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

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 ۱-۱ تنظیمات آزمایشی و پارامترهای محیط شبیه‌سازی معرفی می‌شوند. بخش ۱-۲ به مقایسه مسیرها و فرمان‌های پیشران الگوریتم‌های مختلف در حالت‌های تک‌عاملی و چندعاملی می‌پردازد. ارزیابی مقاومت الگوریتم‌ها در برابر شرایط مختلف اختلال در بخش ۱-۳ بررسی می‌شود. در بخش ۱-۴ مقایسه جامع بین تمام الگوریتم‌ها ارائه می‌گردد. تحلیل پایداری و همگرایی الگوریتم‌ها در بخش ۱-۵ مورد بررسی قرار می‌گیرد و در نهایت در بخش ۱-۶ 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 سه‌جسمی 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.

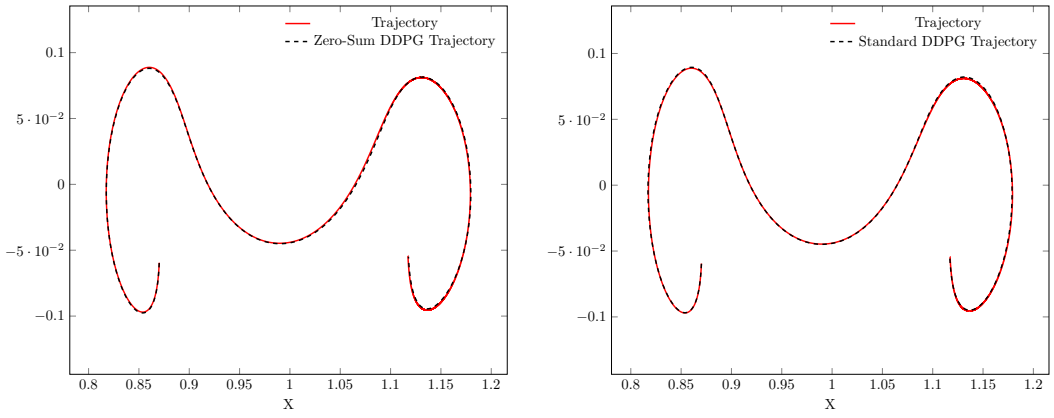
فرآیند آموزش برای هر الگوریتم شامل ۱ میلیون گام تعامل با محیط بوده و اندازه بافر تجربه برای الگوریتم‌های DDPG، SAC و TD3 برابر با ۱۰۰ هزار نمونه تنظیم شده است. هر الگوریتم با ۱۰ مقداردهی اولیه متفاوت آموزش داده شده تا از پایداری نتایج اطمینان حاصل شود.

۱-۲ 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.

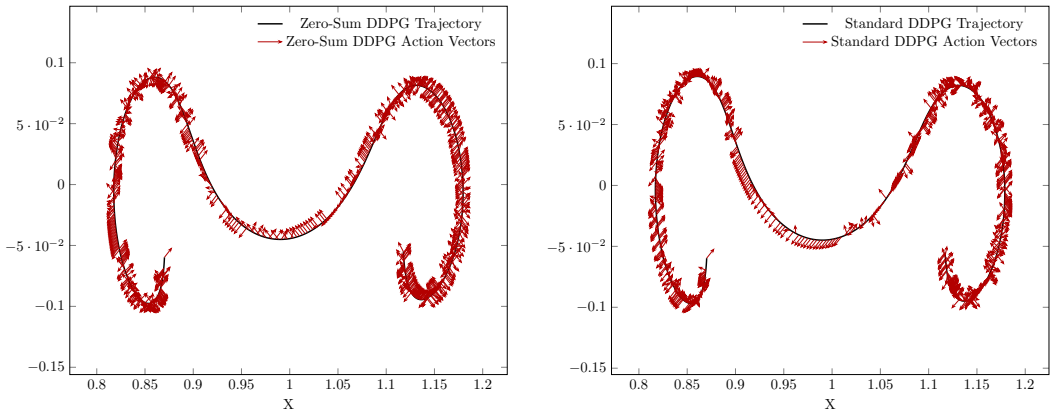
۱-۲-۱ الگوریتم DDPG

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.



(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.



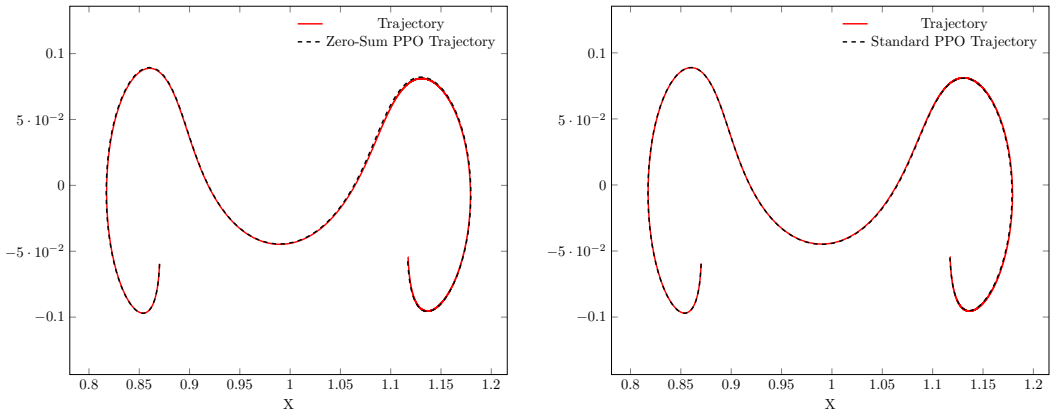
(آ) DDPG استاندارد (ب) DDPG بازی مجموع‌صفر

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.

