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.

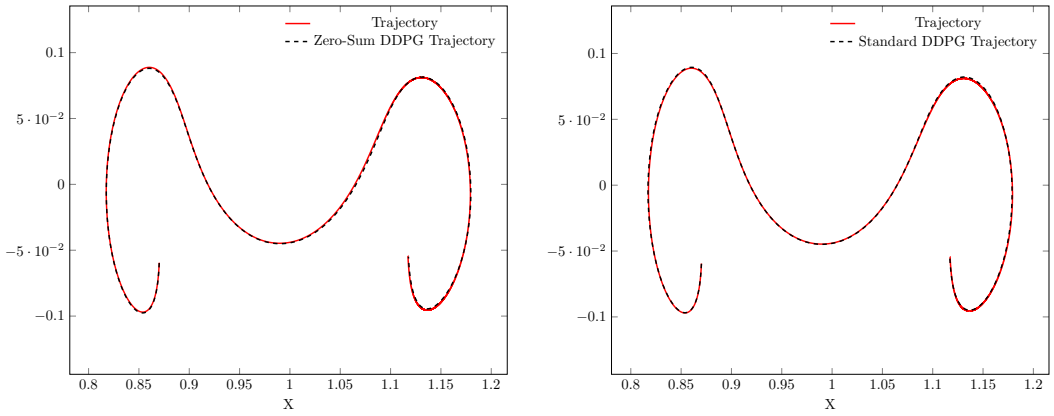
فرآیند آموزش برای هر الگوریتم شامل ۱ میلیون گام تعامل با محیط بوده و اندازه بافر تجربه برای الگوریتم‌های DDPG، SAC و TD3 برابر با ۱۰۰ هزار نمونه تنظیم شده است. هر الگوریتم با ۱۰ مقداردهی اولیه متفاوت آموزش داده شده تا از پایداری نتایج اطمینان حاصل شود.

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.

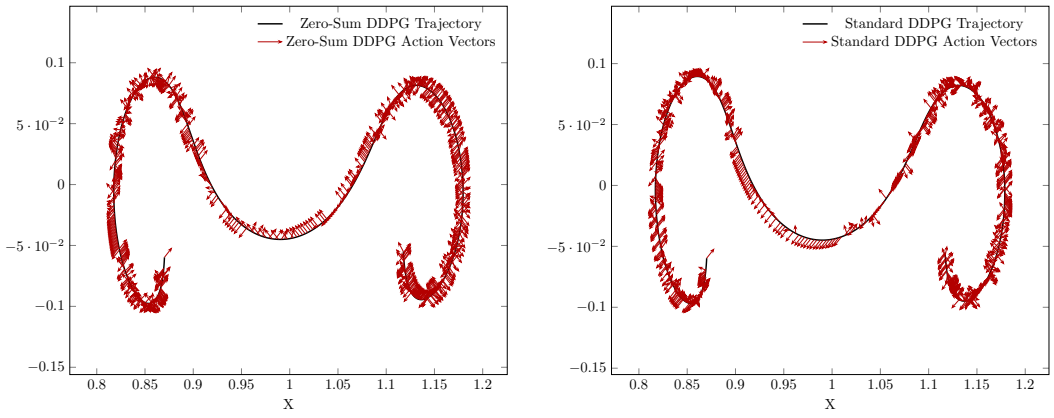
۱-۲-۱ الگوریتم DDPG

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.

