دانشگاه صنعتی شریف
دانشکده‌ی مهندسی هوافضا

پروژه کارشناسی ارشد
مهندسی فضا

عنوان:

هدایت یادگیری تقویتی مقاوم مبتنی بر بازی دیفرانسیلی در محیط‌های پویای چندجسمی با پیشران کم

نگارش:

علی بنی اسد

استاد راهنما:

دکتر هادی نوبهاری

دی ۱۴۰۳


بسم الله الرحمن الرحیم

به نام خدا

دانشگاه صنعتی شریف

دانشکده‌ی مهندسی هوافضا

پروژه کارشناسی ارشد

عنوان: هدایت یادگیری تقویتی مقاوم مبتنی بر بازی دیفرانسیلی در محیط‌های پویای چندجسمی با پیشران کم

نگارش: علی بنی اسد

کمیته‌ی ممتحنین

استاد راهنما: دکتر هادی نوبهاری امضاء:

استاد مشاور: استاد مشاور امضاء:

استاد مدعو: استاد ممتحن امضاء:

تاریخ:

سپاس

از استاد بزرگوارم جناب آقای دکتر نوبهاری که با کمک‌ها و راهنمایی‌های بی‌دریغشان، بنده را در انجام این پروژه یاری داده‌اند، تشکر و قدردانی می‌کنم. از پدر دلسوزم ممنونم که در انجام این پروژه مرا یاری نمود. در نهایت در کمال تواضع، با تمام وجود بر دستان مادرم بوسه می‌زنم که اگر حمایت بی‌دریغش، نگاه مهربانش و دستان گرمش نبود برگ برگ این دست نوشته و پروژه وجود نداشت.

چکیده

در این پژوهش، یک چارچوب هدایت مقاوم برای فضاپیماهای کم‌پیشران در محیط‌های دینامیکی چندجسمی (مدل CRTBP زمین–ماه) ارائه شده است. مسئله به‌صورت بازی دیفرانسیلی مجموع‌صفر بین عامل هدایت (فضاپیما) و عامل مزاحم (عدم قطعیت‌های محیطی) فرمول‌بندی شده و با رویکرد آموزش متمرکز–اجرای توزیع‌شده پیاده‌سازی گردیده است. در این راستا، چهار الگوریتم یادگیری تقویتی پیوسته DDPG، TD3، SAC و PPO به نسخه‌های چندعاملی مجموع‌صفر گسترش یافته‌اند (MA-DDPG، MATD3، MASAC و MAPPO) و جریان آموزش آن‌ها همراه با ساختار شبکه‌ها در قالب ارزش–سیاست مشترک تشریح شده است.

ارزیابی الگوریتم‌ها در سناریوهای متنوع عدم قطعیت شامل شرایط اولیه تصادفی، اغتشاش عملگر، نویز حسگر، تأخیر زمانی و عدم تطابق مدل روی مسیر مدار لیاپانوف زمین–ماه انجام گرفت. نتایج به‌وضوح نشان می‌دهد که نسخه‌های مجموع‌صفر در تمامی معیارهای ارزیابی بر نسخه‌های تک‌عاملی برتری دارند. به‌ویژه الگوریتم MATD3 با حفظ پایداری سیستم، کمترین انحراف مسیر و مصرف سوخت بهینه را حتی در سخت‌ترین سناریوهای آزمون از خود نشان داد.

به‌منظور تسهیل استقرار عملی، سیاست‌های آموخته‌شده روی بستر ROS 2 با بهره‌گیری از کوانتیزاسیون INT8 و تبدیل به فرمت ONNX پیاده‌سازی شدند. این بهینه‌سازی‌ها زمان استنتاج را به ۵٫۸ میلی‌ثانیه و مصرف حافظه را به ۹٫۲ مگابایت کاهش داد که به‌ترتیب بهبود ۴۷ درصدی و ۵۳ درصدی نسبت به مدل FP32 را نشان می‌دهد، در حالی‌که چرخه کنترل ۱۰۰ هرتز بدون هیچ‌گونه نقض زمانی حفظ شد.

در مجموع، چارچوب پیشنهادی نشان می‌دهد که یادگیری تقویتی چندعاملی مبتنی بر بازی دیفرانسیلی می‌تواند بدون نیاز به مدل‌سازی دقیق، هدایت تطبیقی و مقاوم فضاپیماهای کم‌پیشران را در نواحی ذاتاً ناپایدار سیستم‌های سه‌جسمی تضمین کند و برای پیاده‌سازی روی سخت‌افزار در حلقه آماده باشد.

کلیدواژه‌ها: یادگیری تقویتی عمیق، بازی دیفرانسیلی، سیستم‌های چندعاملی، هدایت کم‌پیشران، مسئله محدود سه‌جسمی، کنترل مقاوم.