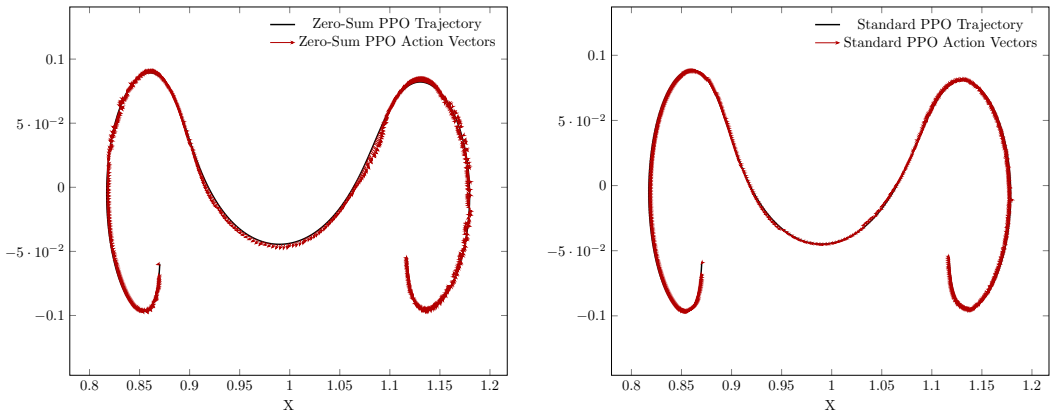
۱-۲-۲ الگوریتم PPO

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.



(آ) PPO استاندارد (ب) PPO بازی مجموع‌صفر

شکل ۱-۳: مقایسه مسیر طی شده در دو الگوریتم تک‌عاملی و چندعاملی PPO. نسخه بازی مجموع‌صفر همگرایی بهتری به مسیر هدف را نشان می‌دهد، به خصوص در مراحل نزدیک شدن به هدف.



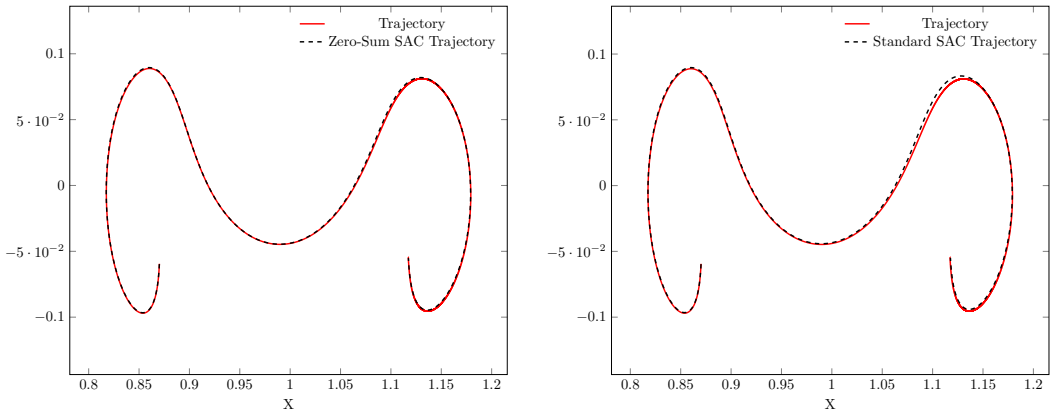
(آ) PPO استاندارد (ب) PPO بازی مجموع‌صفر

شکل ۱-۴: مقایسه مسیر و فرمان پیشران دو الگوریتم تک‌عاملی و چندعاملی PPO. فرمان‌های پیشران در نسخه بازی مجموع‌صفر از نظر توزیع انرژی متوازن‌تر است و نوسانات کمتری را نشان می‌دهد.

نتایج نشان می‌دهد که الگوریتم PPO در حالت بازی مجموع‌صفر عملکرد قابل توجهی دارد، اما تفاوت آن با نسخه استاندارد کمتر از DDPG است. این می‌تواند به دلیل ماهیت ذاتی PPO در ایجاد تعادل بین اکتشاف و بهره‌برداری باشد که آن را در حالت استاندارد نیز نسبتاً مقاوم می‌سازد.

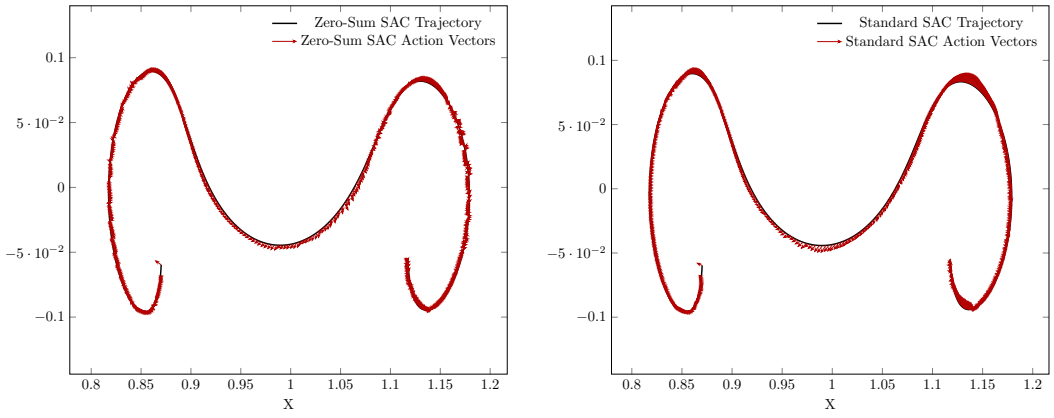
۱-۲-۳ الگوریتم SAC

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.



