



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



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 فضاپیمای کم‌پیشران در محیط‌های دینامیکی چندجسمی (مدل CRTBP زمین-ماه) ارائه شده است. مسئله به صورت بازی دیفرانسیلی مجموع صفر 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 TD3، DDPG، SAC و PPO به نسخه‌های چندعاملی مجموع صفر گسترش یافته‌اند (MASAC، MATD3، MA-DDPG 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 به وضوح نشان می‌دهد که نسخه‌های مجموع صفر 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.

به منظور 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 $5/8$ میلی‌ثانیه و مصرف حافظه را به $9/2$ مگابایت کاهش داد که به ترتیب بهبود ۴۷ درصدی و ۵۳ درصدی نسبت به مدل FP32 را نشان می‌دهد، در حالی که 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 فضاپیمای 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

۱	ارزیابی و نتایج یادگیری	۱
۱-۱	تنظیمات آزمایشی	۱
۲-۱	مقایسه مسیرها و فرمان پیشران	۲
۱-۲-۱	الگوریتم DDPG	۲
۲-۲-۱	الگوریتم PPO	۳
۳-۲-۱	الگوریتم SAC	۴
۴-۲-۱	الگوریتم TD3	۵
۳-۱	ارزیابی مقاومت الگوریتم‌ها	۶
۱-۳-۱	سناریوهای ارزیابی مقاومت	۶
۲-۳-۱	مقایسه الگوریتم‌های تک‌عاملی و چندعاملی DDPG	۸
۳-۳-۱	مقایسه الگوریتم‌های تک‌عاملی و چندعاملی PPO	۱۰
۴-۳-۱	مقایسه الگوریتم‌های تک‌عاملی و چندعاملی SAC	۱۲
۵-۳-۱	مقایسه الگوریتم‌های تک‌عاملی و چندعاملی TD3	۱۴
۴-۱	مقایسه جامع الگوریتم‌ها	۱۴
۱-۴-۱	مقایسه الگوریتم‌های تک‌عاملی	۱۶
۲-۴-۱	مقایسه الگوریتم‌های چندعاملی	۱۷
۵-۱	تحلیل پایداری و همگرایی	۱۷
۶-۱	مقایسه با معیارهای مرجع	۱۸

فهرست جداول

۹	۱-۱ جدول پارامترها و مقادیر پیش فرض الگوریتم DDPG
۱۱	۲-۱ جدول پارامترها و مقادیر پیش فرض الگوریتم PPO
۱۳	۳-۱ جدول پارامترها و مقادیر پیش فرض الگوریتم SAC
۱۵	۴-۱ جدول پارامترها و مقادیر پیش فرض الگوریتم TD3
۱۹	۵-۱ الگوریتم های تک عاملی
۲۰	۶-۱ الگوریتم های چند عاملی

فهرست تصاویر

- ۱-۱ مقایسه مسیر طی شده در دو الگوریتم تک‌عاملی و چندعاملی DDPG. ۲
- ۲-۱ مقایسه مسیر و فرمان پیشران دو الگوریتم تک‌عاملی و چندعاملی DDPG. ۳
- ۳-۱ مقایسه مسیر طی شده در دو الگوریتم تک‌عاملی و چندعاملی PPO. ۳
- ۴-۱ مقایسه مسیر و فرمان پیشران دو الگوریتم تک‌عاملی و چندعاملی PPO. ۴
- ۵-۱ مقایسه مسیر طی شده در دو الگوریتم تک‌عاملی و چندعاملی SAC. ۴
- ۶-۱ مقایسه مسیر و فرمان پیشران دو الگوریتم تک‌عاملی و چندعاملی SAC. ۵
- ۷-۱ مقایسه مسیر طی شده در دو الگوریتم تک‌عاملی و چندعاملی TD3. ۵
- ۸-۱ مقایسه مسیر و فرمان پیشران دو الگوریتم تک‌عاملی و چندعاملی TD3. ۶
- ۹-۱ مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی DDPG در سناریوهای مختلف. ۸
- ۱۰-۱ مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی PPO در سناریوهای مختلف. ۱۰
- ۱۱-۱ مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی SAC در سناریوهای مختلف. ۱۲
- ۱۲-۱ مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی TD3 در سناریوهای مختلف. ۱۴
- ۱۳-۱ مقایسه مجموع پاداش الگوریتم‌های تک‌عاملی در سناریوهای مختلف. ۱۶
- ۱۴-۱ مقایسه مجموع پاداش الگوریتم‌های چندعاملی در سناریوهای مختلف. ۱۷

فهرست الگوریتم‌ها

فصل ۱

ارزیابی و نتایج یادگیری

در این فصل، نتایج حاصل از فرآیند یادگیری تقویتی در محیط سه جسمی ارائه و تحلیل شده است. هدف، بررسی عملکرد الگوریتم‌های استفاده شده و ارزیابی توانایی آن‌ها در دستیابی به اهداف تعیین شده می‌باشد. الگوریتم‌های یادگیری تقویتی مختلف شامل TD3، SAC، PPO، DDPG در دو حالت تک‌عاملی و چندعاملی مبتنی بر بازی مجموع صفر مورد بررسی قرار گرفته‌اند. این فصل به ارائه نتایج عملکردی این الگوریتم‌ها و مقایسه قابلیت‌های آن‌ها در شرایط مختلف می‌پردازد. در بخش ۱-۱ تنظیمات آزمایشی و پارامترهای محیط شبیه‌سازی معرفی می‌شوند. بخش ۲-۱ به مقایسه مسیرها و فرمان‌های پیشران الگوریتم‌های مختلف در حالت‌های تک‌عاملی و چندعاملی می‌پردازد. ارزیابی مقاومت الگوریتم‌ها در برابر شرایط مختلف اختلال در بخش ۳-۱ بررسی می‌شود. در بخش ۴-۱ مقایسه جامع بین تمام الگوریتم‌ها ارائه می‌گردد. تحلیل پایداری و همگرایی الگوریتم‌ها در بخش ۵-۱ مورد بررسی قرار می‌گیرد و در نهایت در بخش ۶-۱ مقایسه با معیارهای مرجع انجام می‌شود.

۱-۱ تنظیمات آزمایشی

تنظیمات شبیه‌سازی، شامل پارامترهای محیط، نرخ یادگیری، و اندازه بافر تجربه، در این بخش تشریح شده است. آزمایش‌ها در محیط سه جسمی پیاده‌سازی شده با استفاده از کتابخانه‌های PyTorch و Gym انجام شده است. برای تمام الگوریتم‌ها، مشخصات یکسانی از شبکه‌های عصبی با ۳ لایه پنهان و ۲۵۶ نورون در هر لایه استفاده شده است. نرخ یادگیری برای تمامی مدل‌ها برابر با 3×10^{-4} تنظیم شده و از بهینه‌ساز Adam برای به‌روزرسانی وزن‌های شبکه استفاده شده است.

