اعمال‌شده به سامانه، $F_s(x,t)$ شامل جمله‌های اسمی دینامیک، مسیر مرجع و عبارت‌های شناخته‌شده مدل، $G_s(x,t)$ ماتریس مؤثر ورودی در دینامیک سطح لغزش، و $d_s(t)$ شامل اغتشاشات، عدم‌قطعیت‌های مدل و خطاهای باقیمانده جبران‌سازی است. فرض می‌شود در ناحیه کاری مورد بررسی، $G_s(x,t)$ کراندار و غیرمنفرد بوده و $d_s(t)$ دارای مشتق کراندار باشد. این فرض با محدودبودن بازه حرکتی شبیه‌ساز، دوربودن از تکینگی‌های نمایش اویلری و کراندار بودن اغتشاشات فیزیکی سامانه سازگار است.

در کنترل‌کننده مقاوم پایه، بخش اسمی دینامیک با استفاده از $F_s(x,t)$ و $G_s(x,t)$ جبران می‌شود تا دینامیک سطح لغزش به فرم مناسب برای اعمال الگوریتم فوق‌پیچشی تبدیل گردد. مؤلفه مجازی الگوریتم فوق‌پیچشی برای هر محور به‌صورت زیر تعریف می‌شود:

$$\tau_{s,STA} = -k_1 \mid s \mid^{1/2} \operatorname{sgn}(s) + v \qquad (۴)$$
$$\dot{v} = -k_2 \operatorname{sgn}(s)$$

که در آن $k_1 > 0$ و $k_2 > 0$ بهره‌های الگوریتم فوق‌پیچشی هستند. بنابراین، فرمان مقاوم متناظر با الگوریتم فوق‌پیچشی به‌صورت زیر نوشته می‌شود:

$$\tau_{STA} = G_s^{-1}(x,t)\left[-F_s(x,t) + \tau_{s,STA}\right] \qquad (۵)$$

در ساختار پیشنهادی، عامل یادگیری تقویتی عمیق در سطح لغزش، متغیر داخلی v، و قانون فوق‌پیچشی وارد نمی‌شود. به بیان دیگر، سطح لغزش و دینامیک داخلی الگوریتم فوق‌پیچشی بدون تغییر باقی می‌مانند و خروجی عامل یادگیری تقویتی عمیق فقط در خروجی کنترل‌کننده با فرمان الگوریتم فوق‌پیچشی جمع می‌شود. بنابراین، فرمان ترکیبی تولیدشده توسط کنترل‌کننده، پیش از ورود اغتشاش فرمانی و اعمال محدودیت نهایی، به‌صورت زیر است:

$$\tau_c = \tau_{STA} + \tau_{DRL} \qquad (۶)$$

در این رابطه، τ_{DRL} گشتاور اصلاحی تولیدشده توسط عامل یادگیری تقویتی عمیق است. نقش این ترم، تضمین پایداری پایه نیست؛ زیرا پایداری توسط هسته مقاوم الگوریتم فوق‌پیچشی تأمین می‌شود. τ_{DRL} به‌عنوان یک مؤلفه کمکی برای اصلاح فرمان نهایی، کاهش تلاش کنترلی، بهبود پاسخ گذرا و افزایش توان مقابله عملی کنترل‌کننده با اغتشاشات به کار می‌رود. از نظر تحلیلی، این تفکیک اهمیت دارد؛ زیرا نشان می‌دهد که اضافه‌شدن یادگیری تقویتی عمیق ساختار پایدارکننده الگوریتم فوق‌پیچشی را تغییر نمی‌دهد، بلکه فقط یک ورودی اصلاحی کراندار به فرمان نهایی اضافه می‌کند.

با توجه به محدودیت فیزیکی عملگرها، خروجی خام عامل یادگیری تقویتی عمیق پیش از اعمال به سامانه محدود می‌شود. بنابراین داریم:

$$\tau_{DRL} = sat_{\tau_{\max}}\left(\bar{\tau}_{DRL}\right) \qquad (۷)$$

که در آن $\bar{\tau}_{DRL}$ خروجی اولیه عامل یادگیری تقویتی عمیق و τ_{DRL} خروجی محدودشده آن است. این محدودسازی باعث می‌شود عامل یادگیری تقویتی عمیق نتواند گشتاوری نامحدود یا خارج از محدوده مجاز عملگر به سامانه اعمال کند.

در این پژوهش، اغتشاشات وارد بر سامانه از دو منشأ در نظر گرفته می‌شوند. دسته نخست، اغتشاشات خارجی و عدم‌قطعیت‌های مدل هستند که در جمله $d_s(t)$ ظاهر می‌شوند. دسته دوم، اغتشاش فرمان گشتاور است که در کانال ورودی عملگر به مجموع فرمان الگوریتم فوق‌پیچشی و یادگیری تقویتی عمیق افزوده می‌شود. بنابراین، قبل از اعمال محدودیت نهایی عملگر، ورودی گشتاوری به‌صورت زیر تعریف می‌شود:

$$\tau_{pre} = \tau_{STA} + \tau_{DRL} + \tau_{dc} \quad (۸)$$

که در آن τ_{dc} اغتشاش فرمان گشتاور است. فرمان واقعی اعمال‌شده به سامانه پس از محدودیت نهایی گشتاور برابر است با:

$$\tau_a = sat_{\tau_{max}}\left(\tau_{pre}\right) \quad (۹)$$

که در این پژوهش $\tau_{\max} = 0.123N.m$ است. این رابطه نشان می‌دهد که محدودیت نهایی گشتاور روی مجموع فرمان مقاوم الگوریتم فوق‌پیچشی، فرمان اصلاحی یادگیری تقویتی عمیق و اغتشاش فرمانی اعمال می‌شود؛ در حالی که اغتشاشات خارجی و عدم‌قطعیت‌های مدل در جمله $d_s(t)$ لحاظ می‌گردند.

برای لحاظ‌کردن اثر محدودیت نهایی عملگر، اختلاف میان فرمان واقعی اعمال‌شده و مجموع ورودی‌های پیش از اشباع به‌صورت زیر تعریف می‌شود:

$$\Delta_{sat} = \tau_a - \tau_{pre} \quad (۱۰)$$

بنابراین:

$$\boldsymbol{\tau}_a = \boldsymbol{\tau}_{STA} + \boldsymbol{\tau}_{DRL} + \boldsymbol{\tau}_{dc} + \Delta_{sat} \quad (۱۱)$$

با جایگذاری این رابطه در دینامیک سطح لغزش داریم:

$$\dot{s} = F_s + G_s\left(\tau_{STA} + \tau_{DRL} + \tau_{dc} + \Delta_{sat}\right) + d_s \quad (۱۲)$$

اکنون با استفاده از تعریف τ_{STA}، رابطه فوق به شکل زیر تبدیل می‌شود:

$$\dot{s} = \tau_{s,STA} + G_s(x,t)\tau_{DRL} + G_s(x,t)\tau_{dc} + G_s(x,t)\Delta_{sat} + d_s(t) \quad (۱۳)$$

رابطه فوق جایگاه واقعی τ_{DRL}، اغتشاش فرمانی و محدودیت نهایی عملگر را در تحلیل پایداری نشان می‌دهد. ترم τ_{DRL} وارد سطح لغزش یا قانون داخلی الگوریتم فوق‌پیچشی نشده است؛ اما چون در خروجی کنترل‌کننده و در کانال گشتاور به فرمان نهایی اضافه می‌شود، اثر آن پس از عبور از ماتریس مؤثر ورودی $G_s(x,t)$ در دینامیک سطح لغزش ظاهر می‌گردد. از آنجا که τ_{DRL} محدود شده و $G_s(x,t)$ در ناحیه کاری کراندار است، سهم $G_s(x,t)\tau_{DRL}$ نیز کراندار خواهد بود.بر این اساس، اغتشاش مؤثر کل به‌صورت زیر تعریف می‌شود:

$$D_T(t) = d_s(t) + G_s(x,t)\tau_{DRL}(t) + G_s(x,t)\tau_{dc}(t) + G_s(x,t)\Delta_{sat}(t) \quad (۱۴)$$

در نتیجه، دینامیک سطح لغزش به فرم زیر بازنویسی می‌شود:

$$\dot{s} = -k_1 \mid s \mid^{1/2} \operatorname{sgn}(s) + v + D_T(t) \quad (۱۵)$$

این بازنویسی به معنای آن نیست که یادگیری تقویتی عمیق بخشی از قانون فوق‌پیچشی یا سطح لغزش شده است. $D_T(t)$ تنها یک بازنویسی تحلیلی از اثر کل جمله‌هایی است که پس از جبران بخش اسمی، در کانال ورودی سطح لغزش باقی می‌مانند. در این بازنویسی، اثر گشتاور اصلاحی یادگیری تقویتی عمیق، اغتشاش فرمان گشتاور، اثر محدودیت نهایی عملگر، اغتشاشات خارجی و عدم‌قطعیت‌های باقیمانده مدل در قالب یک جمله مؤثر گردآوری شده‌اند.

از منظر پایداری، اگر عامل یادگیری تقویتی عمیق عملکرد مناسبی داشته باشد، می‌تواند بخشی از اثر اغتشاش فرمانی یا عدم‌قطعیت‌های مدل را در فرمان نهایی کاهش دهد و در نتیجه بار کنترلی لازم از طرف الگوریتم فوق‌پیچشی را کم کند. با این حال، پایداری حلقه‌بسته به جبران کامل اغتشاش توسط یادگیری تقویتی عمیق وابسته نیست. حتی اگر جبران یادگیری تقویتی عمیق کامل نباشد، چون خروجی آن محدود است، اثر آن از یک کران مشخص فراتر نمی‌رود و می‌تواند در حاشیه مقاومتی الگوریتم فوق‌پیچشی لحاظ شود. بنابراین، پایداری به بهینگی کامل سیاست یادگیری تقویتی عمیق وابسته نیست؛ بلکه به کراندار بودن اثر آن و انتخاب بهره‌های الگوریتم فوق‌پیچشی با حاشیه مقاومتی کافی نسبت به اغتشاش مؤثر کل وابسته است.

فرض می‌شود در ناحیه کاری مورد بررسی، اغتشاش مؤثر کل دارای مشتق کراندار باشد؛ یعنی عدد مثبتی مانند L_T وجود داشته باشد به‌گونه‌ای که:

$$\left\|\dot{D}_T(t)\right\| \le L_T \qquad (۱۶)$$

این فرض با کراندار بودن اغتشاشات خارجی، محدودبودن خروجی یادگیری تقویتی عمیق، کراندار بودن اغتشاش فرمانی، کراندار بودن $G_s(x,t)$، محدودیت نرخ تغییر عملگرها، زمان نمونه‌برداری محدود و قیود فیزیکی سامانه سازگار است. از آنجا که $\Delta_{\text{sat}}(t)$ نیز در تعریف $D_T(t)$ وارد شده است، اثر محدودیت نهایی عملگر در همین کران لحاظ می‌شود. بنابراین، L_T کرانی محافظه‌کارانه برای مجموع اثر اغتشاشات خارجی، عدم‌قطعیت‌های مدل، خروجی کراندار یادگیری تقویتی عمیق، اغتشاش فرمانی و اثر احتمالی محدودیت نهایی عملگر است.

با تعریف متغیر کمکی $z = v + D_T(t)$ دینامیک حلقه‌بسته به‌صورت زیر نوشته می‌شود:

$$\dot{z} = \dot{v} + \dot{D}_T(t) \qquad (۱۷)$$

و در نتیجه با استفاده از قانون فوق‌پیچشی خواهیم داشت:

$$\begin{aligned} \dot{s} &= -k_1 \mid s \mid^{1/2} \operatorname{sgn}(s) + z \\ \dot{z} &= -k_2 \operatorname{sgn}(s) + \rho(t) \end{aligned} \qquad (۱۸)$$

که در آن:

$$\rho(t) = \dot{D}_T(t), \left\|\rho(t)\right\| \le L_T \qquad (۱۹)$$

دستگاه فوق همان فرم استاندارد تحلیل پایداری الگوریتم فوق‌پیچشی در حضور اغتشاش دارای مشتق کراندار است. بنابراین، براساس قضیه پایداری الگوریتم فوق‌پیچشی، اگر بهره‌های k_1 و k_2 با توجه به کران L_T و با حاشیه مقاومتی کافی انتخاب شوند، سطح لغزش در زمان محدود به صفر همگرا می‌شود. در حالت مؤلفه‌ای، برای هر محور می‌توان شرط کافی زیر را در نظر گرفت:

$$\begin{aligned} &k_{2i} > L_{Ti} \\ &k_{1i}^2 > \frac{4L_{Ti}\left(k_{2i} + L_{Ti}\right)}{k_{2i} - L_{Ti}} \end{aligned} \qquad (۲۰)$$

در نتیجه، برای یک زمان محدود t_r داریم:

$$s(t) = 0, t \ge t_r \qquad (۲۱)$$

پس از رسیدن به سطح لغزش، بر اساس رابطه (۱۲)، دینامیک خطای رهگیری به یک دستگاه خطی پایدار تبدیل می‌شود و می‌توان نوشت:

$$\dot{e} = -\Lambda e \qquad (۲۲)$$

از آنجا که Λ مثبت‌تعریف است، خطای رهگیری پس از زمان رسیدن به سطح لغزش به‌صورت نمایی کاهش می‌یابد:

$$e(t) = \exp\left[-\Lambda\left(t - t_r\right)\right] e\left(t_r\right) \qquad (۲۳)$$

بنابراین:

$$\lim_{x \to \infty} e(t) = 0 \qquad (۲۴)$$

نتیجه تحلیل فوق نشان می‌دهد که تضمین پایداری در ساختار پیشنهادی توسط هسته مقاوم الگوریتم فوق‌پیچشی فراهم می‌شود و عامل یادگیری تقویتی عمیق مسئول تضمین پایداری پایه نیست. خروجی یادگیری تقویتی عمیق یک ترم اصلاحی کراندار است که فقط در خروجی کنترل‌کننده با فرمان الگوریتم فوق‌پیچشی جمع می‌شود. اثر این ترم، همراه با اثر اغتشاش فرمانی، اغتشاشات خارجی و اثر احتمالی محدودیت نهایی عملگر، در قالب اغتشاش مؤثر $D_T(t)$ وارد تحلیل می‌شود. بنابراین، تا زمانی که $D_T(t)$ و مشتق آن در حاشیه مقاومتی بهره‌های الگوریتم فوق‌پیچشی باقی بمانند، افزودن مؤلفه یادگیری تقویتی عمیق پایداری حلقه‌بسته را نقض نمی‌کند.