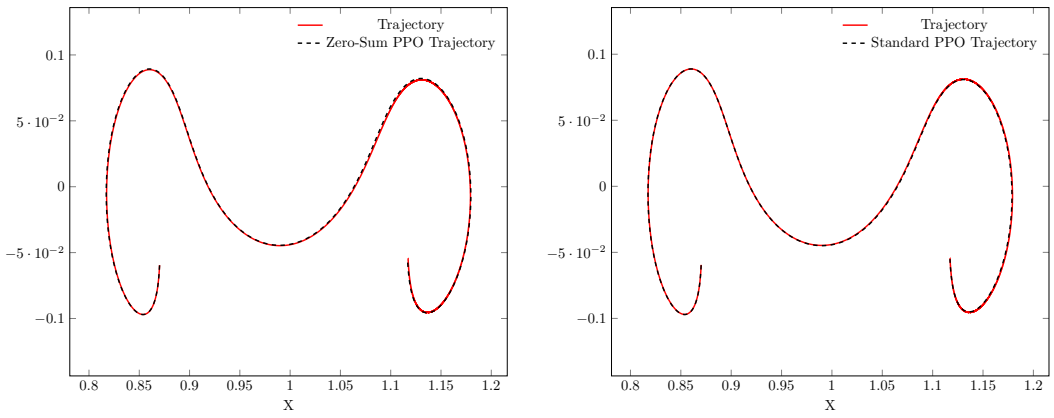


(آ) DDPG استاندارد (ب) DDPG بازی مجموع‌صفر

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

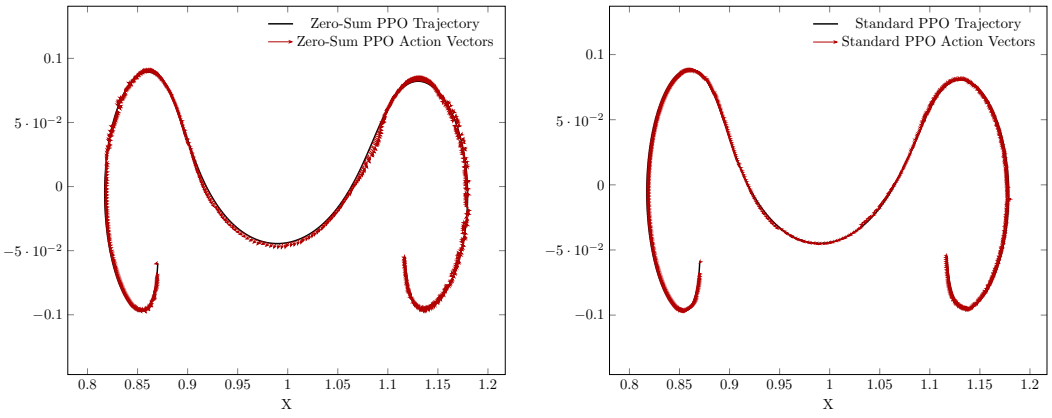
۱-۲-۲ الگوریتم PPO

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.

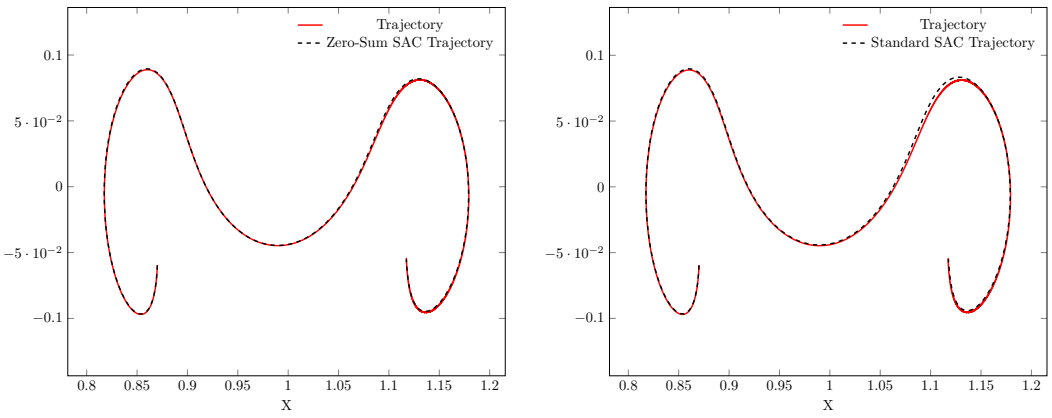


(آ) PPO استاندارد (ب) PPO بازی مجموع‌صفر

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

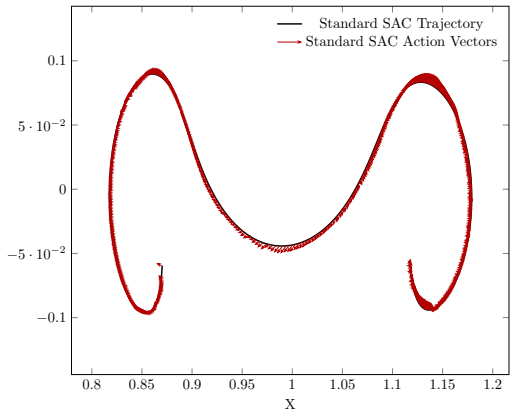
۱-۲-۳ الگوریتم SAC

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.

