



دانشگاه صنعتی شریف
دانشکده‌ی مهندسی هوافضا

پروژه کارشناسی ارشد
مهندسی فضا

عنوان:

هدایت یادگیری تقویتی مقاوم مبتنی بر بازی دیفرانسیلی در محیط‌های پویای چندجسمی با پیشران کم

نگارش:

علی بنی اسد

استاد راهنما:

دکتر هادی نوبهاری

دی ۱۴۰۳



به نام خدا
دانشگاه صنعتی شریف
دانشکده‌ی مهندسی هوافضا

پروژه کارشناسی ارشد

عنوان: هدایت یادگیری تقویتی مقاوم مبتنی بر بازی دیفرانسیلی در محیط‌های پویای چندجسمی
با پیشران کم
نگارش: علی بنی اسد

کمیته‌ی ممتحنین

استاد راهنما: دکتر هادی نوبهاری
امضاء:

استاد مشاور: استاد مشاور
امضاء:

استاد مدعو: استاد ممتحن
امضاء:

تاریخ:

سپاس

از استاد بزرگوارم جناب آقای دکتر نوبهاری که با کمک‌ها و راهنمایی‌های بی‌دریغشان، بنده را در انجام این پروژه یاری داده‌اند، تشکر و قدردانی می‌کنم. از پدر دلسوزم ممنونم که در انجام این پروژه مرا یاری نمود. در نهایت در کمال تواضع، با تمام وجود بر دستان مادرم بوسه می‌زنم که اگر حمایت بی‌دریغش، نگاه مهربانش و دستان گرمش نبود برگ برگ این دست نوشته و پروژه وجود نداشت.

چکیده

در این پژوهش، از یک روش مبتنی بر نظریه بازی^۱ به منظور کنترل وضعیت استند سه درجه آزادی چهارپره استفاده شده است. در این روش بازیکن اول سعی در ردگیری ورودی مطلوب می‌کند و بازیکن دوم با ایجاد اغتشاش سعی در ایجاد خطا در ردگیری بازیکن اول می‌کند. در این روش انتخاب حرکت با استفاده از تعادل نش^۲ که با فرض بدترین حرکت دیگر بازیکن است، انجام می‌شود. این روش نسبت به اغتشاش ورودی و همچنین نسبت به عدم قطعیت مدل‌سازی می‌تواند مقاوم باشد. برای ارزیابی عملکرد این روش ابتدا شبیه‌سازی‌هایی در محیط سیمولینک انجام شده است و سپس، با پیاده‌سازی روی استند سه درجه آزادی صحت عملکرد کنترل‌کننده تایید شده است.

کلیدواژه‌ها: چهارپره، بازی دیفرانسیلی، نظریه بازی، تعادل نش، استند سه درجه آزادی، مدل‌مبنا، تنظیم‌کننده مربعی خطی

¹Game Theory

²Nash Equilibrium

فهرست مطالب

۱	مقدمه	۱
۱	۱-۱ انگیزه پژوهش	۱
۱	۲-۱ تعریف مسئله	۱
۱	۳-۱ اهداف و نوآوری	۱
۱	۴-۱ محتوای گزارش	۱
۲	پیشینه پژوهش	۲
۲	۱-۲ ماموریت‌های بین مداری	۲
۴	۲-۲ بازی دیفرانسیلی	۴
۴	۳-۲ یادگیری تقویتی	۴
۴	۴-۲ یادگیری تقویتی چندعاملی	۴
۵	یادگیری تقویتی	۳
۵	۱-۳ مفاهیم اولیه	۵
۶	۱-۱-۳ حالت و مشاهدات	۶
۶	۲-۱-۳ فضای عمل	۶
۶	۳-۱-۳ سیاست	۶
۷	۴-۱-۳ مسیر	۷
۷	۵-۱-۳ تابع پاداش و بازگشت	۷

۸	۶-۱-۳ ارزش در یادگیری تقویتی
۹	۷-۱-۳ معادلات بلمن
۱۰	۸-۱-۳ تابع مزیت
۱۱	۲-۳ عامل گرادیان سیاست عمیق قطعی
۱۱	۱-۲-۳ یادگیری Q در DDPG
۱۳	۲-۲-۳ سیاست در DDPG
۱۳	۳-۲-۳ اکتشاف و بهره‌برداری در DDPG
۱۳	۴-۲-۳ شبکه‌د DDPG
۱۵	۳-۳ عامل گرادیان سیاست عمیق قطعی تاخیری دوگانه
۱۶	۱-۳-۳ اکتشاف و بهره‌برداری در TD3
۱۶	۲-۳-۳ شبکه‌د TD3
۱۸	۴-۳ عامل عملگر نقاد نرم
۱۸	۱-۴-۳ یادگیری تقویتی تنظیم‌شده با آنتروپی
۱۸	۲-۴-۳ سیاست در SAC
۱۹	۳-۴-۳ تابع ارزش در SAC
۱۹	۴-۴-۳ تابع Q در SAC
۱۹	۵-۴-۳ معادله بلمن در SAC
۲۰	۶-۴-۳ یادگیری Q
۲۰	۷-۴-۳ سیاست در SAC
۲۱	۸-۴-۳ اکتشاف و بهره‌برداری در SAC
۲۲	۹-۴-۳ شبکه‌د SAC
۲۳	۵-۳ عامل بهینه‌سازی سیاست مجاور
۲۴	۱-۵-۳ سیاست در الگوریتم PPO
۲۵	۲-۵-۳ اکتشاف و بهره‌برداری در PPO

۲۵	شبکه‌کد PPO ۳-۵-۳
۲۷	۴ یادگیری تقویتی چند عاملی
۲۷	۱-۴ تعاریف و مفاهیم اساسی
۲۷	۲-۴ اهمیت یادگیری تقویتی چندعاملی
۲۸	۱-۲-۴ نظریه بازی‌ها
۲۸	۲-۲-۴ تعادل نش
۳۰	۳-۲-۴ ایمنی و مقاومت در یادگیری تقویتی چندعاملی
۳۴	۴-۲-۴ الگوریتم‌های یادگیری تقویتی چندعاملی
۳۵	۵ مدل‌سازی محیط یادگیری سه جسمی
۳۶	۱-۵ مسئله سه جسمی محدود دایره‌ای
۳۷	۶ شبیه‌سازی عامل در محیط سه جسمی
۳۷	۱-۶ آموزش عامل در یادگیری تقویتی
۳۷	۱-۱-۶ تعریف عامل و فرآیند آموزش
۳۸	۲-۱-۶ زمان و نوع پردازنده در آموزش عامل
۳۹	۳-۱-۶ نتیجه‌گیری
۴۰	۷ سخت افزار در حلقه عملکرد عامل در محیط

فهرست جداول

فهرست تصاویر

۶	۱-۳ حلقه تعامل عامل و محیط
۳۶	۱-۵ هندسه مسئله سه بدنه محدود

فهرست الگوریتم‌ها

۱۴	گرایان سیاست عمیق قطعی	۱
۱۷	عامل گرایان سیاست عمیق قطعی تاخیری دوگانه	۲
۲۲	عامل عملگرد نقاد نرم	۳
۲۶	بهینه‌سازی سیاست مجاور (PPO-Clip)	۴

فصل ۱

مقدمه

۱-۱ انگیزه پژوهش

۲-۱ تعریف مسئله

در سال‌های اخیر، پیشرفت‌های فناوری در زمینه‌های مختلف، از جمله کنترل پرواز، پردازش سیگنال و هوش مصنوعی، به افزایش کاربردهای ماهواره با پیشران کم در منظومه زمین-ماه کمک کرده است. ماهواره با پیشران کم می‌تواند برای تعقیب ماهواره‌ها، انتقال مداری و استقرار ماهواره‌ها استفاده شود. روش‌های هدایت بهینه قدیمی جهت کنترل ماهواره‌ها اغلب نیازمند فرضیات ساده‌کننده، منابع محاسباتی فراوان و شرایط اولیه مناسب هستند. الگوریتم‌های مبتنی بر یادگیری تقویتی این توانایی را دارند که بدون مشکلات اشاره‌شده هدایت ماهواره را انجام دهند. به همین دلیل، این الگوریتم‌ها می‌توانند امکان محاسبات درونی (On-board Computing) را فراهم می‌کنند.

۳-۱ اهداف و نوآوری

۴-۱ محتوای گزارش

فصل ۲

پیشینه پژوهش

۱-۲ ماموریت‌های بین مداری

هدایت فضاپیماها معمولاً با استفاده از ایستگاه‌های زمینی انجام می‌شود. با این حال، این تکنیک‌ها دارای محدودیت‌هایی از جمله حساسیت به قطع ارتباطات، تاخیرهای زمانی، و محدودیت‌های منابع محاسباتی هستند. الگوریتم‌های یادگیری تقویتی و بازی‌های دیفرانسیلی می‌توانند برای بهبود قابلیت‌های هدایت فضاپیماها، از جمله مقاومت در برابر تغییرات محیطی، کاهش تاخیرهای ناشی از ارتباطات زمینی، و افزایش کارایی محاسباتی، مورد استفاده قرار گیرند.

هدایت فضاپیماها معمولاً پیش از پرواز انجام می‌شود. این روش‌ها می‌توانند از تکنیک‌های بهینه‌سازی فراگیر [۱] یا برنامه‌نویسی غیرخطی برای تولید مسیرها و فرمان‌های کنترلی بهینه استفاده کنند. با این حال، این روش‌ها معمولاً حجم محاسباتی زیادی دارند و برای استفاده درون‌سفینه نامناسب هستند [۲]. یادگیری ماشین می‌تواند برای بهبود قابلیت‌های هدایت فضاپیماها استفاده شود. کنترل‌کننده شبکه عصبی حلقه‌بسته می‌تواند برای محاسبه سریع و خودکار تاریخچه کنترل استفاده شود. یادگیری تقویتی نیز می‌تواند برای یادگیری رفتارهای هدایت بهینه استفاده شود.

روش‌های هدایت و بهینه‌سازی مسیر فضاپیماها به‌طور کلی به راه‌حل‌های اولیه مناسب نیاز دارند. در مسائل چند جسمی، طراحان مسیر اغلب حدس‌های اولیه کم‌هزینه‌ای برای انتقال‌ها با استفاده از نظریه سیستم‌های دینامیکی و منیفولدهای ثابت [۳، ۴] ایجاد می‌کنند.