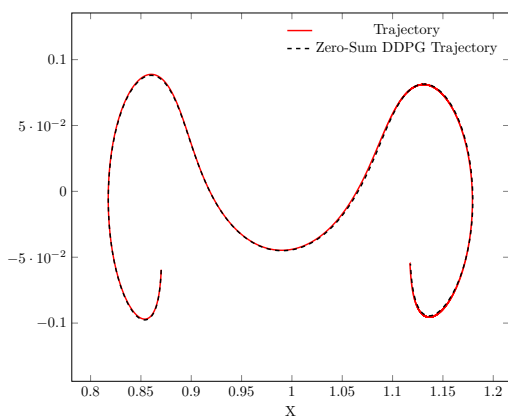
فرآیند آموزش برای هر الگوریتم شامل ۱ میلیون گام تعامل با محیط بوده و اندازه بافر تجربه برای الگوریتم‌های TD3 و SAC برابر با ۱۰۰ هزار نمونه تنظیم شده است. هر الگوریتم با ۱۰ مقداردهی اولیه متفاوت آموزش داده شده تا از پایداری نتایج اطمینان حاصل شود.

۲-۱ مقایسه مسیرها و فرمان پیشران

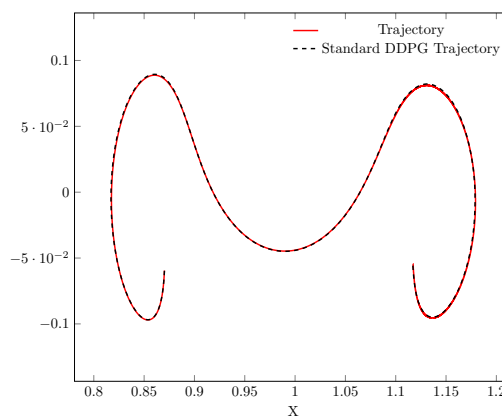
در این بخش، مسیرهای پرواز و فرمان‌های پیشران تولیدشده توسط الگوریتم‌های مختلف یادگیری تقویتی مقایسه شده است. این مقایسه به ما امکان می‌دهد تا تفاوت رفتاری بین روش‌های تک‌عاملی استاندارد و روش‌های چندعاملی مبتنی بر بازی مجموع‌صفر را مشاهده کنیم. هدف اصلی، ارزیابی کیفیت مسیرهای تولیدشده و کارآمدی مصرف سوخت در هر روش است.

۱-۲-۱ 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.

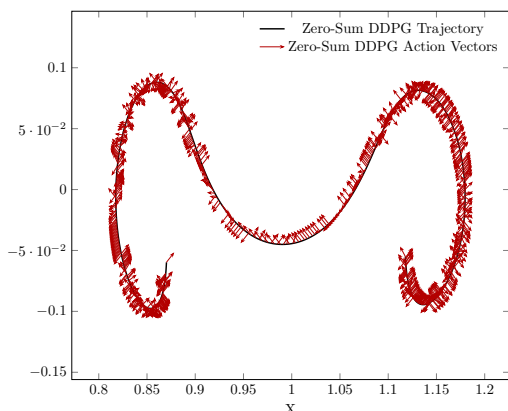


(ب) DDPG بازی مجموع‌صفر

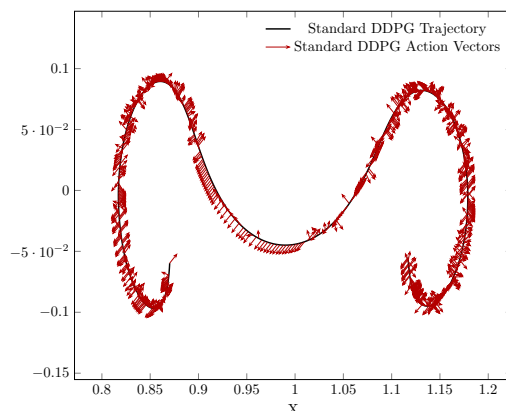


(آ) DDPG استاندارد

شکل ۱-۱: مقایسه مسیر طی شده در دو الگوریتم تک‌عاملی و چندعاملی DDPG.



(ب) DDPG بازی مجموع صفر

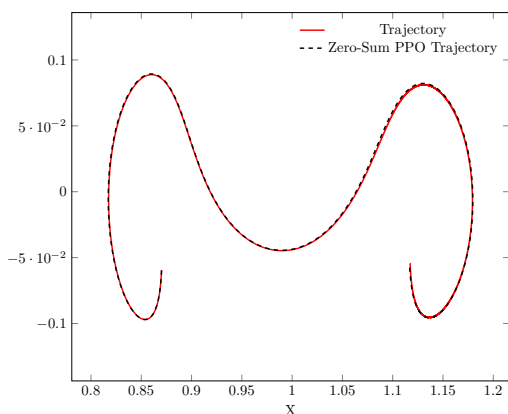


(آ) DDPG استاندارد

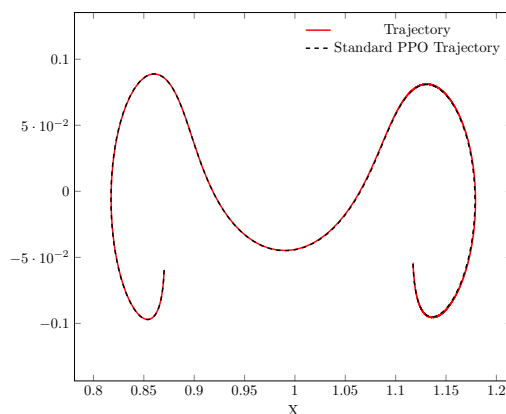
شکل ۱-۲: مقایسه مسیر و فرمان پیشران دو الگوریتم تک عاملی و چندعاملی DDPG.

۲-۲-۱ الگوریتم PPO

الگوریتم PPO از روش‌های نوین سیاست گرادیان است که با محدودسازی میزان تغییرات در هر بروزرسانی، پایداری بیشتری در فرآیند یادگیری ایجاد می‌کند. در ادامه، عملکرد این الگوریتم در دو حالت مورد بررسی قرار گرفته است.