(آ) SAC استاندارد (ب) SAC بازی مجموع‌صفر

شکل ۱-۵: مقایسه مسیر طی شده در دو الگوریتم تک‌عاملی و چندعاملی SAC. مسیرهای تولیدشده توسط هر دو نسخه از کیفیت بالایی برخوردارند، اما نسخه بازی مجموع‌صفر در مناطق با گرادیان جاذبه پیچیده عملکرد پایدارتری را نشان می‌دهد.



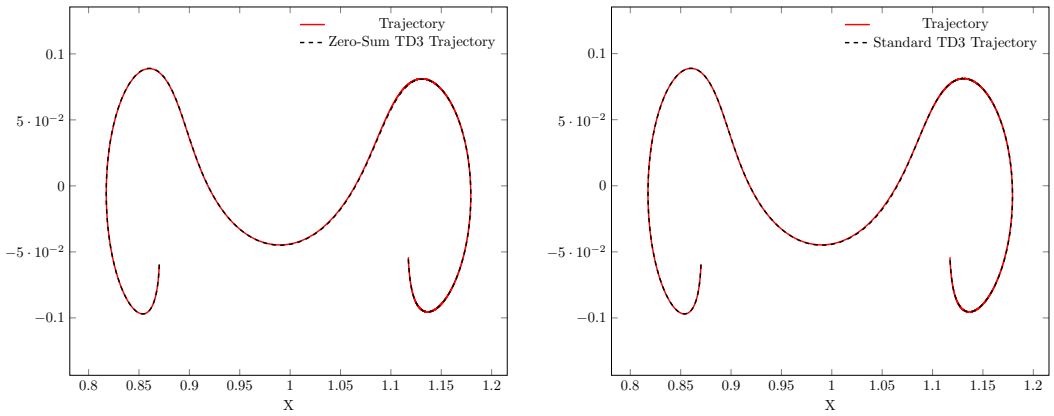
(آ) SAC استاندارد (ب) SAC بازی مجموع‌صفر

شکل ۱-۶: مقایسه مسیر و فرمان پیشران دو الگوریتم تک‌عاملی و چندعاملی SAC. نسخه بازی مجموع‌صفر مصرف سوخت متعادل‌تری را در طول مسیر نشان می‌دهد که می‌تواند منجر به صرفه‌جویی در منابع شود.

الگوریتم SAC در هر دو حالت عملکرد قابل قبولی ارائه می‌دهد. ویژگی خاص این الگوریتم در تنظیم خودکار پارامتر آنتروپی باعث می‌شود که بتواند تعادل مناسبی بین اکتشاف و بهره‌برداری ایجاد کند، اما نسخه بازی مجموع‌صفر آن در شرایط سخت‌تر مقاومت بیشتری نشان می‌دهد.

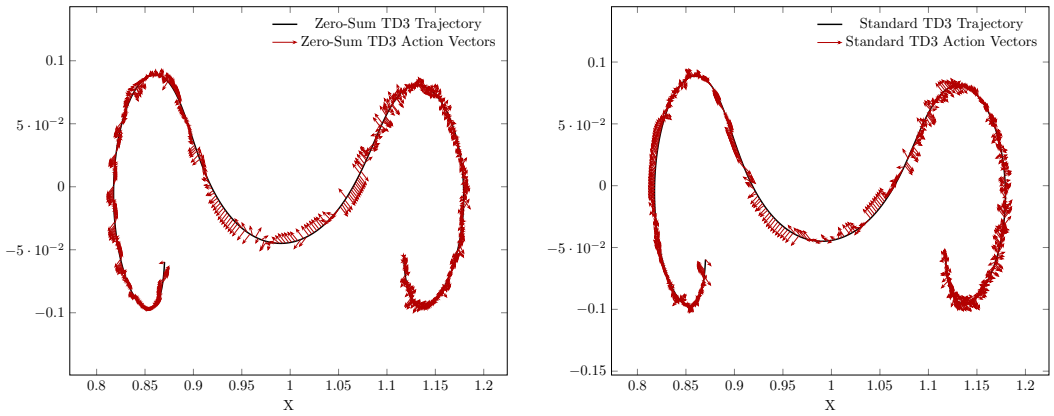
۱-۲-۴ الگوریتم TD3

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



(آ) TD3 استاندارد (ب) TD3 بازی مجموع‌صفر

شکل ۱-۷: مقایسه مسیر طی شده در دو الگوریتم تک‌عاملی و چندعاملی TD3. مسیرهای تولیدشده توسط نسخه بازی مجموع‌صفر نشان‌دهنده کاهش انحراف از مسیر بهینه و همگرایی سریع‌تر به هدف است.



(آ) TD3 استاندارد (ب) TD3 بازی مجموع‌صفر

شکل ۱-۸: مقایسه مسیر و فرمان پیشران دو الگوریتم تک‌عاملی و چندعاملی TD3. فرمان‌های پیشران در نسخه بازی مجموع‌صفر از توزیع یکنواخت‌تری برخوردار است که نشان‌دهنده استفاده بهینه‌تر از منابع پیشرانش می‌باشد.

الگوریتم TD3 در هر دو حالت عملکرد قابل توجهی دارد، اما نسخه بازی مجموع‌صفر آن بهبودهای معناداری در کیفیت مسیر و مصرف سوخت نشان می‌دهد. ثبات بیشتر این الگوریتم در مقایسه با DDPG در هر دو نسخه قابل مشاهده است.

۱-۳ 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.

