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

شبیه‌سازی عامل درمحیط سه جسمی

در این فصل، فرآیند شبیه‌سازی عامل هوشمند کنترل‌کننده فضاپیما در محیط دینامیکی سه جسمی بررسی شده است. در بخش ۱-۱ به طراحی و در بخش ۱-۲ به شبیه‌سازی عامل هدایت‌کننده مبتنی بر یادگیری تقویتی است پرداخته شده است. این عامل طراحی و شبیه‌سازی شده باید توانایی این را داشته باشد که فضاپیما را به‌طور مؤثر به سمت اهداف تعیین‌شده هدایت کند، در حالی که محدودیت‌هایی نظیر مصرف سوخت و وجود اغتشاش دارد.

۱-۱ طراحی عامل

در این زیربخش، معماری عامل هوشمند کنترل‌کننده فضاپیما در محیط سه‌جسمی شرح داده شده است. این معماری شامل تعریف فضای حالت، عمل و تابع پاداش است.

۱-۱-۱ فضای حالت

فضای حالت[1] در این پژوهش به‌گونه‌ای طراحی شده است که وضعیت دینامیکی فضاپیما را نسبت به یک مسیر و سرعت مرجع مشخص می‌کند. این فضا شامل اختلاف‌های موقعیت و سرعت از مسیر و سرعت مرجع است و به‌صورت زیر تعریف شده است:

$$S = \{\delta x, \delta y, \delta \dot{x}, \delta \dot{y}\}$$

که در آن:

[1] State Space

- $\delta x, \delta y$: اختلاف موقعیت فضاپیما نسبت به مسیر مرجع در محورهای x, y .
- $\delta \dot{x}, \delta \dot{y}$: اختلاف سرعت فضاپیما نسبت به سرعت مرجع در محورهای x, y .

هر یک از این متغیرها به‌طور مستقل وضعیت فضاپیما را در یک جهت خاص توصیف می‌کنند و امکان تحلیل دقیق انحرافات را فراهم می‌سازند. استفاده از اختلاف‌های موقعیت و سرعت به جای مقادیر مطلق، به دلایل زیر انجام شده است:

- **تمرکز بر انحرافات:** هدف اصلی سیستم کنترلی، کاهش انحرافات از مسیر و سرعت مطلوب است. با استفاده از اختلاف‌ها، کنترلر می‌تواند به طور مستقیم بر این انحرافات اثر بگذارد و نیازی به محاسبه مقادیر مطلق موقعیت و سرعت ندارد.
- **سازگاری با یادگیری تقویتی:** در الگوریتم‌های یادگیری تقویتی، فضاهای حالت مبتنی بر اختلاف معمولاً دامنه محدودتری دارند که فرآیند یادگیری را سریع‌تر و پایدارتر می‌کند.

۱-۱-۲ فضای عمل

فضای عمل[۲] فضاپیما با پیشران‌کم مجموعه‌ای از عمل‌های پیوسته است که فضاپیما می‌تواند در محیط شبیه‌سازی انجام دهد. این فضا به‌گونه‌ای طراحی شده که امکان اعمال نیرو در جهت‌های مشخص و با مقادیر متناسب با توان واقعی فضاپیماها فراهم شود. به‌طور خاص، فضای اقدام شامل موارد زیر است:

- **نیروی اعمال‌شده در جهت** x: این متغیر پیوسته، مقدار نیرویی را که در جهت محور x به فضاپیما وارد می‌شود، تعیین می‌کند. دامنه این نیرو بر اساس توان پیشرانه‌های موجود در فضاپیماهای واقعی انتخاب شده است. به عبارت دیگر، اگر حداکثر نیروی قابل اعمال در جهت x برابر با $f_{x,\max}$ باشد، این متغیر می‌تواند مقادیری در بازه $[-f_{x,\max}, f_{x,\max}]$ داشته باشد.
- **نیروی اعمال‌شده در جهت** y: این متغیر پیوسته، مقدار نیرویی را که در جهت محور y به فضاپیما وارد می‌شود، مشخص می‌کند. مشابه جهت x، دامنه این نیرو نیز بر اساس توان پیشرانه‌های موجود تعیین شده و می‌تواند در بازه $[-f_{y,\max}, f_{y,\max}]$ قرار گیرد.