فهرست مطالب

فهرست جداول

فهرست تصاویر

فهرست الگوریتم‌ها

فصل ۱

مقدمه

در سال‌های آغازین عصر فضا، فرایند هدایت فضاپیماها عمدتاً بر مبانی دینامیک کلاسیک و کنترل خطی استوار بوده است. با این حال، پیچیدگی روزافزون مأموریت‌های کنونی مانند سفرهای میان‌سیاره‌ای با پیشران‌کم و شبکه‌های انبوه ماهواره‌ای در مدار زمین موجب دوچندان شدن ضرورت بهره‌گیری از روش‌های هوشمند و تطبیق‌پذیر شده است. در ادامه، انگیزه‌ی پژوهش در بخش ۱-۱ و تعریف دقیق مسئله در بخش ۱-۲ آمده است. سپس، مروری کوتاه بر مبانی یادگیری تقویتی و نسخه‌ی چندعاملی آن در بخش‌های ۱-۳ و ۱-۴ ارائه شده و در نهایت، ساختار کل گزارش در بخش ۱-۵ تشریح شده است.

۱-۱ انگیزه پژوهش

در دو دهه‌ی اخیر، به‌دلیل کوچک‌سازی سامانه‌ها، توسعه‌ی الکترونیک مقرون‌به‌صرفه و افزایش ظرفیت‌های پرتاب، تحولات بنیادینی در مأموریت‌های فضایی تجربه شده است. از پروژه‌های علمیِ بین‌سیاره‌ای تا منظومه‌های انبوه ماهواره‌ای در مدارهای پایین زمین، مواجهه با چالشِ فراگیرِ هدایتِ بهینه در حضور عدم‌قطعیت‌ها به‌طور گسترده گزارش شده است. در مسیرهای فرا-قمری[۱] و به‌طور خاص در ناحیه‌های ناپایدارِ نقاط لاگرانژ در چارچوب مسئله‌ی سه‌جسمیِ کرویِ محدودِ دایروی[۲]، طراحی سامانه‌ی کنترل مستلزم تضمین هم‌زمانِ پایداریِ ایستا و بهره‌وریِ سوخت با پیشران‌کم[۳] است.

هم‌راستا با این تحولات، ظهور و گسترشِ الگوریتم‌های یادگیری تقویتیِ عمیق[۴] امکانات نوینی برای طراحی

[1] Trans-lunar
[2] Circular Restricted Three-Body Problem (CRTBP)
[3] Low-thrust
[4] Deep Reinforcement Learning (DRL)

کنترل‌کننده‌های تطبیقی فراهم آورده است؛ با این حال، غالبِ رویکردهای رایج بر سناریوهای تک‌عاملی و اتکا به مدل‌های دینامیکیِ دقیق استوار شده‌اند. غیابِ یک راهبردِ مقاوم در برابر اغتشاشاتِ مدل و تغییرات محیطی—از جمله خطای تراستِ پیشران و تأخیر حسگر—به ایجادِ فاصله‌ی معنادار میان عملکرد واقعی و پیش‌بینی‌های شبیه‌سازیِ ایده‌آل منجر شده است. در این پژوهش، این شکاف با بهره‌گیری از چارچوبِ یادگیریِ تقویتیِ چندعاملیِ مقاوم پُر می‌شود و اطمینانِ هدایتِ پیشران‌کم در CRTBP ارتقا داده می‌شود. در ادامه، تعریف دقیق مسئله و سپس اهداف و نوآوری‌های پژوهش ارائه می‌شود.

۱-۲ تعریف مسئله

در سال‌های اخیر، پیشرفت‌های فناوری درکنترل پرواز، پردازش و هوش مصنوعی به‌گسترشِ کاربردِ فضاپیماهای پیشران‌کم در منظومه‌ی زمین–ماه انجامیده است؛ از تعقیب و انتقال مداری تا استقرار و نگه‌داری. روش‌های هدایتِ بهینه‌ی کلاسیک، هرچند قدرتمند، عموماً به ساده‌سازی‌های بسیار، منابعِ محاسباتیِ زیاد و شرایط اولیه‌ی مناسب متکی بوده‌اند؛ در مقابل، بخشی از این محدودیت‌ها با الگوریتم‌های مبتنی بر یادگیری تقویتی و تکیه بر تعامل و امکان محاسباتِ درون‌برد[۵] برطرف می‌شود.