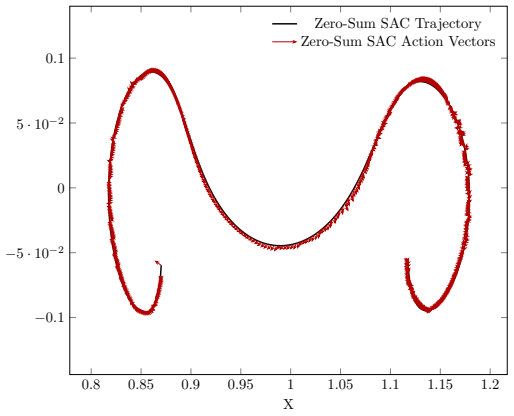


(آ) SAC استاندارد (ب) SAC بازی مجموع‌صفر

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



(آ) SAC استاندارد

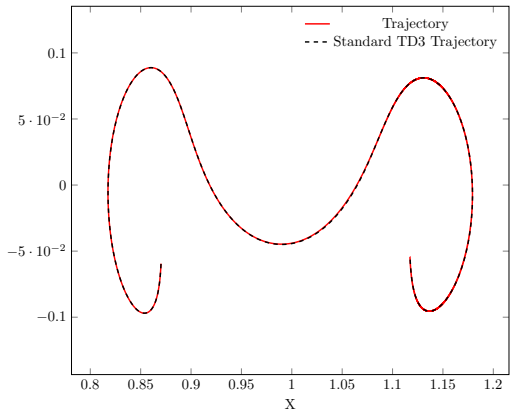


(ب) SAC بازی مجموع‌صفر

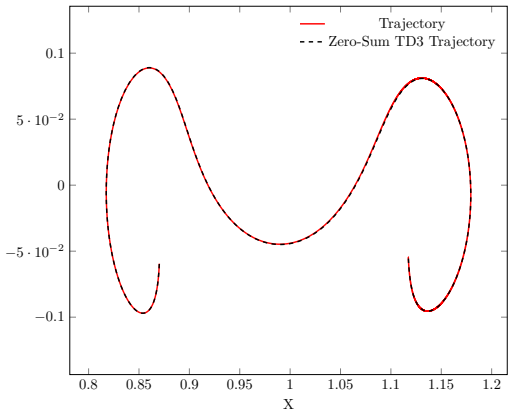
شکل ۱-۶: مقایسه مسیر و فرمان پیشران دو الگوریتم تک‌عاملی و چندعاملی SAC.

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.

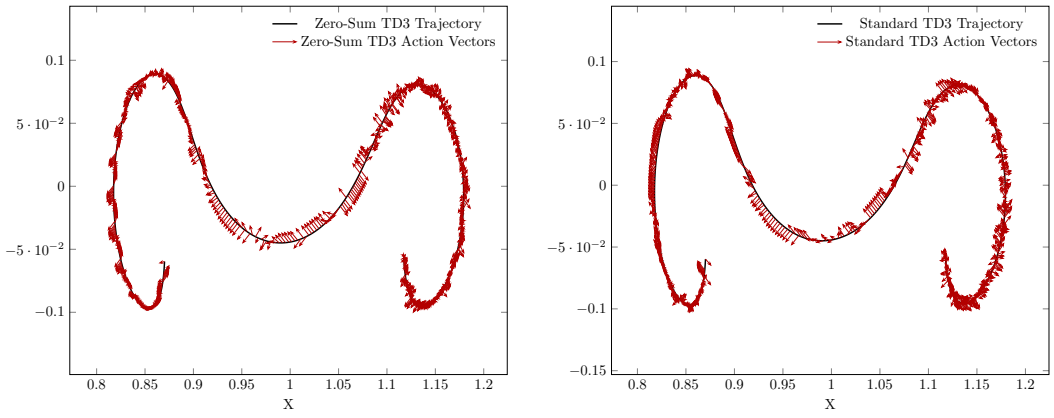


(آ) TD3 استاندارد



(ب) TD3 بازی مجموع‌صفر

شکل ۱-۷: مقایسه مسیر طی شده در دو الگوریتم تک‌عاملی و چندعاملی TD3.



(آ) TD3 استاندارد (ب) TD3 بازی مجموع‌صفر

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

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.