انتخاب این نیروها بر اساس ویژگی‌های واقعی فضاپیماها، به‌ویژه توان و محدودیت‌های پیشرانه‌های آن‌ها، صورت گرفته است. این امر اطمینان می‌دهد که شبیه‌سازی تا حد ممکن به شرایط واقعی نزدیک باشد و نتایج

[۲] Action Space

به‌دست‌آمده قابلیت تعمیم به کاربردهای عملی را داشته باشند. همچنین، تعریف فضای اقدام به‌صورت پیوسته، امکان کنترل دقیق و انعطاف‌پذیر بر حرکت فضاپیما را فراهم می‌کند، که برای دستیابی به اهداف کنترلی در محیط‌های دینامیکی پیچیده ضروری است. به‌طور خلاصه، فضای اقدام به‌صورت زیر تعریف می‌شود:

$$a = \{f_x, f_y \mid f_x \in [-f_{x,\max}, f_{x,\max}],\ f_y \in [-f_{y,\max}, f_{y,\max}]\}$$

انطباق بازه‌ی فضای عمل با داده‌های واقعی

برای هم‌تراز کردن شبیه‌سازی با سخت‌افزارهای واقعی، از بیشینه‌ی نیروی بی‌بُعدِ پیشران‌ها استفاده می‌شود. جدول زیر نمونه‌هایی از فضاپیماهای مجهز به پیشران‌های یونی/الکتریکی را نشان می‌دهد که مبنای انتخاب بازه‌ی نیروی عمل قرارگرفته شده‌اند. با توجه به برداری‌بودن عمل $\boldsymbol{a} = [f_x\ f_y]$، کران‌ها را به دو صورت اعمال شده است:

$$|a| \leqslant f_{\text{nondim max}}, \quad \text{یا} \quad f_{x,\max} = f_{y,\max} = f_{\text{nondim max}}.$$

با استناد به جدول ۱-۱، مقدار نمونه‌ی 4×10^{-2} شبیه‌سازی شده با Psyche هم‌مرتبه و کمتر از DS1 است که باعث شده‌است بازه‌ی عمل را در چارچوب پیشران‌های کم‌تراست واقع‌گرایانه نگه داشته شود.

جدول ۱-۱: قابلیت‌های بی‌بعد پیشران‌کم‌تراستِ فضاپیماهای مختلف در سامانه‌ی زمین–ماه [۶۱].

$F_{\max}$ (mN)	$M_{3,0}$ (kg)	$f_{\max,\ \text{nondim}}$	نام فضاپیما	نام اختصار
92.0	486.3	$6.940 \cdot 10^{-2}$	Deep Space 1	DS1
279.3	2464	$4.158 \cdot 10^{-2}$	Psyche	Psyche
91.0	1217.8	$2.741 \cdot 10^{-2}$	Dawn	Dawn
1.25	14	$3.276 \cdot 10^{-2}$	Lunar IceCube	LIC
22.8	510	$1.640 \cdot 10^{-2}$	Hayabusa 1	H1
27.0	608.6	$1.628 \cdot 10^{-2}$	Hayabusa 2	H2
–	–	$4 \cdot 10^{-2}$	فضاپیمای نمونه	s/c

۱-۱-۳ تابع پاداش

تابع پاداش[3] به‌منظور هدایت رفتار عامل طراحی شده و شامل سه بخش اصلی در طول شبیه‌سازی و یک پاداش نهایی در هنگام پایان است:

- **پاداش نهایی برای دستیابی به هدف:** در صورت رسیدن به مدار هدف، شبیه‌سازی پایان یافته و یک پاداش بزرگ مثبت به عامل داده می‌شود.
- **جریمه نهایی برای دور شدن بیش‌ازحد:** اگر عامل از محدوده مجاز فاصله بگیرد، شبیه‌سازی خاتمه یافته و یک جریمه بزرگ منفی اعمال می‌گردد.
- **جریمه برای مصرف سوخت:** در طول مسیر، استفاده بیش‌ازحد از پیشرانه با جریمه همراه است.
- **جریمه برای انحراف از مسیر مرجع:** در طول مسیر، انحراف از مسیر مرجع باعث دریافت جریمه متناسب می‌شود.