هدف، طراحیِ سیاستِ کنترلی برای فضاپیمایی با جرم m در میدانِ گرانشِ سامانه‌ی زمین–ماه (مدل دوبعدی در چارچوبِ چرخان) است. ویژگی‌ها به‌اختصار:

- **پویایی‌ها:** معادلاتِ حرکت در چارچوب مرجعِ چرخان به‌صورت $\dot{\boldsymbol{x}} = f(\boldsymbol{x}) + g(\boldsymbol{x})\,\boldsymbol{a}$ با $\boldsymbol{x} = [x,\, y,\, \dot{x},\, \dot{y}]^\top$ و کنترلِ پیوسته‌ی $\boldsymbol{a} \in \mathcal{A}$ تعریف می‌شود، به‌طوری‌که کران $|\boldsymbol{a}| \leqslant a_{\max}$ برقرار است.
- **عدم‌قطعیت‌ها:** شرایطِ اولیه‌ی تصادفی، اغتشاش‌های عملگر، عدم‌تطابقِ مدل (در پارامترهای جرم)، مشاهده‌ی ناقص، نویز حسگر و تأخیر زمانی، که بر پایداری و کارایی اثرگذارند.
- **صورت‌بندیِ بازیِ دیفرانسیلی (جمع‌صفر):** فضاپیما و طبیعت (اغتشاشات) به‌ترتیب به‌عنوان عاملِ کنترل و حریفِ مزاحم مدل می‌شوند؛ با افق زمانیِ محدود t_f، هدف، دستیابی به سیاستی مقاوم در برابر بدترین سناریو است.

صورتِ فشرده‌ی بهینه‌سازی به‌صورت کمینه–بیشینه است:

$$\min_{\pi}\ \max_{\omega}\ \mathbb{E}_{p,\pi,\omega}\left[\ \sum_{t=0}^{T} r\left(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{\delta}_t\right)\ \right], \tag{۱-۱}$$

[5] On-board Computing

که در آن، پاداشِ r به‌عنوان تابعی از مصرفِ سوخت، انحراف از مسیر یا مدار نامی و قیودِ مسئله تعریف می‌شود. خروجیِ مورد انتظار، سیاستی سبک و غیرمتمرکز برای اجرای درون‌برد مدنظر است.

۱-۳ یادگیری تقویتی

یادگیری تقویتی[۶] شاخه‌ای از یادگیریِ ماشین است که در آن توالیِ اقدام‌ها $\boldsymbol{a}_t \in \mathcal{A}$ به‌گونه‌ای انتخاب می‌شود که بازدهِ تجمعیِ آینده بیشینه شود. یک فرایندِ تصمیم‌گیریِ مارکوف[۷] به‌صورت $\langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$ تعریف می‌شود که در آن:

- $\mathcal{S}$: مجموعه‌ی حالات،
- $p(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})$: دینامیکِ انتقال،
- $r(s, \boldsymbol{a})$: پاداشِ آنی،
- $\gamma \in [0, 1)$: ضریبِ تنزیل.

سیاست[۸] $\pi(a|s)$ به‌عنوان احتمالِ انتخابِ اقدامِ $\boldsymbol{a}$ در وضعیتِ $\boldsymbol{s}$ بیان می‌شود. هدف، بیشینه‌سازیِ برگشت[۹] است:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \tag{۲-۱}$$

روش‌های RL معمولاً در دو دسته‌ی ارزش‌محور (مانند Q-learning و DQN) و سیاست‌محور (مانند Reinforce) جای می‌گیرند؛ ترکیبِ این دو به چارچوبِ Actor–Critic منتهی می‌شود که در آن، یک بازیگر (Actor) سیاست را به‌روزرسانی می‌کند و یک منتقد (Critic) ارزش یا Q برآورد می‌شود [۱].

در حضورِ فضاهای پیوسته‌ی حالت–عمل، الگوریتم‌های DDPG، TD3، SAC و PPO با تکیه بر شبکه‌های عصبی به‌عنوان تقریب‌گر توابع، کارایی بالایی نشان داده‌اند. در این پژوهش، خانواده‌ی Actor–Critic به‌عنوان پایه‌ی توسعه‌ی کنترل‌کننده پیشنهاد شده است و در ادامه، به نسخه‌ی چندعاملیِ آن در بخش ۱-۴ پیوند داده می‌شود.

[6] Reinforcement Learning (RL)
[7] Markov Decision Process (MDP)
[8] Policy
[9] Return

۱-۴ یادگیری تقویتی چندعاملی

در یادگیری تقویتی چندعاملی[۱۰]، فضای تصمیم‌گیری به‌صورت یک بازی مارکفی[۱۱] با مجموعه‌ی عامل‌ها $\mathcal{N} = \{1, \ldots, N\}$ مدل شده که در آن هر عامل با سیاست π_i به‌دنبال بیشینه‌سازی بازده تجمعی خود است. در سناریوهای رقابتیِ دونفره‌ی جمع‌صفر[۱۲]، مفهوم تعادل نش[۱۳] به‌عنوان معیار پایداری سیاست‌ها در نظر گرفته می‌شود.