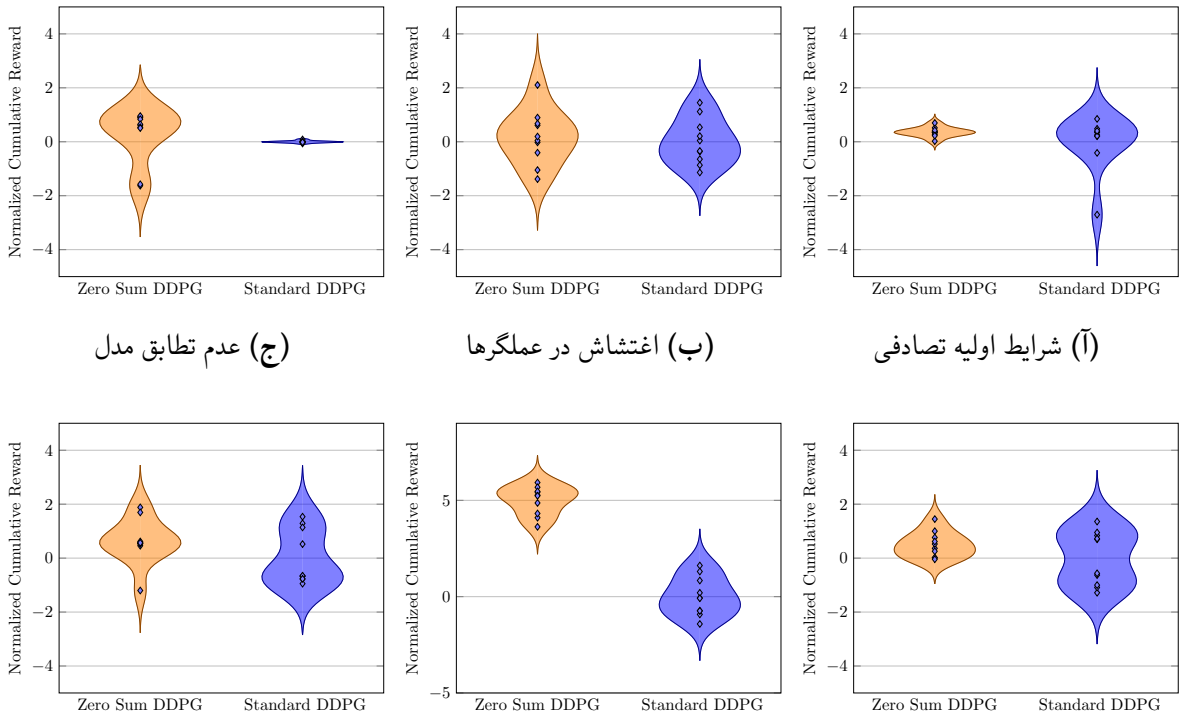
نویز حسگر

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

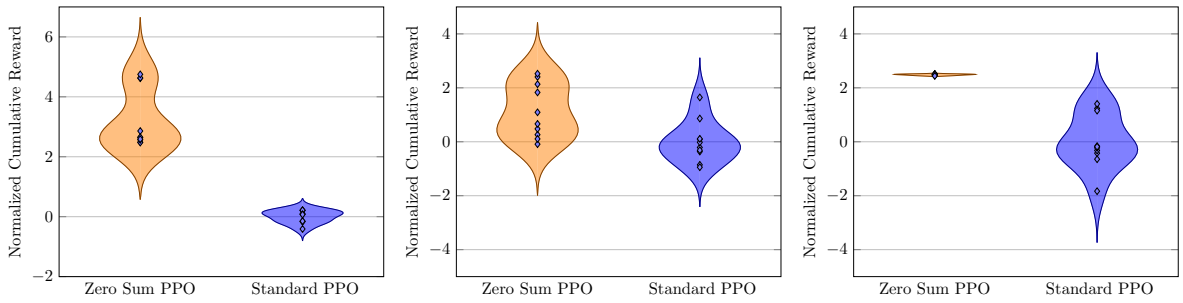
شکل ۱-۹: مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی DDPG در سناریوهای مختلف.

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.

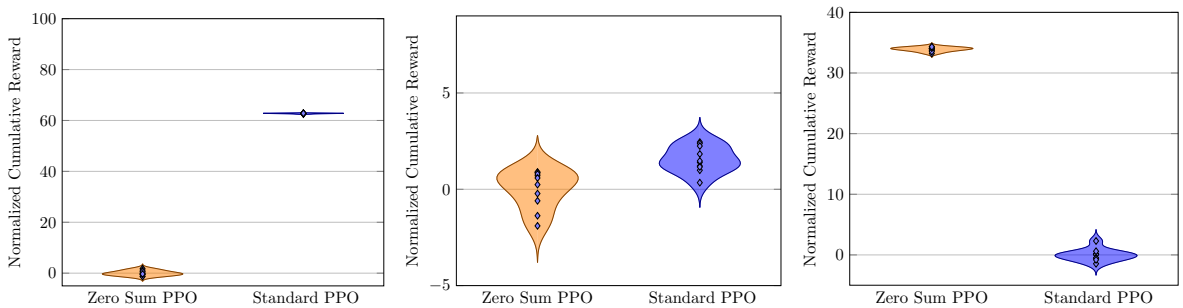
سناریو	پاداش تجمعی		مجموع خطای مسیر		مجموع تلاش کنترلی		احتمال شکست	
	DDPG	MA-DDPG	DDPG	MA-DDPG	DDPG	MA-DDPG	DDPG	MA-DDPG
شرایط اولیه تصادفی	−4.17	−3.63	0.40	0.63	5.60	5.60	1.00	1.00
اغتشاش در عملگرها	−1.93	−1.96	7.56	7.94	5.60	5.59	0.90	0.30
عدم تطابق مدل	−3.24	−2.70	0.70	0.76	5.57	5.57	1.00	1.00
مشاهده 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: جدول پارامترها و مقادیر پیش‌فرض الگوریتم DDPG

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

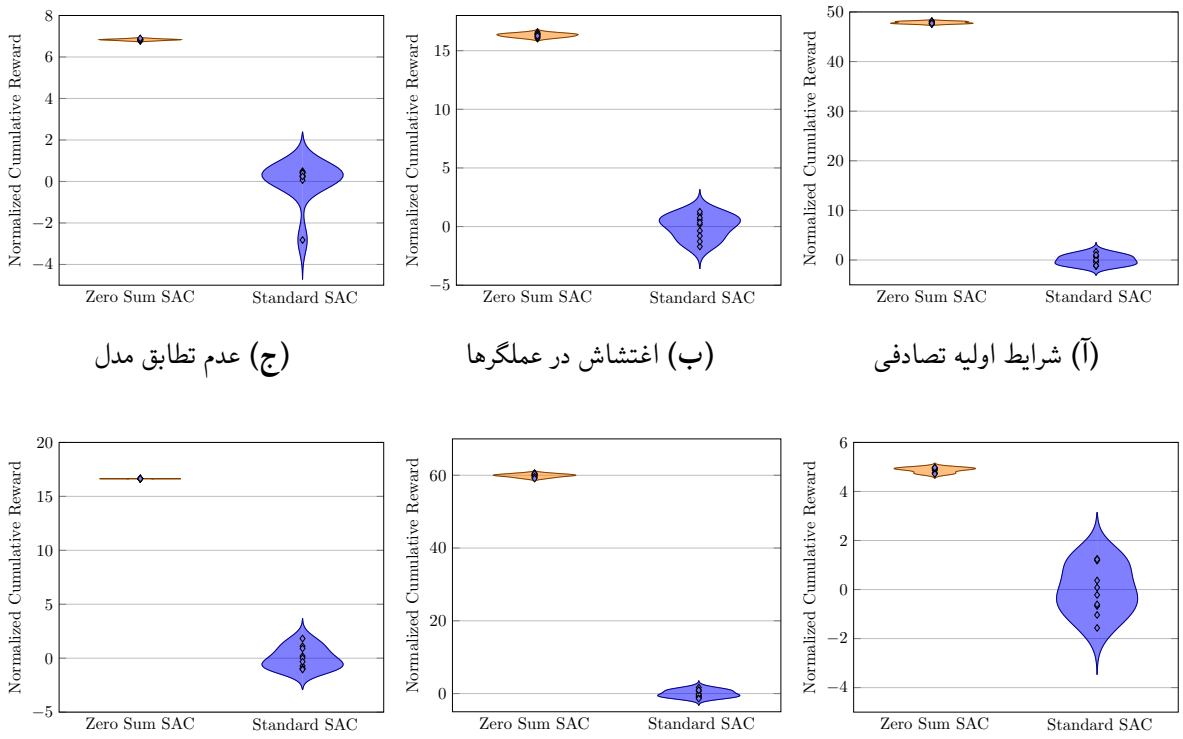
شکل ۱-۱۰: مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی PPO در سناریوهای مختلف.