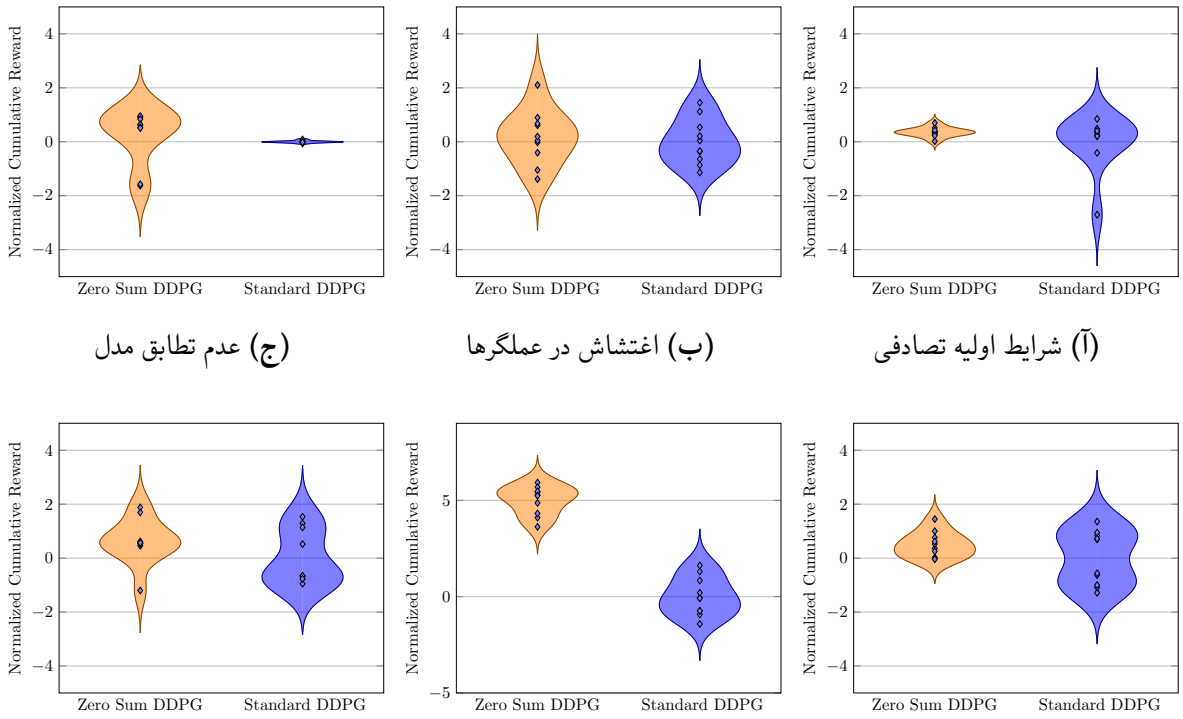
نویز حسگر

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.

۱-۳-۲ 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 ۱-۹: 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.

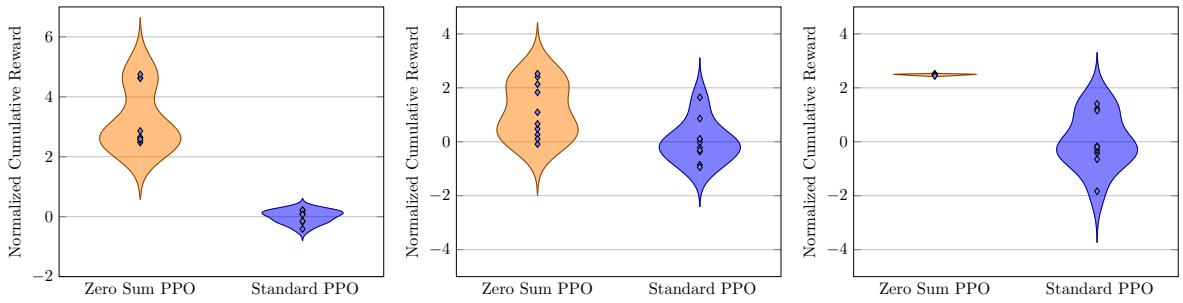
نتایج نشان می‌دهد که الگوریتم DDPG مبتنی بر بازی مجموع‌صفر در اکثر سناریوهای چالش‌برانگیز، عملکرد بهتری نسبت به نسخه استاندارد دارد. این برتری به خصوص در شرایط نویز حسگر و شرایط اولیه تصادفی مدل قابل توجه است، که نشان می‌دهد رویکرد چندعاملی توانایی بیشتری در مقابله با عدم قطعیت‌های سیستم دارد.

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

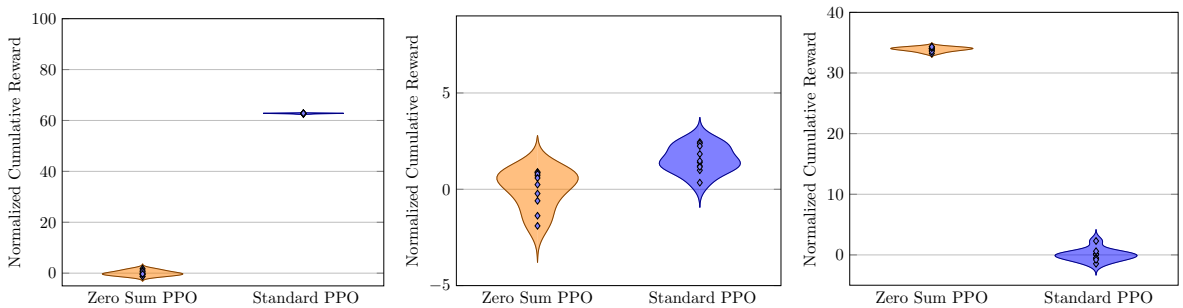
جدول ۱-۱: مقایسه عملکرد الگوریتم‌های تک‌عاملی DDPG و چندعاملی MA-DDPG در سناریوهای مختلف. مقادیر بهتر در هر دسته با رنگ پررنگ مشخص شده‌اند.