شبکه‌های عصبی ویژگی‌های جذابی برای فعال‌سازی هدایت در فضاپیما دارند. به‌عنوان مثال، شبکه‌های عصبی می‌توانند به‌طور مستقیم از تخمین‌های وضعیت به دستورهای پیش‌ران کنترلی که با محدودیت‌های مأموریت

سازگار است، برسند. عملکرد هدایت شبکه‌های عصبی در مطالعاتی مانند فرود بر سیارات [۵]، عملیات نزدیکی به سیارات [۶] و کنترل فضاپیما با پیشران ازدست‌رفته [۷] نشان داده شده است. تازه‌ترین پیشرفت‌های تکنیک‌های یادگیری ماشین در مسائل خودکارسازی درونی به‌طور گسترده‌ای مورد مطالعه قرار گرفته‌اند؛ از پژوهش‌های اولیه تا توانایی‌های پیاده‌سازی. به‌عنوان مثال، الگوریتم‌های یادگیری ماشین ابتدایی در فضاپیماهای مریخی نبرد برای کمک به شناسایی ویژگی‌های زمین‌شناسی تعبیه شده‌اند. الگوریتم AEGIS توانایی انتخاب خودکار هدف توسط یک دوربین در داخل فضاپیماهای Spirit، Opportunity و Curiosity را فعال دارد [۸]. در کامپیوتر پرواز اصلی، فرآیند دقت‌افزایی (Refinement Process) نیاز به ۹۴ تا ۹۶ ثانیه دارد [۹]، که به‌طور قابل‌توجهی کمتر از زمان مورد نیاز برای ارسال تصاویر به زمین و انتظار برای انتخاب دستی توسط دانشمندان است. برنامه‌های آینده برای کاربردهای یادگیری ماشین درون‌سفینه شامل توانایی‌های رباتیکی درون‌سفینه برای فضاپیمای Perseverance [۱۰، ۱۱] و شناسایی عیب برای Europa Clipper [۱۲] می‌شود. الگوریتم‌های یادگیری ماشین پتانسیلی برای سهم مهمی در مأموریت‌های اتوماسیون آینده دارند. علاوه بر رباتیک سیاره‌ای، پژوهش‌های مختلفی به استفاده از تکنیک‌های مختلف یادگیری ماشین در مسائل نجومی پرداخته‌اند. در طراحی مسیر عملکرد رگرسیون معمولاً مؤثرتر هست. به‌عنوان مثال، از یک شبکه عصبی (NN) در بهینه‌سازی مسیرهای رانشگر کم‌پیشران استفاده شده است [۱۳]. پژوهش‌های جدید شامل شناسایی انتقال‌های هتروکلینیک [۱۴]، اصلاح مسیر رانشگر کم‌پیشران [۱۵] و تجزیه و تحلیل مشکلات ازدست‌رفتن رانشگر [۷] می‌شود.

تکنیک‌های یادگیری نظارتی می‌توانند نتایج مطلوبی تولید کنند؛ اما، دارای محدودیت‌های قابل‌توجهی هستند. یکی از این محدودیت‌ها این است که این رویکردها بر وجود دانش پیش از فرآیند تصمیم‌گیری متکی هستند. این امر مستلزم دقیق‌بودن داده‌های تولیدشده توسط کاربر برای نتایج مطلوب و همچنین وجود تکنیک‌های موجود برای حل مشکل کنونی و تولید داده است.

در سال‌های اخیر، قابلیت یادگیری تقویتی (RL) در دستیابی به عملکرد بهینه در دامنه‌هایی با ابهام محیطی قابل‌توجه، به اثبات رسیده است [۱۶، ۱۷]. هدایت انجام‌شده توسط RL را می‌توان به‌صورت گسترده بر اساس فاز پرواز دسته‌بندی کرد. مسائل فرود [۱۸، ۱۹] و عملیات در نزدیکی اجسام کوچک [۶، ۵]، از حوزه‌های پژوهشی هستند که از RL استفاده می‌کنند. تحقیقات دیگر شامل مواجهه تداخل خارجی جوی [۲۰]، نگهداری ایستگاهی [۲۱] و هدایت به‌صورت جلوگیری از شناسایی [۲۲] است. مطالعاتی که فضاپیماهای رانشگر کم‌پیشران را در یک چارچوب دینامیکی چند بدنی با استفاده از RL انجام‌شده است، شامل طراحی انتقال با استفاده از Q-learning [۲۳]، Proximal Policy Optimization [۲۴] و هدایت نزدیکی مدار [۲۵] است.

۲-۲ بازی دیفرانسیلی

بازی دیفرانسیلی زیر مجموعه‌ای از نظریه بازی است. نظریه بازی با استفاده از مدل‌های ریاضی به تحلیل روش‌های همکاری یا رقابت موجودات منطقی و هوشمند می‌پردازد. نظریه بازی، شاخه‌ای از ریاضیات کاربردی است که در علوم اجتماعی و به ویژه در اقتصاد، زیست‌شناسی، مهندسی، علوم سیاسی، روابط بین‌الملل، علوم رایانه، بازاریابی و فلسفه مورد استفاده قرار می‌گیرد. نظریه بازی در تلاش است تا به وسیله ریاضیات، رفتار را در شرایط راهبردی یا در یک بازی که در آن موفقیت فرد در انتخاب کردن، وابسته به انتخاب دیگران می‌باشد، برآورد کند. در سال ۱۹۹۴ جان فوربز نش به همراه جان هارسانی و راینهارد سیلتن به خاطر مطالعات خلاقانه‌ی خود در زمینه‌ی نظریه بازی، برنده‌ی جایزه نوبل اقتصاد شدند. در سال‌های پس از آن نیز بسیاری از برندگان جایزه‌ی نوبل اقتصاد از میان متخصصین نظریه بازی انتخاب شدند. آخرین آن‌ها، ژان تیرول فرانسوی است که در سال ۲۰۱۴ این جایزه را کسب کرد [۲۶].

پژوهش‌ها در این زمینه اغلب بر مجموعه‌ای از راهبردهای شناخته شده به عنوان تعادل در بازی‌ها استوار است. این راهبردها به طور معمول از قواعد عقلانی به نتیجه می‌رسند. مشهورترین تعادل‌ها، تعادل نش است. تعادل نش در بازی‌هایی کاربرد دارد در آن فرض شده‌است که هر بازیکن به راهبرد تعادل دیگر بازیکنان آگاه است. بر اساس نظریه‌ی تعادل نش، در یک بازی که هر بازیکن امکان انتخاب‌های گوناگون دارد اگر بازیکنان به روش منطقی راهبردهای خود را انتخاب کنند و به دنبال حداکثر سود در بازی باشند، دست کم یک راهبرد برای به دست آوردن بهترین نتیجه برای هر بازیکن وجود دارد و چنانچه بازیکن راهکار دیگری را انتخاب کند، نتیجه‌ی بهتری به دست نخواهد آورد.

۳-۲ یادگیری تقویتی

۴-۲ یادگیری تقویتی چندعاملی

فصل ۳

یادگیری تقویتی

۱-۳ مفاهیم اولیه

دو بخش اصلی یادگیری تقویتی^۱ شامل عامل^۲ و محیط^۳ است. عامل در محیط قرار دارد و با آن تعامل دارد. در هر مرحله از تعامل بین عامل و محیط، عامل یک مشاهده جزئی از وضعیت محیط انجام می‌دهد و سپس در مورد اقدامی که باید انجام دهد تصمیم می‌گیرد. وقتی عامل بر روی محیط عمل می‌کند، محیط تغییر می‌کند، اما ممکن است محیط به تنهایی نیز تغییر کند. عامل همچنین یک سیگنال پاداش^۴ از محیط دریافت می‌کند، سیگنالی که به آن می‌گویند وضعیت تعامل فعلی عامل در محیط چقدر خوب یا بد است. هدف عامل به حداکثر رساندن پاداش انباشته خود است که بازگشت^۵ نام دارد. یادگیری تقویتی به روش‌هایی گفته می‌شود که در آن‌ها عامل رفتارهای مناسب برای رسیدن به هدف خود را می‌آموزد. در شکل ۱-۳ تعامل بین محیط و عامل نشان داده شده است.

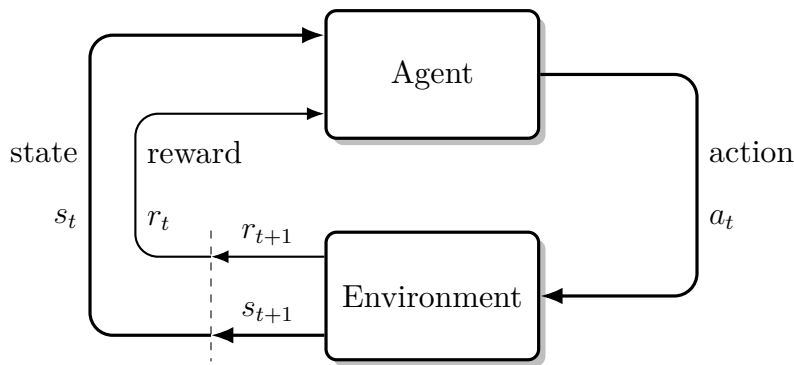
^۱ Reinforcement Learning (RL)

^۲ Agent

^۳ Environment

^۴ Reward

^۵ Return



شکل ۱-۳: حلقه تعامل عامل و محیط

۱-۱-۳ حالت و مشاهدات

حالت^۶ (s) توصیف کاملی از وضعیت محیط است. همه‌ی اطلاعات محیط در حالت وجود دارد. مشاهده^۷ (o) یک توصیف جزئی از حالت است که ممکن است تمامی اطلاعات نباشد. در این پژوهش مشاهده توصیف کاملی از محیط هست در نتیجه حالت و مشاهده برابر هستند.

۲-۱-۳ فضای عمل

فضای عمل (a) در یادگیری تقویتی، مجموعه‌ای از تمام اقداماتی است که یک عامل می‌تواند در محیط انجام دهد. این فضا می‌تواند گسسته^۸ یا پیوسته^۹ باشد. در این پژوهش فضای عمل پیوسته و محدود به یک بازه مشخص است.