در مورد محدودیت نهایی عملگر، اثر آن از طریق Δ_{sat} در اغتشاش مؤثر لحاظ شده است. در طراحی و پیاده‌سازی این پژوهش، خروجی یادگیری تقویتی عمیق و فرمان واقعی اعمال‌شده به سامانه در محدوده مجاز گشتاور محدود شده‌اند تا اثر محدودیت عملگر

در ناحیه کاری مورد مطالعه کراندار باقی بماند. در صورت فعال‌شدن گذرای محدودیت عملگر، اثر آن در کران L_T منظور می‌شود و شرط پایداری با حاشیه مقاومتی افزایش‌یافته بیان می‌گردد. در مقابل، اگر محدودیت عملگر به‌صورت شدید و طولانی‌مدت فعال شود، اختیار کنترلی مؤثر کاهش یافته و تضمین نظری می‌تواند به یک تضمین عملی و محلی محدود شود. ازاین‌رو، انتخاب بهره‌های الگوریتم فوق‌پیچشی و محدودسازی خروجی یادگیری تقویتی عمیق باید به‌گونه‌ای انجام شود که در شرایط کاری مورد بررسی، اغتشاش مؤثر در حاشیه مقاومتی کنترل‌کننده باقی بماند. بنابراین، نتیجه پایداری ارائه‌شده به‌صورت یک تضمین مقاوم در ناحیه کاری مورد بررسی و تحت فرض کراندار بودن اغتشاش مؤثر و مشتق آن تفسیر می‌شود.

۳-۶- ملاحظات پیچیدگی محاسباتی و قابلیت اجرای برخط روش پیشنهادی

در تحلیل پیچیدگی محاسباتی ساختار پیشنهادی، لازم است میان مرحله آموزش عامل یادگیری تقویتی عمیق و مرحله اجرای کنترل‌کننده تفکیک قائل شد. بخش عمده بار محاسباتی روش پیشنهادی مربوط به مرحله آموزش یادگیری تقویتی عمیق است. در این مرحله، عامل یادگیرنده طی اپیزودهای متعدد با محیط تعامل می‌کند، دینامیک غیرخطی ماهواره شبیه‌سازی می‌شود، کنش کنترلی تولید می‌گردد، مقدار پاداش محاسبه می‌شود و وزن‌های شبکه‌های عصبی بازیگر و منتقد به‌روزرسانی می‌شوند. ازاین‌رو، هزینه محاسباتی آموزش به عواملی مانند تعداد اپیزودهای آموزشی، طول هر اپیزود، اندازه مینی‌بچ، تعداد و اندازه لایه‌های شبکه عصبی و نوع الگوریتم یادگیری تقویتی وابسته است.

با وجود این، مرحله آموزش در ساختار پیشنهادی به‌صورت برون‌خط انجام می‌شود و بنابراین بار محاسباتی آن مستقیماً به حلقه کنترل بلادرنگ تحمیل نمی‌گردد. پس از پایان آموزش، عامل یادگیری تقویتی عمیق در زمان اجرا تنها به‌صورت یک نگاشت آموزش‌دیده از بردار مشاهده به فرمان کنترلی اصلاحی عمل می‌کند. در این مرحله، به‌روزرسانی وزن‌ها، نمونه‌برداری آموزشی یا محاسبه گرادیان انجام نمی‌شود و محاسبات عامل یادگیری تقویتی عمیق به ارزیابی شبکه آموزش‌دیده محدود می‌گردد.

از سوی دیگر، الگوریتم فوق‌پیچشی نیز در مرحله اجرا از بار محاسباتی محدودی برخوردار است. در هر گام زمانی، خطای رهگیری و سطح لغزش محاسبه شده و فرمان کنترلی بر اساس بهره‌های ثابت، تابع علامت و جمله انتگرالی الگوریتم تولید می‌شود. این محاسبات عمدتاً از نوع عملیات جبری برداری و ماتریسی با ابعاد محدود هستند. در سامانه سه‌درجه‌آزادی مورد بررسی، بردارهای حالت و کنترل ابعاد کوچکی دارند و ماتریس‌های بهره نیز به‌صورت قطری در نظر گرفته شده‌اند؛ ازاین‌رو، اجرای الگوریتم فوق‌پیچشی نیازمند حل مسئله بهینه‌سازی تکراری یا محاسبات سنگین درون حلقه کنترل نیست.

در ساختار ترکیبی پیشنهادی، فرمان نهایی کنترل از جمع مؤلفه مقاوم الگوریتم فوق‌پیچشی و مؤلفه اصلاحی یادگیری تقویتی عمیق به دست می‌آید. بنابراین، هزینه اجرای کنترل‌کننده شامل محاسبه جبری فرمان الگوریتم فوق‌پیچشی و ارزیابی شبکه آموزش‌دیده یادگیری تقویتی عمیق است. بر این اساس، اگرچه آموزش و تنظیم فراپارامترهای یادگیری تقویتی عمیق از نظر محاسباتی پرهزینه است، اما این هزینه به مرحله طراحی و آموزش برون‌خط منتقل شده و اجرای نهایی کنترل‌کننده دارای بار محاسباتی محدود و قابل مدیریت است. همچنین، استفاده از آرایه متعامد L_{27} در روش تاگوچی باعث کاهش تعداد آموزش‌های لازم برای تنظیم فراپارامترها شده و هزینه محاسباتی مرحله طراحی را نیز به‌طور قابل‌توجهی کاهش داده است.

در پیاده‌سازی آزمایشگاهی این پژوهش، حلقه کنترل با نرخ ۱۰ هرتز و با تأخیر ارتباطی کمتر از ۵۰ میلی‌ثانیه اجرا شده است. همچنین، بستر پیاده‌سازی شامل سکوی شبیه‌ساز وضعیت ماهواره، حسگر[1]، ارتباط میان لب‌ویو[2] و متلب/سیمولینک[3] و اجرای عامل یادگیری تقویتی عمیق در محیط سیمولینک بوده است. نتایج پیاده‌سازی نشان می‌دهد که ساختار پیشنهادی در این بستر آزمایشگاهی قادر به اجرای برخط و حفظ پایداری وضعیت بوده است. با این حال، تعمیم این نتیجه به سامانه‌های پیچیده‌تر باید با توجه به محدودیت‌های سخت‌افزاری، نرخ نمونه‌برداری مورد نیاز، ابعاد شبکه یادگیری تقویتی عمیق، تعداد درجات آزادی، قیود عملگر و سطح اغتشاشات محیطی انجام شود.

[1] Attitude and Heading Reference System (AHRS)
[2] LABVIEW
[3] MATLAB/Simulink

برای سامانه‌هایی با دینامیک پیچیده‌تر، درجات آزادی بیشتر، قیود سخت‌تر عملگر یا نرخ نمونه‌برداری بالاتر، لازم است زمان اجرای سیاست یادگیری تقویتی عمیق، اندازه شبکه عصبی، اشباع عملگرها و پایداری حلقه بسته پیش از پیاده‌سازی نهایی به‌صورت همزمان بررسی شوند. همچنین، استفاده از آزمون‌های سخت‌افزار-در-حلقه و انتخاب ساختار فشرده‌تر برای شبکه یادگیری تقویتی عمیق می‌تواند برای اطمینان از عملکرد بلادرنگ در سامانه‌های پیچیده‌تر ضروری باشد. بر این اساس، روش پیشنهادی از نظر معماری قابلیت توسعه به سامانه‌های پیچیده‌تر را دارد؛ اما اجرای بلادرنگ آن باید متناسب با محدودیت‌های سخت‌افزاری و دینامیکی هر سامانه ارزیابی شود.

۴- نتایج

در این بخش، نتایج حاصل از ارزیابی چارچوب کنترلی پیشنهادی ارائه و تحلیل می‌شود. ابتدا، نتایج بهینه‌سازی فراپارامترهای الگوریتم‌های یادگیری تقویتی عمیق با استفاده از روش طرح‌ریزی آزمایش تاگوچی گزارش می‌گردد و اثر عوامل کلیدی بر شاخص عملکرد چندهدفه مورد بررسی قرار می‌گیرد. سپس، عملکرد روش‌های کنترلی در محیط شبیه‌سازی عددی شامل کنترل‌کننده پایه مبتنی بر الگوریتم فوق‌پیچشی و ساختار ترکیبی STA–DRL برای الگوریتم‌های DDPG، TD3 و PPO با یکدیگر مقایسه می‌شود. در نهایت، نتایج پیاده‌سازی عملی بر روی شبیه‌ساز وضعیت ماهواره آزمایشگاه فضایی دانشگاه اصفهان ارائه شده و تطابق نتایج تجربی با نتایج شبیه‌سازی مورد بحث قرار می‌گیرد.

به‌منظور افزایش شفافیت و امکان بازتولیدپذیری نتایج، مشخصات رایانه‌ای که فرایندهای آموزش عامل‌های یادگیری تقویتی عمیق و شبیه‌سازی‌های عددی بر روی آن انجام شده‌اند، در جدول ۳ ارائه شده است.

جدول ۳ . پیکربندی سخت‌افزاری و نرم‌افزاری رایانه مورد استفاده برای آموزش و شبیه‌سازی

Table 3. Hardware and software configuration of the computer used for training and simulation

مؤلفه	مشخصات
CPU	Intel Core i7-9700K
RAM	16GB DDR4 @ 3000MHz
GPU	NVIDIA GTX 1080Ti (11GB)
OS	Windows 11
Software	MATLAB (R2023b)

۴-۱- نتایج بهینه‌سازی فراپارامترها با روش تاگوچی

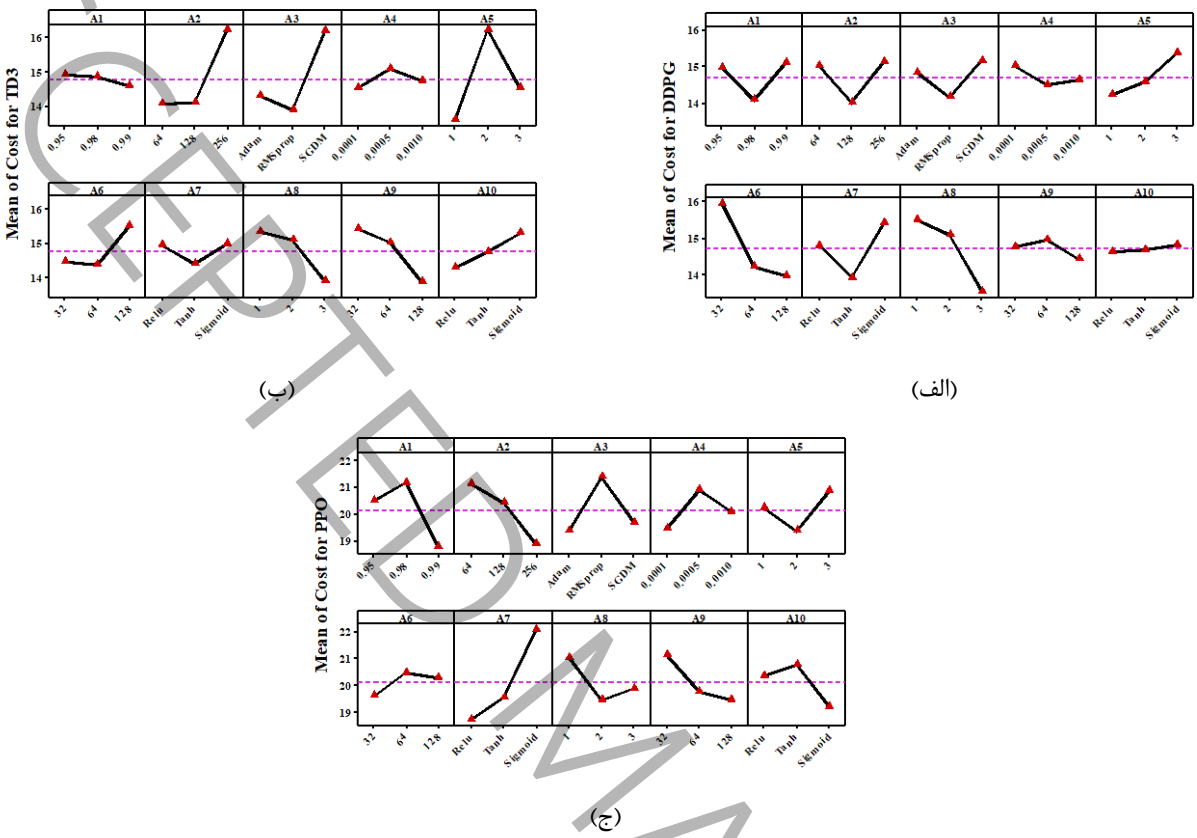
در شکل ۳، نتایج حاصل از اجرای روش طرح‌ریزی آزمایش تاگوچی برای تنظیم فراپارامترهای الگوریتم‌های DDPG، TD3 و PPO به‌ترتیب در زیرشکل‌های (الف)، (ب) و (ج) ارائه شده‌اند. به‌منظور تحلیل دقیق‌تر روند بهینه‌سازی و استخراج تنظیمات نهایی منتخب برای هر الگوریتم، مقادیر بهینه فراپارامترها به‌صورت خلاصه در جدول ۴ گزارش شده است.