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: جدول پارامترها و مقادیر پیش‌فرض الگوریتم PPO

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

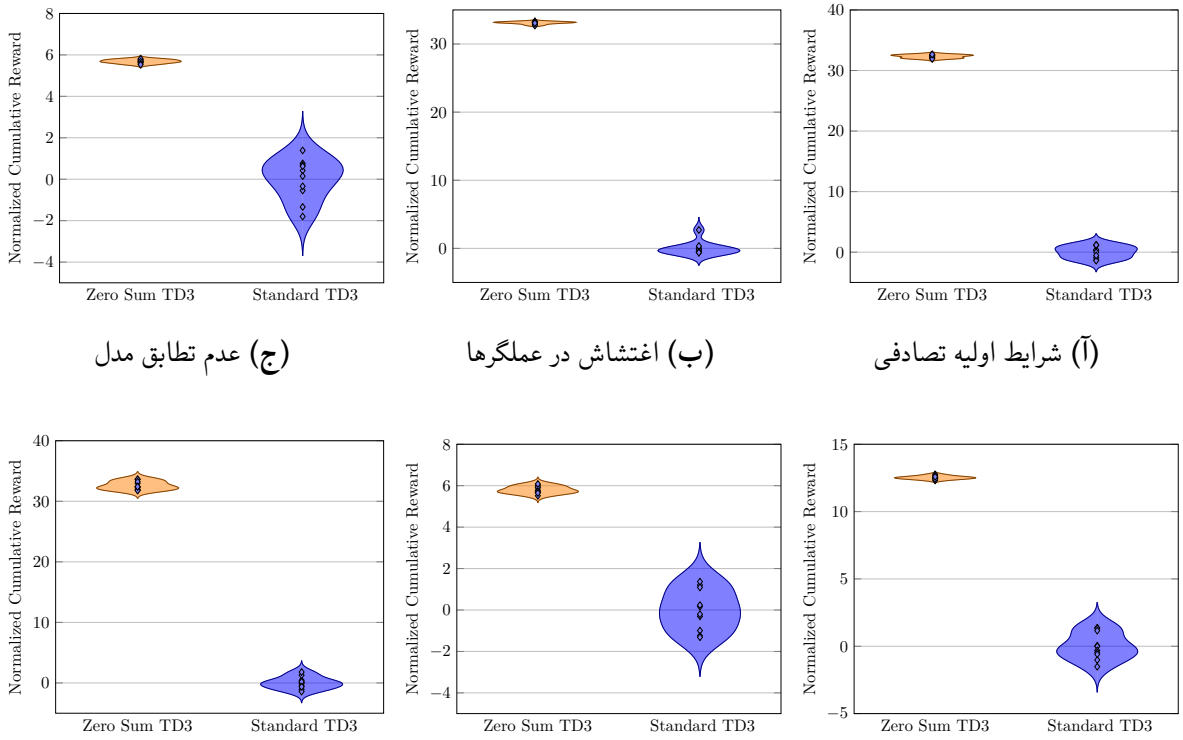
شکل ۱-۱۱: مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی SAC در سناریوهای مختلف.

