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

شهریور ۱۴۰۴


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 ممتحن: دکتر سیدعلی امامی خوانساری امضاء:

استاد ممتحن: دکتر علیرضا باصحبت نوین زاده 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

ارزیابی و نتایج یادگیری

در این فصل، نتایج حاصل از فرآیند یادگیری تقویتی در محیط سه‌جسمی ارائه و تحلیل شده است. هدف، بررسی عملکرد الگوریتم‌های استفاده‌شده و ارزیابی توانایی آن‌ها در دستیابی به اهداف تعیین‌شده می‌باشد. الگوریتم‌های یادگیری تقویتی مختلف شامل DDPG، PPO، SAC و TD3 در دو حالت تک‌عاملی و چندعاملی مبتنی بر بازی مجموع‌صفر مورد بررسی قرار گرفته‌اند. این فصل به ارائه نتایج عملکردی این الگوریتم‌ها و مقایسه قابلیت‌های آن‌ها در شرایط مختلف می‌پردازد. در بخش ؟؟ تنظیمات آزمایشی و پارامترهای محیط شبیه‌سازی معرفی می‌شوند. بخش ؟؟ به مقایسه مسیرها و فرمان‌های پیشران الگوریتم‌های مختلف در حالت‌های تک‌عاملی و چندعاملی می‌پردازد. ارزیابی مقاومت الگوریتم‌ها در برابر شرایط مختلف اختلال در بخش ۱-۱ بررسی می‌شود. در بخش ؟؟ مقایسه جامع بین تمام الگوریتم‌ها ارائه می‌گردد. تحلیل پایداری و همگرایی الگوریتم‌ها در بخش ؟؟ مورد بررسی قرار می‌گیرد و در نهایت در بخش ؟؟ مقایسه با معیارهای مرجع انجام می‌شود.

۱-۱ ارزیابی مقاومت الگوریتم‌ها

در این بخش، مقاومت الگوریتم‌های یادگیری در برابر شرایط مختلف اختلال مورد بررسی قرار گرفته است. این ارزیابی شامل شش سناریوی چالش‌برانگیز می‌شود: (۱) شرایط اولیه تصادفی، (۲) اغتشاش در عملگرها، (۳) عدم تطابق مدل، (۴) مشاهده ناقص، (۵) نویز حسگر و (۶) تأخیر زمانی. هدف، بررسی توانایی الگوریتم‌ها در حفظ کارایی خود در شرایط غیرایده‌آل و نزدیک به واقعیت است.

۱-۱-۱ سناریوهای ارزیابی مقاومت

در این بخش، سناریوهای مختلفی که برای ارزیابی مقاومت الگوریتم‌ها طراحی شده‌اند، با جزئیات کامل توضیح داده می‌شوند. هدف از این سناریوها بررسی عملکرد الگوریتم‌ها در شرایط غیرایده‌آل و چالش‌برانگیز است. این سناریوها شامل موارد زیر هستند:

شرایط اولیه تصادفی

در این سناریو، شرایط اولیه محیط به صورت تصادفی تغییر داده می‌شود. برای این منظور، به هر متغیر حالت اولیه نویز گوسی با میانگین صفر و انحراف معیار $\sigma = 0.1$ اضافه می‌شود. این تغییرات به منظور بررسی توانایی الگوریتم‌ها در سازگاری با تغییرات اولیه طراحی شده است.

اغتشاش در عملگرها

در این سناریو، نویز گوسی با انحراف معیار $\sigma = 0.05$ به اعمال نیروها اضافه می‌شود. علاوه بر این، نویز سنسور با انحراف معیار $\sigma = 0.02$ اعمال می‌شود. این تنظیمات برای شبیه‌سازی اغتشاشات در عملگرها و ارزیابی مقاومت الگوریتم‌ها در برابر این اغتشاشات استفاده شده است.

عدم تطابق مدل

در این سناریو، دینامیک محیط به صورت تصادفی تغییر داده می‌شود. برای این منظور، به پارامترهای محیط در طول انتقال نویز گوسی با انحراف معیار $\sigma = 0.05$ اضافه می‌شود. این تغییرات برای شبیه‌سازی عدم تطابق مدل و بررسی توانایی الگوریتم‌ها در مقابله با این شرایط طراحی شده است.

مشاهده ناقص

در این سناریو، بخشی از اطلاعات مشاهده‌شده توسط عامل حذف می‌شود. به طور خاص، 50% از متغیرهای حالت به صورت تصادفی پنهان شده و مقدار آن‌ها صفر می‌شود. این سناریو برای ارزیابی عملکرد الگوریتم‌ها در شرایط مشاهده ناقص طراحی شده است.