با توجه به اینکه هدف بهینه‌سازی، کمینه‌سازی تابع هزینه تعریف‌شده است، سطوح بهینه فراپارامترها برای هر الگوریتم بر اساس انتخاب مقادیری تعیین شده‌اند که کمترین مقدار شاخص عملکرد را در نمودارهای شکل ۳ ایجاد می‌کنند؛ بر این اساس، ترکیب نهایی پارامترهای منتخب برای هر یک از روش‌ها در جدول ۴ ارائه شده است.

برای تکمیل گزارش فرایند آموزش، تنظیمات عمومی و اختصاصی آموزش عامل‌های یادگیری تقویتی عمیق در جدول ۵ ارائه شده است. تنظیمات عمومی آموزش برای هر سه الگوریتم یکسان در نظر گرفته شد تا مقایسه عملکرد DDPG، TD3 و PPO تحت شرایط آموزشی مشابه انجام شود. همچنین، مقدار سید[1] در تمام آموزش‌ها برابر ۴۲ تنظیم شد تا اثر تغییرات تصادفی کنترل شده و نتایج قابل بازتولید باشند. بر اساس تنظیمات جدول ۵ ، روند آموزش سه عامل برای ترکیب نهایی تاگوچی در شکل ۴ نشان داده شده

1 Seed

است. در این شکل، خطوط کم‌رنگ بیانگر پاداش هر اپیزود و خطوط پررنگ بیانگر میانگین متحرک پاداش با پنجره ۲۰ اپیزودی هستند. انتخاب پنجره ۲۰ اپیزودی با طول پنجره میانگین‌گیری پاداش در معیار توقف آموزش هماهنگ است و روند کلی یادگیری را نسبت به پاداش خام اپیزودی خواناتر نشان می‌دهد.



شکل ۳ . نتایج بهینه‌سازی فراپارامترهای الگوریتم‌های یادگیری تقویتی عمیق با روش تاگوچی: (الف) DDPG، (ب) TD3، (ج) PPO

Fig. 3. Hyperparameter optimization results of deep reinforcement learning algorithms using the Taguchi method: (a) DDPG, (b) TD3, (c) PPO

جدول ۴ . مقادیر بهینه فراپارامترهای منتخب الگوریتم‌های DDPG، TD3 و PPO بر اساس روش تاگوچی

Table 4. Optimal hyperparameter values of the selected DDPG, TD3, and PPO algorithms based on the Taguchi method

روش DRL	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
DDPG	۰/۹۸	۱۲۸	RMSProp	۰/۰۰۰۵	۱	۱۲۸	Tanh	۳	۱۲۸	Relu
TD3	۰/۹۹	۶۴	RMSProp	۰/۰۰۰۱	۱	۱۲۸	Tanh	۳	۱۲۸	Relu
PPO	۰/۹۹	۲۵۶	Adam	۰/۰۰۰۱	۱	۳۲	Relu	۲	۱۲۸	Sigmoid

مطابق شکل ۴، رفتار آموزشی سه الگوریتم یکسان نیست. DDPG پس از چند اپیزود اولیه به‌سرعت از ناحیه پاداش‌های منفی خارج شده و وارد ناحیه پاداش‌های مثبت می‌شود. TD3 نیز با وجود نوسانات اولیه، به‌تدریج روند پایدارتر و افزایشی‌تری پیدا کرده و در انتهای آموزش در ناحیه پاداش‌های مثبت باقی می‌ماند. در مقابل، PPO در تنظیمات آموزشی مورد استفاده، نوسانات شدیدتری نشان داده و در بخش عمده‌ای از آموزش در ناحیه پاداش‌های پایین قرار می‌گیرد.

این تفاوت می‌تواند به ساختار یادگیری الگوریتم‌ها مربوط باشد. DDPG و TD3 از نوع برون‌سیاستی[1] بوده و با استفاده از حافظه تجربه، امکان بهره‌برداری مجدد از داده‌های آموزشی را دارند؛ بنابراین، در مسئله کنترل پیوسته و کراندار حاضر، کارایی نمونه بهتری نشان می‌دهند. در مقابل، PPO یک الگوریتم درون‌سیاستی[2] است و وابستگی بیشتری به داده‌های تازه تولیدشده در هر مرحله آموزش دارد. ازاین‌رو، در این مسئله و با بودجه آموزشی یکسان، حساسیت بیشتری نسبت به فرایند اکتشاف نشان داده و به روند یادگیری مطلوب DDPG و TD3 نرسیده است.

جدول ۵ . تنظیمات آموزش عامل‌های DDPG، TD3 و PPO

Table 5. Training settings of the DDPG, TD3, and PPO agents

تنظیمات آموزش	DDPG	TD3	PPO
زمان نمونه‌برداری (ثانیه)	۰/۱	۰/۱	۰/۱
بیشینه تعداد اپیزودها	۳۰۰	۳۰۰	۳۰۰
بیشینه تعداد گام در هر اپیزود	۱۰۰۰	۱۰۰۰	۱۰۰۰
طول پنجره میانگین‌گیری پاداش	۲۰	۲۰	۲۰
معیار توقف آموزش	Average Reward	Average Reward	Average Reward
مقدار توقف آموزش	۱۰۰	۱۰۰	۱۰۰
سید تصادفی	۴۲	۴۲	۴۲
راهبرد اکتشاف / نویز سیاست	Ornstein–Uhlenbeck، $\sigma = 0.3$، نرخ کاهش 10^{-4}	Gaussian، $\sigma = 0.2$، کران ± 0.5	سیاست تصادفی گاوسی
طول حافظه تجربه	۱۰۰۰۰۰۰	۱۰۰۰۰۰۰	—
ضریب نرم‌سازی هدف	۰/۰۰۱	۰/۰۰۱	—
پارامترهای اختصاصی PPO	—	—	Clip=0.2، GAE^{3}=0.95، Entropy=0.01، Horizon=512، Epoch=3

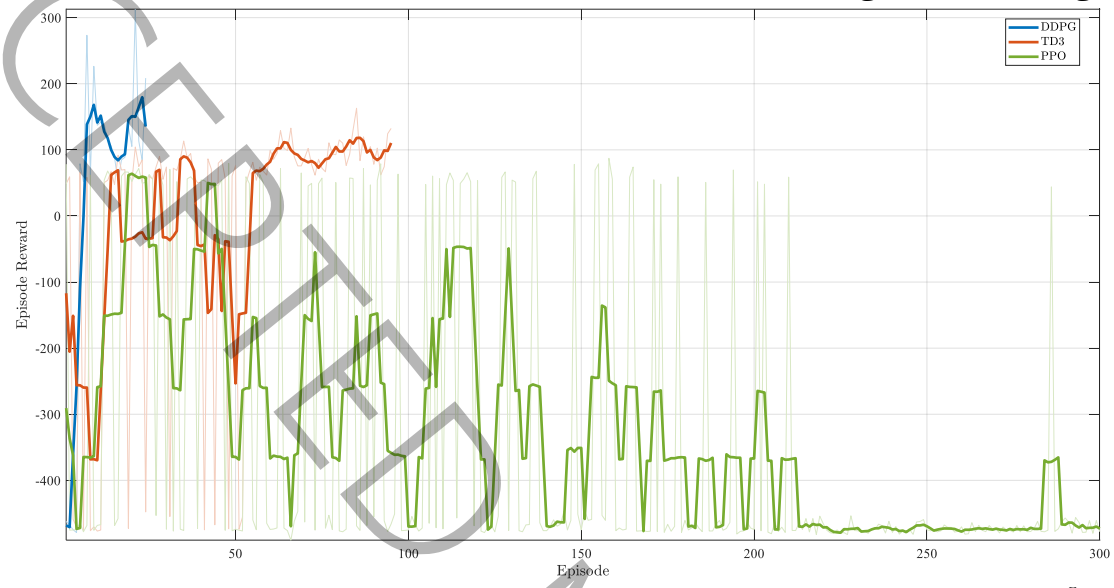
با این حال، انتخاب کنترل‌کننده نهایی صرفاً بر اساس مقدار پاداش آموزشی انجام نشده است. ارزیابی نهایی بر مبنای عملکرد حلقه‌بسته، دقت رهگیری، تلاش کنترلی و مقاومت در برابر اغتشاش صورت گرفته است. بر این اساس، اگرچه DDPG نیز در مرحله آموزش به پاداش‌های مثبت دست یافته است، نتایج عملکردی نشان می‌دهد که TD3 رفتار متوازن‌تر و مقاوم‌تری در ساختار ترکیبی STA–DRL ایجاد کرده است.

۴-۲- نتایج شبیه‌سازی

در این بخش، به‌منظور ارائه‌ی یک ارزیابی دقیق و قابل اتکا از عملکرد راهبردهای کنترلی، نتایج شبیه‌سازی در دو سناریوی مکمل گزارش می‌شود. سناریوی نخست به‌عنوان حالت مرجع، بدون اعمال اغتشاش خارجی در نظر گرفته شده است تا رفتار ذاتی کنترل‌کننده‌ها از منظر دقت رهگیری، سرعت همگرایی، میزان فراجهش/کمینه‌جهش و مصرف تلاش کنترلی به‌صورت شفاف مقایسه شود. سناریوی دوم با هدف بررسی مقاوم بودن[4] و توانایی کاهش اغتشاشات، شامل اعمال یک اغتشاش گشتاوری از نوع «اغتشاش در فرمان[5]» است که به‌صورت $\tau_{\text{command disturbance}}(t) = 0.1 \cdot (1+\sin(t/100)) \cdot [1 \quad 1 \quad 1]^T$ تعریف می‌شود.

[1] Off-Policy
[2] On-Policy
[3] Generalized Advantage Estimation
[4] Robustness
5 Command disturbance

این اغتشاش با دامنه‌ی ثابت و مؤلفه‌ی سینوسیِ آهسته‌تغییر، به‌طور همزمان بر هر سه محور اعمال می‌گردد و به‌عنوان یک ورودی مزاحم هدفمند، شرایطی نزدیک به واقعیت را برای سنجش تاب‌آوری کنترل‌کننده‌ها در برابر عدم‌قطعیت‌ها و تحریکات خارجی فراهم می‌کند. بر همین اساس، مقایسه‌ی نتایج دو حالت «بدون اغتشاش» و «با اغتشاش» امکان تحلیل هم‌زمان دقت عملکرد و پایداری/سازگاری‌پذیری را مهیا کرده و دید روشنی از موازنه‌ی میان کیفیت پاسخ، نرمی سیگنال کنترلی (کاهش نوسانات و لرزش) و هزینه‌ی کنترلی در هر راهبرد ارائه می‌دهد.



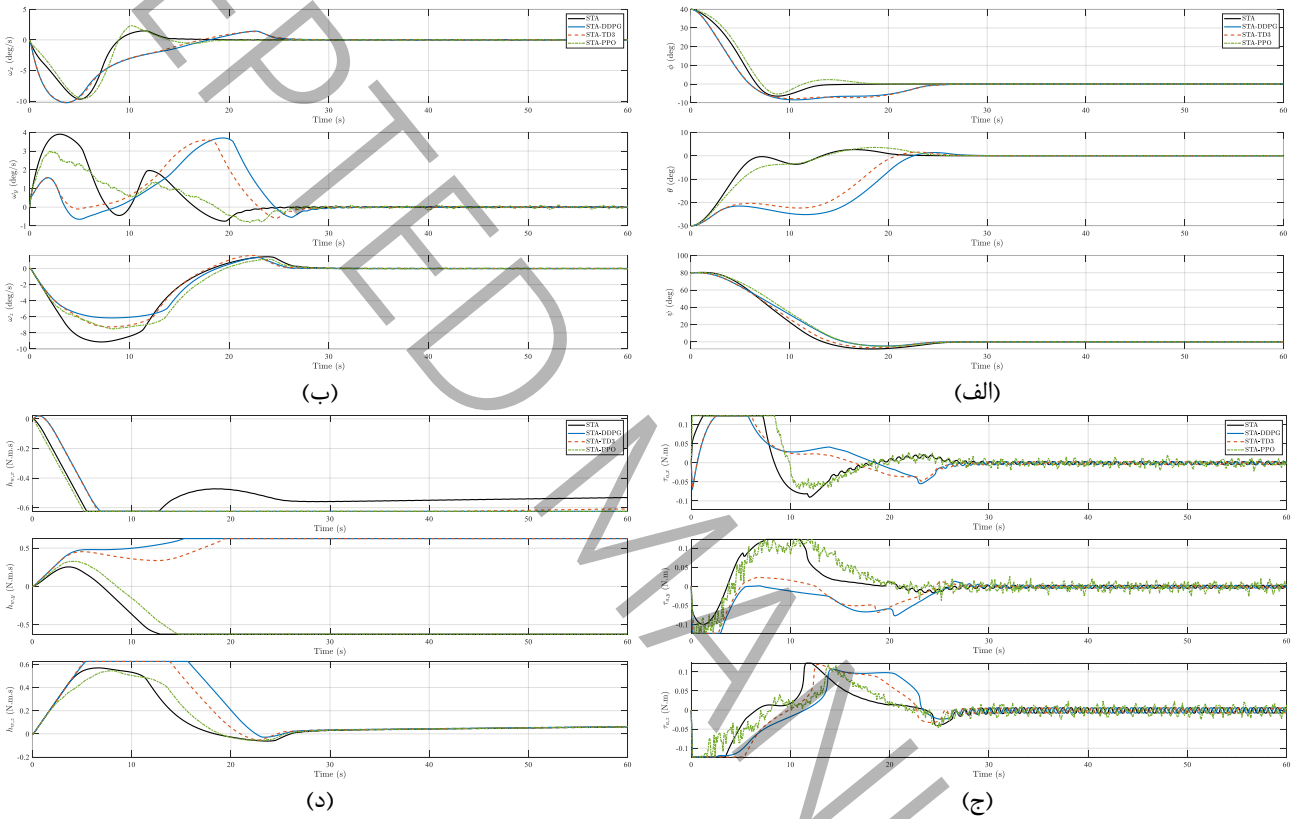
شکل ۴ . روند آموزش عامل‌های DDPG، TD3 و PPO برای ترکیب نهایی تاگوچی؛ خطوط کم‌رنگ پاداش اپیزودی و خطوط پررنگ میانگین متحرک پاداش با پنجره ۲۰ اپیزودی را نشان می‌دهند.