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								
اغتشاش در عملگرها								
عدم تطابق مدل								
مشاهده ناقص								
نویز حسگر								
تأخیر زمانی								

جدول ۱-۳: جدول پارامترها و مقادیر پیش‌فرض الگوریتم SAC

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

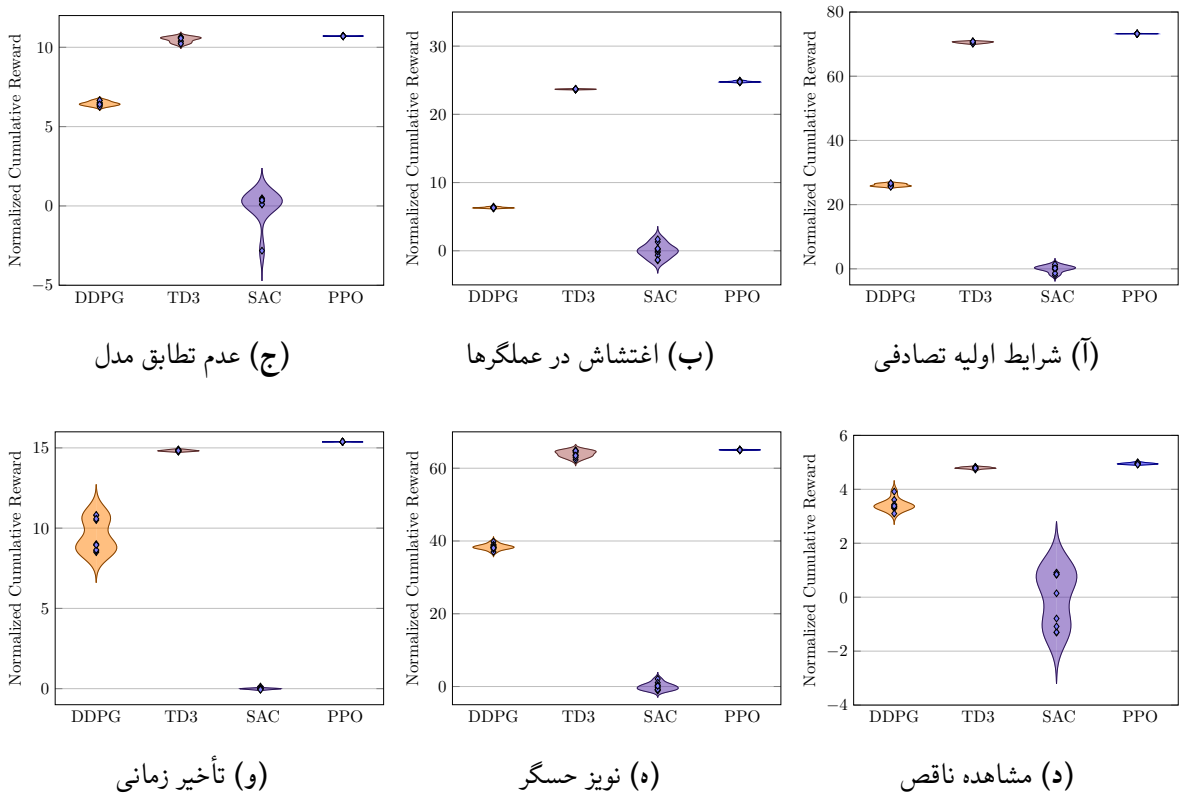
شکل ۱-۱۲: مقایسه مجموع پاداش دو الگوریتم تک‌عاملی و چندعاملی TD3 در سناریوهای مختلف.

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.

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



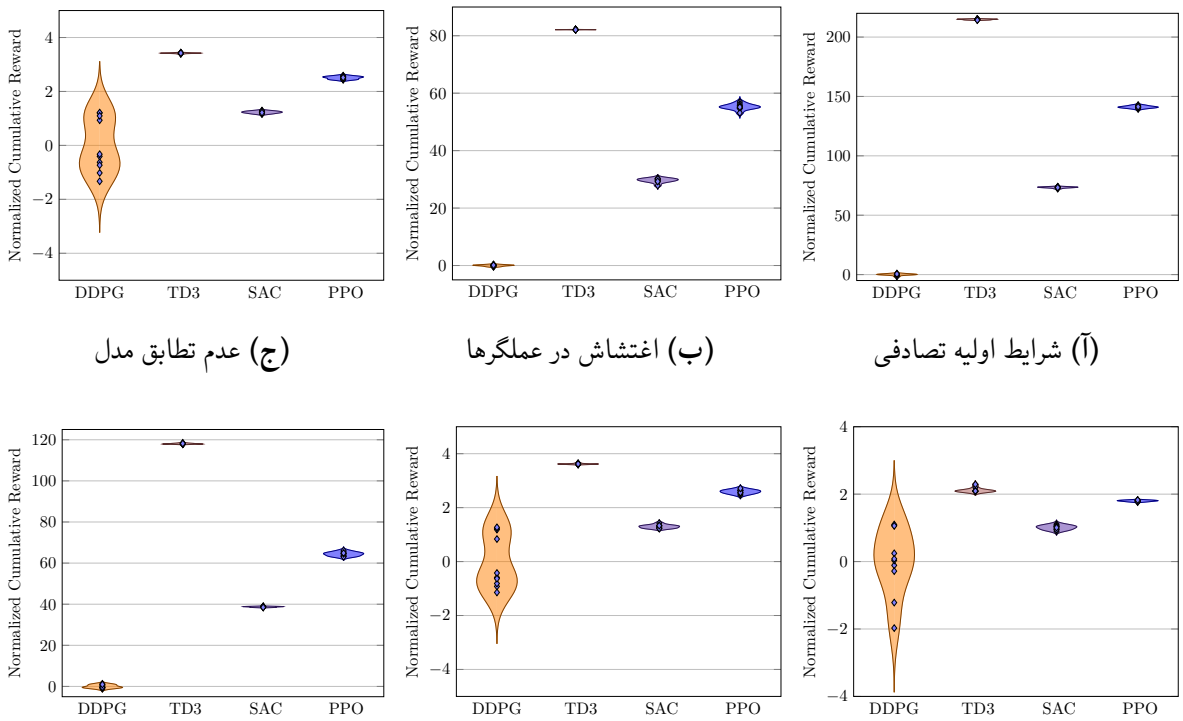
(آ) شرایط اولیه تصادفی (ب) اغتشاش در عملگرها (ج) عدم تطابق مدل

(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.