۳-۱-۳ سیاست

یک سیاست^{۱۰} قاعده‌ای است که یک عامل برای تصمیم‌گیری در مورد اقدامات خود استفاده می‌کند. در این پژوهش به تناسب الگوریتم پیاده‌سازی شده از سیاست قطعی^{۱۱} یا تصادفی^{۱۲} استفاده شده‌است، که به دو صورت

⁶State

⁷Observation

⁸Discrete

⁹Continuous

¹⁰Policy

¹¹Deterministic

¹²Stochastic

زیر نشان داده می‌شود:

$$a_t = \mu(s_t) \quad (۱-۳)$$

$$a_t \sim \pi(\cdot | s_t) \quad (۲-۳)$$

که زیروند t بیانگر زمان است. در یادگیری تقویتی عمیق از سیاست‌های پارامتری شده استفاده می‌شود. خروجی این سیاست‌ها تابعی از مجموعه‌ای از پارامترها (برای مثال وزن‌ها و بایاس‌های یک شبکه عصبی) هستند که می‌توان از الگوریتم‌های بهینه‌سازی جهت تعیین پارامترها استفاده کرد. در این پژوهش پارامترهای سیاست با θ نشان داده شده‌است و سپس نماد آن به عنوان زیروند سیاست مانند معادله (۳-۳) نشان داده شده‌است.

$$a_t = \mu_\theta(s_t) \quad (۳-۳)$$

$$a_t \sim \pi_\theta(\cdot | s_t)$$

۴-۱-۳ مسیر

یک مسیر^{۱۳} توالی‌ای از حالت‌ها و عمل‌ها در محیط است.

$$\tau = (s_0, a_0, s_1, a_1, \dots) \quad (۴-۳)$$

گذار حالت^{۱۴} به اتفاقاتی که در محیط بین زمان t در حالت s_t و زمان $t + 1$ در حالت s_{t+1} رخ می‌دهد، گفته می‌شود. این گذارها توسط قوانین طبیعی محیط انجام می‌شوند و تنها به آخرین اقدام انجام شده توسط عامل (a_t) بستگی دارند. گذار حالت را می‌توان به صورت زیر تعریف کرد.

$$s_{t+1} = f(s_t, a_t) \quad (۵-۳)$$

۵-۱-۳ تابع پاداش و بازگشت

تابع پاداش^{۱۵} به حالت فعلی محیط، آخرین عمل انجام شده و حالت بعدی محیط بستگی دارد. تابع پاداش را می‌توان به صورت زیر تعریف کرد.

$$r_t = R(s_t, a_t, s_{t+1}) \quad (۶-۳)$$

¹³Trajectory

¹⁴State Transition

¹⁵Reward Function

در این پژوهش پاداش تنها تابعی از جفت حالت-عمل ($r_t = R(s_t, a_t)$) است. هدف عامل این است که مجموع پاداش‌های به‌دست‌آمده در طول یک مسیر را به حداکثر برساند. در این پژوهش مجموع پاداش‌ها در طول یک مسیر را با نماد $R(\tau)$ نشان داده شده‌است و به آن تابع بازگشت^{۱۶} گفته می‌شود. یکی از انواع بازگشت، بازگشت بدون تنزیل^{۱۷} با افق محدود^{۱۸} است که مجموع پاداش‌های به‌دست‌آمده در یک بازه زمانی ثابت و از مسیر τ است که در معادله (۷-۳) نشان داده شده‌است.

$$R(\tau) = \sum_{t=0}^T r_t \quad (۷-۳)$$

نوع دیگری از بازگشت، بازگشت تنزیل‌شده با افق نامحدود^{۱۹} است که مجموع همه پاداش‌هایی است که تا به حال توسط عامل به‌دست آمده‌است. اما، فاصله زمانی تا دریافت پاداش باعث تنزیل ارزش آن می‌شود. این معادله بازگشت (۸-۳) شامل یک فاکتور تنزیل^{۲۰} با نماد γ است که عددی بین صفر و یک است.

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t \quad (۸-۳)$$

۳-۱-۶ ارزش در یادگیری تقویتی

در یادگیری تقویتی، دانستن ارزش^{۲۱} یک حالت یا جفت حالت-عمل ضروری است. منظور از ارزش، بازگشت مورد انتظار^{۲۲} است. یعنی اگر از آن حالت یا جفت حالت-عمل شروع شود و سپس برای همیشه طبق یک سیاست خاص عمل شود، به طور میانگین چه مقدار پاداش دریافت خواهد کرد. توابع ارزش تقریباً در تمام الگوریتم‌های یادگیری تقویتی به کار می‌روند. در اینجا به چهار تابع مهم اشاره شده‌است.

۱. تابع ارزش تحت سیاست^{۲۳} ($V^\pi(s)$): خروجی این تابع بازگشت مورد انتظار است در صورتی که از حالت s شروع شود و همیشه طبق سیاست π عمل شود و به‌صورت زیر بیان می‌شود:

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi} [R(\tau) | s_0 = s] \quad (۹-۳)$$

۲. تابع ارزش-عمل تحت سیاست^{۲۴} ($Q^\pi(s, a)$): خروجی این تابع بازگشت مورد انتظار است در صورتی که از حالت s شروع شود، یک اقدام دلخواه a (که ممکن است از سیاست π نباشد) انجام شود و سپس

¹⁶Return

¹⁷Discount

¹⁸Finite-Horizon Undiscounted Return

¹⁹Infinite-Horizon Discounted Return

²⁰Discount Factor

²¹Value

²²Expected Return

²³On-Policy Value Function

²⁴On-Policy Action-Value Function

برای همیشه طبق سیاست π عمل شود و به صورت زیر بیان می شود:

$$Q^\pi(s, a) = \mathbb{E}_{\tau \sim \pi} [R(\tau) | s_0 = s, a_0 = a] \quad (10-3)$$

۳. تابع ارزش بهینه^{۲۵} ($V^*(s)$): خروجی این تابع بازگشت مورد انتظار است در صورتی که از حالت s شروع شود و همیشه طبق سیاست بهینه در محیط عمل شود و به صورت زیر بیان می شود:

$$V^*(s) = \max_{\pi} (V^\pi(s)) \quad (11-3)$$

۴. تابع ارزش-عمل بهینه^{۲۶} ($Q^*(s, a)$): خروجی این تابع بازگشت مورد انتظار است در صورتی که از حالت s شروع شود، یک اقدام دلخواه a انجام شود و سپس برای همیشه طبق سیاست بهینه در محیط عمل شود و به صورت زیر بیان می شود:

$$Q^*(s, a) = \max_{\pi} (Q^\pi(s, a)) \quad (12-3)$$

۷-۱-۳ معادلات بلمن

توابع ارزش اشاره شده از معادلات خاصی که به آن ها معادلات بلمن گفته می شود، پیروی می کنند. ایده اصلی پشت معادلات بلمن است که ارزش نقطه شروع برابر است با پاداشی است که انتظار دارید از آنجا دریافت کنید، به علاوه ارزش مکانی که بعداً به آنجا می رسید. معادلات بلمن برای توابع ارزش سیاست محور به شرح زیر هستند:

$$V^\pi(s) = \mathbb{E}_{\substack{a \sim \pi \\ s' \sim P}} [r(s, a) + \gamma V^\pi(s')] \quad (13-3)$$

$$Q^\pi(s, a) = \mathbb{E}_{s' \sim P} \left[r(s, a) + \gamma \mathbb{E}_{a' \sim \pi} [Q^\pi(s', a')] \right] \quad (14-3)$$

که در آن $V^\pi(s)$ تابع ارزش حالت s تحت سیاست π است؛ $Q^\pi(s, a)$ تابع ارزش عمل a در حالت s تحت سیاست π است؛ $R(s, a)$ پاداش دریافتی پس از انجام عمل a در حالت s است؛ γ ضریب تخفیف است که ارزش پاداش های آینده را کاهش می دهد؛ $s' \sim P(\cdot | s, a)$ نشان می دهد که حالت بعدی s' از توزیع انتقال محیط P با شرط های s و a نمونه برداری می شود؛ و $a' \sim \pi(\cdot | s')$ نشان می دهد که عمل بعدی a' از سیاست

²⁵Optimal Value Function

²⁶Optimal Action-Value Function

π با شرط حالت جدید s' نمونه‌برداری می‌شود. این معادلات بیانگر این هستند که ارزش یک حالت یا عمل، مجموع پاداش مورد انتظار آن و ارزش حالت بعدی است که بر اساس سیاست فعلی تعیین می‌شود. معادلات بلمن برای توابع ارزش بهینه به شرح زیر هستند:

$$V^*(s) = \max_a \mathbb{E}_{s' \sim P} [r(s, a) + \gamma V^*(s')] \quad (۱۵-۳)$$

$$Q^*(s, a) = \mathbb{E}_{s' \sim P} \left[r(s, a) + \gamma \max_{a'} Q^*(s', a') \right] \quad (۱۶-۳)$$

تفاوت حیاتی بین معادلات بلمن برای توابع ارزش سیاست محور و توابع ارزش بهینه، عدم حضور یا حضور عملگر max بر روی اعمال است. حضور آن منعکس‌کننده این است که هرگاه عامل بتواند عمل خود را انتخاب کند، برای عمل بهینه، باید هر عملی را که منجر به بالاترین ارزش می‌شود انتخاب کند.

۳-۱-۸ تابع مزیت

گاهی در یادگیری تقویتی، نیازی به توصیف میزان خوبی یک عمل به صورت مطلق نیست، بلکه تنها می‌خواهیم بدانیم که چه مقدار بهتر از سایر اعمال به طور متوسط است. به عبارت دیگر، مزیت نسبی آن عمل مورد بررسی قرار می‌گیرد. این مفهوم با تابع مزیت^{۲۷} توضیح داده می‌شود.

تابع مزیت $A^\pi(s, a)$ که مربوط به سیاست π است، توصیف می‌کند که انجام یک عمل خاص a در حالت s چقدر بهتر از انتخاب تصادفی یک عمل بر اساس $\pi(\cdot|s)$ است، با فرض اینکه شما برای همیشه پس از آن مطابق با π عمل می‌کنید. به صورت ریاضی، تابع مزیت به صورت زیر تعریف می‌شود:

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$$

که در آن $A^\pi(s, a)$ تابع مزیت برای عمل a در حالت s است. $Q^\pi(s, a)$ تابع ارزش عمل a در حالت s تحت سیاست π است و $V^\pi(s)$ تابع ارزش حالت s تحت سیاست π است. این تابع مزیت نشان می‌دهد که انجام عمل a در حالت s نسبت به میانگین اعمال تحت سیاست π چقدر مزیت دارد. اگر $A^\pi(s, a)$ مثبت باشد، نشان‌دهنده این است که عمل a بهتر از میانگین اعمال است و اگر منفی باشد، نشان‌دهنده کمتر بودن عملکرد آن نسبت به میانگین است.

²⁷ Advantage Function

۲-۳ عامل گرادیان سیاست عمیق قطعی

گرادیان سیاست عمیق قطعی^{۲۸} الگوریتمی است که همزمان یک تابع Q و یک سیاست را یاد می‌گیرد. این الگوریتم برای یادگیری تابع Q از داده‌های غیرسیاست محور^{۲۹} و معادله بلمن استفاده می‌کند. این الگوریتم برای یادگیری سیاست نیز از تابع Q استفاده می‌کند.