Fig. 4. Training progress of the DDPG, TD3, and PPO agents for the final Taguchi combination; faint lines represent episodic rewards and bold lines represent the 20-episode moving average of rewards

در شکل ۵، پاسخ سامانه در حالت بدون اغتشاش نشان می‌دهد که هر چهار راهبرد کنترلی پایدار بوده و خطای نهایی در مرتبه بسیار کوچک باقی مانده است؛ به‌طوری‌که مقدار MSE برای الگوریتم فوق‌پیچشی برابر $۱/۷۱۹\times۱۰^{-۸}$، برای STA-DDPG برابر $۳/۷۱۹\times۱۰^{-۸}$، برای STA-TD3 برابر $۴/۵۶۷\times۱۰^{-۸}$ و برای STA-PPO برابر $۶/۰۰۵\times۱۰^{-۸}$ به‌دست آمده است. با وجود دقت بالاتر الگوریتم فوق‌پیچشی از نظر شاخص‌های خطا، روش‌های ترکیبی از نظر سرعت و تلاش کنترلی عملکرد بهتری دارند؛ زمان نشست از ۲۶ ثانیه در الگوریتم فوق‌پیچشی به ۲۳ ثانیه در STA-DDPG و ۲۲ ثانیه در STA-TD3 کاهش یافته است، و تلاش کنترلی نیز از ۲۴/۲ در الگوریتم فوق‌پیچشی به ۲۳/۵۸و ۲۳/۱ در این دو روش رسیده است. بنابراین، STA-TD3 نسبت به الگوریتم فوق‌پیچشی حدود ٪۱۵/۴ کاهش زمان نشست و حدود ٪۴/۵ کاهش تلاش کنترلی ایجاد کرده است. نوسانات کوچک مشاهده‌شده در ورودی کنترلی، با توجه به زمان نمونه‌برداری ۰/۱ ثانیه و نرخ اجرای ۱۰ هرتز، عمدتا ناشی از ماهیت گسسته پیاده‌سازی و ساختار سوئیچینگ کنترل مقاوم است و به‌معنای ناپایداری پاسخ نیست؛ زیرا همه روش‌ها همگرایی زوایا و سرعت‌های زاویه‌ای را حفظ کرده‌اند. انتظار می‌رود در نرخ‌های نمونه‌برداری بالاتر و سخت‌افزار سریع‌تر، این نوسانات کاهش یابد و فرمان کنترلی هموارتر شود. نمودارهای مومنتوم زاویه‌ای نیز نشان می‌دهند که چرخ‌های واکنشی پس از تغییر اولیه برای انتقال سامانه به ناحیه مطلوب، با مومنتومی تقریباً ثابت و اصلاحات محدود، پایداری وضعیت را حفظ می‌کنند.

در شکل ۶، پاسخ سامانه در حضور اغتشاش نشان می‌دهد که هر چهار روش پایداری نهایی را حفظ کرده‌اند، اما دامنه گذرا و زمان فروکش‌کردن نوسانات در آن‌ها متفاوت است. در پاسخ زاویه‌ای، روش الگوریتم فوق‌پیچشی در محورهای ϕ و θ پس از افت اولیه به‌ترتیب تا حدود ۵- و سپس حدود ۸+ تا ۱۲+ درجه نوسان می‌کند، در حالی‌که STA-DDPG و STA-TD3 دامنه نوسانات کوچک‌تری داشته و در حدود ۲۰ تا ۲۵ ثانیه به نزدیکی صفر می‌رسند. در مقابل، STA-PPO در محور ϕ یک فراجهش مثبت حدود ۲۵ درجه در بازه ۲۰ تا ۳۰ ثانیه ایجاد می‌کند و در محور ψ نیز کندتر از سایر روش‌ها به مقدار مرجع نزدیک می‌شود. در

سرعت‌های زاویه‌ای نیز الگوریتم فوق‌پیچشی در برخی محورها دامنه‌هایی در حدود ۶ تا ۸ درجه بر ثانیه ایجاد می‌کند، در حالی‌که STA-DDPG و STA-TD3 عمدتاً نوسانات را در محدوده کوچک‌تری، حدود ۳ تا ۵ درجه بر ثانیه، نگه می‌دارند. ورودی‌های کنترلی همه روش‌ها در محدوده مجاز گشتاور باقی مانده‌اند، اما STA-PPO بیشترین پراکندگی و نوسان ریز را در فرمان نشان می‌دهد؛ در مقابل، STA-DDPG و STA-TD3 فرمان‌های منظم‌تری تولید کرده‌اند. نمودارهای مومنتوم زاویه‌ای نیز نشان می‌دهند که چرخ‌ها در برخی محورها تا نزدیکی کران مومنتوم حرکت می‌کنند، اما پس از گذرا واگرایی یا تجمع نامطلوب مومنتوم مشاهده نمی‌شود. بنابراین، شکل ۶ نشان می‌دهد که در حضور اغتشاش، STA-DDPG و به‌ویژه STA-TD3 نسبت به STA و STA-PPO پاسخ گذرای کنترل‌شده‌تر، دامنه نوسان کمتر در نرخ‌های زاویه‌ای و رفتار کنترلی منظم‌تری ایجاد می‌کنند، در حالی‌که پایداری نهایی سامانه در هر چهار روش حفظ می‌شود.

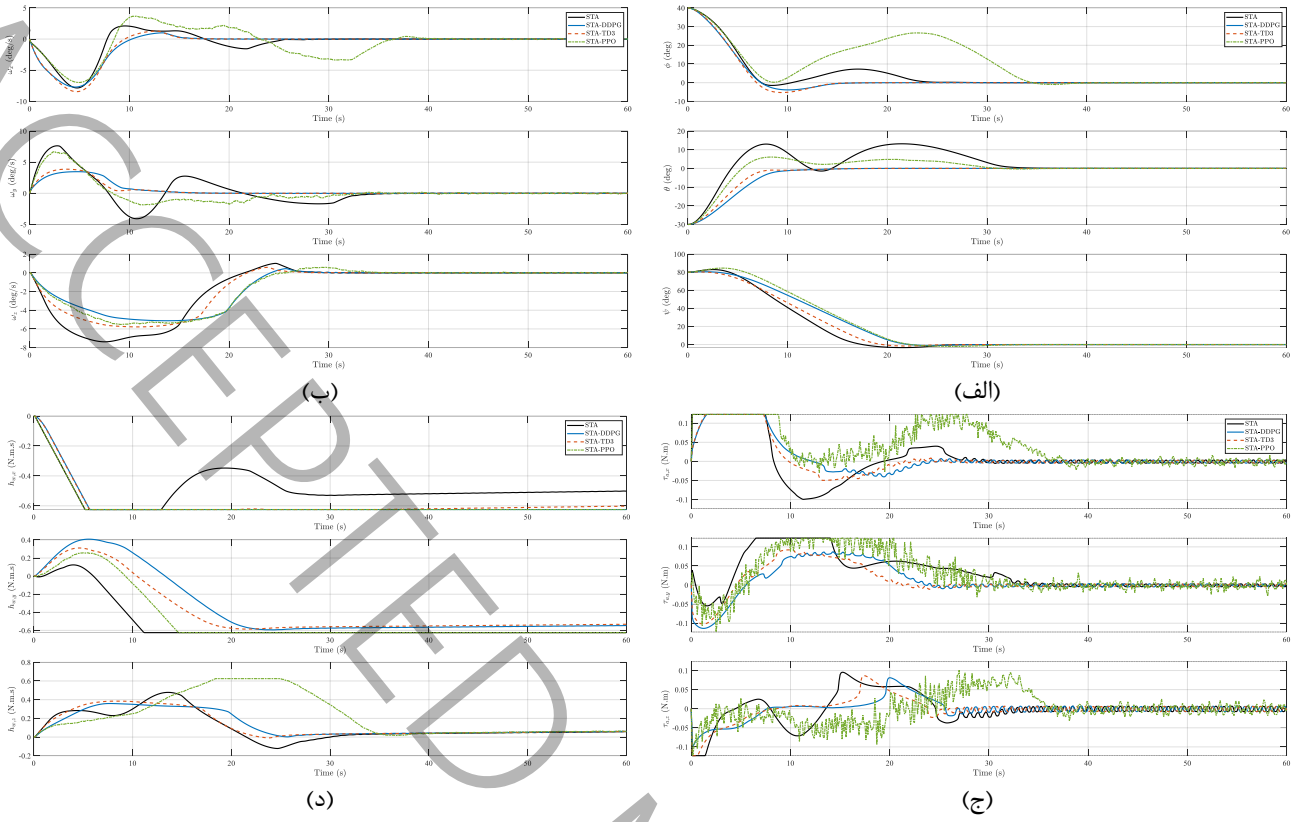


شکل ۵ . مقایسه پاسخ‌های شبیه‌سازی چهار روش کنترلی: (الف) زوایا، (ب) سرعت زاویه‌ای، (ج) ورودی کنترلی، (د) مومنتوم زاویه‌ای

Fig. 5. Comparison of simulation responses of four control methods: (a) angles, (b) angular velocity, (c) control input, (d) angular momentum

در گام نهایی مقایسه، هر چهار راهبرد کنترلی در شبیه‌سازی و بر اساس شاخص‌های عملکردی خلاصه‌شده در جدول ۶ ارزیابی شده‌اند. این شاخص‌ها شامل میانگین مربعات خطا[1]، انتگرال مربعات خطا[2]، انتگرال خطای مربعی وزن‌دهی‌شده با زمان، تلاش کنترلی، زمان نشست و بیشینه سطح اغتشاش قابل تحمل هستند. برای ارزیابی مقاومتی جامع‌تر، سه سناریوی اغتشاش در نظر گرفته شده است.

[1] Mean Squared Error (MSE)
[2] Integral of Squared Error (ISE)



شکل ۶ . مقایسه پاسخ‌های شبیه‌سازی چهار روش کنترلی در حضور اغتشاش: (الف) زوایا، (ب) سرعت زاویه‌ای، (ج) ورودی کنترلی، (د) مومنتوم زاویه‌ای

Fig. 6. Comparison of simulation responses of four control methods in the presence of disturbance: (a) angles, (b) angular velocity, (c) control input, (d) angular momentum

سناریوی نخست، اغتشاش فرمان گشتاور سینوسی ـ ثابت است که پیش از اشباع نهایی به مسیر ورودی کنترلی اعمال می‌شود و به‌صورت زیر تعریف می‌گردد:

$$\boldsymbol{\tau}_{\text{command disturbance}}(t) = A \cdot \left(1 + \sin\left(\frac{t}{100}\right)\right) \cdot \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T \qquad (۲۵)$$

در این رابطه، A دامنه اغتشاش فرمان گشتاور است. سناریوی دوم، اغتشاش تصادفی فرمان گشتاور است. در این حالت، خروجی بلوک نویز سفید با زمان نمونه‌برداری ۰/۱ ثانیه به مسیر ورودی کنترلی و پیش از اشباع نهایی افزوده شده است. مقدار گزارش‌شده برای این سناریو در جدول ۶ ، توان نویز سفید اعمال‌شده در بلوک «نویز سفید» است. بنابراین، دو سناریوی نخست هر دو از نوع اغتشاش در مسیر فرمان گشتاور هستند و پیش از اعمال اشباع نهایی وارد ورودی کنترلی می‌شوند.