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								
اغتشاش در عملگرها								
عدم تطابق مدل								
مشاهده ناقص								
نویز حسگر								
تأخیر زمانی								

جدول ۱-۴: جدول پارامترها و مقادیر پیش‌فرض الگوریتم PPO

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.

Bibliography

[1] J. Achiam. Spinning Up in Deep Reinforcement Learning. 2018.

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

[3] M. A. Vavrina, J. A. Englander, S. M. Phillips, and K. M. Hughes. Global, multi-objective trajectory optimization with parametric spreading. In *AAS AIAA Astrodynamics Specialist Conference 2017*, 2017. Tech. No. GSFC-E-DAA-TN45282.

[4] C. Ocampo. Finite burn maneuver modeling for a generalized spacecraft trajectory design and optimization system. *Annals of the New York Academy of Sciences*, 1017:210–233, 2004.

[5] B. G. Marchand, S. K. Scarritt, T. A. Pavlak, and K. C. Howell. A dynamical approach to precision entry in multi-body regimes: Dispersion manifolds. *Acta Astronautica*, 89:107–120, 2013.

[6] A. F. Haapala and K. C. Howell. A framework for constructing transfers linking periodic libration point orbits in the spatial circular restricted three-body problem. *International Journal of Bifurcation and Chaos*, 26(05):1630013, 2016.

[7] B. Gaudet, R. Linares, and R. Furfaro. Six degree-of-freedom hovering over an asteroid with unknown environmental dynamics via reinforcement learning. In *20th AIAA Scitech Forum*, Orlando, Florida, 2020.

[8] B. Gaudet, R. Linares, and R. Furfaro. Terminal adaptive guidance via reinforcement meta-learning: Applications to autonomous asteroid close-proximity operations. *Acta Astronautica*, 171:1–13, 2020.

[9] A. Rubinsztejn, R. Sood, and F. E. Laipert. Neural network optimal control in astrodynamics: Application to the missed thrust problem. *Acta Astronautica*, 176:192–203, 2020.

[10] T. A. Estlin, B. J. Bornstein, D. M. Gaines, R. C. Anderson, D. R. Thompson, M. Burl, R. Castaño, and M. Judd. Aegis automated science targeting for the mer opportunity rover. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3:1–19, 2012.

[11] R. Francis, T. Estlin, G. Doran, S. Johnstone, D. Gaines, V. Verma, M. Burl, J. Frydenvang, S. Montano, R. Wiens, S. Schaffer, O. Gasnault, L. Deflores, D. Blaney, and B. Bornstein. Aegis autonomous targeting for chemcam on mars science laboratory: Deployment and results of initial science team use. *Science Robotics*, 2, 2017.

[12] S. Higa, Y. Iwashita, K. Otsu, M. Ono, O. Lamarre, A. Didier, and M. Hoffmann. Vision-based estimation of driving energy for planetary rovers using deep learning and terramechanics. *IEEE Robotics and Automation Letters*, 4:3876–3883, 2019.

[13] B. Rothrock, J. Papon, R. Kennedy, M. Ono, M. Heverly, and C. Cunningham. Spoc: Deep learning-based terrain classification for mars rover missions. In *AIAA Space and Astronautics Forum and Exposition, SPACE 2016*. American Institute of Aeronautics and Astronautics Inc, AIAA, 2016.

[14] K. L. Wagstaff, G. Doran, A. Davies, S. Anwar, S. Chakraborty, M. Cameron, I. Daubar, and C. Phillips. Enabling onboard detection of events of scientific interest for the europa clipper spacecraft. In *25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2191–2201, Anchorage, Alaska, 2019.

[15] B. Dachwald. Evolutionary neurocontrol: A smart method for global optimization of low-thrust trajectories. In *AIAA/AAS Astrodynamics Specialist Conference and Exhibit*, pages 1–16, Providence, Rhode Island, 2004.

[16] S. D. Smet and D. J. Scheeres. Identifying heteroclinic connections using artificial neural networks. *Acta Astronautica*, 161:192–199, 2019.

[17] N. L. O. Parrish. *Low Thrust Trajectory Optimization in Cislunar and Translunar Space*. PhD thesis, University of Colorado Boulder, 2018.

[18] N. Heess, D. TB, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. M. A. Eslami, M. A. Riedmiller, and D. Silver. Emergence of locomotion behaviours in rich environments. *CoRR*, abs/1707.02286, 2017.

[19] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den

Driessche, T. Graepel, and D. Hassabis. Mastering the game of go without human knowledge. *Nature*, 550, 2017.