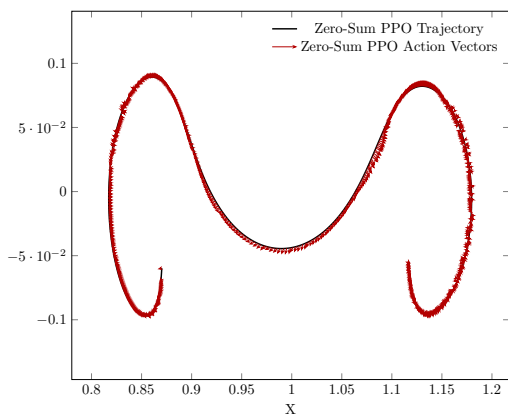


(ب) PPO بازی مجموع صفر

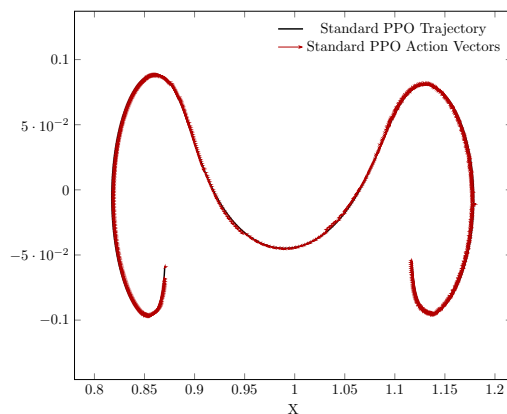


(آ) PPO استاندارد

شکل ۱-۳: مقایسه مسیر طی شده در دو الگوریتم تک عاملی و چندعاملی PPO.



(ب) PPO بازی مجموع صفر

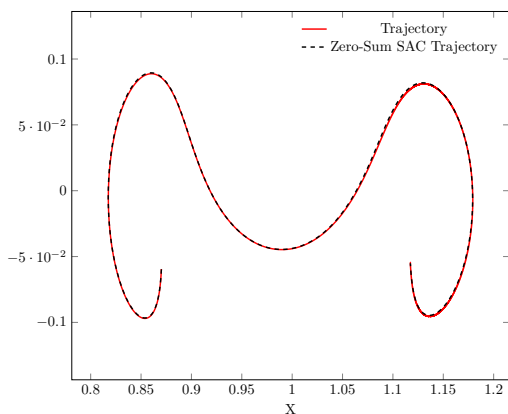


(آ) PPO استاندارد

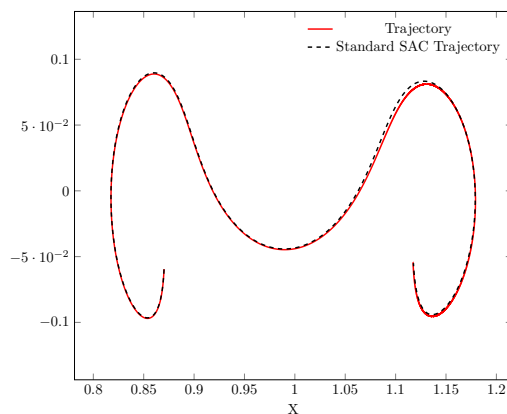
شکل ۱-۴: مقایسه مسیر و فرمان پیشران دو الگوریتم تک‌عاملی و چندعاملی PPO.

۳-۲-۱ الگوریتم SAC

الگوریتم SAC از روش‌های نوین یادگیری تقویتی است که با استفاده از مفهوم آنتروپی، تعادل بهتری بین اکتشاف و بهره‌برداری ایجاد می‌کند. این الگوریتم در شرایط فضاهای پیوسته عملکرد قابل توجهی دارد.

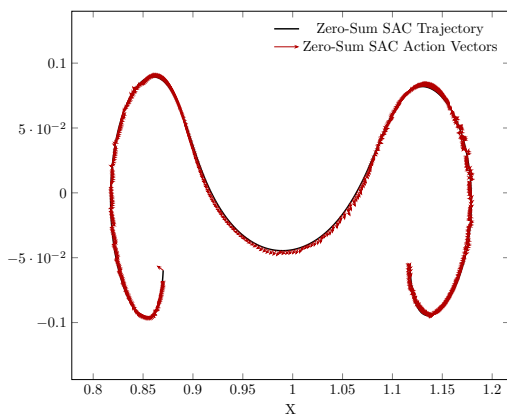


(ب) SAC بازی مجموع صفر

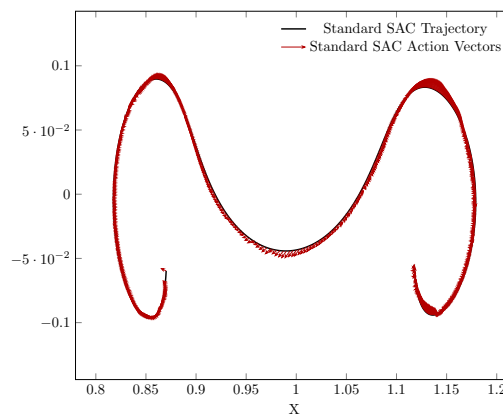


(آ) SAC استاندارد

شکل ۱-۵: مقایسه مسیر طی شده در دو الگوریتم تک‌عاملی و چندعاملی SAC.



(ب) بازی مجموع صفر SAC

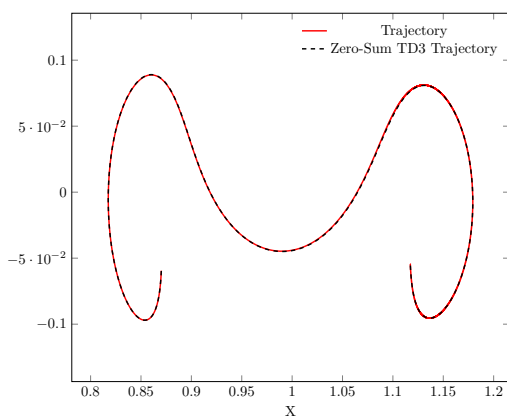


(آ) استاندارد SAC

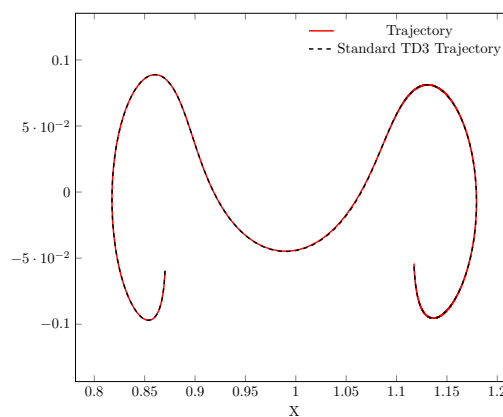
شکل ۱-۶: مقایسه مسیر و فرمان پیشران دو الگوریتم تک عاملی و چند عاملی SAC.

۴-۲-۱ الگوریتم TD3

الگوریتم TD3 (یادگیری تفاضل زمانی سه گانه عمیق) نسخه بهبود یافته DDPG است که با استفاده از تکنیک‌های جدید مانند شبکه‌های دو گانه منتقد و تأخیر در بروزرسانی سیاست، مشکلات تخمین بیش از حد را کاهش می‌دهد.