سناریوی سوم، اغتشاش ضربه‌ای وارد بر بدنه است. برخلاف دو سناریوی قبل، این اغتشاش در مسیر فرمان کنترلی اعمال نشده، بلکه مستقیماً در مدل دینامیکی بدنه لحاظ شده است. بر اساس رابطه (۸)، گشتاورهای اختلال خارجی از طریق جمله $d(t)$ وارد دینامیک دورانی بدنه می‌شوند. در مدل پایه، این جمله مطابق رابطه (۵)، ناشی از عدم‌تعادل سکو و جابه‌جایی مرکز جرم نسبت به مرکز هندسی در نظر گرفته شده است. برای سناریوی ضربه‌ای، جمله اغتشاش خارجی به‌صورت زیر توسعه داده می‌شود:

$$d_{\text{new}}(t) = d(t) + d_{\text{imp}}(t) \qquad (۲۶)$$

که در آن $d(t)$ همان گشتاور اختلال تعریف‌شده در رابطه (۵) و $d_{\text{imp}}(t)$ گشتاور ضربه‌ای خارجی وارد بر بدنه است. این گشتاور ضربه‌ای به‌صورت زیر تعریف شده است:

$$d_{imp}(t) = \begin{cases} \tau_{imp}\begin{bmatrix}1 & 1 & 1\end{bmatrix}^T, t_{imp} \le t < t_{imp} + T_s \\ 0, \text{ otherwise} \end{cases} \quad (۲۷)$$

در این رابطه، τ_{imp} دامنه گشتاور ضربه‌ای، t_{imp} زمان اعمال ضربه و ۰٫۲ ثانیه مدت اعمال آن است. در نتیجه، اغتشاش ضربه‌ای برخلاف اغتشاش فرمان گشتاور و نویز سفید، از مسیر اشباع فرمان عبور نمی‌کند و مستقیماً روی دینامیک دورانی بدنه اثر می‌گذارد. سه ستون پایانی جدول ۶ به‌ترتیب بیشینه اغتشاش فرمان گشتاور، بیشینه گشتاور ضربه‌ای وارد بر بدنه و بیشینه توان نویز سفید قابل تحمل را برای هر روش نشان می‌دهند.

جدول ۶ . مقایسه عملکرد سناریوهای کنترلی

Table 6. Performance comparison of control scenarios

روش	MSE (میانگین مربعات خطا)	ISE (انتگرال مربعات خطا)	ITSE (انتگرال خطای مربعی وزن‌دهی‌شده با زمان)	تلاش کنترلی	زمان نشست (ثانیه)	اغتشاش ($N.m$) فرمان گشتاور (A)	اغتشاش ($N.m$) ضربه‌ای وارد بر دینامیک بدنه τ_{imp}	اغتشاش ($N.m$) نویز سفید فرمان گشتاور P_w
STA	۱/۷۱۹۱×۱۰$^{-۸}$	۲/۵۵	۶۲۹/۹	۲۴/۲	۲۶	۰/۱۷	۳/۱۵	۰/۰۰۱
STA-DDPG	۳/۷۱۹۰×۱۰$^{-۸}$	۲/۹۳۷	۹۷۱	۲۳/۵۸	۲۳	۰/۲۵	۳/۲۵	۰/۰۰۳
STA-TD3	۴/۵۶۷۰×۱۰$^{-۸}$	۲/۷۵۷	۸۷۵	۲۳/۱۰	۲۲	۰/۲۵۵	۳	۰/۰۰۳
STA-PPO	۶/۰۰۵۰×۱۰$^{-۸}$	۲/۸۲۴	۶۷۷/۲	۲۶/۲۳	۲۴	۰/۱۳۵	۲/۸	۰/۰۰۲

توضیح: در ستون «اغتشاش فرمان گشتاور»، مقدار A دامنه اغتشاش سینوسی ـ ثابت اعمال‌شده پیش از اشباع نهایی است. ستون «گشتاور ضربه‌ای وارد بر بدنه» مقدار دامنه گشتاور ضربه‌ای τ_{imp} را نشان می‌دهد که مستقیماً در جمله $d(t)$ معادله دینامیکی بدنه و به‌مدت ۰/۲ ثانیه اعمال شده است. ستون «نویز سفید فرمان گشتاور» مقدار توان نویز سفید P_w در بلوک «نویز سفید» را بیان می‌کند؛ این نویز با زمان نمونه‌برداری ۰/۱ ثانیه، پیش از اشباع نهایی به مسیر ورودی کنترلی افزوده شده است.

در جدول ۶ ، هر چهار روش از نظر دقت رهگیری عملکرد بسیار بالایی دارند و مقدار MSE در همه حالت‌ها در مرتبه 10^{-8} باقی مانده است؛ با این حال، تفاوت اصلی روش‌ها در موازنه میان دقت، سرعت پاسخ، تلاش کنترلی و بیشینه سطح اغتشاش قابل تحمل ظاهر می‌شود. از نظر شاخص‌های خطا، الگوریتم فوق‌پیچشی دقیق‌ترین پاسخ را ارائه می‌دهد؛ زیرا کمترین مقدار MSE و همچنین کمترین مقادیر ISE و ITSE را ثبت کرده است. این نتیجه نشان می‌دهد که خطا، هم در کل بازه زمانی و هم با وزن‌دهی زمانی، در روش الگوریتم فوق‌پیچشی کمتر باقی می‌ماند. با وجود این، الگوریتم فوق‌پیچشی کندترین پاسخ را دارد و زمان نشست آن ۲۶ ثانیه است. در مقابل، روش‌های ترکیبی STA-DDPG و به‌ویژه STA-TD3 از نظر اجرایی کاراتر ظاهر شده‌اند؛ زیرا سریع‌تر همگرا می‌شوند، به‌ترتیب با زمان نشست ۲۳ و ۲۲ ثانیه، و تلاش کنترلی کمتری نیز نیاز دارند، به‌ترتیب ۲۳/۵۸ و ۲۳/۱۰. البته این بهبود در سرعت و کاهش تلاش کنترلی با افزایش محدود شاخص‌های خطا همراه است؛ به‌ویژه مقدار ITSE در STA-DDPG و STA-TD3 به‌ترتیب به ۹۷۱ و ۸۷۵ رسیده است، که نشان می‌دهد بخشی از خطا در زمان‌های دیرتر نسبت به الگوریتم فوق‌پیچشی بیشتر باقی مانده یا دیرتر حذف شده است. روش STA-PPO اگرچه از نظر ITSE مقدار نسبتاً مناسبی دارد، اما به‌دلیل بیشترین تلاش کنترلی، یعنی ۲۶٫۲۳، و همچنین MSE بالاتر، از نظر کارایی عملی نسبت به STA-DDPG و STA-TD3 برتری نشان نمی‌دهد.

سه ستون پایانی جدول ۶ ، برخلاف شاخص‌های خطا، بیانگر دقت پاسخ نیستند؛ بلکه بیشینه سطح اغتشاش قابل تحمل را در سه سناریوی متفاوت نشان می‌دهند. در سناریوی اغتشاش فرمان گشتاور، STA-TD3 با مقدار ۰/۲۵۵ نیوتن‌متر بیشترین سطح قابل

تحمل را دارد و پس از آن STA-DDPG با مقدار ۰/۲۵ نیوتن‌متر قرار می‌گیرد. این مقادیر نسبت به الگوریتم فوق‌پیچشی با ۰/۱۷ نیوتن‌متر و STA-PPO با ۰/۱۳۵ نیوتن‌متر بالاتر هستند. بنابراین، در برابر اغتشاشاتی که در مسیر فرمان گشتاور و پیش از اشباع نهایی اعمال می‌شوند، افزودن مؤلفه یادگیری تقویتی عمیق، به‌ویژه TD3، باعث افزایش مقاومت کنترل‌کننده شده است. در سناریوی نویز سفید فرمان گشتاور نیز STA-DDPG و STA-TD3 هر دو توان نویز ۰/۰۰۳ را تحمل کرده‌اند، در حالی که مقدار قابل تحمل برای الگوریتم فوق‌پیچشی برابر ۰/۰۰۱ و برای STA-PPO برابر ۰/۰۰۲ است. این نتیجه نشان می‌دهد که ساختارهای ترکیبی مبتنی بر DDPG و TD3 نسبت به STA تنها، ظرفیت بیشتری برای مقابله با اغتشاشات تصادفی وارد بر مسیر فرمان دارند.

در سناریوی گشتاور ضربه‌ای وارد بر بدنه، روند نتایج متفاوت است. در این حالت، اغتشاش از مسیر فرمان کنترلی عبور نمی‌کند، بلکه مستقیماً به جمله اختلال خارجی $d(t)$ در مدل دینامیکی بدنه افزوده می‌شود؛ ازاین‌رو، اثر آن بیشتر به مقاومت ذاتی دینامیک حلقه‌بسته در برابر ضربه خارجی مربوط است تا تعدیل فرمان پیش از اشباع. در این سناریو، STA-DDPG با مقدار ۳/۲۵ نیوتن‌متر بیشترین تحمل را نشان داده و پس از آن الگوریتم فوق‌پیچشی با مقدار ۳/۱۵ نیوتن‌متر قرار گرفته است. STA-TD3 نیز مقدار ۳ نیوتن‌متر را تحمل کرده که همچنان مقدار قابل قبولی است، اما در برابر ضربه مستقیم وارد بر بدنه، بهترین مقدار مطلق را ندارد. این نتیجه نشان می‌دهد که برتری هر روش به نوع اغتشاش وابسته است: STA و STA-DDPG در برابر ضربه مستقیم وارد بر دینامیک بدنه عملکرد رقابتی‌تری دارند، در حالی که STA-TD3 در اغتشاش فرمان گشتاور و نویز سفید فرمانی مقاوم‌تر است.

در مجموع، اگر معیار اصلی فقط کمینه‌سازی خطا باشد، الگوریتم فوق‌پیچشی گزینه دقیق‌تری است؛ و اگر معیار فقط تحمل ضربه مستقیم وارد بر بدنه باشد، STA-DDPG بهترین مقدار را ثبت می‌کند. با این حال، از دیدگاه عملکرد کلی، یعنی ترکیب زمان نشست کمتر، تلاش کنترلی پایین‌تر، مقاومت بالاتر در برابر اغتشاش فرمان گشتاور و تحمل مناسب نویز سفید، STA-TD3 متوازن‌ترین رفتار را ارائه می‌دهد. بنابراین، نتایج جدول ۶ نشان می‌دهد که ساختار STA-TD3 بهترین مصالحه میان سرعت پاسخ، هزینه کنترلی و مقاومت در برابر اغتشاشات فرمانی و تصادفی را ایجاد کرده و در عین حال، در برابر اغتشاش ضربه‌ای مستقیم نیز پایداری و عملکرد قابل قبول خود را حفظ می‌کند.

برای روشن‌تر شدن نقش عامل TD3 در ساختار پیشنهادی، یک آزمون تکمیلی تحت اغتشاش فرمان گشتاور انجام شد. در این آزمون، دو ساختار STA و STA–TD3 تحت شرایط اولیه یکسان، قیود عملگر یکسان و اغتشاش فرمان گشتاور مشترک به مدت ۱۰۰ ثانیه ارزیابی شدند. اغتشاش اعمال‌شده به‌صورت $\tau_{\text{command disturbance}}(t) = 0.17(1+\sin(t/100))\begin{bmatrix}1 & 1 & 1\end{bmatrix}^T$ تعریف شد. هدف از این آزمون، تفکیک نقش مؤلفه TD3 از مؤلفه الگوریتم فوق‌پیچشی و بررسی اثر آن بر پاسخ حالت، سرعت زاویه‌ای، تلاش کنترلی، خطای اشباع نهایی و تکانه زاویه‌ای چرخ‌ها است.

برای کمی‌سازی نتایج شکل ۶، از شاخص ریشهٔ میانگین مربعات[1] استفاده شده است. برای یک سیگنال سه‌محوره y، مقدار RMS در بازه زمانی شبیه‌سازی به‌صورت زیر محاسبه می‌شود:

$$RMS(y) = \sqrt{\frac{1}{3N}\sum_{k=1}^{N}\sum_{i=1}^{3} y_i^2(k)} \qquad (۱)$$

که در آن N تعداد نمونه‌های زمانی، i شماره محور و $y_i(k)$ مقدار مؤلفه i-ام سیگنال در نمونه زمانی k است. همچنین، برای ارزیابی میزان انرژی فرمان کنترلی، شاخص تلاش کنترلی به‌صورت زیر تعریف می‌شود:

$$E_\tau = \int_0^T \tau_a(t)^2\, dt \qquad (۲)$$

که در آن $\tau_a(t)$ گشتاور واقعی اعمال‌شده به سامانه پس از اشباع نهایی است.