تابع پاداش به‌صورت زیر تعریف می‌شود:

$$r(s,a) = r_{\text{thrust}}(a) + r_{\text{reference}}(s) + r_{\text{terminal}}(s)$$

که در آن مؤلفه‌ها عبارتند از:

$$r_{\text{thrust}}(a) = -k_1 \cdot |a| \tag{۱-۱}$$

$$r_{\text{reference}}(s) = -k_2 \cdot d(s, s_{\text{reference}}) \tag{۲-۱}$$

$$r_{\text{terminal}}(s) = \begin{cases} +R_{\text{goal}} & \text{if } s \in S_{\text{goal}} \\ -R_{\text{fail}} & \text{if } d(s, s_{\text{reference}}) > \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{۳-۱}$$

در این رابطه:

۱. R_{goal}: یک پاداش بزرگ مثبت برای دستیابی به هدف است.

۲. R_{fail}: یک جریمه بزرگ منفی برای خروج از محدوده مجاز است.

۳. $d(s, s')$: فاصله بین دو وضعیت بوده و معمولاً به‌صورت فاصله اقلیدسی محاسبه می‌شود.

[3] Reward Function

ضرایب k_1, k_2 برای تنظیم تعادل بین بهینه‌سازی مصرف سوخت و حفظ نزدیکی به مسیر مرجع استفاده می‌شوند. انتخاب مناسب مقادیر این ضرایب نقش کلیدی در سرعت همگرایی و پایداری الگوریتم یادگیری تقویتی دارد.

۱-۲ شبیه‌سازی عامل

در این زیربخش، فرآیند شبیه‌سازی و آموزش عامل با استفاده از الگوریتم‌های یادگیری تقویتی پیشرفته ارائه شده است. تمرکز بر طراحی شبکه‌ها، منطق انتخاب الگوریتم‌ها، فراپارامترهای کلیدی و ملاحظات پایداری در حین آموزش است تا تکرارپذیری و دقت نتایج تضمین شود.

۱-۲-۱ پارامترهای یادگیری و منطق انتخاب الگوریتم‌ها

الگوریتم‌های DDPG، TD3، SAC و PPO به دلیل کارایی در فضاهای کنش پیوسته و عملکرد پایدار در محیط‌های پیچیده انتخاب شده‌اند. به‌طور خلاصه:

- DDPG: سیاست قطعی با شبکه‌های هدف و میانگین پلیاک؛ مناسب محیط‌های پیوسته با هزینه محاسباتی پایین‌تر، اما حساس به نویز.

جدول ۱-۲: جدول پارامترها و مقادیر پیش‌فرض الگوریتم DDPG [۶۲]

نام پارامتر	مقدار	نام پارامتر	مقدار
گام در هر دوره یادگیری	30 000	تعداد دوره‌های یادگیری	100
اندازه‌ی مخزنِ تجربه	10^6	ضریب تنزیل (γ)	0.99
ضریب میانگین پلیاک	0.995	نرخِ یادگیریِ سیاست	10^{-3}
نرخِ یادگیریِ Q	10^{-3}	اندازه‌ی دسته	1024
گام شروعِ استفاده از سیاست	5 000	گام شروعِ به‌روزرسانی	1 000
فاصله‌ی به‌روزرسانی	2 000	نویز عمل	0.1
حداکثر طولِ رخداد	6 000	دستگاه	Cuda
اندازه شبکه‌ی Actor	$(2^5, 2^5)$	تابع فعال‌سازی Actor	ReLU
اندازه شبکه‌ی Critic	$(2^5, 2^5)$	تابع فعال‌سازی Critic	ReLU

- TD3: بهبود DDPG با دو Critic، هموارسازی سیاست هدف و به‌روزرسانی تأخیری سیاست؛ کاهش بیش‌براوردی Q و پایداری بیشتر.