تحلیل نتایج جدول ۱-۱ نشان می‌دهد که الگوریتم چندعاملی MA-DDPG در اکثر سناریوها عملکرد بهتری نسبت به نسخه تک‌عاملی دارد. به طور خاص، در سناریوی اغتشاش در عملگرها، احتمال شکست در نسخه چندعاملی به میزان قابل توجهی کاهش یافته است (از 0.90 به 0.30). همچنین در سناریوی تأخیر زمانی، پاداش تجمعی در نسخه چندعاملی به میزان 40% بهبود یافته است (از -3.20 به -1.91). این بهبودها نشان‌دهنده مقاومت بیشتر الگوریتم چندعاملی در برابر شرایط نامطلوب است. با این حال، در برخی موارد مانند مجموع خطای مسیر، نسخه تک‌عاملی عملکرد بهتری داشته است.

۱-۳-۳ مقایسه الگوریتم‌های تک‌عاملی و چندعاملی PPO



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



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

شکل ۱-۱۰: مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی PPO در سناریوهای مختلف. نسخه بازی مجموع‌صفر در سناریوهای تأخیر زمانی و نویز حسگر برتری قابل توجهی نشان می‌دهد.

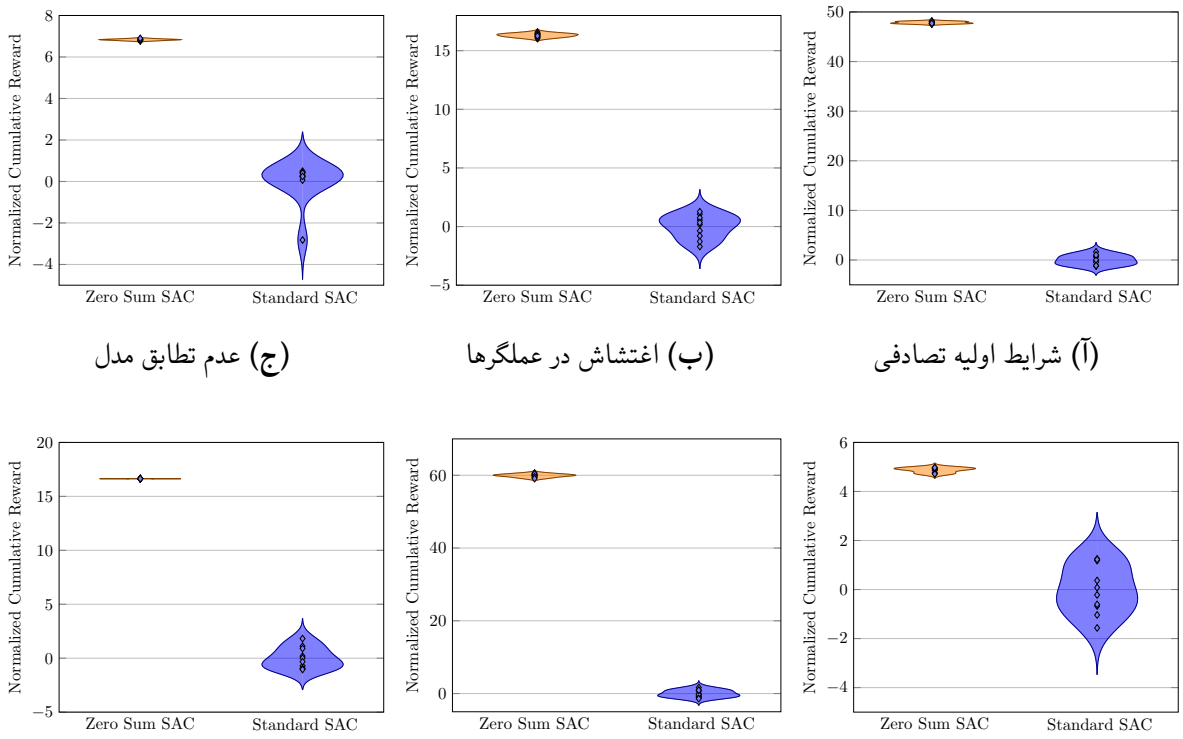
الگوریتم PPO در حالت بازی مجموع‌صفر در اکثر سناریوها عملکرد بهتری نشان می‌دهد، به خصوص در شرایط تأخیر زمانی و نویز حسگر. این می‌تواند نشان‌دهنده توانایی روش چندعاملی در مدیریت بهتر شرایط دارای عدم قطعیت در ورودی‌ها باشد. با این حال، تفاوت در برخی از سناریوها کمتر از DDPG است.

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

جدول ۱-۲: مقایسه عملکرد الگوریتم‌های تک‌عاملی PPO و چندعاملی MA-PPO در سناریوهای مختلف. مقادیر بهتر در هر مقایسه با رنگ پررنگ مشخص شده‌اند.

بررسی جدول ۱-۲ نشان می‌دهد که الگوریتم چندعاملی MA-PPO در سناریوهای شرایط اولیه تصادفی، اغتشاش در عملگرها و مشاهده ناقص عملکرد بهتری نسبت به نسخه تک‌عاملی دارد. به طور مشخص، در سناریوی شرایط اولیه تصادفی، نسخه چندعاملی به پاداش مثبت 0.46 دست یافته و احتمال شکست را به صفر کاهش داده است، در حالی که نسخه تک‌عاملی پاداش منفی -1.85 و احتمال شکست 0.70 را ثبت کرده است. با این حال، در سناریوی تأخیر زمانی، نسخه تک‌عاملی عملکرد بسیار بهتری داشته است. این نتایج نشان می‌دهد که الگوریتم PPO چندعاملی برای شرایط با عدم قطعیت اولیه بسیار مناسب است، اما در مواجهه با تأخیرهای زمانی آسیب‌پذیر می‌باشد.