مطابق شکل ۷، هر دو ساختار STA و STA–TD3 قادر به پایدارسازی وضعیت و بازگرداندن زوایا به حوالی صفر هستند. با این حال، اثر TD3 در پاسخ زاویه‌ای بیشتر به‌صورت کاهش محدود در مقدار RMS ظاهر شده است. مقدار RMS زوایا از ۱۷/۶۵۹۹

[1] Root Mean Square (RMS)

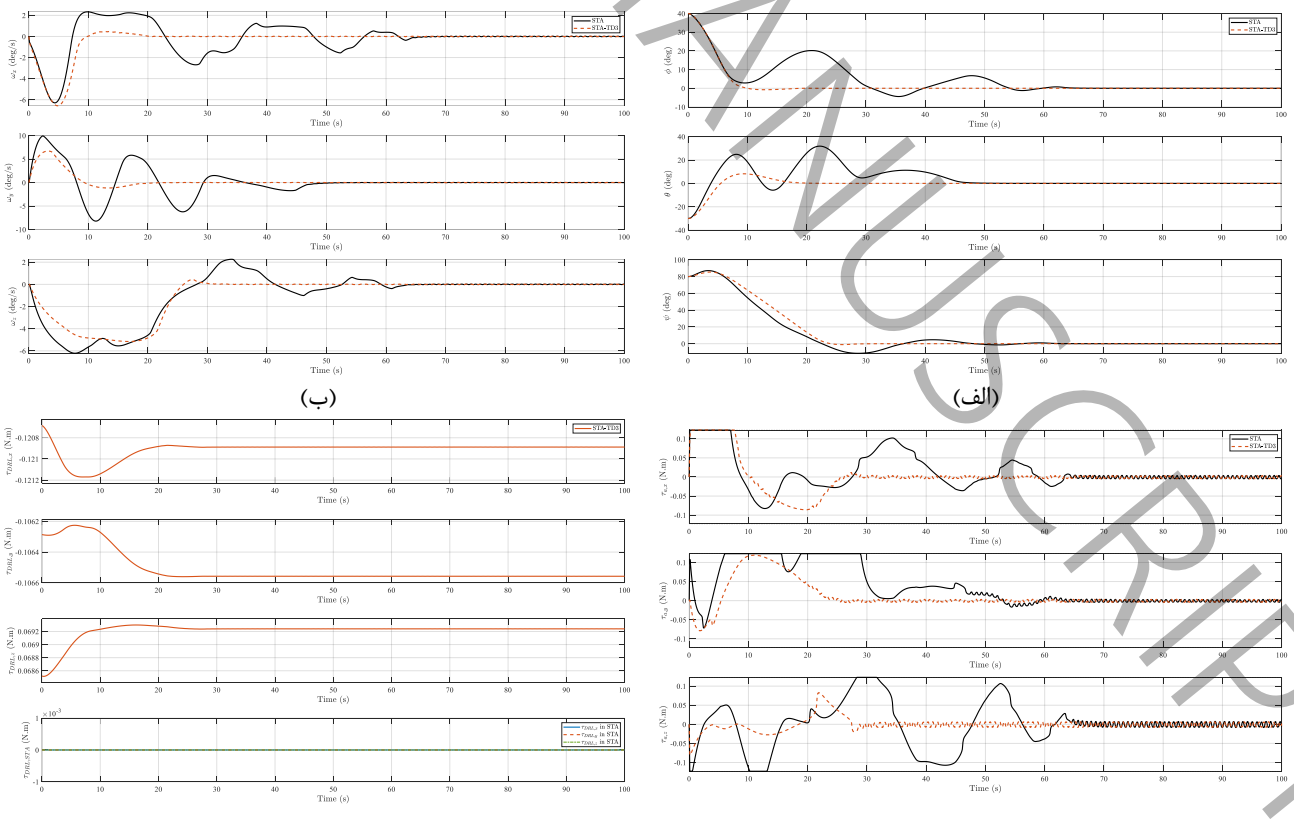
درجه در الگوریتم فوق‌پیچشی به ۱۷/۲۱۶۲ درجه در STA–TD3 کاهش یافته است. این کاهش نشان می‌دهد که اثر TD3 بر دامنه کلی خطای زاویه‌ای مثبت است، اما به دلیل بزرگی شرایط اولیه، بهبود در شاخص زاویه‌ای محدود باقی می‌ماند.

اثر TD3 در سرعت‌های زاویه‌ای آشکارتر است. مقدار RMS سرعت زاویه‌ای از ۲/۳۹۵۷ درجه بر ثانیه در STA به ۱/۶۲۶۷ درجه بر ثانیه در STA–TD3 کاهش یافته است که معادل ٪۳۲/۱۰ کاهش است. این نتیجه نشان می‌دهد که TD3 بیش از آنکه صرفاً مقدار بیشینه زاویه را کاهش دهد، رفتار گذرای سامانه را آرام‌تر کرده و دامنه نوسان‌های سرعت زاویه‌ای را کاهش داده است. از این جهت، نقش TD3 در کاهش نوسانات دینامیکی و بهبود میرایی پاسخ حلقه‌بسته قابل توجه است.

از نظر گشتاور اعمال‌شده، مقدار بیشینه گشتاور در هر دو روش برابر ۰/۱۲۳ نیوتن‌متر است؛ زیرا فرمان نهایی توسط محدودیت عملگر اشباع می‌شود. بنابراین، کاهش مقدار بیشینه گشتاور شاخص مناسبی برای مقایسه دو روش نیست. در مقابل، مقدار RMS گشتاور اعمال‌شده از ۰/۰۵۵۳ نیوتن‌متر در STA به ۰/۰۳۴۰ نیوتن‌متر در STA–TD3 کاهش یافته است. همچنین شاخص تلاش گشتاور اعمال‌شده از ۰/۹۱۹۴ به ۰/۳۴۷۴کاهش یافته است. این اعداد نشان می‌دهند که ساختار STA–TD3 با وجود رسیدن هر دو روش به حد اشباع در برخی لحظات، در مجموع فرمانی کم‌انرژی‌تر و مؤثرتر به سامانه اعمال می‌کند.

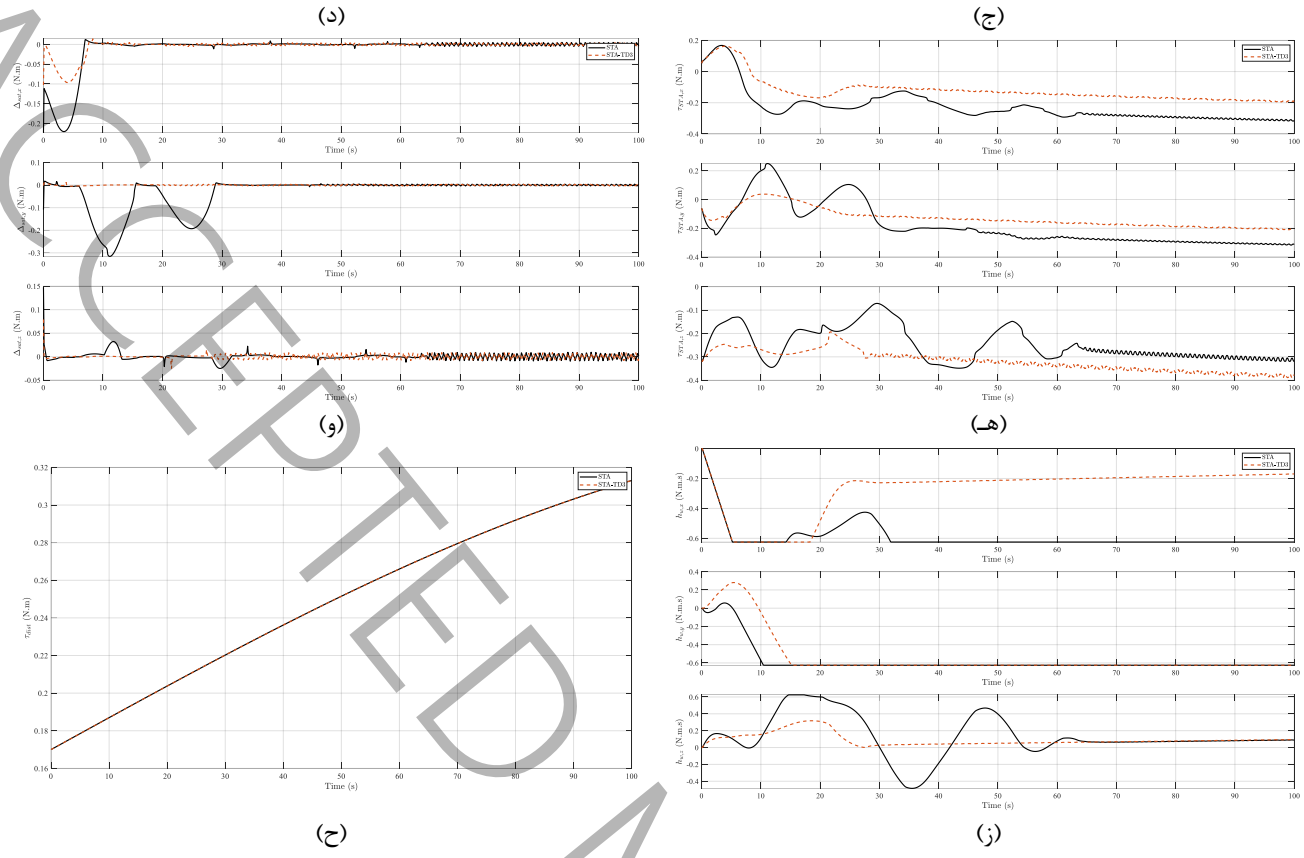
گشتاور اصلاحی یادگیری تقویتی عمیق فقط در ساختار STA–TD3 فعال است و در حالت STA مقدار آن صفر است. این موضوع نشان می‌دهد که عامل TD3 به‌صورت یک مؤلفه مستقل جایگزین STA نشده، بلکه به‌عنوان یک ترم اصلاحی کراندار به خروجی کنترل‌کننده افزوده شده است. مقدار RMS این مؤلفه در آزمون حاضر ۰/۱۰۱۳ نیوتن‌متر است. بنابراین، نقش TD3 را باید به‌عنوان اصلاح عملکرد فرمان نهایی تفسیر کرد، نه به‌عنوان جایگزین کنترل‌کننده مقاوم یا تضمین‌کننده اصلی پایداری.

بررسی مؤلفه گشتاور STA نیز این برداشت را تأیید می‌کند. مقدار RMS گشتاور STA از ۰/۲۴۷۲ نیوتن‌متر در ساختار STA به ۰/۲۲۰۴ نیوتن‌متر در ساختار STA–TD3 کاهش یافته است. همچنین تلاش مؤلفه STA از ۱۸/۳۳۳۷ به ۱۴/۵۷۰۰ کاهش یافته است. بنابراین، اگرچه ممکن است در برخی لحظات مقدار بیشینه گشتاور STA در ساختار ترکیبی افزایش یابد، اما در مقیاس متوسط و انرژی، بار وارد بر مؤلفه STA کاهش پیدا کرده است. این نکته نشان می‌دهد که TD3 بخشی از بار عملکردی کنترل‌کننده مقاوم را کاهش می‌دهد، بدون آنکه نقش پایدارکننده STA را حذف کند.



(الف)

(ب)



شکل ۷ . تحلیل تفکیکی عملکرد STA و STA–TD3 تحت اغتشاش فرمان گشتاور: (الف) زوایا، (ب) سرعت‌های زاویه‌ای، (ج) گشتاور اعمال‌شده پس از اشباع، (د) گشتاور DRL، (هـ) گشتاور STA، (و) خطای اشباع، (ز) تکانه زاویه‌ای چرخ‌ها، (ح) اغتشاش اعمال‌شده

Fig. 7. Disaggregated performance analysis of STA and STA–TD3 under torque command disturbance: (a) angles, (b) angular velocities, (c) applied torque after saturation, (d) DRL torque, (e) STA torque, (f) saturation error, (g) wheel angular momentum, (h) applied disturbance

یکی از نتایج مهم آزمون، کاهش خطای ناشی از اشباع نهایی است. مقدار RMS خطای اشباع از ۰/۰۵۱۸ نیوتن‌متر در STA به ۰/۰۱۱۱ نیوتن‌متر در STA–TD3 کاهش یافته است. این کاهش معادل ٪۷۸/۴۹ است و نشان می‌دهد که ساختار ترکیبی اختلاف کمتری میان گشتاور پیش از اشباع و گشتاور واقعی اعمال‌شده ایجاد می‌کند. بنابراین، TD3 باعث می‌شود فرمان نهایی با محدودیت عملگر سازگارتر شود و اثر عملی اشباع کاهش یابد.

از نظر تکانه زاویه‌ای چرخ‌ها نیز مقدار RMS از ۰/۵۱۳۸ نیوتن‌مترثانیه در STA به ۰/۳۸۹۴ نیوتن‌مترثانیه در STA–TD3 کاهش یافته است. این کاهش ٪۲۴/۲۲ نشان می‌دهد که ساختار ترکیبی، در مقیاس متوسط، بار تکانه‌ای چرخ‌ها را کاهش می‌دهد. هرچند بیشینه تکانه چرخ‌ها در هر دو روش به مقدار مشابهی می‌رسد، شاخص RMS نشان می‌دهد که TD3 باعث توزیع آرام‌تر و کم‌بارتر تکانه در طول زمان شده است.

در نهایت، اغتشاش اعمال‌شده در هر دو آزمون کاملاً یکسان است و مقدار بیشینه آن برابر ۰/۳۱۳۱ نیوتن‌متر است. بنابراین، تفاوت‌های مشاهده‌شده ناشی از ساختار کنترل‌کننده است و نه تفاوت در ورودی اغتشاشی. بر اساس این نتایج، TD3 در ساختار پیشنهادی به‌عنوان یک مؤلفه اصلاحی عملکردمحور عمل می‌کند. اثر اصلی آن در کاهش RMS سرعت زاویه‌ای، کاهش RMS و تلاش گشتاور اعمال‌شده، کاهش تلاش مؤلفه STA، کاهش خطای اشباع نهایی و کاهش RMS تکانه زاویه‌ای چرخ‌ها ظاهر می‌شود؛ در حالی که نقش پایدارکننده اصلی همچنان توسط STA حفظ می‌گردد.