جدول ۱-۳: جدول پارامترها و مقادیر پیش‌فرض الگوریتم TD3 [۶۲]

نام پارامتر	مقدار	نام پارامتر	مقدار
گام در هر دوره یادگیری	30 000	تعداد دوره‌های یادگیری	100
اندازه‌ی مخزنِ تجربه	10^6	ضریب تنزیل (γ)	0.99
ضریب میانگین پلیاک	0.995	نرخِ یادگیریِ سیاست	10^{-3}
نرخِ یادگیریِ Q	10^{-3}	اندازه‌ی دسته	1024
گام شروعِ استفاده از سیاست	5 000	گام شروعِ به‌روزرسانی	1 000
فاصله‌ی به‌روزرسانی	2 000	نویز عمل	0.1
نویز هدف	0.2	برش نویز	0.5
تأخیر در به‌روزرسانی سیاست	2	حداکثر طولِ رخداد	30 000
اندازه شبکه‌ی Actor	$(2^5, 2^5)$	تابع فعال‌سازی Actor	ReLU
اندازه شبکه‌ی Critic	$(2^5, 2^5)$	تابع فعال‌سازی Critic	ReLU

- SAC: سیاست تصادفی بیشینه‌ساز آنتروپی با دمای α؛ کاوش مؤثرتر و همگرایی پایدارتر در محیط‌های نویزی.

جدول ۱-۴: جدول پارامترها و مقادیر پیش‌فرض الگوریتم SAC [۶۲]

نام پارامتر	مقدار	نام پارامتر	مقدار
گام در هر دوره یادگیری	30 000	تعداد دوره‌های یادگیری	100
اندازه‌ی مخزنِ تجربه	10^6	ضریب تنزیل (γ)	0.99
ضریب میانگین پلیاک	0.995	نرخِ یادگیری	10^{-3}
نرخ دمای آلفا	0.2	اندازه‌ی دسته	1024
گام شروعِ استفاده از سیاست	5 000	گام شروعِ به‌روزرسانی	1 000
تعداد به‌روزرسانی در هر مرحله	10	فاصله‌ی به‌روزرسانی	2 000
تعداد اپیزودهای آزمون	10	حداکثر طولِ رخداد	30 000
اندازه شبکه‌ی Actor	$(2^5, 2^5)$	تابع فعال‌سازی Actor	ReLU
اندازه شبکه‌ی Critic	$(2^5, 2^5)$	تابع فعال‌سازی Critic	ReLU

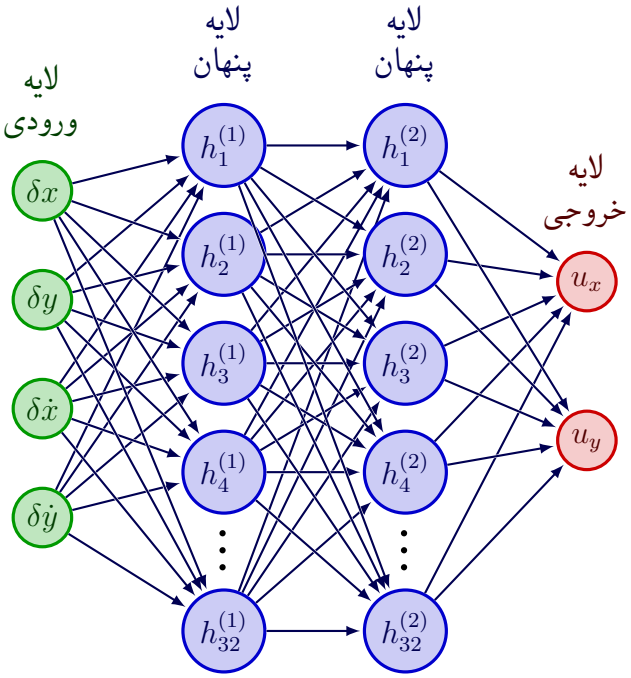
- PPO: روش مبتنی بر سیاست با برش نسبت احتمال؛ به‌روزرسانی‌های ایمن و پیاده‌سازی ساده با کارایی تجربی بالا.

جدول ۱-۵: جدول پارامترها و مقادیر پیش‌فرض الگوریتم PPO [۶۲]