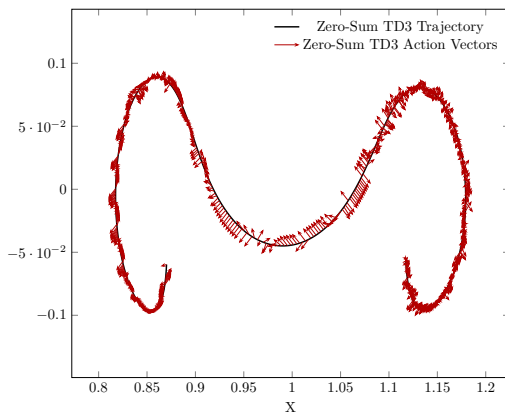


(ب) بازی مجموع صفر TD3

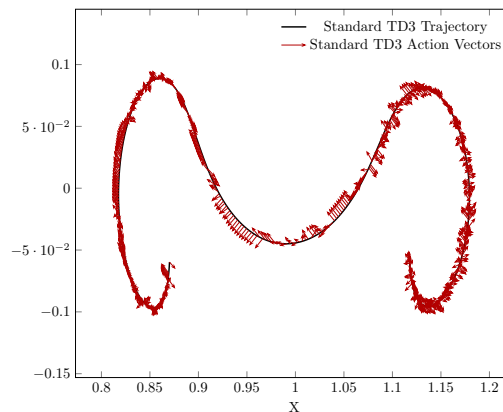


(آ) استاندارد TD3

شکل ۱-۷: مقایسه مسیر طی شده در دو الگوریتم تک عاملی و چند عاملی TD3.



TD3 بازی مجموع صفر (ب)



TD3 استاندارد (آ)

شکل ۸-۱: مقایسه مسیر و فرمان پیشران دو الگوریتم تک‌عاملی و چندعاملی TD3.

۳-۱ 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.

۱-۳-۱ 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.

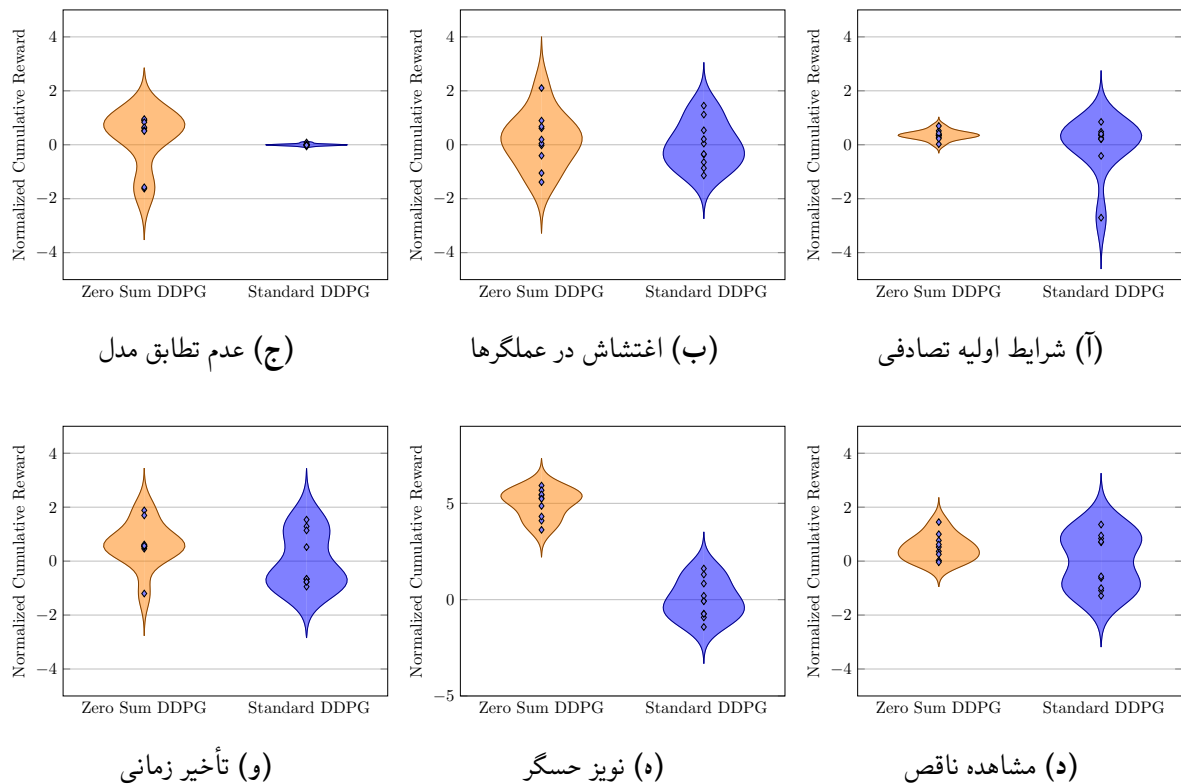
نویز حسگر

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.