جدول ۷ . شاخص‌های اصلی مقایسه STA و STA–TD3 تحت اغتشاش فرمان گشتاور

Table 7. Key performance indices comparing STA and STA–TD3 under torque command disturbance

شاخص اصلی	واحد	STA	STA–TD3	نتیجه
RMS زوایا	درجه	۱۷/۶۵۹۹	۱۷/۲۱۶۲	۲/۵۱ ٪ کاهش
RMS سرعت‌های زاویه‌ای	درجه بر ثانیه	۲/۳۹۵۷	۱/۶۲۶۷	۳۲/۱۰ ٪ کاهش
RMS گشتاور اعمال‌شده پس از اشباع	نیوتن‌متر	۰/۰۵۵۳	۰/۰۳۴۰	۳۸/۵۳ ٪ کاهش
تلاش گشتاور اعمال‌شده پس از اشباع	—	۰/۹۱۹۴	۰/۳۴۷۴	۶۲/۲۲ ٪ کاهش
RMS گشتاور STA	نیوتن‌متر	۰/۲۴۷۲	۰/۲۲۰۴	۱۰/۸۵ ٪ کاهش
تلاش گشتاور STA	—	۱۸/۳۳۳۷	۱۴/۵۷۰۰	۲۰/۵۳ ٪ کاهش
RMS گشتاور DRL	نیوتن‌متر	۰	۰/۱۰۱۳	فقط در STA–TD3 فعال است
RMS خطای اشباع نهایی	نیوتن‌متر	۰/۰۵۱۸	۰/۰۱۱۱	۷۸/۴۹ ٪ کاهش
RMS تکانه زاویه‌ای چرخ‌ها	نیوتن‌مترثانیه	۰/۵۱۳۸	۰/۳۸۹۴	۲۴/۲۲ ٪ کاهش
بیشینه اغتشاش اعمال‌شده	نیوتن‌متر	۰/۳۱۳۱	۰/۳۱۳۱	یکسان در دو آزمون

۴-۳- نتایج پیاده‌سازی

در ابتدای این بخش، به‌منظور یکسان‌سازی شرایط آزمون‌ها و امکان بازتولید نتایج پیاده‌سازی، پارامترهای مدل ماهواره و محدودیت‌های عملگرها که در تمام سناریوها ثابت در نظر گرفته شده‌اند، در جدول ۸ خلاصه شده است. همچنین تنظیمات اجرای پیاده‌سازی شامل نرخ حلقه کنترل، تأخیر ارتباطی و مشخصات واحد حسگر و بستر نرم‌افزاری به‌کاررفته برای اجرای عامل‌های یادگیری تقویتی عمیق، در جدول ۹ ارائه می‌شود. ادامه‌ی این بخش به ارائه نمودارهای زوایا، سرعت زاویه‌ای، ورودی کنترلی و مومنتوم زاویه‌ای در شرایط نامی و اغتشاشی اختصاص دارد.

در این پژوهش، سیاست‌های یادگیری تقویتی عمیق پس از آموزش در محیط شبیه‌سازی، بدون به‌روزرسانی برخط وزن‌ها، روی سکوی آزمایشگاهی اجرا شده‌اند. این رویکرد به‌دلیل محدودیت‌های عملی بستر آزمایش، از جمله زمان محدود شارژ باتری، فرسایش عملگرها و ریسک تولید فرمان‌های نامطلوب در مرحله اکتشاف، برای ارزیابی ایمن‌تر سیاست‌های آموزش‌دیده انتخاب شده است. از سوی دیگر، نتایج شبیه‌سازی و پیاده‌سازی به‌صورت جداگانه ارائه شده‌اند تا خوانایی نمودارهای چندمحوره حفظ شود. در ارزیابی عملی، اثراتی مانند نویز حسگر، تأخیر ارتباطی، کابل‌های متصل به سکو، خطاهای کوچک عملگر و عدم‌تعادل‌های مکانیکی به‌صورت طبیعی حضور دارند؛ ازاین‌رو، نتایج این بخش ارزیابی تجربی عملکرد کنترل‌کننده‌ها در شرایط واقعی‌تر از شبیه‌سازی ایده‌آل محسوب می‌شود.

در شکل ۸، مشاهده می‌شود که زوایا در هر چهار راهبرد کنترلی به مقدار مرجع همگرا شده و پس از گذرای اولیه، پایداری وضعیت برقرار می‌گردد. در نمودارهای سرعت زاویه‌ای نیز نرخ‌ها به حوالی صفر می‌رسند که بیانگر میرایی مناسب و حذف حرکت دورانی ناخواسته است. با وجود نویزها و اغتشاشات ذاتی بستر آزمایش، کنترل‌کننده‌ها توانسته‌اند خطای ماندگار را کوچک نگه دارند و همگرایی را حفظ کنند. در ورودی کنترلی، به‌ویژه در ناحیه ماندگار، لرزش مشاهده می‌شود که هم به ساختار STA و هم به محدودیت‌ها و گسستگی‌های اجرای دیجیتال مرتبط است. در ساختارهای ترکیبی نیز ترم افزوده‌شده توسط عامل‌های یادگیری تقویتی عمیق فرمان نهایی را تعدیل کرده و از نظر تلاش کنترلی، روش‌های STA-DDPG و به‌ویژه STA-TD3 نسبت به STA تنها ورودی ملایم‌تر و کم‌هزینه‌تری در طول اجرا تولید می‌کنند. همچنین نمودارهای مومنتوم زاویه‌ای نشان می‌دهند چرخ‌ها در ابتدای مانور تغییر اولیه لازم برای انتقال سیستم به ناحیه مطلوب را ایجاد کرده و پس از استقرار، با ریزتنظیم‌های کوچک برای جبران اغتشاشات ذاتی، وضعیت پایدار را حفظ می‌کنند.

جدول ۸ . پارامترهای مدل شبیه‌سازی و مشخصات بدنه ماهواره

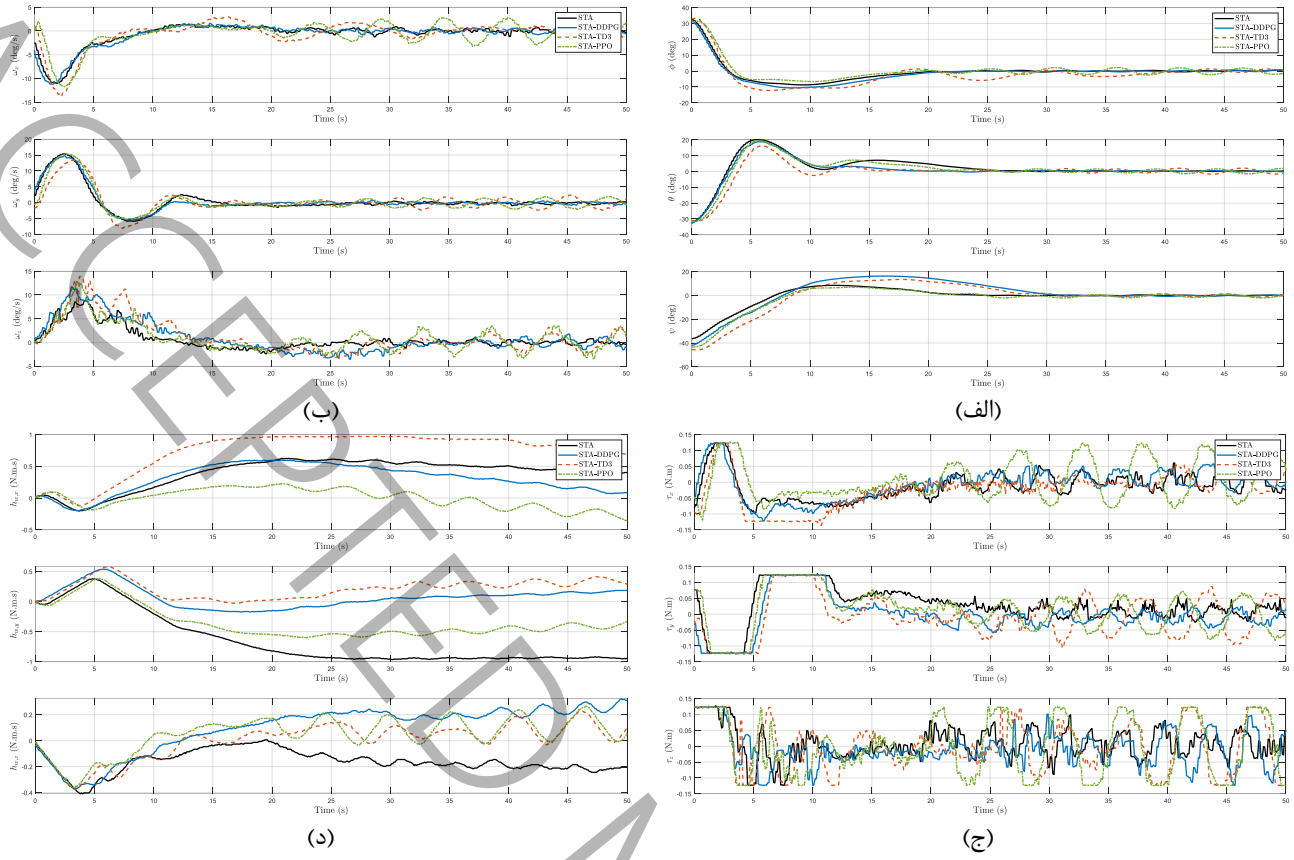
Table 8. Simulation model parameters and satellite body specifications

پارامتر	مقدار
زمان نمونه‌برداری	0.1 s
مدت زمان شبیه‌سازی	100 s
سید تصادفی	42
جرم (m)	40 kg
شتاب گرانش (g)	9.81 m/s²
گشتاور اغتشاش $\left(mgr_x, mgr_y, mgr_z\right)$	$[0, 0, 0.326]\mathrm{N}\cdot\mathrm{m}$
ماتریس ممان اینرسی (J)	$J = \begin{bmatrix} 1.8140 & -0.1185 & 0.0275 \\ -0.1185 & 1.7350 & 0.0169 \\ 0.0275 & 0.0169 & 3.4320 \end{bmatrix} kg\cdot\mathrm{m}^2$
ممان اینرسی چرخ‌های عکس‌العمل $\left(I_w\right)$	$I_w = \begin{bmatrix} 0.0027 & 0 & 0 \\ 0 & 0.0027 & 0 \\ 0 & 0 & 0.0027 \end{bmatrix} kg\cdot\mathrm{m}^2$
بازه زوایای اویلر	$\phi, \theta \in \left[-45^\circ, 45^\circ\right], \psi \in \left[-180^\circ, 180^\circ\right]$
PSD زوایای اویلر	$3.0625\times10^{-8}\, rad^2 / Hz$
PSD سرعت زاویه‌ای	$3.23\times10^{-7}\,(rad/\mathrm{s})^2 / Hz$
حد گشتاور کنترلی	$\pm0.123\ \mathrm{N}\cdot\mathrm{m}$
محدوده گشتاور خروجی DRL	$\left\|\tau_{DRL}\right\| \le 0.123$
محدوده گشتاور خروجی STA	$\left\|\tau_{STA}\right\| \le 0.123$
حد تکانه زاویه‌ای	$\pm0.625\ \mathrm{N}\cdot\mathrm{m}\cdot\mathrm{s}$

جدول ۹ . پارامترهای پیاده‌سازی

Table 9. Implementation parameters

مؤلفه	مشخصات
سخت‌افزار هدف	سکوی شبیه‌ساز ماهواره به‌همراه حسگرهای AHRS
نرخ حلقه کنترل	10 Hz
تأخیر ارتباطی	< 50 ms
واحد حسگر	AHRS مبتنی بر MEMS (زاویه اویلر + نرخ زاویه‌ای)
دقت زاویه‌ای ایستا	$\le 0.5°$
دقت زاویه‌ای پویا	$\le 2°$
مدل نویز حسگر	نویز سفید + بایاس (بر اساس PSD)
نرم‌افزار رابط	MATLAB/Simulink + LabVIEW (یکپارچه‌سازی UDP/TCP)
چارچوب اجرای RL	بلوک MATLAB RL Agent (در Simulink)



شکل ۸ . نتایج پیاده‌سازی مقایسه‌ای چهار روش کنترلی در حضور اغتشاشات ذاتی سکوی آزمایش: (الف) زوایا، (ب) سرعت زاویه‌ای، (ج) ورودی کنترلی، (د) مومنتوم زاویه‌ای

Fig. 8. Comparative implementation results of four control methods in the presence of inherent disturbances of the test platform: (a) angles, (b) angular velocity, © control input, (d) angular momentum

۵- نتیجه‌گیری و پیشنهادات

در این پژوهش، یک چارچوب کنترل ترکیبی برای کنترل وضعیت شبیه‌ساز ماهواره سه‌درجه‌آزادی ارائه شد که در آن، الگوریتم فوق‌پیچشی به‌عنوان هسته مقاوم و پایدارکننده حفظ شده و عامل‌های DDPG، TD3 و PPO به‌صورت مؤلفه اصلاحی کراندار در کنار آن قرار گرفتند. در این ساختار، یادگیری تقویتی عمیق جایگزین الگوریتم فوق‌پیچشی نمی‌شود، بلکه با اصلاح فرمان نهایی، کاهش تلاش کنترلی و بهبود رفتار گذرا، عملکرد عملی کنترل‌کننده را ارتقا می‌دهد. همچنین، فراپارامترهای عامل‌های یادگیری تقویتی عمیق با روش تاگوچی و آرایه متعامد L_{27} تنظیم شدند تا فرآیند انتخاب پارامترها به‌صورت نظام‌مند، تکرارپذیر و کم‌هزینه‌تر انجام شود.