رویکرد آموزشِ متمرکز، اجرایِ توزیع‌شده[۱۴] با جدا کردن مرحله‌ی آموزش که در آن اطلاعات خصوصیِ همه‌ی عامل‌ها برای منتقدها در دسترس است و مرحله‌ی اجرا در آن هر عامل صرفاً بر مشاهده‌ی محلی اتکا می‌کند که باعث تعادل میان کارایی، مقیاس‌پذیری و هزینه‌ی ارتباطی برقرار شده است.

در این پایان‌نامه، یک صورت‌بندیِ دوعاملیِ جمع‌صفر اتخاذ شده است که در آن سیاست هدایت توسط عاملِ کنترل آموخته می‌شود و اغتشاشات یا نامعینی‌ها توسط عاملِ مزاحم مدل‌سازی می‌شوند تا سیاستی مقاوم حاصل شود.

- DDPG: الگوریتم مبتنی بر گرادیان سیاستِ قطعی برای فضاهای کنشِ پیوسته،
- TD3: نسخه‌ی بهبودیافته‌ی DDPG با برآورد دوسویه‌ی Q برای کاهش تورشِ بیش‌براورد[۱۵]،
- PPO: الگوریتم سیاست احتمالیِ پایدار با قیود نسبت احتمال و بهبود تدریجی سیاست،
- SAC: الگوریتم حداکثرسازی آنتروپی که تعادل میان بهره‌برداری و اکتشاف به‌طور ذاتی برقرار می‌شود.

تابع پاداشِ طراحی‌شده، مصالحه‌ی سوخت، انحراف و قیود منعکس می‌شود و مبنایی برای ارزیابیِ کیفیتِ سیاست‌های مختلف فراهم می‌گردد.

۱-۵ ساختار گزارش

در **فصل دوم** مروری انتقادی بر کارهای مرتبط در هدایت پیشران‌کم و یادگیری تقویتی (تک‌عاملی و چندعاملی) ارائه می‌شود. **فصل سوم** به مبانی یادگیری تقویتی اختصاص داده شده و الگوریتم‌های DDPG، TD3، SAC

[10] Multi-Agent Reinforcement Learning (MARL)
[11] Markov Games (MG)
[12] Zero-Sum
[13] Nash Equilibrium
[14] Centralized Training with Decentralized Execution (CTDE)
[15] Overestimation Bias

و PPO مرور می‌شوند. در **فصل چهارم** چارچوب یادگیری تقویتی چندعاملی و رویکرد CTDE تشریح می‌شود و پیوند آن با بازی‌های جمع‌صفر و تعادل نش بیان می‌گردد. در **فصل پنجم** مدل‌سازی محیط آزمایش بر پایه‌ی CRTBP ارائه می‌شود. در **فصل ششم** طراحی عامل‌ها، فضای حالت/عمل، تابع پاداش و جزئیات آموزش توضیح داده می‌شود. در **فصل هفتم** چارچوب «سخت‌افزار در حلقه» و ارزیابی زمان‌واقعی گزارش می‌شود. سرانجام، در **فصل هشتم** نتایج، مقایسه با معیارهای مرجع و مسیرهای آینده‌ی پژوهش جمع‌بندی می‌شود.

Bibliography

[1] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, second edition, 2018.

Abstract

This thesis proposes a robust guidance framework for low-thrust spacecraft operating in multi-body dynamical environments modeled by the Earth–Moon circular restricted three-body problem (CRTBP). The guidance task is cast as a zero-sum differential game between a controller agent (spacecraft) and an adversary agent (environmental disturbances), implemented under a centralized-training/ decentralized-execution paradigm. Four continuous-control reinforcement-learning algorithms—DDPG, TD3, SAC, and PPO—are extended to their multi-agent zero-sum counterparts (MA-DDPG, MATD3, MASAC, MAPPO); their actor–critic network structures and training pipelines are detailed.

The policies are trained and evaluated on transfers to the Earth–Moon lyapunov orbit under five uncertainty scenarios: random initial states, actuator perturbations, sensor noise, communication delays, and model mismatch. Zero-sum variants consistently outperform their single-agent baselines, with MATD3 delivering the best trade-off between trajectory accuracy and propellant consumption while maintaining stability in the harshest conditions.

The results demonstrate that the proposed multi-agent, game-theoretic reinforcement-learning framework enables adaptive and robust low-thrust guidance in unstable three-body regions without reliance on precise dynamics models, and is ready for hardware-in-the-loop implementation.

Keywords: Deep Reinforcement Learning; Differential Game; Multi-Agent; Low-Thrust Guidance; Three-Body Problem; Robustness.

Sharif University of Technology

Department of Aerospace Engineering

Master Thesis

Robust Reinforcement Learning Differential Game Guidance in Low-Thrust, Multi-Body Dynamical Environments

By:

Ali BaniAsad

Supervisor:

Dr.Hadi Nobahari

December 2024