این رویکرد وابستگی نزدیکی به یادگیری Q دارد. اگر تابع ارزش-عمل بهینه مشخص باشد، در هر حالت داده‌شده عمل بهینه را می‌توان با حل معادله (۱۷-۳) به دست آورد.

$$a^*(s) = \arg \max_a Q^*(s, a) \quad (۱۷-۳)$$

الگوریتم DDPG ترکیبی از یادگیری تقریبی برای $Q^*(s, a)$ و یادگیری تقریبی برای $a^*(s)$ است و به صورتی طراحی شده است که برای محیط‌هایی با فضاها عمل پیوسته مناسب باشد. آنچه این الگوریتم را برای فضای عمل پیوسته مناسب می‌کند، روش محاسبه $a^*(s)$ است. فرض می‌شود که تابع $Q^*(s, a)$ نسبت به آرگومان عمل مشتق‌پذیر است. مشتق‌پذیری این امکان را می‌دهد که یک روش یادگیری مبتنی بر گرادیان برای سیاست $\mu(s)$ استفاده شود. سپس، به جای اجرای یک بهینه‌سازی زمان‌بر در هر بار محاسبه $\max_a Q(s, a)$ ، می‌توان آن را با رابطه $\max_a Q(s, a) \approx Q(s, \mu(s))$ تقریب زد.

۱-۲-۳ یادگیری Q در DDPG

معادله بلمن که تابع ارزش عمل بهینه $(Q^*(s, a))$ را توصیف می‌کند، در پایین آورده شده است.

$$Q^*(s, a) = \mathbb{E}_{s' \sim P} \left[r(s, a) + \gamma \max_{a'} Q^*(s', a') \right] \quad (۱۸-۳)$$

عبارت $s' \sim P$ به این معنی است که وضعیت بعدی یعنی s' از توزیع احتمال $P(\cdot | s, a)$ نمونه‌گرفته می‌شود. در معادله بلمن نقطه شروع برای یادگیری $Q^*(s, a)$ یک مقداردهی تقریبی است. پارامترهای شبکه عصبی $Q_\phi(s, a)$ با علامت ϕ نشان داده شده است. مجموعه \mathcal{D} شامل اطلاعات جمع‌آوری شده تغییر از یک حالت به حالت دیگر (s, a, r, s', d) (که d نشان می‌دهد که آیا وضعیت s' پایانی است یا خیر) است. در بهینه‌سازی از تابع خطای میانگین مربعات بلمن^{۳۰} (MSBE) استفاده شده است که معیاری برای نزدیکی Q_ϕ به حالت بهینه برای برآورده کردن معادله بلمن است.

²⁸Deep Deterministic Policy Gradient (DDPG)

²⁹Off-Policy

³⁰Mean Squared Bellman Error

$$L(\phi, \mathcal{D}) = \mathbb{E}_{(s,a,r,s',d) \sim \mathcal{D}} \left[\left(Q_{\phi}(s, a) - \left(r + \gamma(1-d) \max_{a'} Q_{\phi}(s', a') \right) \right)^2 \right] \quad (19-3)$$

در الگوریتم DDPG دو ترفند برای عملکرد بهتر استفاده شده است که در ادامه به بررسی آن پرداخته شده است.

- بافرهای تکرار بازی

الگوریتم‌های یادگیری تقویتی جهت آموزش یک شبکه عصبی عمیق برای تقریب $Q^*(s, a)$ از بافرهای تکرار بازی^{۳۱} تجربه شده استفاده می‌کنند. این مجموعه \mathcal{D} شامل تجربیات قبلی عامل است. برای داشتن رفتار پایدار در الگوریتم، بافر تکرار بازی باید به اندازه کافی بزرگ باشد تا شامل یک دامنه گسترده از تجربیات شود. انتخاب داده‌های بافر به دقت انجام شده است چرا که اگر فقط از داده‌های بسیار جدید استفاده شود، بیش‌برازش^{۳۲} رخ می‌دهد و اگر از تجربه بیش از حد استفاده شود، ممکن است فرآیند یادگیری کند شود.

- شبکه‌های هدف

الگوریتم‌های یادگیری Q از شبکه‌های هدف استفاده می‌کنند. اصطلاح زیر به عنوان هدف شناخته می‌شود.

$$r + \gamma(1-d) \max_{a'} Q_{\phi}(s', a') \quad (20-3)$$

در هنگام کمینه کردن تابع خطای میانگین مربعات بلمن، سعی شده است تا تابع Q شبیه‌تر به هدف یعنی رابطه (۲۰-۳) شود. اما مشکل این است که هدف بستگی به پارامترهای در حال آموزش ϕ دارد. این باعث ایجاد ناپایداری در کمینه کردن تابع خطای میانگین مربعات بلمن می‌شود. راه حل آن استفاده از یک مجموعه پارامترهایی است که با تأخیر زمانی به ϕ نزدیک می‌شوند. به عبارت دیگر، یک شبکه دوم ایجاد می‌شود که به آن شبکه هدف گفته می‌شود. شبکه هدف با تأخیر پارامترهای شبکه اول را دنبال می‌کند. پارامترهای شبکه هدف با نشان ϕ_{targ} نشان داده می‌شوند. در الگوریتم DDPG، شبکه هدف در هر به‌روزرسانی شبکه اصلی، با میانگین‌گیری پولیاک^{۳۳} به صورت زیر به‌روزرسانی می‌شود.

$$\phi_{\text{targ}} \leftarrow \rho \phi_{\text{targ}} + (1 - \rho) \phi \quad (21-3)$$

در رابطه بالا ρ یک ابرپارامتر^{۳۴} است که بین صفر و یک انتخاب می‌شود. در این پژوهش این مقدار نزدیک به یک در نظر گرفته شده است.

³¹Replay Buffers

³²Overfit

³³Polyak Averaging

³⁴Hyperparameter

الگوریتم DDPG نیاز به یک شبکه سیاست هدف ($\mu_{\theta_{\text{targ}}}$) برای محاسبه عمل‌هایی که به‌طور تقریبی بیشینه $Q_{\phi_{\text{targ}}}$ را حاصل کند، را دارد. برای رسیدن به این شبکه سیاست هدف از همان روشی که تابع Q به دست می‌آید یعنی با میانگین‌گیری پولیاک از پارامترهای سیاست در طول زمان آموزش استفاده می‌شود.

با در نظر گرفتن موارد اشاره‌شده، یادگیری Q در DDPG با کمینه‌کردن تابع خطای میانگین مربعات بلمن (MSBE) یعنی معادله (۲۲-۳) با استفاده از کاهش گرادیان تصادفی^{۳۵} انجام می‌شود.

$$L(\phi, \mathcal{D}) = \mathbb{E}_{(s,a,r,s',d) \sim \mathcal{D}} \left[\left(Q_{\phi}(s, a) - (r + \gamma(1-d)Q_{\phi_{\text{targ}}}(s', \mu_{\theta_{\text{targ}}}(s'))) \right)^2 \right] \quad (22-3)$$

۲-۲-۳ سیاست در DDPG

در این بخش یک سیاست تعیین‌شده $\mu_{\theta}(s)$ یاد گرفته می‌شود تا عملی را انجام می‌دهد که بیشینه $Q_{\phi}(s, a)$ رخ دهد. از آنجا که فضای عمل پیوسته است و فرض شده‌است که تابع Q نسبت به عمل مشتق‌پذیر است، رابطه زیر با استفاده از صعود گرادیان^{۳۶} (تنها نسبت به پارامترهای سیاست) بیشینه می‌شود.

$$\max_{\theta} \mathbb{E}_{s \sim \mathcal{D}} [Q_{\phi}(s, \mu_{\theta}(s))] \quad (23-3)$$

۳-۲-۳ اکتشاف و بهره‌برداری در DDPG

برای بهبود اکتشاف^{۳۷} در سیاست‌های DDPG، در زمان آموزش نویز به عمل‌ها اضافه می‌شود. نویسندگان مقاله DDPG [۲۷] توصیه کرده‌اند که نویز OU^{۳۸} با هم‌بندی زمانی^{۳۹} اضافه شود. در زمان بهره‌برداری^{۴۰} سیاست، از آنچه یاد گرفته است، نویز به عمل‌ها اضافه نمی‌شود.

۴-۲-۳ شبکه‌کد DDPG

در این بخش، شبکه‌کد الگوریتم DDPG پیاده‌سازی شده آورده شده‌است. در این پژوهش الگوریتم ۱ در محیط پایتون با استفاده از کتابخانه TensorFlow [۲۸] پیاده‌سازی شده‌است.

³⁵Stochastic Gradient Descent

³⁶Gradient Ascent

³⁷Exploration

³⁸Ornstein-Uhlenbeck

³⁹Time-Correlated

⁴⁰Exploitation

ورودی: پارامترهای اولیه سیاست (θ) ، پارامترهای تابع $Q(\phi)$ ، بافر تکرار بازی خالی (\mathcal{D})

۱: پارامترهای هدف را برابر با پارامترهای اصلی قرار دهید $\phi_{\text{targ}} \leftarrow \phi, \theta_{\text{targ}} \leftarrow \theta$

۲: تا وقتی همگرایی رخ دهد:

۳: وضعیت s را مشاهده کرده و عمل $a = \text{clip}(\mu_{\theta}(s) + \epsilon, a_{\text{Low}}, a_{\text{High}})$ را انتخاب کنید به طوری که $\epsilon \sim \mathcal{N}$ است.

۴: عمل a را در محیط اجرا کنید.

۵: وضعیت بعدی s' ، پاداش r و سیگنال پایان d را مشاهده کنید تا نشان دهد آیا s' پایانی است یا خیر.

۶: اگر s' پایانی است، وضعیت محیط را بازنشانی کنید.

۷: اگر زمان بهروزرسانی فرا رسیده است:

۸: به ازای هر تعداد بهروزرسانی:

۹: یک دسته تصادفی گذر از یک حالت به حالت دیگر، $B = \{(s, a, r, s', d)\}$ ، از \mathcal{D} نمونه‌گیری شود.

۱۰: هدف را محاسبه کنید:

$$y(r, s', d) = r + \gamma(1 - d)Q_{\phi_{\text{targ}}}(s', \mu_{\theta_{\text{targ}}}(s'))$$

۱۱: تابع Q را با یک مرحله از نزول گرادیان با استفاده از رابطه زیر بهروزرسانی کنید:

$$\nabla_{\phi} \frac{1}{|B|} \sum_{(s, a, r, s', d) \in B} (Q_{\phi}(s, a) - y(r, s', d))^2$$

۱۲: سیاست را با یک مرحله از صعود گرادیان با استفاده از رابطه زیر بهروزرسانی کنید:

$$\nabla_{\theta} \frac{1}{|B|} \sum_{s \in B} Q_{\phi}(s, \mu_{\theta}(s))$$

۱۳: شبکه‌های هدف را با استفاده از معادلات زیر بهروزرسانی کنید:

$$\phi_{\text{targ}} \leftarrow \rho \phi_{\text{targ}} + (1 - \rho) \phi$$

$$\theta_{\text{targ}} \leftarrow \rho \theta_{\text{targ}} + (1 - \rho) \theta$$

۳-۳ عامل گرادیان سیاست عمیق قطعی تاخیری دوگانه

عامل گرادیان سیاست عمیق قطعی تاخیری دوگانه^{۴۱} یکی از الگوریتم‌های یادگیری تقویتی است که برای حل مسائل کنترل در محیط‌های پیوسته طراحی شده‌است. این الگوریتم بر اساس الگوریتم DDPG توسعه یافته و با استفاده از تکنیک‌های مختلف، پایداری و کارایی یادگیری را بهبود می‌بخشد. در حالی که DDPG گاهی اوقات می‌تواند عملکرد بسیار خوبی داشته باشد، اما اغلب نسبت به ابرپارامترها و سایر انواع تنظیمات یادگیری حساس است. یک حالت رایج شکست عامل DDPG در یادگیری این است که تابع Q یادگرفته شده شروع به بیش برآورد مقادیر Q می‌کند که منجر به واگرایی سیاست می‌شود. واگرایی به این دلیل رخ می‌دهد که در فرایند یادگیری سیاست از تخمین تابع Q استفاده می‌شود که افزایش خطای تابع Q منجر به ناپایداری در یادگیری سیاست می‌شود.