نویز حسگر

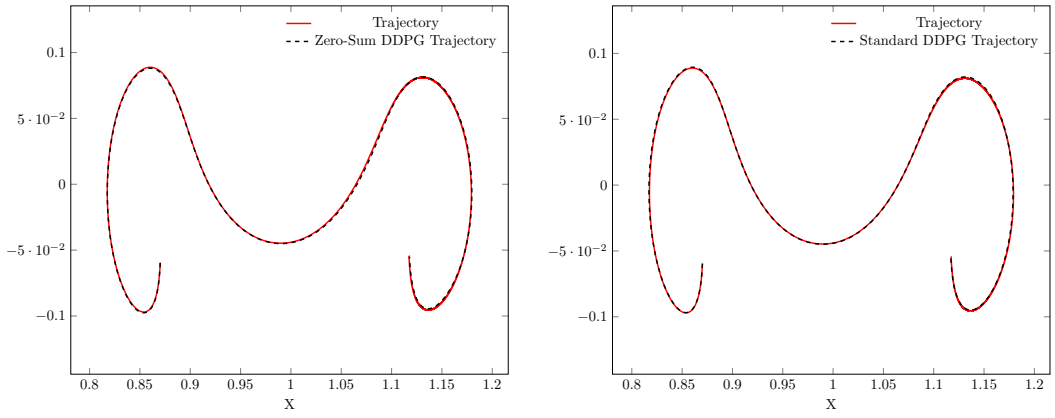
در این سناریو، نویز گوسی با انحراف معیار $\sigma = 0.05$ به مشاهدات حسگر اضافه می‌شود. این نویز به صورت ضربی به هر متغیر حالت اعمال می‌شود تا مقاومت الگوریتم‌ها در برابر نویز حسگر بررسی شود.

تأخیر زمانی

در این سناریو، تأخیر زمانی در اعمال اقدامات عامل به محیط شبیه‌سازی می‌شود. به طور خاص، اقدامات عامل با تأخیر 10 گام زمانی اعمال می‌شوند. علاوه بر این، نویز گوسی با انحراف معیار $\sigma = 0.05$ به اقدامات تأخیری اضافه می‌شود. این سناریو برای بررسی توانایی الگوریتم‌ها در مدیریت تأخیر زمانی طراحی شده است.

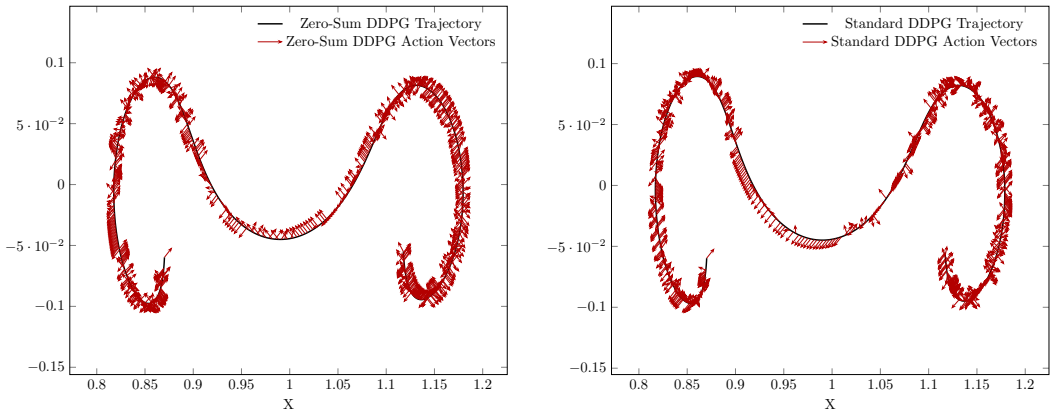
۱-۱-۲ الگوریتم DDPG

الگوریتم DDPG از جمله روش‌های یادگیری خارج از سیاست است که از دو شبکه عصبی برای بازیگر و منتقد استفاده می‌کند. در اینجا، عملکرد نسخه استاندارد و نسخه مبتنی بر بازی مجموع‌صفر این الگوریتم در کنترل فضاپیما مقایسه شده است.



(آ) DDPG استاندارد (ب) DDPG بازی مجموع‌صفر

شکل ۱-۱: مقایسه مسیر طی شده در دو الگوریتم تک‌عاملی و چندعاملی DDPG. مشاهده می‌شود که نسخه بازی مجموع‌صفر مسیر مستقیم‌تری را با انحراف کمتر از مسیر بهینه طی می‌کند.