۱-۳-۴ مقایسه الگوریتم‌های تک‌عاملی و چندعاملی SAC



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

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

شکل ۱-۱۱: مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی SAC در سناریوهای مختلف. هر دو نسخه عملکرد نسبتاً خوبی دارند، اما نسخه بازی مجموع‌صفر در شرایط عدم تطابق مدل و مشاهده ناقص برتری بیشتری نشان می‌دهد.

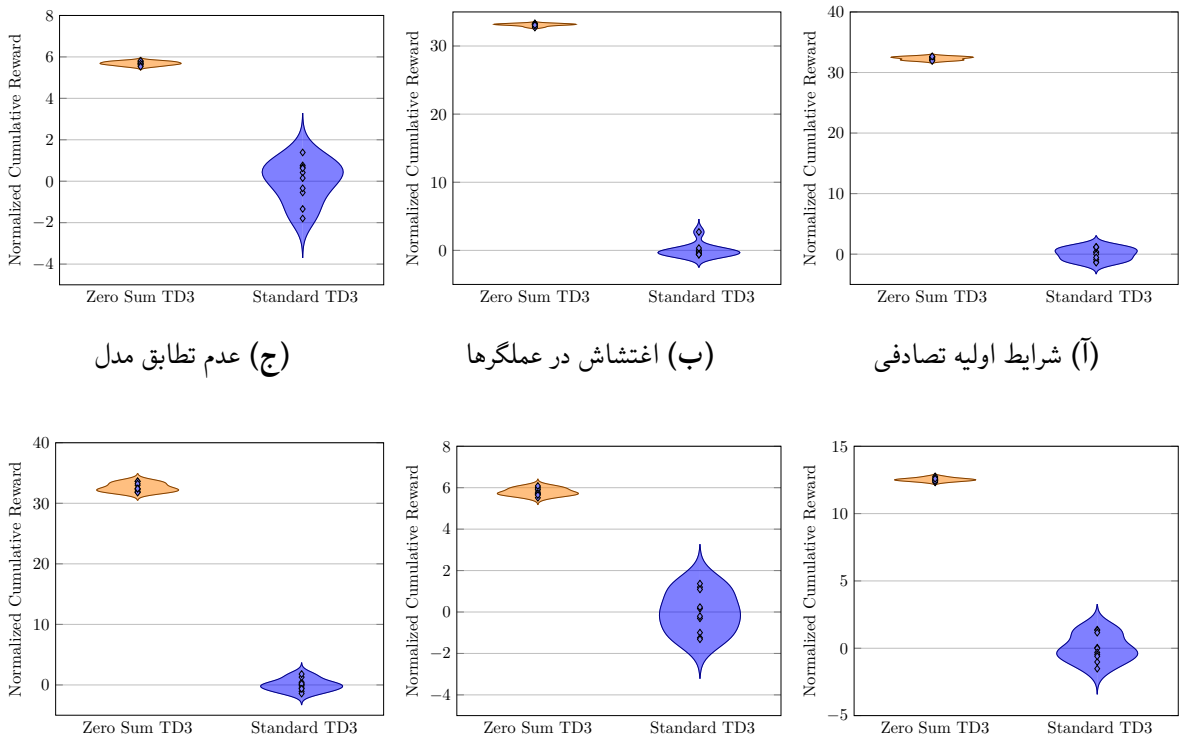
الگوریتم SAC در هر دو حالت عملکرد نسبتاً خوبی در سناریوهای مختلف نشان می‌دهد. این می‌تواند به دلیل استفاده از مکانیزم آنتروپی باشد که به صورت ذاتی اکتشاف بیشتری را تشویق می‌کند. با این حال، نسخه بازی مجموع‌صفر در تمامی سناریوها برتری معناداری دارد که نشان‌دهنده مقاومت بیشتر آن در شرایط با اطلاعات محدود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

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

جدول ۱-۳: مقایسه عملکرد الگوریتم‌های تک‌عاملی SAC و چندعاملی MA-SAC در سناریوهای مختلف. مقادیر بهتر در هر مقایسه با رنگ پررنگ مشخص شده‌اند.

تحلیل داده‌های جدول ۱-۳ نشان‌دهنده برتری قابل توجه نسخه چندعاملی MA-SAC در تمامی سناریوهای آزمایش است. به ویژه در سناریوی مشاهده ناقص، بهبود چشمگیری در پاداش تجمعی (از -3.63 به -0.44) و مجموع خطای مسیر (از 1.95 به 0.07) مشاهده می‌شود. همچنین در سناریوی تأخیر زمانی، نسخه چندعاملی توانسته احتمال شکست را از 1.00 به 0.00 کاهش دهد و همزمان پاداش تجمعی را به میزان قابل توجهی افزایش دهد (از -4.14 به -0.05). این نتایج نشان می‌دهد که رویکرد چندعاملی در الگوریتم SAC به طور مؤثری می‌تواند با شرایط عدم قطعیت و مشاهده ناقص مقابله کند و پایداری سیستم را افزایش دهد.

۱-۳-۵ مقایسه الگوریتم‌های تک‌عاملی و چندعاملی TD3



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

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

شکل ۱-۱۲: مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی TD3 در سناریوهای مختلف. نسخه بازی مجموع‌صفر در تمام سناریوها عملکرد بهتری را نشان می‌دهد، با برتری قابل توجه در سناریوهای اغتشاش در عملگرها و نویز حسگر.