الگوریتم TD3 (Twin Delayed DDPG) از دو طرفند زیر جهت بهبود مشکلات اشاره شده استفاده می‌کند.

- یادگیری دوگانه‌ی محدود شده^{۴۲}: الگوریتم TD3 به جای یک تابع Q ، دو تابع Q_{ϕ_1} و Q_{ϕ_2} را یاد می‌گیرد (از این رو دوگانه^{۴۳} نامیده می‌شود) و از کوچک‌ترین مقدار این دو Q_{ϕ_1} و Q_{ϕ_2} در تابع بلمن استفاده می‌شود. نحوه محاسبه هدف بر اساس دو تابع Q اشاره شده در رابطه (۲۴-۳) آورده شده‌است.

$$y(r, s', d) = r + \gamma(1 - d) \min_{i=1,2} Q_{\phi_i, \text{targ}}(s', a'(s')) \quad (24-3)$$

سپس، در هر دو تابع Q_{ϕ_1} و Q_{ϕ_2} یادگیری انجام می‌شود.

$$L(\phi_1, \mathcal{D}) = \mathbb{E}_{(s,a,r,s',d) \sim \mathcal{D}} \left(Q_{\phi_1}(s, a) - y(r, s', d) \right)^2 \quad (25-3)$$

$$L(\phi_2, \mathcal{D}) = \mathbb{E}_{(s,a,r,s',d) \sim \mathcal{D}} \left(Q_{\phi_2}(s, a) - y(r, s', d) \right)^2 \quad (26-3)$$

- به‌روزرسانی‌های تاخیری سیاست^{۴۴}: الگوریتم TD3 سیاست را با تاخیر بیشتری نسبت به تابع Q به‌روزرسانی می‌کند. در مرجع [۲۹] توصیه شده‌است که برای هر دو به‌روزرسانی تابع Q ، یک به‌روزرسانی سیاست انجام شود.

⁴¹Twin Delayed Deep Deterministic Policy Gradient (TD3)

⁴²Clipped Double-Q Learning

⁴³twin

⁴⁴Delayed Policy Updates

این دو ترفند منجر به بهبود قابل توجه عملکرد TD3 نسبت به DDPG پایه می‌شوند. در نهایت سیاست با به حداکثر رساندن Q_{ϕ_1} آموخته می‌شود:

$$\max_{\theta} E_{s \sim \mathcal{D}} [Q_{\phi_1}(s, \mu_{\theta}(s))] \quad (27-3)$$

۱-۳-۳ اکتشاف و بهره‌برداری در TD3

الگوریتم TD3 یک سیاست قطعی را به صورت غیر سیاست محور آموزش می‌دهد. از آنجایی که سیاست قطعی است، در ابتدا عامل تنوع کافی از اعمال را برای یافتن روش‌های مفید امتحان نمی‌کند. برای بهبود اکتشاف سیاست‌های TD3، در زمان آموزش نویز به عمل‌ها اضافه می‌شود، در این پژوهش نویز گاوسی با میانگین صفر بدون همبندی زمانی اعمال شده‌است. شدت نویز جهت بهره‌برداری بهتر در طول زمان کاهش می‌یابد.

۲-۳-۳ شبه‌کد TD3

در این بخش الگوریتم TD3 پیاده‌سازی شده آورده شده‌است. در این پژوهش الگوریتم ۴ در محیط پایتون با استفاده از کتابخانه PyTorch [۳۰] پیاده‌سازی شده‌است.

الگوریتم ۲ عامل گرادیان سیاست عمیق قطعی تاخیری دوگانه

ورودی: پارامترهای اولیه سیاست (θ) ، پارامترهای تابع Q (ϕ_1, ϕ_2) ، بافر بازی خالی (\mathcal{D})

۱: پارامترهای هدف را برابر با پارامترهای اصلی قرار دهید $\phi_{\text{targ},2} \leftarrow \phi_2, \phi_{\text{targ},1} \leftarrow \phi_1, \theta_{\text{targ}} \leftarrow \theta$

۲: تا وقتی همگرایی رخ دهد:

۳: وضعیت (s) را مشاهده کرده و عمل $a = \text{clip}(\mu_\theta(s) + \epsilon, a_{\text{Low}}, a_{\text{High}})$ را انتخاب کنید، به طوری که $\epsilon \sim \mathcal{N}$ است.

۴: عمل a را در محیط اجرا کنید.

۵: وضعیت بعدی s' ، پاداش r و سیگنال پایان d را مشاهده کنید تا نشان دهد آیا s' پایانی است یا خیر.

۶: اگر s' پایانی است، وضعیت محیط را بازنشانی کنید.

۷: اگر زمان به روزرسانی فرا رسیده است:

۸: به ازای j در هر تعداد به روزرسانی:

۹: یک دسته تصادفی گذر از یک حالت به حالت دیگر، $B = \{(s, a, r, s', d)\}$ ، از \mathcal{D} نمونه‌گیری شود.

۱۰: هدف را محاسبه کنید:

$$y(r, s', d) = r + \gamma(1 - d) \min_{i=1,2} Q_{\phi_{\text{targ},i}}(s', a'(s'))$$

۱۱: تابع Q را با یک مرحله از نزول گرادیان با استفاده از رابطه زیر به روزرسانی کنید:

$$\nabla_{\phi_i} \frac{1}{|B|} \sum_{(s,a,r,s',d) \in B} (Q_{\phi_i}(s, a) - y(r, s', d))^2 \quad \text{for } i = 1, 2$$

۱۲: اگر باقیمانده j بر تاخیر سیاست برابر ۰ باشد:

۱۳: سیاست را با یک مرحله از صعود گرادیان با استفاده از رابطه زیر به روزرسانی کنید:

$$\nabla_{\theta} \frac{1}{|B|} \sum_{s \in B} Q_{\phi_1}(s, \mu_\theta(s))$$

۱۴: شبکه‌های هدف را با استفاده از معادلات زیر به روزرسانی کنید:

$$\phi_{\text{targ},i} \leftarrow \rho \phi_{\text{targ},i} + (1 - \rho) \phi_i \quad \text{for } i = 1, 2$$

$$\theta_{\text{targ}} \leftarrow \rho \theta_{\text{targ}} + (1 - \rho) \theta$$

۴-۳ عامل عملگر نقاد نرم

عملگر نقاد نرم^{۴۵} الگوریتمی است که یک سیاست تصادفی را به صورت سیاست محور بهینه می‌کند و پلی بین بهینه‌سازی سیاست تصادفی و رویکردهای مانند DDPG ایجاد می‌کند. این الگوریتم جانشین مستقیم TD3 نیست (زیرا تقریباً همزمان منتشر شده است)، اما ترفند یادگیری دوگانه محدود شده را در خود جای داده است و به دلیل سیاست تصادفی SAC، از چیزی روشی به نام صاف کردن سیاست هدف^{۴۶} استفاده شده است. یکی از ویژگی‌های اصلی SAC، تنظیم آنتروپی است. آنتروپی معیاری از تصادفی بودن انتخاب عمل در سیاست است. سیاست به گونه ای آموزش داده می‌شود که حداکثر سازی تعادل بین بازده مورد انتظار و آنتروپی را بهینه کند. این شرایط ارتباط نزدیکی با تعادل اکتشاف-بهره‌برداری دارد. افزایش آنتروپی منجر به اکتشاف بیشتر می‌شود که می‌تواند یادگیری را در مراحل بعدی تسریع کند. همچنین می‌تواند از همگرایی زودهنگام سیاست به یک بهینه محلی بد جلوگیری کند. برای توضیح SAC، ابتدا باید بررسی یادگیری تقویتی تنظیم‌شده با آنتروپی^{۴۷} را پرداخته شود. در RL تنظیم‌شده با آنتروپی، روابط تابع ارزش کمی متفاوت است.

۱-۴-۳ یادگیری تقویتی تنظیم‌شده با آنتروپی

آنتروپی کمیتی است که به طور کلی می‌گوید که یک متغیر تصادفی چقدر تصادفی است. اگر وزن یک سکه به گونه‌ای باشد که تقریباً همیشه نتیجه یک سمت آن باشد، آنتروپی پایینی دارد. اگر به طور مساوی وزن داشته باشد و شانس هر طرف سکه نصف باشد، آنتروپی بالایی دارد. فرض کنید x یک متغیر تصادفی با تابع چگالی احتمال P باشد. آنتروپی H متغیر x از توزیع آن P مطابق با رابطه زیر محاسبه می‌شود:

$$H(P) = \mathbb{E}_{x \sim P} [-\log P(x)] \quad (۲۸-۳)$$

۲-۴-۳ سیاست در SAC

در یادگیری تقویتی تنظیم‌شده با آنتروپی، عامل در هر مرحله زمانی متناسب با آنتروپی سیاست در آن مرحله زمانی پاداش دریافت می‌کند. بر اساس توضیحات اشاره شده روابط یادگیری تقویتی به صورت زیر می‌شود.

^{۴۵}Soft Actor Critic (SAC)

^{۴۶}Target Policy Smoothing

^{۴۷}Entropy-Regularized Reinforcement Learning

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{\tau \sim \pi} \sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t, s_{t+1}) + \alpha H(\pi(\cdot|s_t)) \right) \quad (29-3)$$

که در آن $(\alpha > 0)$ ضریب مبادله^{۴۸} است.

۳-۴-۳ تابع ارزش در SAC

اکنون می‌توان تابع ارزش کمی متفاوت را بر اساس این مفهوم تعریف کرد. V^π به گونه‌ای تغییر می‌کند که پاداش‌های آنتروپی را از هر مرحله زمانی شامل می‌شود.