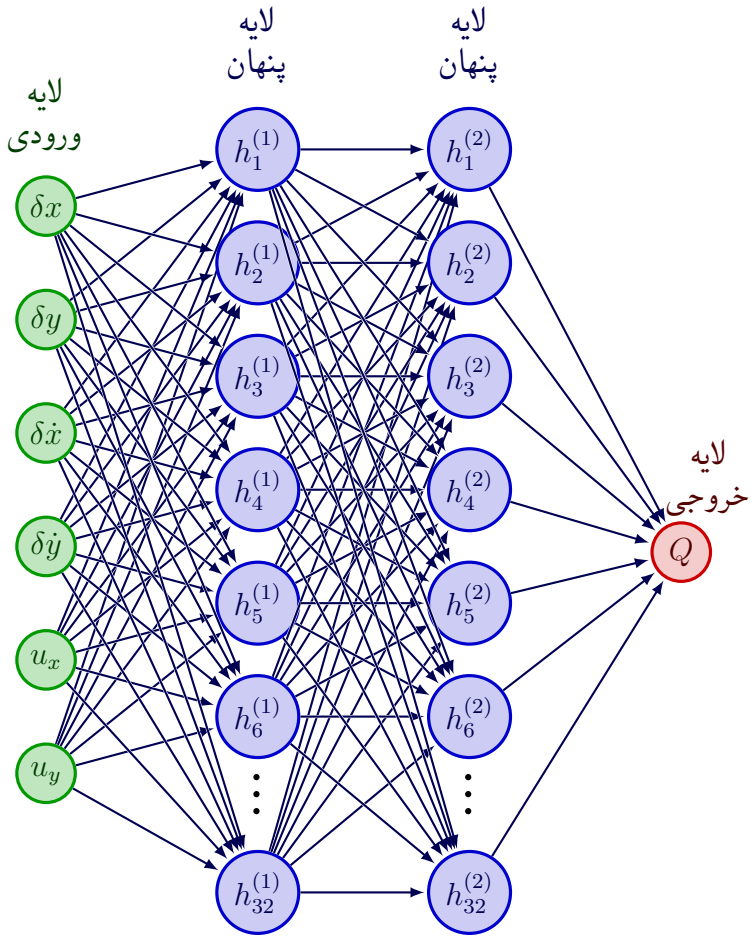
نام پارامتر	مقدار	نام پارامتر	مقدار
گام در هر دوره یادگیری	30 000	تعداد دوره‌های یادگیری	100
ضریب تنزیل (γ)	0.99	ضریب برش ratio clip	0.2
نرخِ یادگیریِ سیاست	3×10^{-4}	نرخِ یادگیریِ تابع ارزش	10^{-3}
تعداد تکرار آموزش سیاست	80	تعداد تکرار آموزش ارزش	80
اندازه شبکه‌ی Actor	$(2^5, 2^5)$	تابع فعال‌سازی Actor	ReLU
اندازه شبکه‌ی Critic	$(2^5, 2^5)$	تابع فعال‌سازی Critic	ReLU

این الگوریتم‌ها به دلیل توانایی در مدیریت فضاهای پیوسته و عملکرد مؤثر در محیط‌های پیچیده انتخاب شده‌اند.

در شکل‌های ۱-۱ و ۱-۲ ساختار شبکه‌های Actor و Critic آورده شده است.



شکل ۱-۱: ساختار شبکه عصبی سیاست



شکل ۱-۲: ساختار شبکه عصبی نقاد

۱-۲-۲ فرآیند آموزش

رویه آموزش با PyTorch و اجرای Cuda به‌صورت زیر انجام شده است:

۱. گردآوری تجربه‌ی اولیه با سیاست تصادفی تا رسیدن به گام شروع به‌روزرسانی برای پرشدن اولیه‌ی مخزن تجربه.

۲. حلقه‌ی یادگیری: در هر گام، اجرای کنش، ذخیره‌ی چهارتایی‌ها (s, a, r, s') (و در صورت نیاز d برای پایان اپیزود) در مخزن تجربه با ظرفیت 10^6.

۳. نمونه‌گیری دسته داده و به‌روزرسانی Criticها با هدف‌های حاوی شبکه‌های هدف و میانگین پلیاک؛ در TD3 استفاده از دو شبکه Q مستقل و هدف‌های کمینه‌شده.

۴. به‌روزرسانی Actor: در TD3/DDPG بیشینه‌سازی $\mathbb{E}_s[Q(s, \pi_\theta(s))]$ و در SAC بیشینه‌سازی بازگشت

انتروپی‌دار؛ در PPO به‌روزرسانی برش‌خورده با نسبت احتمال.

۵. تکنیک‌های پایداری: Target networks با پلیاک، reward/observation normalization، هموارسازی هدف TD3، gradient clipping در صورت نیاز، و بذردهی ثابت برای تکرارپذیری.