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

سناریو	پاداش تجمعی		مجموع خطای مسیر		مجموع تلاش کنترلی		احتمال شکست	
	TD3	MA-TD3	TD3	MA-TD3	TD3	MA-TD3	TD3	MA-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

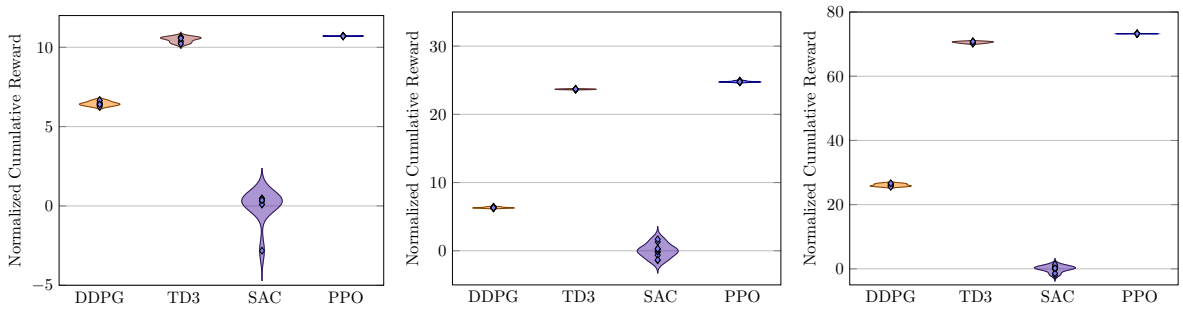
جدول ۱-۴: مقایسه عملکرد الگوریتم‌های تک‌عاملی TD3 و چندعاملی MA-TD3 در سناریوهای مختلف. مقادیر بهتر در هر مقایسه با رنگ پررنگ مشخص شده‌اند.

بررسی جدول ۱-۴ نشان می‌دهد که الگوریتم چندعاملی MA-TD3 در اکثر سناریوها (به جز نویز حسگر) عملکرد بهتری نسبت به نسخه تک‌عاملی دارد. به طور خاص، در سناریوی شرایط اولیه تصادفی، نسخه چندعاملی توانسته است پاداش تجمعی را به میزان قابل توجهی بهبود بخشد (از -2.95 به -0.26) و همچنین احتمال شکست را از 1.00 به 0.30 کاهش دهد. در سناریوی مشاهده ناقص نیز، نسخه چندعاملی پاداش تجمعی بالاتری کسب کرده است (0.71 در مقابل 0.21). با این حال، در سناریوی نویز حسگر، نسخه تک‌عاملی عملکرد بسیار بهتری داشته است، که نشان می‌دهد الگوریتم TD3 چندعاملی ممکن است نسبت به نویزهای سنسوری حساسیت بیشتری داشته باشد.

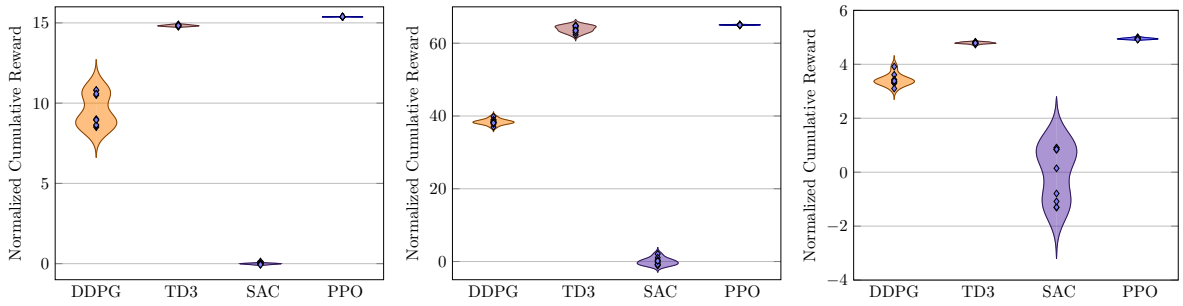
۱-۴ مقایسه جامع الگوریتم‌ها

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.