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t, s_{t+1}) + \alpha H(\pi(\cdot|s_t)) \right) \middle| s_0 = s \right] \quad (30-3)$$

۴-۴-۳ تابع Q در SAC

تابع Q^π به گونه‌ای تغییر می‌کند که پاداش‌های آنتروپی را از هر مرحله زمانی به جز مرحله اول شامل می‌شود.

$$Q^\pi(s, a) = \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) + \alpha \sum_{t=1}^{\infty} \gamma^t H(\pi(\cdot|s_t)) \middle| s_0 = s, a_0 = a \right] \quad (31-3)$$

با این تعاریف رابطه V^π و Q^π به صورت زیر است.

$$V^\pi(s) = \mathbb{E}_{a \sim \pi} [Q^\pi(s, a)] + \alpha H(\pi(\cdot|s)) \quad (32-3)$$

۵-۴-۳ معادله بلمن در SAC

معادله بلمن در حالت تنظیم‌شده با آنتروپی به صورت زیر ارائه می‌شود.

$$Q^\pi(s, a) = \mathbb{E}_{\substack{s' \sim P \\ a' \sim \pi}} [R(s, a, s') + \gamma (Q^\pi(s', a') + \alpha H(\pi(\cdot|s')))] \quad (33-3)$$

$$= \mathbb{E}_{s' \sim P} [R(s, a, s') + \gamma V^\pi(s')] \quad (34-3)$$

⁴⁸Trade-Off

۶-۴-۳ یادگیری Q

با در نظر گرفتن موارد اشاره شده، یادگیری Q در SAC با کمینه کردن تابع خطای میانگین مربعات بلمن (MSBE) یعنی معادله (۳۵-۳) با استفاده از کاهش گرادیان انجام می شود.

$$L(\phi_i, \mathcal{D}) = \mathbb{E}_{(s,a,r,s',d) \sim \mathcal{D}} \left[\left(Q_{\phi_i}(s, a) - y(r, s', d) \right)^2 \right] \quad (۳۵-۳)$$

در معادله (۳۵-۳) تابع هدف برای روش یادگیری تقویتی SAC به صورت زیر تعریف می شود.

$$y(r, s', d) = r + \gamma(1 - d) \left(\min_{j=1,2} Q_{\phi_{\text{targ},j}}(s', \tilde{a}') - \alpha \log \pi_{\theta}(\tilde{a}'|s') \right), \quad \tilde{a}' \sim \pi_{\theta}(\cdot|s') \quad (۳۶-۳)$$

نماد عمل بعدی را به جای a' به \tilde{a}' تغییر داده شده تا مشخص شود که عمل های بعدی باید آخرین سیاست نمونه برداری شوند در حالی که r و s باید از بافر تکرار بازی آمده باشند.

۷-۴-۳ سیاست در SAC

سیاست باید در هر وضعیت برای به حداکثر رساندن بازگشت مورد انتظار آینده به همراه آنتروپی مورد انتظار آینده عمل کند. یعنی باید $V^{\pi}(s)$ را به حداکثر برساند، بسط تابع ارزش در ادامه آمده است.

$$V^{\pi}(s) = \mathbb{E}_{a \sim \pi} [Q^{\pi}(s, a)] + \alpha H(\pi(\cdot|s)) \quad (۳۷-۳)$$

$$= \mathbb{E}_{a \sim \pi} [Q^{\pi}(s, a) - \alpha \log \pi(a|s)] \quad (۳۸-۳)$$

روش بهینه سازی سیاست از ترفند پارامترسازی مجدد^{۴۹} استفاده می کند، که در آن نمونه ای از $\pi_{\theta}(\cdot|s)$ با محاسبه یک تابع قطعی از وضعیت، پارامترهای سیاست و نویز مستقل استخراج می شود. در این پژوهش مانند نویسندگان مقاله SAC [۳۱]، از یک سیاست گاوسی^{۵۰} فشرده استفاده شده است. بر اساس این روش نمونه ها مطابق با رابطه زیر بدست می آیند:

$$\tilde{a}_{\theta}(s, \xi) = \tanh(\mu_{\theta}(s) + \sigma_{\theta}(s) \odot \xi), \quad \xi \sim \mathcal{N}(0, I) \quad (۳۹-۳)$$

تابع \tanh در سیاست SAC تضمین می کند که اعمال در یک محدوده متناهی محدود شوند. این مورد در سیاست های VPG، TRPO و PPO وجود ندارد. همچنین اعمال این تابع توزیع را از حالت گاوسی تغییر

می دهد.

⁴⁹Reparameterization

⁵⁰Squashed Gaussian Policy

در الگوریتم SAC با استفاده از ترفند پارامتری‌سازی مجدد، عمل‌ها از یک توزیع نرمال به‌وسیله نویز تصادفی تولید شده و به این ترتیب امکان محاسبه مشتق‌ها به‌طور مستقیم از طریق تابع توزیع فراهم می‌شود، که باعث ثبات و کارایی بیشتر در آموزش می‌شود. اما در حالت بدون پارامتری‌سازی مجدد، عمل‌ها مستقیماً از توزیع سیاست نمونه‌برداری می‌شوند و محاسبه گرادیان نیازمند استفاده از ترفند نسبت احتمال^{۵۱} است که معمولاً باعث افزایش واریانس و ناپایداری در آموزش می‌شود.

$$\mathbb{E}_{a \sim \pi_\theta} [Q^{\pi_\theta}(s, a) - \alpha \log \pi_\theta(a|s)] = \mathbb{E}_{\xi \sim \mathcal{N}} [Q^{\pi_\theta}(s, \tilde{a}_\theta(s, \xi)) - \alpha \log \pi_\theta(\tilde{a}_\theta(s, \xi)|s)] \quad (۴۰-۳)$$

برای به‌دست آوردن تابع هزینه سیاست، گام نهایی این است که باید Q^{π_θ} را با یکی از تخمین‌زننده‌های تابع خود جایگزین کنیم. برخلاف TD3 که از Q_{ϕ_1} (فقط اولین تخمین‌زننده Q) استفاده می‌کند، SAC از $\min_{j=1,2} Q_{\phi_j}$ (کمینه‌ی دو تخمین‌زننده Q) استفاده می‌کند. بنابراین، سیاست طبق رابطه زیر بهینه‌سازی می‌شود:

$$\max_{\theta} \mathbb{E}_{\substack{s \sim \mathcal{D} \\ \xi \sim \mathcal{N}}} \left[\min_{j=1,2} Q_{\phi_j}(s, \tilde{a}_\theta(s, \xi)) - \alpha \log \pi_\theta(\tilde{a}_\theta(s, \xi)|s) \right] \quad (۴۱-۳)$$

که تقریباً مشابه بهینه‌سازی سیاست در DDPG و TD3 است، به جز ترفند min-double-Q، تصادفی بودن و عبارت آنتروپی.

۸-۴-۳ اکتشاف و بهره‌برداری در SAC

الگوریتم SAC یک سیاست تصادفی با تنظیم‌سازی آنتروپی آموزش می‌دهد و به صورت سیاست محور به اکتشاف می‌پردازد. ضریب تنظیم آنتروپی α به طور صریح تعادل بین اکتشاف و بهره‌برداری را کنترل می‌کند، به‌طوری‌که مقادیر بالاتر α به اکتشاف بیشتر و مقادیر پایین‌تر α به بهره‌برداری بیشتر منجر می‌شود. مقدار بهینه α (که به یادگیری پایدارتر و پاداش بالاتر منجر می‌شود) ممکن است در محیط‌های مختلف متفاوت باشد و نیاز به تنظیم دقیق داشته باشد. در زمان آزمایش، برای ارزیابی میزان بهره‌برداری سیاست از آنچه یاد گرفته است، تصادفی بودن را حذف کرده و از عمل میانگین به جای نمونه‌برداری از توزیع استفاده می‌کنیم. این روش معمولاً عملکرد را نسبت به سیاست تصادفی بهبود می‌بخشد.

⁵¹Likelihood Ratio Trick

۳-۴-۹ شبکه‌کد SAC

در این بخش الگوریتم SAC پیاده‌سازی شده آورده شده است. در این پژوهش الگوریتم ۳ در محیط پایتون با استفاده از کتابخانه PyTorch [۳۰] پیاده‌سازی شده است.

الگوریتم ۳ عامل عملگرد نقاد نرم

ورودی: پارامترهای اولیه سیاست (θ) ، پارامترهای تابع Q (ϕ_1, ϕ_2) ، بافر بازی خالی (\mathcal{D})

۱: پارامترهای هدف را برابر با پارامترهای اصلی قرار دهید $\theta_{\text{targ}} \leftarrow \theta$ ، $\phi_{\text{targ},1} \leftarrow \phi_1$ ، $\phi_{\text{targ},2} \leftarrow \phi_2$

۲: تا وقتی همگرایی رخ دهد:

۳: وضعیت (s) را مشاهده کرده و عمل $a \sim \pi_{\theta}(\cdot|s)$ را انتخاب کنید.

۴: عمل a را در محیط اجرا کنید.

۵: وضعیت بعدی s' ، پاداش r و سیگنال پایان d را مشاهده کنید تا نشان دهد آیا s' پایانی است یا

خیر.

۶: اگر s' پایانی است، وضعیت محیط را بازنشانی کنید.

۷: اگر زمان به‌روزرسانی فرا رسیده است:

۸: به ازای j در هر تعداد به‌روزرسانی:

۹: یک دسته تصادفی گذر از یک حالت به حالت دیگر، $B = \{(s, a, r, s', d)\}$ ، از \mathcal{D}

نمونه‌گیری شود.