۶. ارزیابی دوره‌ای: اجرای چند اپیزود آزمون بدون نویز کنش و ثبت بازگشت، نرخ موفقیت و واریانس.

برای جلوگیری از بیش‌برازش و همگرایی زودرس، از نویز کاوش کنش و هموارسازی سیاست هدف (در TD3) استفاده شده است. معیار توقف زمانی فعال می‌شود که نرخ موفقیت آزمون در چند پنجره‌ی پیاپی از ۹۰٪ عبور کند و واریانس بازگشت کاهش یابد.

بهینه‌سازی و پس‌انتشار گرادیان

محاسبه‌ی گرادیان‌ها با autograd انجام شده است. به‌روزرسانی پارامترها با Adam [۶۳] بوده است که در عمل نسبت به گرادیان نزولی ساده پایدارتر است:

$$g_t = \nabla_w L_t, \quad m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \quad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$
$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t}, \quad w_{t+1} = w_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{۴-۱}$$

که در آن η نرخ یادگیری، β_1, β_2 ضرایب مومنتوم $(0.9,\ 0.999)$ و ϵ برای پایدارسازی عددی است. به‌صورت مفهومی، زنجیره گرادیان نیز برقرار است:

$$\nabla_w L = \frac{\partial L}{\partial y} \frac{\partial y}{\partial w} \tag{۵-۱}$$

در این رابطه:

- L_t: مقدار تابع هزینه (Loss) در گام زمانی t.
- w_t: بردار وزن‌ها یا پارامترهای مدل در گام t.
- $g_t = \nabla_w L_t$: گرادیان تابع هزینه نسبت به پارامترها در زمان t.
- m_t: میانگین نمایی گرادیان‌ها (مومنتوم مرتبه اول) که حافظه‌ای از جهت گرادیان‌ها ایجاد می‌کند.
- v_t: میانگین نمایی مربعات گرادیان‌ها (مومنتوم مرتبه دوم) که بزرگی تغییرات گرادیان را ثبت می‌کند.
- $\hat{m}_t,\ \hat{v}_t$: نسخه‌های اصلاح‌شده‌ی بایاس برای m_t و v_t به‌منظور پایدارسازی در مراحل اولیه.

- η: نرخ یادگیری (Learning Rate) که اندازه‌ی گام به‌روزرسانی وزن‌ها را مشخص می‌کند.
- β_1, β_2: ضرایب کاهش (Decay Rates) برای میانگین‌گیری نمایی؛ مقادیر معمول آن‌ها به‌ترتیب 0.9 و 0.999 است.
- ϵ: یک مقدار بسیار کوچک (معمولاً 10^{-8}) برای جلوگیری از تقسیم بر صفر و افزایش پایداری عددی.

الگوریتم Adam به این صورت عمل می‌کند که همزمان از میانگین مرتبه‌ی اول (m_t) برای جهت حرکت و از میانگین مرتبه‌ی دوم (v_t) برای تنظیم نرخ یادگیری هر پارامتر استفاده می‌کند. در نتیجه هم از نوسانات شدید جلوگیری می‌شود و هم فرآیند همگرایی سرعت می‌گیرد.

از دیدگاه محاسبه‌ی گرادیان، زنجیره‌ی مشتق‌گیری (قاعده‌ی زنجیره‌ای) نیز برقرار است:

$$\nabla_w L = \frac{\partial L}{\partial y} \frac{\partial y}{\partial w} \tag{۶-۱}$$

که در آن y خروجی لایه یا شبکه است. این فرمول مبنای پس‌انتشار خطا (Backpropagation) در شبکه‌های عصبی محسوب می‌شود و باعث می‌گردد که گرادیان تابع هزینه نسبت به تمامی پارامترها به‌صورت کارآمد محاسبه شود.

Bibliography

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

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

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

[56] 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.

[57] 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.

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

[59] 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. *NeurIPS Autodiff Workshop*, 2017.

[60] 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.

[61] N. B. LaFarge, D. Miller, K. C. Howell, and R. Linares. Autonomous closed-loop guidance using reinforcement learning in a low-thrust, multi-body dynamical environment. *Acta Astronautica*, 186:1–23, 2021.

[62] J. Achiam. Spinning Up in Deep Reinforcement Learning. *OpenAI*, 2018.

[63] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

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