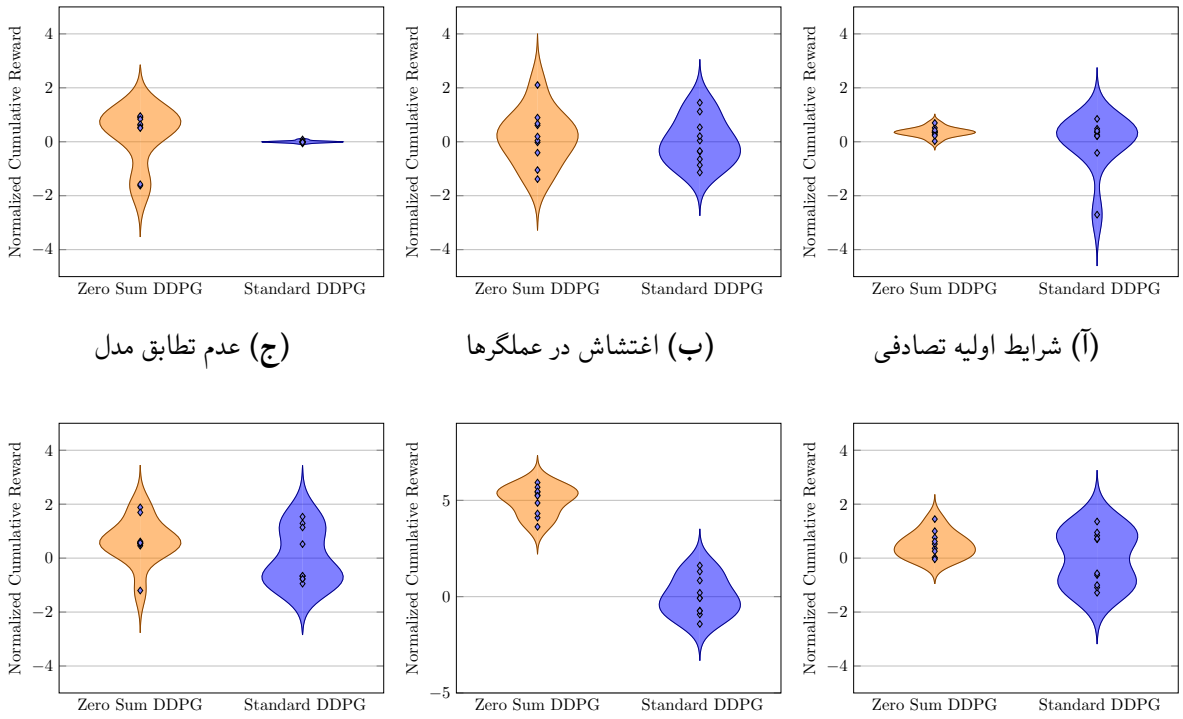


(آ) DDPG استاندارد (ب) DDPG بازی مجموع‌صفر

شکل ۱-۲: مقایسه مسیر و فرمان پیشران دو الگوریتم تک‌عاملی و چندعاملی DDPG. نمودارهای پایین نشان‌دهنده فرمان پیشران در طول زمان است که در نسخه بازی مجموع‌صفر، الگوی منظم‌تری را نشان می‌دهد و اوج‌های پیشران کمتری دارد.

همانطور که در شکل‌ها مشاهده می‌شود، الگوریتم DDPG مبتنی بر بازی مجموع‌صفر مسیر مستقیم‌تری را طی می‌کند و از نظر مصرف سوخت نیز بهینه‌تر عمل می‌کند. این بهبود عملکرد را می‌توان به ماهیت رقابتی بازی مجموع‌صفر و قابلیت آن در مقابله با عدم قطعیت‌های محیطی نسبت داد.

۱-۱-۳ مقایسه الگوریتم‌های تک‌عاملی و چندعاملی DDPG



(آ) شرایط اولیه تصادفی (ب) اغتشاش در عملگرها (ج) عدم تطابق مدل

(د) مشاهده ناقص (ه) نویز حسگر (و) تأخیر زمانی

شکل ۱-۳: مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی DDPG در سناریوهای مختلف. نسخه بازی مجموع‌صفر در اکثر سناریوها، به خصوص در شرایط اغتشاش در عملگرها و عدم تطابق مدل، عملکرد بهتری را نشان می‌دهد.

نتایج نشان می‌دهد که الگوریتم DDPG مبتنی بر بازی مجموع‌صفر در اکثر سناریوهای چالش‌برانگیز، عملکرد بهتری نسبت به نسخه استاندارد دارد. این برتری به خصوص در شرایط نویز حسگر و شرایط اولیه تصادفی مدل قابل توجه است، که نشان می‌دهد رویکرد چندعاملی توانایی بیشتری در مقابله با عدم قطعیت‌های سیستم دارد.

سناریو	پاداش تجمعی		مجموع خطای مسیر		مجموع تلاش کنترلی		احتمال شکست	
	DDPG	MA-DDPG	DDPG	MA-DDPG	DDPG	MA-DDPG	DDPG	MA-DDPG
شرایط اولیه تصادفی	-4.17	-3.63	0.40	0.63	5.60	5.60	1.00	1.00
اغتشاش در عملگرها	-1.93	-1.96	7.56	7.94	5.60	5.59	0.90	0.30
عدم تطابق مدل	-3.24	-2.70	0.70	0.76	5.57	5.57	1.00	1.00
مشاهده ناقص	-3.28	-2.89	0.68	0.75	5.57	5.57	0.60	0.80
نویز حسگر	-1.07	-0.47	0.10	0.15	5.54	5.54	0.00	0.00
تأخیر زمانی	-3.20	-1.91	1.74	2.43	5.61	5.61	0.70	0.70

جدول ۱-۱: مقایسه عملکرد DDPG و MA-DDPG در سناریوهای مختلف مقاومت

Bibliography

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

September 2025