۱۰: هدف را محاسبه کنید:

$$y(r, s', d) = r + \gamma(1 - d) \left(\min_{i=1,2} Q_{\phi_{\text{targ},i}}(s', \tilde{a}') - \alpha \log \pi_{\theta}(\tilde{a}'|s') \right), \quad \tilde{a}' \sim \pi_{\theta}(\cdot|s')$$

۱۱: تابع Q را با یک مرحله از نزول گرادیان با استفاده از رابطه زیر به‌روزرسانی کنید:

$$\nabla_{\phi_i} \frac{1}{|B|} \sum_{(s,a,r,s',d) \in B} (Q_{\phi_i}(s, a) - y(r, s', d))^2 \quad \text{for } i = 1, 2$$

۱۲: سیاست را با یک مرحله از صعود گرادیان با استفاده از رابطه زیر به‌روزرسانی کنید:

$$\nabla_{\theta} \frac{1}{|B|} \sum_{s \in B} \left(\min_{i=1,2} Q_{\phi_i}(s, \tilde{a}_{\theta}(s)) - \alpha \log \pi_{\theta}(\tilde{a}_{\theta}(s)|s) \right)$$

۱۳: شبکه‌های هدف را با استفاده از معادلات زیر به‌روزرسانی کنید:

$$\phi_{\text{targ},i} \leftarrow \rho \phi_{\text{targ},i} + (1 - \rho) \phi_i \quad \text{for } i = 1, 2$$

۵-۳ عامل بهینه‌سازی سیاست مجاور

الگوریتم بهینه‌سازی سیاست مجاور^{۵۲} یک الگوریتم بهینه‌سازی سیاست مبتنی بر گرادینان است که برای حل مسائل کنترل مسئله‌های یادگیری تقویتی استفاده می‌شود. این الگوریتم از الگوریتم TRPO^{۵۳} الهام گرفته شده است و با اعمال تغییراتی بر روی آن، سرعت و کارایی آن را افزایش داده است. در این بخش به بررسی این الگوریتم و نحوه عملکرد آن می‌پردازیم. الگوریتم PPO همانند سایر الگوریتم‌های یادگیری تقویتی، به دنبال یافتن بهترین گام ممکن برای بهبود عملکرد سیاست با استفاده از داده‌های موجود است. این الگوریتم تلاش می‌کند تا از گام‌های بزرگ که می‌توانند منجر به افت ناگهانی عملکرد شوند، اجتناب کند. برخلاف روش‌های پیچیده‌تر مرتبه دوم مانند TRPO، PPO از مجموعه‌ای از روش‌های مرتبه اول ساده‌تر برای حفظ نزدیکی سیاست‌های جدید به سیاست‌های قبلی استفاده می‌کند. این سادگی در پیاده‌سازی، PPO را به روشی کارآمدتر تبدیل می‌کند، در حالی که از نظر تجربی نشان داده شده است که عملکردی حداقل به اندازه TRPO دارد. از جمله ویژگی‌های مهم این الگوریتم می‌توان به سیاست محور بودن آن اشاره کرد. این الگوریتم برای عامل‌های یادگیری تقویتی که سیاست‌های پیوسته و گسسته دارند، مناسب است.

الگوریتم PPO دارای دو گونه اصلی PPO-Clip و PPO-Penalty است. در ادامه به بررسی هر یک از این دو گونه پرداخته شده است.

- **روش PPO-Penalty:** روش PPO-Penalty به دنبال حل تقریبی و به‌روزرسانی با محدودیت واگرایی کولباک-لیبلر^{۵۴} است، مشابه روشی که در الگوریتم TRPO استفاده شده است. با این حال، به جای اعمال یک محدودیت سخت^{۵۵}، PPO-Penalty واگرایی KL را در تابع هدف جریمه می‌کند. این جریمه به طور خودکار در طول آموزش تنظیم می‌شود تا از افت ناگهانی عملکرد جلوگیری کند.

- **روش PPO-Clip:** در این روش، هیچ عبارت واگرایی KL در تابع هدف وجود ندارد و هیچ محدودیتی اعمال نمی‌شود. در عوض، PPO-Clip از یک عملیات بریدن^{۵۶} خاص در تابع هدف استفاده می‌کند تا انگیزه سیاست جدید برای دور شدن از سیاست قبلی را از بین ببرد.

در این پژوهش از روش PPO-Clip برای آموزش عامل‌های یادگیری تقویتی استفاده شده است.

⁵²Proximal Policy Optimization (PPO)

⁵³Trust Region Policy Optimization

⁵⁴Kullback-Leibler (KL) Divergence

⁵⁵Hard Constraint

⁵⁶Clipping

۳-۵-۱ سیاست در الگوریتم PPO

تابع سیاست در الگوریتم PPO به صورت یک شبکه عصبی پیچیده پیاده‌سازی شده است. این شبکه عصبی ورودی‌های محیط را دریافت کرده و اقدامی را که باید عامل انجام دهد را تولید می‌کند. این شبکه عصبی می‌تواند شامل چندین لایه پنهان با توابع فعال‌سازی مختلف باشد. در این پژوهش از یک شبکه عصبی با سه لایه پنهان و تابع فعال‌سازی \tanh استفاده شده است. تابع سیاست در الگوریتم PPO به صورت زیر به‌روزرسانی می‌شود:

$$\theta_{k+1} = \arg \max_{\theta} \mathbb{E}_{s,a \sim \pi_{\theta_k}} [L(s, a, \theta_k, \theta)] \quad (۴۲-۳)$$

در این پژوهش برای به حداکثر رساندن تابع هدف، چندین گام بهینه‌سازی گرادیان کاهشی تصادفی^{۵۷} اجرا شده است. در معادله بالا L به صورت زیر تعریف شده است:

$$L(s, a, \theta_k, \theta) = \min \left(\frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s, a), \text{clip} \left(\frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)}, 1 - \epsilon, 1 + \epsilon \right) A^{\pi_{\theta_k}}(s, a) \right) \quad (۴۳-۳)$$

که در آن ϵ یک فرامتر است که مقدار آن معمولاً کوچک است. این فرامتر مشخص می‌کند که چقدر اندازه گام بهینه‌سازی باید محدود شود. در این پژوهش مقدار $\epsilon = 0.2$ انتخاب شده است. جهت سادگی در پیاده‌سازی معادله (۴۳-۳) به معادله تغییر داده شده است.

$$L(s, a, \theta_k, \theta) = \min \left(\frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s, a), g(\epsilon, A^{\pi_{\theta_k}}(s, a)) \right) \quad (۴۴-۳)$$

که تابع g به صورت زیر تعریف شده است.

$$g(\epsilon, A) = \begin{cases} (1 + \epsilon)A & A \geq 0 \\ (1 - \epsilon)A & A < 0 \end{cases} \quad (۴۵-۳)$$

در حالی که این نوع محدود کردن (PPO-Clip) تا حد زیادی به اطمینان از به‌روزرسانی‌های معقول سیاست کمک می‌کند، همچنان ممکن است با سیاست به‌دست آید که بیش از حد از سیاست قدیمی دور باشد. برای جلوگیری از این امر، پیاده‌سازی‌های مختلف PPO از مجموعه‌ای از ترفندها استفاده می‌کنند. در پیاده‌سازی این پژوهش، از روشی ساده به نام **توقف زودهنگام**^{۵۸} استفاده شده است. اگر میانگین واگرایی کولباک-لیبلر (KL) خط‌مشی جدید از خط‌مشی قدیمی از یک آستانه فراتر رود، گام‌های گرادیان (بهینه‌سازی) را متوقف می‌شوند.

⁵⁷Stochastic Gradient Descent (SGD)

⁵⁸Early Stopping

۳-۵-۲ اکتشاف و بهره‌برداری در PPO

الگوریتم PPO از یک سیاست تصادفی به صورت سیاست محور برای آموزش استفاده می‌کند. این به این معنی است که اکتشاف محیط با نمونه‌گیری عمل‌ها بر اساس آخرین نسخه از این سیاست تصادفی انجام می‌شود. میزان تصادفی بودن انتخاب عمل به شرایط اولیه و فرآیند آموزش بستگی دارد.

در طول آموزش، سیاست به طور کلی به تدریج کمتر تصادفی می‌شود، زیرا قانون به‌روزرسانی آن را تشویق می‌کند تا از پاداش‌هایی که قبلاً پیدا کرده است، بهره‌برداری کند. البته این موضوع می‌تواند منجر به گیر افتادن خط‌مشی در بهینه‌های محلی^{۵۹} شود.

۳-۵-۳ شبه‌کد PPO

در این بخش الگوریتم PPO پیاده‌سازی شده آورده شده است. در این پژوهش الگوریتم^۴ در محیط پایتون با استفاده از کتابخانه PyTorch [۳۰] پیاده‌سازی شده است.

⁵⁹Local Optima

الگوریتم ۴ بهینه‌سازی سیاست مجاور (PPO-Clip)

ورودی: پارامترهای اولیه سیاست (θ_0) ، پارامترهای تابع ارزش (ϕ_0)

۱: به ازای $k = 0, 1, 2, \dots$:

۲: مجموعه‌ای از مسیرها به نام $\mathcal{D}_k = \{\tau_i\}$ با اجرای سیاست $\pi_k = \pi(\theta_k)$ در محیط جمع‌آوری شود.

۳: پاداش‌های باقی‌مانده (\hat{R}_t) محاسبه شود.

۴: برآوردهای مزیت را محاسبه کنید، \hat{A}_t (با استفاده از هر روش تخمین مزیت) بر اساس تابع ارزش فعلی V_{ϕ_k} .

۵: سیاست را با به حداکثر رساندن تابع هدف PPO-Clip به‌روزرسانی کنید:

$$\theta_{k+1} = \arg \max_{\theta} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^T \min \left(\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} A^{\pi_{\theta_k}}(s_t, a_t), g(\epsilon, A^{\pi_{\theta_k}}(s_t, a_t)) \right)$$

معمولاً از طریق گرادیان افزایشی تصادفی Adam.

۶: برازش تابع ارزش با رگرسیون بر روی میانگین مربعات خطا:

$$\phi_{k+1} = \arg \min_{\phi} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^T \left(V_{\phi}(s_t) - \hat{R}_t \right)^2$$

معمولاً از طریق برخی از الگوریتم‌های کاهشی گرادیان.

فصل ۴

یادگیری تقویتی چند عاملی

کاربردهای پیچیده در یادگیری تقویتی نیازمند اضافه کردن چندین عامل^۱ برای انجام همزمان وظایف مختلف هستند. با این حال، افزایش تعداد عامل‌ها چالش‌هایی در مدیریت تعاملات میان آن‌ها به همراه دارد. در این فصل، بر اساس مسئله بهینه‌سازی برای هر عامل، مفهوم تعادل^۲ معرفی شده تا رفتارهای توزیعی چندعاملی را تنظیم کنند. روابط همکاری و رقابت میان عامل‌ها را در سناریوهای مختلف تحلیل شده و آن‌ها با الگوریتم‌های معمول یادگیری تقویتی چندعاملی ترکیب شده است. بر اساس انواع تعاملات، یک چارچوب نظریه بازی برای مدل‌سازی عمومی در سناریوهای چندعاملی استفاده شده است. با تحلیل بهینه‌سازی و وضعیت تعادل برای هر بخش از چارچوب، سیاست بهینه یادگیری تقویتی چندعاملی برای هر عامل بررسی شد.

۴-۱ تعاریف و مفاهیم اساسی

۴-۲ اهمیت یادگیری تقویتی چندعاملی

یادگیری تقویتی چندعاملی به دلیل قابلیت‌های بالقوه‌اش در مدل‌سازی و حل مسائل پیچیده و پویا، اهمیت زیادی در حوزه‌های مختلف علمی و صنعتی دارد. در این بخش، به بررسی اهمیت MARL در زمینه‌های مختلف پرداخته و نقش آن را در توسعه سیستم‌های هوشمند متعدد بررسی می‌کنیم.