۲-۳-۱ مقایسه الگوریتم‌های تک‌عاملی و چندعاملی DDPG



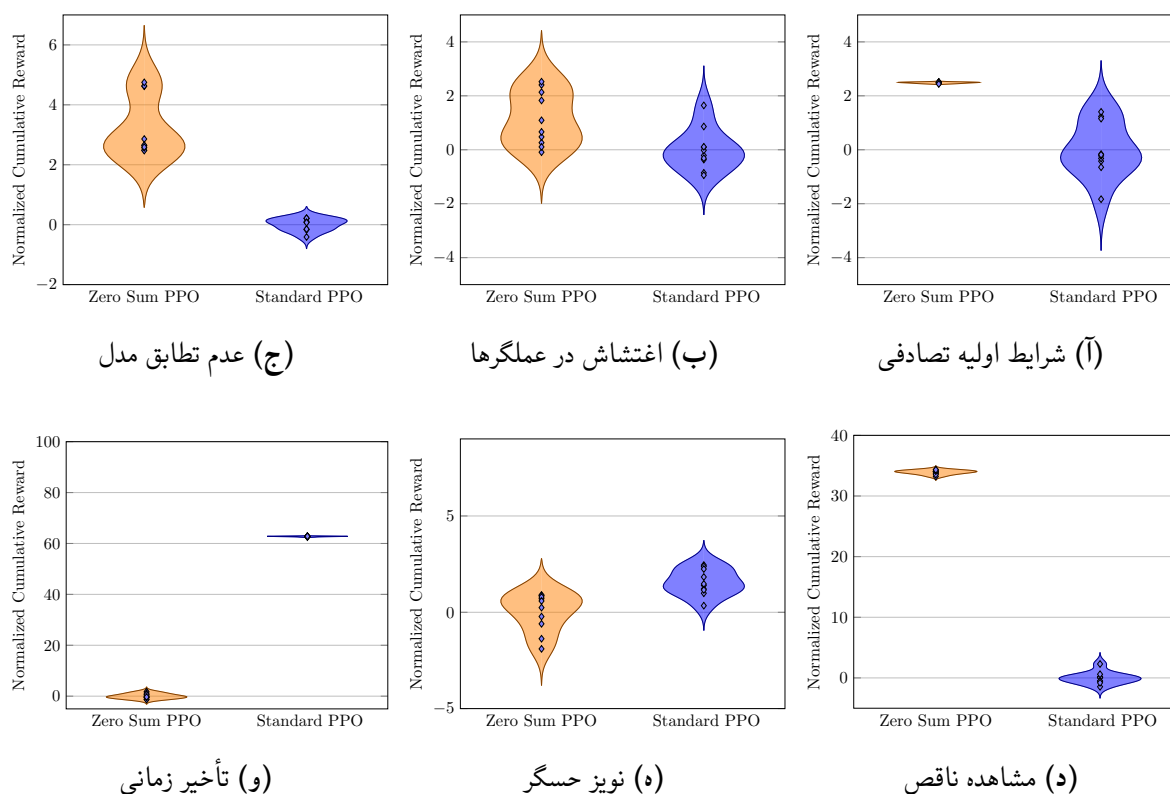
شکل ۹-۱: مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی DDPG در سناریوهای مختلف.

نتایج نشان می‌دهد که الگوریتم DDPG مبتنی بر بازی مجموع صفر 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					
	MA-DDPG	DDPG	MA-DDPG	DDPG	MA-DDPG	DDPG	MA-DDPG	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: جدول پارامترها و مقادیر پیش فرض الگوریتم DDPG

۳-۳-۱ مقایسه الگوریتم‌های تک‌عاملی و چندعاملی PPO



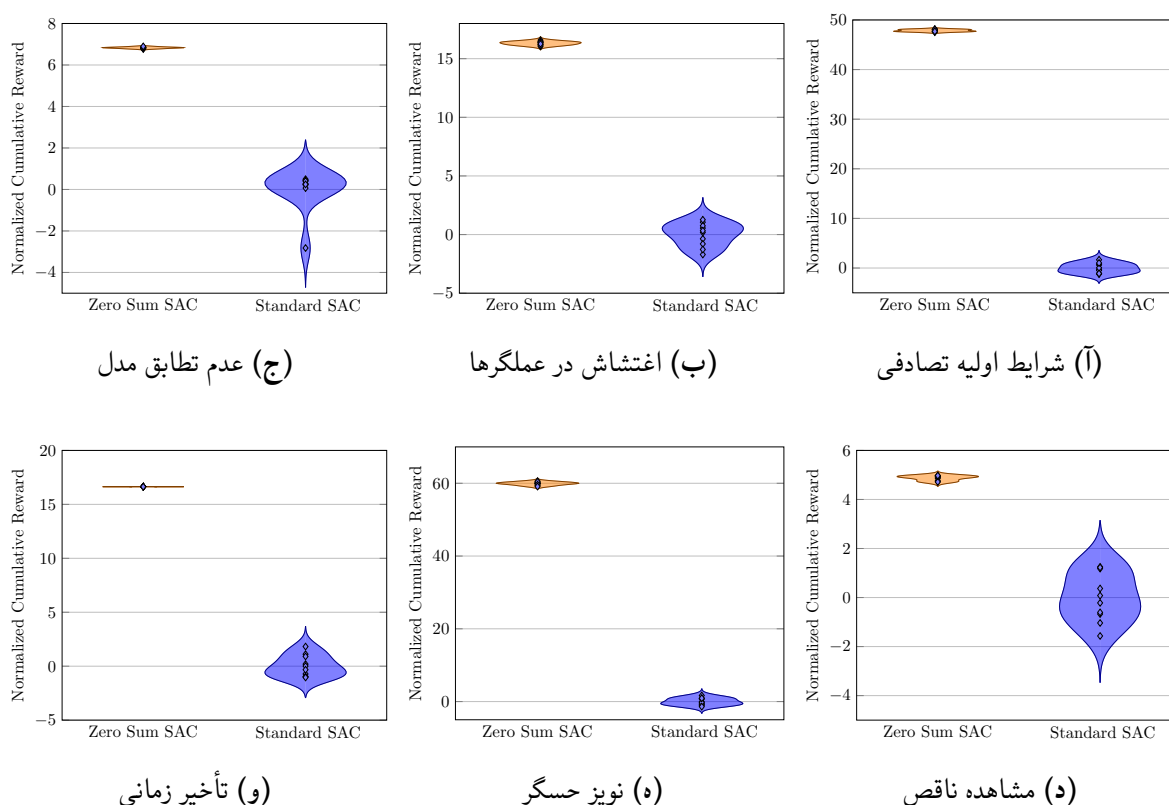
شکل ۱-۱۰: مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی PPO در سناریوهای مختلف.

الگوریتم PPO در حالت بازی مجموع صفر در اکثر سناریوها عملکرد بهتری نشان می‌دهد، به خصوص در شرایط تأخیر زمانی و نویز حسگر. این می‌تواند نشان‌دهنده توانایی روش چندعاملی در مدیریت بهتر شرایط دارای عدم قطعیت در ورودی‌ها باشد. با این حال، تفاوت در برخی از سناریوها کمتر از DDPG است.

سناریو	پاداش تجمعی		مجموع خطای مسیر مجموع تلاش کنترلی		احتمال شکست	
	MA-PPO	PPO	MA-PPO	PPO	MA-PPO	PPO
شرایط اولیه تصادفی	0.46	-1.85	0.14	0.22	1.98	0.70
اغتشاش در عملگرها	-1.91	-1.97	7.50	8.33	3.42	1.00
عدم تطابق مدل	0.30	0.46	0.08	0.07	1.13	0.00
مشاهده ناقص	-1.81	-3.60	2.06	2.34	2.15	1.00
نویز حسگر	0.48	0.52	0.15	0.13	2.08	0.00
تأخیر زمانی	-2.44	0.58	2.49	0.03	2.56	0.00