مهم‌ترین یافته‌های کمی پژوهش عبارت‌اند از:

- در مقایسه کلی چهار روش، مقدار MSE در همه کنترل‌کننده‌ها در مرتبه 10^{-8} باقی ماند که نشان‌دهنده حفظ دقت رهگیری در تمام ساختارها است. با این حال، الگوریتم فوق‌پیچشی کمترین مقادیر MSE، ISE و ITSE را ثبت کرد و از نظر کمینه‌سازی خطا دقیق‌ترین پاسخ را ارائه داد.
- روش STA-TD3 بهترین توازن میان سرعت پاسخ و تلاش کنترلی را نشان داد. زمان نشست از ۲۶ ثانیه در STA به ۲۲ ثانیه در STA-TD3 کاهش یافت و تلاش کنترلی نیز از ۲۴/۲ به ۲۳/۱۰ رسید؛ یعنی به‌ترتیب حدود ۱۵/۴ درصد و ۴/۵ درصد بهبود نسبت به STA.

- در مقایسه مستقیم STA و STA-TD3 تحت اغتشاش فرمان گشتاور یکسان، RMS سرعت‌های زاویه‌ای از ۲/۳۹۵۷ به ۱/۶۲۶۷ درجه بر ثانیه کاهش یافت که معادل ۳۲/۱۰ درصد کاهش است. همچنین RMS گشتاور اعمال‌شده پس از اشباع از ۰/۰۵۵۳ به ۰/۰۳۴۰ نیوتن‌متر رسید که نشان‌دهنده ۳۸/۵۳ درصد کاهش است.
- خطای اشباع نهایی از ۰/۰۵۱۸ به ۰/۰۱۱۱ نیوتن‌متر کاهش یافت که معادل ۷۸/۴۹ درصد کاهش است. همچنین RMS تکانه زاویه‌ای چرخ‌ها از ۰/۵۱۳۸ به ۰/۳۸۹۴ نیوتن‌مترثانیه رسید که بیانگر ۲۴/۲۲ درصد کاهش بار مومنتومی چرخ‌ها است.
- از نظر مقاومت در برابر اغتشاش فرمان گشتاور، STA-TD3 بیشترین سطح قابل تحمل را با ۰/۲۵۵ نیوتن‌متر ثبت کرد؛ در حالی که این مقدار برای STA برابر ۰/۱۷ نیوتن‌متر بود. در سناریوی نویز سفید فرمان گشتاور نیز STA-DDPG و STA-TD3 هر دو توان نویز ۰/۰۰۳ را تحمل کردند، در حالی که مقدار متناظر برای STA برابر ۰/۰۰۱ بود.

بر اساس این نتایج، STA از نظر کمینه‌سازی خطا همچنان دقیق‌ترین گزینه است، اما STA-TD3 از نظر موازنه کلی میان سرعت همگرایی، کاهش تلاش کنترلی، کاهش اثر اشباع، کاهش مومنتوم چرخ‌ها و افزایش مقاومت در برابر اغتشاشات فرمانی و تصادفی عملکرد مناسب‌تری ارائه می‌دهد. نتایج PPO نیز نشان داد که افزودن یک عامل یادگیری تقویتی عمیق به‌تنهایی تضمین‌کننده بهبود عملکرد نیست و سازگاری الگوریتم، فراپارامترها و تابع پاداش با دینامیک مسئله نقش تعیین‌کننده دارد.

محدودیت‌های اصلی پژوهش عبارت‌اند از:

- عامل‌های یادگیری تقویتی عمیق به‌صورت برون‌خط و در محیط شبیه‌سازی آموزش داده شدند و در مرحله پیاده‌سازی، وزن‌های شبکه به‌صورت برخط به‌روزرسانی نشدند.
- عملکرد یادگیری تقویتی عمیق به ساختار تابع پاداش، فراپارامترهای آموزشی، کیفیت مدل شبیه‌سازی و شرایط اولیه آزمون وابسته است؛ بنابراین تعمیم نتایج به سامانه‌های دیگر نیازمند تنظیم و ارزیابی مجدد است.
- نتایج آموزشی و عملکردی عامل‌های یادگیری تقویتی عمیق بر اساس یک مقدار سید ثابت گزارش شده‌اند. اگرچه این انتخاب به بازتولیدپذیری آزمایش‌ها کمک می‌کند، اما ارزیابی چندبذری و گزارش میانگین و انحراف معیار عملکرد می‌تواند در مطالعات آینده برای تقویت اعتبار آماری نتایج تکمیل شود.
- همه اثرات سخت‌افزاری، از جمله نویزهای پیچیده حسگر، اصطکاک باقیمانده، اثر کابل‌ها، عدم‌تعادل‌های مکانیکی و تأخیرهای اجرایی، به‌صورت کامل و هم‌زمان در شبیه‌سازی مدل نشده‌اند؛ بخشی از این اثرات در پیاده‌سازی عملی به‌صورت طبیعی حضور داشته‌اند.
- نرخ اجرای حلقه کنترل در بستر آزمایشگاهی ۱۰ هرتز بوده است؛ بنابراین ارزیابی همین چارچوب روی سخت‌افزار سریع‌تر یا نرخ نمونه‌برداری بالاتر می‌تواند از نظر هموای فرمان و شدت لرزش نتایج متفاوتی ایجاد کند.

برای ادامه‌ی پژوهش پیشنهاد می‌شود:

- کاهش لرزش با اعمال قیود عملگر و جریمه‌کردن در تابع پاداش، و هموارسازی فرمان (فیلتر/محدودکننده نرخ تغییر/پیوسته‌سازی الگوریتم فوق‌پیچشی).
- مدلسازی دقیق‌تر عملگرها در شبیه‌سازی و آموزش: اشباع، تأخیر، اصطکاک و مدیریت مومنتوم.
- تعمیم چارچوب به سناریوهای تحمل‌خطا: افت کارایی/خرابی چرخ‌ها، بازپیکربندی و استفاده از چرخ افزونه.
- استفاده از نمایش کواترنیون به‌جای اویلر برای حذف تکینگی و بهبود پایداری عددی در مانورهای بزرگ.
- بررسی حساسیت ضرایب تابع پاداش و تنظیم اختصاصی آن‌ها برای هر الگوریتم به‌منظور بهبود دقت رهگیری و کاهش تلاش کنترلی.

Robust Attitude Control of a Three-Degree-of-Freedom Satellite via Integration of the Super-Twisting Algorithm and Deep Reinforcement Learning with Hyperparameter Tuning Using Taguchi Design of Experiments

Mostafa Sarjoughian, Hojat Taei*

Department of Aerospace Engineering, Faculty of Engineering, University of Isfahan, Isfahan, Iran

* h.taei@eng.ui.ac.ir

ABSTRACT

This study presents a hybrid control framework for the attitude regulation of a three-degree-of-freedom satellite subject to parametric uncertainties, external disturbances, actuator constraints, and implementation imperfections. The core robust controller is formulated using the Super-Twisting Algorithm, which guarantees finite-time convergence and robustness while effectively suppressing the high-frequency chattering typically associated with conventional sliding mode control. To enhance tracking precision and improve adaptability under nonlinear and uncertain conditions, deep reinforcement learning is incorporated as an adaptive compensator within the control loop. Three representative algorithms, namely Deep Deterministic Policy Gradient, Twin Delayed Deep Deterministic Policy Gradient, and Proximal Policy Optimization, are investigated and comparatively evaluated in terms of stability, convergence behavior, and control efficiency. To systematically tune the learning hyperparameters and reduce the computational burden associated with manual trial-and-error procedures, the Taguchi design of experiments method is employed to perform multi-objective optimization considering both tracking performance and control effort. The performance index is defined as a composite measure that combines time-weighted tracking error and control energy. Numerical simulations together with experimental validation on a satellite attitude simulator demonstrate that the proposed hybrid control architecture reduces settling time and control effort while improving disturbance rejection capability, without compromising stability or steady-state tracking accuracy.

KEYWORDS

Satellite Attitude Control; Super-Twisting Algorithm; Deep Reinforcement Learning; Taguchi Design of Experiments.

۶- مراجع

[1] K. Lu, Y. Xia, Finite-time attitude control for rigid spacecraft-based on adaptive super-twisting algorithm, IET Control Theory & Applications, 8(15) (2014) 1465-1477.

[2] Y. Su, S. Shen, Adaptive predefined-time fault-tolerant attitude tracking control for rigid spacecraft with guaranteed performance, Acta Astronautica, 214 (2024) 677-688.

[3] C. Xiao, Y. Guo, C.-q. Xie, A.-j. Li, C.-q. Wang, Adaptive super-twisting sliding mode attitude coordination control for spacecraft formation flying with actuator saturation, Advances in Space Research, 72(10) (2023) 4244-4255.

[4] M. Khodaverdian, M. Malekzadeh, Fault-tolerant model predictive sliding mode control with fixed-time attitude stabilization and vibration suppression of flexible spacecraft, Aerospace Science and Technology, 139 (2023) 108381.

[5] X.-N. Shi, W. Chen, R. Li, Z.-G. Zhou, K. Wen, Prescribed Performance Attitude Tracking Control for Spacecraft under Multi-Constraint, in: 2020 39th Chinese Control Conference (CCC), IEEE, 2020, pp. 270-275.

[6] M. Tipaldi, R. Iervolino, P.R. Massenio, Reinforcement learning in spacecraft control applications: Advances, prospects, and challenges, Annual Reviews in Control, 54 (2022) 1-23.

[7] W. Retagne, J. Dauer, G. Waxenegger-Wilfing, Adaptive satellite attitude control for varying masses using deep reinforcement learning, Frontiers in Robotics and AI, 11 (2024) 1402846.

[8] K.-H. Lee, S. Lim, D.-H. Cho, H.-D. Kim, Development of fault detection and identification algorithm using deep learning for nanosatellite attitude control system, International Journal of Aeronautical and Space Sciences, 21(2) (2020) 576-585.

[9] S. Oghim, J. Park, H. Bang, H. Leeghim, Deep reinforcement learning-based attitude control for spacecraft using control moment gyros, Advances in Space Research, 75(1) (2025) 1129-1144.

[10] M. Wu, K. Guo, X. Li, Z. Lin, Y. Wu, T.A. Tsiftsis, H. Song, Deep reinforcement learning-based energy efficiency optimization for RIS-aided integrated satellite-aerial-terrestrial relay networks, IEEE Transactions on Communications, 72(7) (2024) 4163-4178.

[11] N.A. Mosali, S.S. Shamsudin, O. Alfandi, R. Omar, N. Al-Fadhali, Twin delayed deep deterministic policy gradient-based target tracking for unmanned aerial vehicle with achievement rewarding and multistage training, IEEE Access, 10 (2022) 23545-23559.

[12] M. Ran, J. Li, L. Xie, Reinforcement-learning-based disturbance rejection control for uncertain nonlinear systems, IEEE Transactions on Cybernetics, 52(9) (2021) 9621-9633.

[13] X. Zhang, X. Chen, L. Yao, C. Ge, M. Dong, Deep neural network hyperparameter optimization with orthogonal array tuning, in: International conference on neural information processing, Springer, 2019, pp. 287-295.

[14] J.B. Chandar, M. Sivakumar, N. Lenin, R. Čep, S. Salunkhe, E.A. Nasr, C. Rathinasuriyan, A novel predictive model for abrasive waterjet deep hole drilling on AL7075 T6 using machine learning and evolutionary algorithmic approach, Scientific Reports, 15(1) (2025) 43951.

[15] N. Ranković, D. Ranković, GOAT method: Green Orthogonal Array Tuning method, Alexandria Engineering Journal, 133 (2025) 13-41.

[16] P. Arevalo, A. Cano, O. Fedoseienko, F. Jurado, A data-driven approach to microgrid fault detection and classification using Taguchi-optimized CNNs and wavelet transform, Applied Soft Computing, 170 (2025) 112667.

[17] L. Grbcic, M. Park, J. Müller, V. Zorba, W.A. de Jong, Artificial intelligence driven laser parameter search: Inverse design of photonic surfaces using greedy surrogate-based optimization, Engineering Applications of Artificial Intelligence, 143 (2025) 109971.

[18] C.-J. Lin, S.-Y. Jeng, C.-L. Lee, Hyperparameter Optimization of Deep Learning Networks for Classification of Breast Histopathology Images, Sensors & Materials, 33 (2021).
[19] M. Sarjoughian, M. Malekzadeh, N. Sayyaf, Hybrid Control of Spacecraft: Super-Twisting Algorithm Based on Taguchi-Driven Deep Reinforcement Learning, Results in Engineering, (2026) 110530.
[20] S. Jamshidi, M. Mirzaei, M. Malekzadeh, Applied optimal control of spacecraft simulator subject to failures of reaction wheels, Arabian Journal for Science and Engineering, 49(2) (2024) 1697-1712.
[21] R.S. Sutton, A.G. Barto, Reinforcement learning: An introduction, MIT press Cambridge, 1998.
[22] S. Fujimoto, H. Hoof, D. Meger, Addressing function approximation error in actor-critic methods, in: International conference on machine learning, PMLR, 2018, pp. 1587-1596.
[23] A. Surriani, O. Wahyunggoro, A trajectory control for bipedal walking robot using stochastic-based continuous deep reinforcement learning, (2023).
[24] S. Qi, L. Lu, F. Ziruo, B. Xingzi, L. Huaqiu, C. Wen, Y. Jinpei, Efficient and fair PPO-based integrated scheduling method for multiple tasks of SATech-01 satellite, Chinese Journal of Aeronautics, 37(2) (2024) 417-430.