۱-۴-۱ مقایسه الگوریتم‌های تک‌عاملی



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

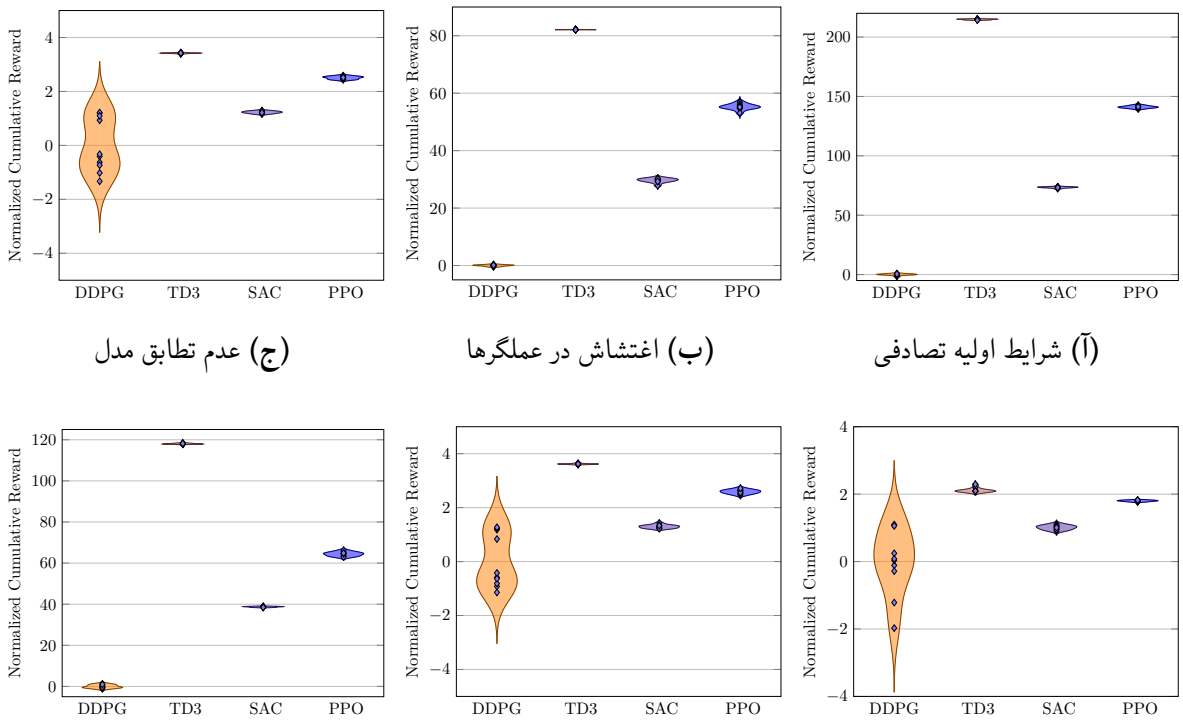


(د) مشاهده ناقص (ه) نویز حسگر (و) تأخیر زمانی

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.

۱-۴-۲ مقایسه الگوریتم‌های چندعاملی



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

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

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.

۱-۵ 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 مجموع‌صفر 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 مجموع‌صفر، 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 مجموع‌صفر 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 مجموع‌صفر، مقاومت و انعطاف‌پذیری بیشتری نشان می‌دهند.

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

پاداش تجمعی

سناریو	DDPG	PPO	SAC	TD3
شرایط اولیه تصادفی	−0.27	0.61	−0.76	0.56
اغتشاش در عملگرها	−0.38	0.61	−0.72	0.55
عدم تطابق مدل	−0.84	0.58	−2.98	0.51
مشاهده ناقص	−0.88	0.36	−3.65	0.23
نویز حسگر	−0.85	0.58	−2.90	0.52
تأخیر زمانی	−0.76	0.61	−2.98	0.48

مجموع خطای مسیر

سناریو	DDPG	PPO	SAC	TD3
شرایط اولیه تصادفی	3.30	2.56	8.06	0.72
اغتشاش در عملگرها	3.74	2.58	7.91	0.77
عدم تطابق مدل	10.87	3.06	17.12	1.09
مشاهده ناقص	8.18	3.34	15.47	1.77
نویز حسگر	11.04	3.08	16.81	1.02
تأخیر زمانی	8.95	2.27	15.70	0.81

مجموع تلاش کنترلی

سناریو	DDPG	PPO	SAC	TD3
شرایط اولیه تصادفی	5.11	0.77	1.76	3.31
اغتشاش در عملگرها	4.89	0.77	1.71	3.07
عدم تطابق مدل	5.48	0.86	2.37	4.32
مشاهده ناقص	5.37	1.03	2.33	4.10
نویز حسگر	5.48	0.86	2.37	4.30
تأخیر زمانی	5.51	0.76	2.11	5.12

احتمال شکست

سناریو	DDPG	PPO	SAC	TD3
شرایط اولیه تصادفی	0.00	0.00	0.00	0.00
اغتشاش در عملگرها	0.00	0.00	0.00	0.00
عدم تطابق مدل	0.00	0.00	1.00	0.00
مشاهده ناقص	0.00	0.00	1.00	0.00
نویز حسگر	0.00	0.00	1.00	0.00
تأخیر زمانی	0.00	0.00	1.00	0.00

جدول ۱-۵: الگوریتم‌های تک‌عاملی

پاداش تجمعی

سناریو	DDPG	PPO	SAC	TD3
شرایط اولیه تصادفی	−0.41	0.34	−0.02	0.74
اغتشاش در عملگرها	−0.44	0.35	−0.02	0.73
عدم تطابق مدل	−0.63	0.38	−0.13	0.75
مشاهده ناقص	−1.52	0.40	−0.44	0.71
نویز حسگر	−0.60	0.37	−0.12	0.75
تأخیر زمانی	−1.19	0.17	−0.05	0.67

مجموع خطای مسیر

سناریو	DDPG	PPO	SAC	TD3
شرایط اولیه تصادفی	4.42	4.30	4.02	1.22
اغتشاش در عملگرها	4.39	4.38	4.01	1.26
عدم تطابق مدل	8.85	3.57	4.78	1.25
مشاهده ناقص	9.65	2.44	5.17	1.09
نویز حسگر	9.12	3.58	4.66	1.25
تأخیر زمانی	6.73	4.53	4.12	1.21

مجموع تلاش کنترلی

سناریو	DDPG	PPO	SAC	TD3
شرایط اولیه تصادفی	5.11	0.77	1.76	3.31
اغتشاش در عملگرها	4.89	0.77	1.71	3.07
عدم تطابق مدل	5.48	0.86	2.37	4.32
مشاهده ناقص	5.37	1.03	2.33	4.10
نویز حسگر	5.48	0.86	2.37	4.30
تأخیر زمانی	5.51	0.76	2.11	5.12

احتمال شکست

سناریو	DDPG	PPO	SAC	TD3
شرایط اولیه تصادفی	0.00	0.00	0.00	0.00
اغتشاش در عملگرها	0.00	0.00	0.00	0.00
عدم تطابق مدل	0.00	0.00	1.00	0.00
مشاهده ناقص	0.00	0.00	1.00	0.00
نویز حسگر	0.00	0.00	1.00	0.00
تأخیر زمانی	0.00	0.00	1.00	0.00

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.

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