جدول ۱-۲: جدول پارامترها و مقادیر پیش فرض الگوریتم PPO

۴-۳-۱ مقایسه الگوریتم‌های تک‌عاملی و چندعاملی SAC



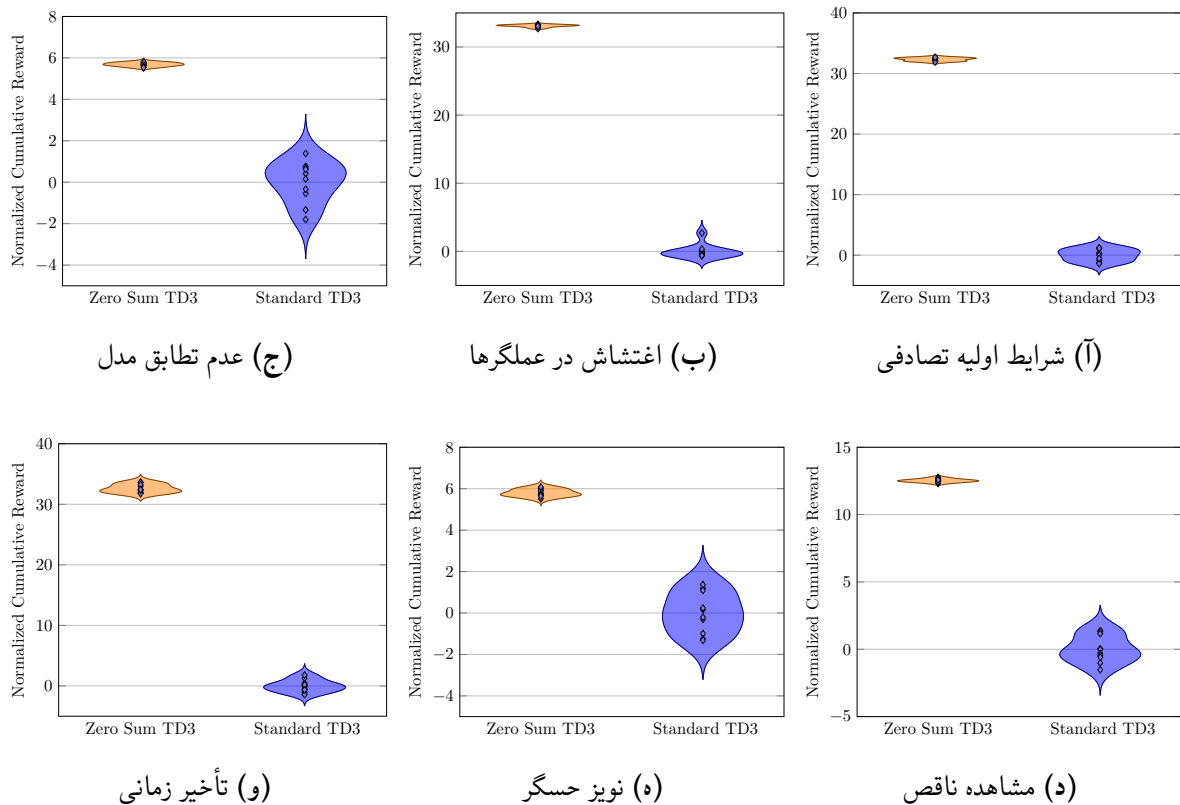
شکل ۱-۱۱: مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی SAC در سناریوهای مختلف.

الگوریتم SAC در هر دو حالت عملکرد نسبتاً خوبی در سناریوهای مختلف نشان می‌دهد. این می‌تواند به دلیل استفاده از مکانیزم آنتروپی باشد که به صورت ذاتی اکتشاف بیشتری را تشویق می‌کند. با این حال، نسخه بازی مجموع صفر در تمامی سناریوها برتری معناداری دارد که نشان‌دهنده مقاومت بیشتر آن در شرایط با اطلاعات محدود است.

سناریو	پاداش تجمعی								مجموع خطای مسیر مجموع تلاش کنترلی احتمال شکست							
	MA-SAC		SAC		MA-SAC		SAC		MA-SAC		SAC		MA-SAC		SAC	
شرایط اولیه تصادفی	1.00	1.00	1.37	1.37	0.26	0.29	-2.98	-4.69	1.00	1.00	1.37	1.37	0.26	0.29	-2.98	-4.69
اغتشاش در عملگرها	1.00	1.00	3.09	3.09	7.72	8.02	-1.93	-1.95	1.00	1.00	3.09	3.09	7.72	8.02	-1.93	-1.95
عدم تطابق مدل	1.00	1.00	1.16	1.16	0.26	0.38	-4.35	-4.89	1.00	1.00	1.16	1.16	0.26	0.38	-4.35	-4.89
مشاهده ناقص	0.00	1.00	1.99	1.99	0.07	1.95	-0.44	-3.63	0.00	1.00	1.99	1.99	0.07	1.95	-0.44	-3.63
نویز حسگر	0.00	0.00	1.86	1.86	0.12	0.12	0.12	-0.89	0.00	0.00	1.86	1.86	0.12	0.12	0.12	-0.89
تأخیر زمانی	0.00	1.00	1.25	1.25	0.01	1.87	-0.05	-4.14	0.00	1.00	1.25	1.25	0.01	1.87	-0.05	-4.14

جدول ۱-۳: جدول پارامترها و مقادیر پیش فرض الگوریتم SAC

۵-۳-۱ مقایسه الگوریتم‌های تک‌عاملی و چندعاملی TD3



شکل ۱-۱۲: مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی TD3 در سناریوهای مختلف.

الگوریتم TD3 مبتنی بر بازی مجموع صفر در تمامی سناریوها نشان می‌دهد. این نتایج نشان می‌دهد که ترکیب مکانیزم‌های پایدارسازی TD3 با رویکرد بازی مجموع صفر می‌تواند منجر به مقاومت قابل توجهی در برابر شرایط نامطلوب شود.