[20] R. Furfaro, A. Scorsoglio, R. Linares, and M. Massari. Adaptive generalized zem-zev feedback guidance for planetary landing via a deep reinforcement learning approach. *Acta Astronautica*, 171:156–171, 2020.

[21] B. Gaudet, R. Linares, and R. Furfaro. Deep reinforcement learning for six degrees of freedom planetary landing. *Advances in Space Research*, 65:1723–1741, 2020.

[22] B. Gaudet, R. Furfaro, and R. Linares. Reinforcement learning for angle-only intercept guidance of maneuvering targets. *Aerospace Science and Technology*, 99, 2020.

[23] D. Guzzetti. Reinforcement learning and topology of orbit manifolds for station-keeping of unstable symmetric periodic orbits. In *AAS/AIAA Astrodynamics Specialist Conference*, Portland, Maine, 2019.

[24] J. A. Reiter and D. B. Spencer. Augmenting spacecraft maneuver strategy optimization for detection avoidance with competitive coevolution. In *20th AIAA Scitech Forum*, Orlando, Florida, 2020.

[25] A. Das-Stuart, K. C. Howell, and D. C. Folta. Rapid trajectory design in complex environments enabled by reinforcement learning and graph search strategies. *Acta Astronautica*, 171:172–195, 2020.

[26] D. Miller and R. Linares. Low-thrust optimal control via reinforcement learning. In *29th AAS/AIAA Space Flight Mechanics Meeting*, Ka'anapali, Hawaii, 2019.

[27] C. J. Sullivan and N. Bosanac. Using reinforcement learning to design a low-thrust approach into a periodic orbit in a multi-body system. In *20th AIAA Scitech Forum*, Orlando, Florida, 2020.

[28] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb. 2015.

[29] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1889–1897, 2015.

[30] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1928–1937, 2016. arXiv:1602.01783.

[31] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning, 2019.

[32] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint*, arXiv:1707.06347, 2017.

[33] S. Fujimoto, H. V. Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1587–1596, 2018.

[34] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1861–1870, 2018.

[35] A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 1179–1191, 2020.

[36] K. Prudencio, J. L. Xiang, and A. T. Cemgil. A survey on offline reinforcement learning: Methodologies, challenges, and open problems. *arXiv preprint*, arXiv:2203.01387, 2022.

[37] J. GarcÃa and F. FernÃ¡ndez. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.

[38] F. Ghazalpour, S. Samangouei, and R. Vaughan. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys*, 54(12):1–35, 2021.

[39] K. Song, J. Zhu, Y. Chow, D. Psomas, and M. Wainwright. A survey on multi-agent reinforcement learning: Foundations, advances, and open challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. In press, arXiv:2401.01234.

[40] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. V. D. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach,

K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[41] O. Vinyals, I. Babuschkin, W. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[42] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, S. Legg, and K. Kavukcuoglu. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1407–1416, 2018.

[43] M. Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the 10th International Conference on Machine Learning (ICML)*, pages 330–337, 1993.

[44] L. Panait and S. Luke. Cooperative multi-agent learning: The state of the art. *Autonomous Robots*, 8(3):355–377, 2005.

[45] L. Buşoniu, R. Babuška, and B. D. Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(2):156–172, 2008.

[46] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 6379–6390, 2017.

[47] P. Sunehag, G. Lever, A. Gruslys, W. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel. Value-decomposition networks for cooperative multi-agent learning. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2018. arXiv:1706.05296.

[48] T. Rashid, M. Samvelyan, C. S. de Witt, G. Farquhar, J. Foerster, and S. Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 4292–4301, 2018.

[49] M. Samvelyan, T. Rashid, C. S. de Witt, G. Farquhar, J. Foerster, N. Nardelli, T. G. J. Rudner, and et al. The starcraft multi-agent challenge. *arXiv preprint*, arXiv:1902.04043, 2019.

[50] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5887–5896, 2019.

[51] A. Mahajan, T. Rashid, M. Samvelyan, and S. Whiteson. Maven: Multi-agent variational exploration. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 7611–7622, 2019.

[52] T. Wang, Y. Jiang, T. Da, W. Zhang, and J. Wang. Roma: Multi-agent reinforcement learning with emergent roles. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 9876–9886, 2020.

[53] K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of RL and Control*, 2021. arXiv:2106.05230.

[54] A. Mitriakov, P. Papadakis, J. Kerdreux, and S. Garlatti. Reinforcement learning based, staircase negotiation learning: Simulation and transfer to reality for articulated tracked robots. *IEEE Robotics & Automation Magazine*, 28(4):10–20, 2021.

[55] Y. Yu et al. Heterogeneous-agent reinforcement learning: An overview. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. In press, arXiv:2203.00596.

[56] D. Vallado and W. McClain. *Fundamentals of Astrodynamics and Applications*. Fundamentals of Astrodynamics and Applications. Microcosm Press, 2001.

[57] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. Pmlr, 2014.

[58] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[59] S. Fujimoto, H. van Hoof, and D. Meger. Addressing function approximation error in actor-critic methods, 2018.

[60] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[61] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *CoRR*, abs/1801.01290, 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.

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