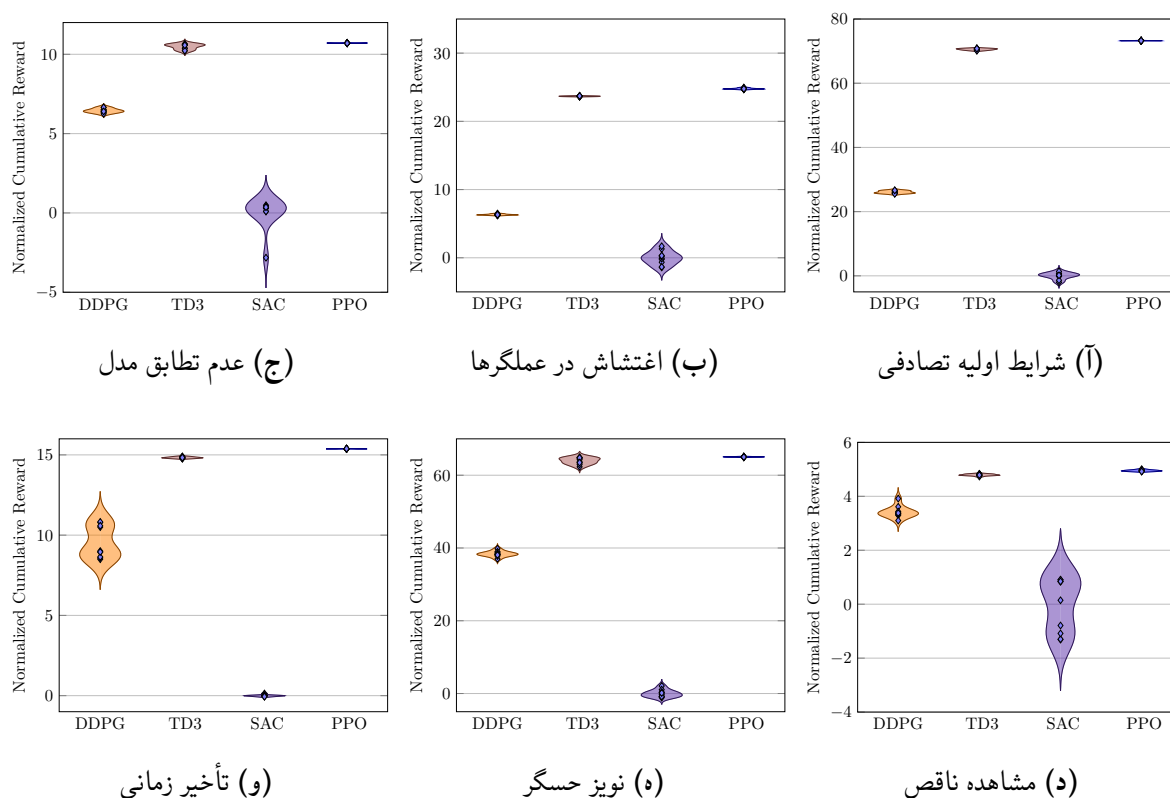
۴-۱ مقایسه جامع الگوریتم‌ها

در این بخش، مقایسه جامعی بین تمام الگوریتم‌ها در دو حالت تک‌عاملی و چندعاملی ارائه شده است. هدف، تعیین بهترین الگوریتم برای هر سناریوی خاص و درک بهتر نقاط قوت و ضعف هر روش است.

سناریو	پاداش تجمعی		مجموع خطای مسیر مجموع تلاش کنترلی احتمال شکست					
	MA-TD3	TD3	MA-TD3	TD3	MA-TD3	TD3	MA-TD3	TD3
شرایط اولیه تصادفی	-2.95	-0.26	0.39	0.14	4.57	4.57	1.00	0.30
اغتشاش در عملگرها	0.56	0.73	0.02	0.00	2.66	2.66	0.00	0.00
عدم تطابق مدل	-4.73	-3.30	0.47	0.73	5.41	5.41	1.00	1.00
مشاهده ناقص	0.21	0.71	0.02	0.01	3.18	3.18	0.00	0.00
نویز حسگر	-0.08	-2.93	0.11	3.19	5.50	5.50	0.00	1.00
تأخیر زمانی	0.55	0.67	0.01	0.01	4.57	4.57	0.00	0.00

جدول ۱-۴: جدول پارامترها و مقادیر پیش فرض الگوریتم TD3

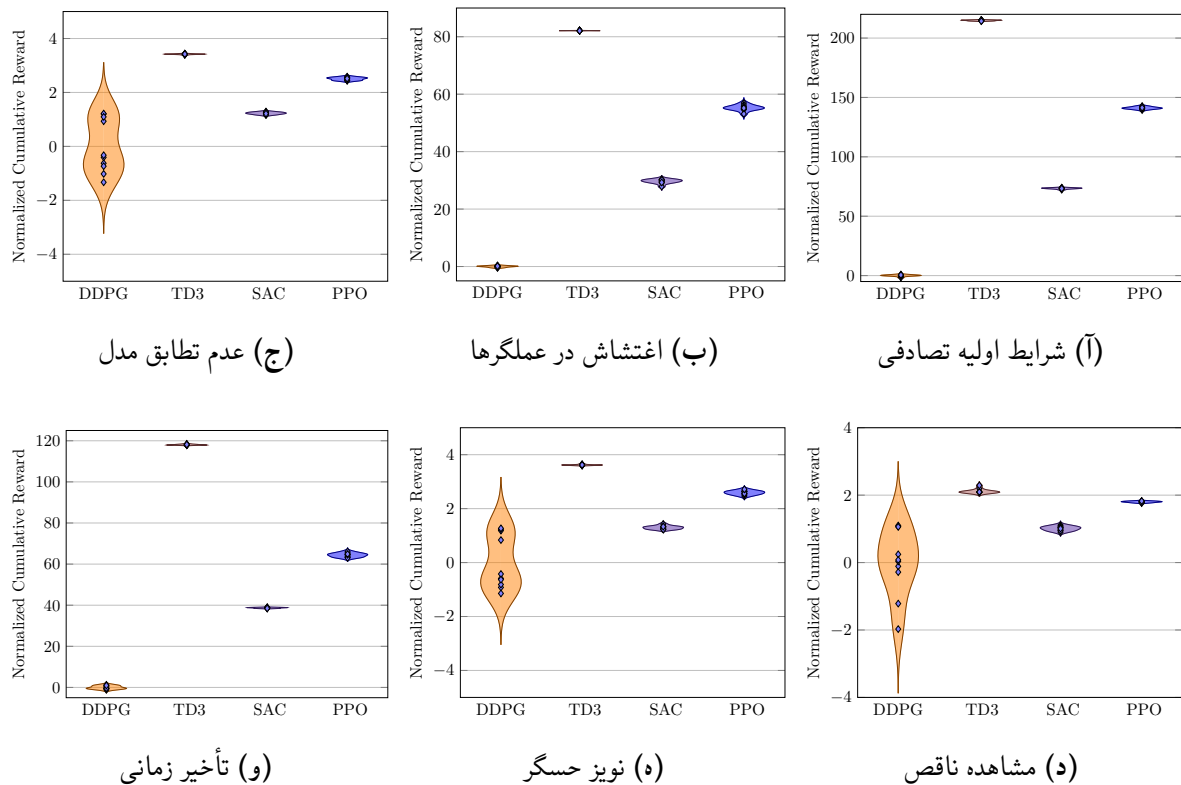
۱-۴-۱ مقایسه الگوریتم‌های تک‌عاملی



شکل ۱-۱۳: مقایسه مجموع پاداش الگوریتم‌های تک‌عاملی در سناریوهای مختلف.

در میان الگوریتم‌های تک‌عاملی، PPO و TD3 در اکثر سناریوها عملکرد بهتری نسبت به DDPG و SAC نشان می‌دهند.

۲-۴-۱ مقایسه الگوریتم‌های چندعاملی



شکل ۱-۱۴: مقایسه مجموع پاداش الگوریتم‌های چندعاملی در سناریوهای مختلف.

در میان الگوریتم‌های تک‌عاملی، PPO و TD3 در اکثر سناریوها عملکرد بهتری نسبت به DDPG و SAC نشان می‌دهند.

۵-۱ تحلیل پایداری و همگرایی

پایداری و سرعت همگرایی فرآیند یادگیری با استفاده از نمودارهای پاداش و معیارهای عددی مورد بررسی قرار گرفته است. نتایج نشان می‌دهد که الگوریتم‌های مبتنی بر بازی مجموع صفر در اکثر موارد همگرایی پایدارتری را نسبت به نسخه‌های استاندارد نشان می‌دهند. این پایداری به خصوص در TD3 و PPO قابل توجه است.

تحلیل نرخ همگرایی نشان می‌دهد که PPO در هر دو نسخه استاندارد و بازی مجموع صفر، سریع‌ترین همگرایی را دارد، در حالی که DDPG کندترین نرخ را نشان می‌دهد. با این حال، کیفیت نهایی سیاست آموخته‌شده در TD3 مبتنی بر بازی مجموع صفر بالاترین است.

۱-۶ مقایسه با معیارهای مرجع

عملکرد الگوریتم‌ها با روش‌های مرجع مانند کنترل بهینه کلاسیک و کنترل پیش‌بین مدل مقایسه شده تا برتری‌ها و محدودیت‌های آن‌ها مشخص گردد. نتایج نشان می‌دهد که در شرایط ایده‌آل، روش‌های کنترل بهینه کلاسیک دقت بالاتری دارند، اما در حضور عدم قطعیت‌ها و اختلالات، الگوریتم‌های یادگیری تقویتی به خصوص نسخه‌های مبتنی بر بازی مجموع صفر، مقاومت و انعطاف‌پذیری بیشتری نشان می‌دهند.

در مجموع، الگوریتم TD3 مبتنی بر بازی مجموع صفر بهترین تعادل بین دقت، کارایی و مقاومت را در مقایسه با سایر روش‌ها و معیارهای مرجع ارائه می‌دهد.

سناریو	DDPG	PPO	SAC	TD3	سناریو	DDPG	PPO	SAC	TD3
شرایط اولیه تصادفی	3.30	2.56	8.06	0.72	شرایط اولیه تصادفی	-0.27	0.61	-0.76	0.56
اغتشاش در عملگرها	3.74	2.58	7.91	0.77	اغتشاش در عملگرها	-0.38	0.61	-0.72	0.55
عدم تطابق مدل	10.87	3.06	17.12	1.09	عدم تطابق مدل	-0.84	0.58	-2.98	0.51
مشاهده ناقص	8.18	3.34	15.47	1.77	مشاهده ناقص	-0.88	0.36	-3.65	0.23
نویز حسگر	11.04	3.08	16.81	1.02	نویز حسگر	-0.85	0.58	-2.90	0.52
تأخیر زمانی	8.95	2.27	15.70	0.81	تأخیر زمانی	-0.76	0.61	-2.98	0.48
مجموع خطای مسیر					پاداش تجمعی				
سناریو	DDPG	PPO	SAC	TD3	سناریو	DDPG	PPO	SAC	TD3
شرایط اولیه تصادفی	0.00	0.00	0.00	0.00	شرایط اولیه تصادفی	5.11	0.77	1.76	3.31
اغتشاش در عملگرها	0.00	0.00	0.00	0.00	اغتشاش در عملگرها	4.89	0.77	1.71	3.07
عدم تطابق مدل	0.00	0.00	1.00	0.00	عدم تطابق مدل	5.48	0.86	2.37	4.32
مشاهده ناقص	0.00	0.00	1.00	0.00	مشاهده ناقص	5.37	1.03	2.33	4.10
نویز حسگر	0.00	0.00	1.00	0.00	نویز حسگر	5.48	0.86	2.37	4.30
تأخیر زمانی	0.00	0.00	1.00	0.00	تأخیر زمانی	5.51	0.76	2.11	5.12
احتمال شکست					مجموع تلاش کنترلی				

جدول ۱-۵: الگوریتم‌های تک‌عاملی

سناریو	DDPG	PPO	SAC	TD3	سناریو	DDPG	PPO	SAC	TD3
شرایط اولیه تصادفی	-1.95	-1.95	-2.26	-1.97	شرایط اولیه تصادفی	28.78	24.97	22.78	32.17
اغتشاش در عملگرها	-1.95	-1.95	-2.26	-1.97	اغتشاش در عملگرها	28.91	24.99	22.70	32.12
عدم تطابق مدل	-1.94	-1.95	-2.25	-1.96	عدم تطابق مدل	29.32	25.68	23.93	32.66
مشاهده ناقص	-1.93	-1.95	-2.23	-1.94	مشاهده ناقص	23.40	20.78	19.40	25.12
نویز حسگر	-1.94	-1.95	-2.25	-1.96	نویز حسگر	29.55	25.71	23.74	32.47
تأخیر زمانی	-1.94	-1.95	-2.25	-1.98	تأخیر زمانی	28.65	24.93	22.75	31.43
پاداش تجمعی					مجموع خطای مسیر				
سناریو	DDPG	PPO	SAC	TD3	سناریو	DDPG	PPO	SAC	TD3
شرایط اولیه تصادفی	1.89	0.53	2.68	3.60	شرایط اولیه تصادفی	1.00	1.00	1.00	1.00
اغتشاش در عملگرها	1.90	0.53	2.70	3.59	اغتشاش در عملگرها	1.00	1.00	1.00	1.00
عدم تطابق مدل	1.81	0.51	2.52	3.36	عدم تطابق مدل	1.00	1.00	1.00	1.00
مشاهده ناقص	1.75	0.50	2.34	3.03	مشاهده ناقص	1.00	1.00	1.00	1.00
نویز حسگر	1.80	0.51	2.54	3.38	نویز حسگر	1.00	1.00	1.00	1.00
تأخیر زمانی	1.89	0.53	2.70	3.59	تأخیر زمانی	1.00	1.00	1.00	1.00
مجموع تلاش کنترلی					احتمال شکست				

جدول ۱-۶: الگوریتم‌های چندعاملی

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.

For real-time deployment, the learned networks are quantized to INT8 and exported to ONNX for execution on a ROS 2 hardware-in-the-loop platform. Inference latency is reduced to 5.8 ms and memory footprint to 9.2 MB—improvements of 47 % and 53 % over the FP32 models—while sustaining a 100 Hz control loop with no deadline misses.

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