^۱Multi-Agent

^۲Equilibrium

۴-۲-۱ نظریه بازی‌ها

۴-۲-۲ تعادل نش

تعادل نش (Nash Equilibrium) یکی از مفاهیم بنیادی در نظریه بازی‌ها است که به تحلیل تعاملات میان عوامل در محیط‌های رقابتی و همکاری می‌پردازد. این مفهوم به ما امکان می‌دهد تا نقاط تعادلی را شناسایی کنیم که در آن هیچ کدام از عوامل نمی‌توانند با تغییر یکجانبه استراتژی خود سود بیشتری کسب کنند، مشروط بر این که سایر عوامل استراتژی‌های خود را ثابت نگه دارند. در این بخش، به دو نوع اصلی تعادل نش، یعنی تعادل نش خالص و تعادل نش ترکیبی، پرداخته می‌شود.

تعادل نش خالص (Pure Strategy Nash Equilibrium)

تعادل نش خالص به حالتی از بازی گفته می‌شود که در آن هر عامل یک استراتژی مشخص و غیر تصادفی را انتخاب کرده است و هیچ کدام از عوامل نمی‌توانند با تغییر استراتژی خود بدون تغییر استراتژی‌های سایر عوامل سود بیشتری کسب کنند. به عبارت دیگر، در یک تعادل نش خالص، هیچ عاملی انگیزه‌ای برای انحراف از استراتژی انتخابی خود ندارد.

تعریف ریاضی فرض کنید یک بازی با n بازیکن داریم که هر بازیکن i دارای مجموعه‌ای از استراتژی‌ها S_i است. تابع سود هر بازیکن i به صورت $u_i : S_1 \times S_2 \times \dots \times S_n \rightarrow \mathbb{R}$ تعریف می‌شود. یک پروفایل استراتژی $(s_1^*, s_2^*, \dots, s_n^*)$ تعادل نش خالص است اگر برای هر بازیکن i و برای هر استراتژی $s_i \in S_i$ ، شرط زیر برقرار باشد:

$$u_i(s_i^*, s_{-i}^*) \geq u_i(s_i, s_{-i}^*)$$

که در آن s_{-i}^* نشان‌دهنده مجموعه استراتژی‌های سایر بازیکنان است.

مثال یک مثال کلاسیک از تعادل نش خالص، بازی Prisoner's Dilemma است. در این بازی، هر دو بازیکن انتخاب می‌کنند که همکاری کنند یا خیانت کنند. اگر هر دو بازیکن همکاری کنند، هر کدام دو سال زندانی می‌شوند. اگر یکی خیانت کند و دیگری همکاری کند، خیانتکار بی‌نیازی و همکار یک سال زندانی می‌شود. اگر هر دو خیانت کنند، هر کدام سه سال زندانی می‌شوند. تعادل نش خالص در این بازی زمانی رخ می‌دهد که هر دو بازیکن انتخاب به خیانت کردن هستند، زیرا هیچ یک از بازیکنان نمی‌توانند با تغییر استراتژی خود سود بیشتری کسب کنند.

تعادل نش ترکیبی (Mixed Strategy Nash Equilibrium)

تعادل نش ترکیبی به حالتی از بازی گفته می‌شود که در آن هر عامل می‌تواند با احتمال معینی بین استراتژی‌های مختلف خود انتخاب کند. در این نوع تعادل نش، هیچ کدام از عوامل نمی‌توانند با تغییر احتمالات انتخاب استراتژی خود بدون تغییر احتمالات انتخاب استراتژی‌های سایر عوامل سود بیشتری کسب کنند.

تعریف ریاضی در بازی‌هایی که تعادل نش خالص وجود ندارد، ممکن است تعادل نش ترکیبی وجود داشته باشد. فرض کنید مجموعه استراتژی‌های بازیکن i شامل $S_i = \{s_i^1, s_i^2, \dots, s_i^m\}$ است. یک تعادل نش ترکیبی شامل توزیع‌های احتمالاتی σ_i^* برای هر بازیکن i است که در آن هیچ بازیکن نمی‌تواند با تغییر توزیع احتمالات انتخاب استراتژی خود سود بیشتری کسب کند. به طور ریاضی، برای هر بازیکن i و هر استراتژی $s_i^j \in S_i$ شرط زیر برقرار است:

$$\sum_{s_i \in S_i} \sigma_i^*(s_i) u_i(s_i, \sigma_{-i}^*) \geq \sum_{s_i \in S_i} \sigma_i(s_i) u_i(s_i, \sigma_{-i}^*)$$

که در آن σ_{-i}^* توزیع‌های احتمالاتی سایر بازیکنان است.

مثال یک مثال از تعادل نش ترکیبی، بازی Rock-Paper-Scissors است. در این بازی، هر بازیکن می‌تواند سنگ، کاغذ یا قیچی انتخاب کند. اگر هر دو بازیکن انتخاب یک گزینه را داشته باشند، بازی مساوی است. اگر یک بازیکن سنگ انتخاب کند و دیگری قیچی، سنگ برقی قیچی را شکست می‌دهد، و به همین ترتیب. در این بازی، تعادل نش ترکیبی زمانی رخ می‌دهد که هر بازیکن با احتمال برابر $1/3$ سنگ، کاغذ و قیچی را انتخاب کند. در این حالت، هیچ کدام از بازیکنان نمی‌توانند با تغییر احتمالات انتخاب استراتژی خود سود بیشتری کسب کنند.

پیش‌بینی و استراتژی‌ها در تعادل نش ترکیبی، بازیکنان به گونه‌ای استراتژی‌های خود را انتخاب می‌کنند که بازیکنان دیگر نیز مجبور به انتخاب استراتژی‌های ترکیبی خود می‌شوند تا تعادل حفظ شود. این نوع تعادل نش به ویژه در بازی‌هایی که تعادل نش خالص وجود ندارد یا تعادل نش خالص ناامن است، اهمیت دارد.

چالش‌ها و فرصت‌ها یکی از چالش‌های اصلی در تعادل نش ترکیبی، محاسبه و پیش‌بینی توزیع‌های احتمالاتی مناسب برای هر بازیکن است. این امر به ویژه در بازی‌های با تعداد زیاد استراتژی‌ها و بازیکنان پیچیده‌تر است. با این حال، تعادل نش ترکیبی فرصت‌های زیادی را برای تحلیل و بهبود استراتژی‌های بازیکنان فراهم می‌کند، به ویژه در بازی‌هایی که نیاز به تعادل میان استراتژی‌های مختلف دارند.

کاربردها تعادل نش ترکیبی در بسیاری از حوزه‌ها کاربرد دارد، از جمله اقتصاد، سیاست، و هوش مصنوعی. در اقتصاد، این مفهوم می‌تواند برای تحلیل رقابت‌های بازار و تصمیم‌گیری‌های استراتژیک شرکت‌ها مورد استفاده قرار گیرد. در هوش مصنوعی، تعادل نش ترکیبی می‌تواند به طراحی الگوریتم‌های هوشمند برای بازی‌ها و سیستم‌های چندعاملی کمک کند که قادر به تعامل بهینه و تعادلی با سایر عوامل هستند.

نتیجه‌گیری تعادل نش، چه خالص و چه ترکیبی، ابزار قدرتمندی در تحلیل تعاملات میان عوامل در محیط‌های رقابتی و همکاری فراهم می‌کند. فهم و کاربرد این تعادل‌ها به محققان و توسعه‌دهندگان امکان می‌دهد تا الگوریتم‌های هوشمندتری طراحی کنند که قادر به اتخاذ تصمیمات بهینه در مواجهه با استراتژی‌های متغیر و پیچیده دیگر عوامل هستند. با ادامه تحقیقات در این زمینه، انتظار می‌رود که تعادل نش نقش مهم‌تری در بهبود عملکرد سیستم‌های چندعاملی ایفا کند.

۴-۲-۳ ایمنی و مقاومت در یادگیری تقویتی چندعاملی

در استفاده از یادگیری تقویتی چندعاملی (Multi-Agent Reinforcement Learning - MARL)، مسائل مربوط به ایمنی و مقاومت در برابر اختلالات یکی از چالش‌های اساسی مطرح می‌گردد. به منظور اطمینان از عملکرد قابل اعتماد و ایمن الگوریتم‌های یادگیری تقویتی چندعاملی، نیازمند توسعه روش‌هایی هستیم که بتوانند در مواجهه با رفتارهای غیرمنتظره یا مخرب سایر عوامل، پایداری و ایمنی سیستم را حفظ نمایند. در این بخش، به بررسی مفاهیم ایمنی و مقاومت در MARL پرداخته شده و چگونگی افزایش مقاومت الگوریتم‌ها از طریق در نظر گرفتن عوامل به عنوان اختلالات مورد بحث قرار گرفته است.

ایمنی در MARL به معنای تضمین این است که تعاملات میان عوامل منجر به نتایج نامطلوب یا خطرناک نشوند. برای دستیابی به ایمنی، روش‌هایی نظیر محدود کردن فضای عملیاتی، اعمال قیود بر سیاست‌های یادگیری و استفاده از الگوریتم‌های مقاوم در برابر خطا به کار گرفته شده‌اند. یکی از رویکردهای موثر در افزایش مقاومت سیستم، فرض کردن یکی از عوامل به عنوان اختلال (Disturbance) در محیط است. با این فرض، الگوریتم‌ها قادر خواهند بود تا به گونه‌ای طراحی شوند که در حضور اختلالات احتمالی، عملکرد سیستم همچنان قابل اعتماد باقی بماند.

فرض کردن اختلال به عنوان عامل

در محیط‌های چندعاملی، برخی از عوامل ممکن است رفتارهای مخرب یا غیرمنتظره‌ای را از خود نشان دهند که می‌تواند به عملکرد کلی سیستم آسیب برساند. برای مقابله با این مسئله، فرض می‌شود که یک یا چند عامل

به عنوان اختلالات در نظر گرفته شوند. این اختلالات می‌توانند به صورت عمدی یا غیرعمدی ایجاد شوند و هدف آن‌ها کاهش کارایی سیستم است. با فرض کردن این اختلالات، الگوریتم‌های MARL قادر خواهند بود تا سیاست‌هایی را یاد بگیرند که در مواجهه با این اختلالات نیز عملکرد بهینه و ایمنی را حفظ کنند.

تعریف مقاومت و ایمنی در MARL

مقاومت در MARL به معنای توانایی الگوریتم در حفظ عملکرد مطلوب در حضور اختلالات و تغییرات محیطی است. این مقاومت می‌تواند از طریق طراحی سیاست‌های بهینه که به گونه‌ای تنظیم شده‌اند که تأثیر اختلالات را به حداقل برسانند، به دست آید. به علاوه، ایمنی می‌تواند از طریق تضمین عدم وقوع رفتارهای خطرناک و حفظ تعادل سیستم در مواجهه با رفتارهای مخرب حاصل شود.

تعریف ریاضی مقاومت فرض کنید یک محیط چندعاملی با مجموعه‌ای از عوامل $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$ وجود دارد که در آن یک عامل A_d به عنوان اختلال تعریف شده است. هدف این است که الگوریتم MARL به گونه‌ای طراحی شود که سیاست‌های یادگرفته شده $\pi = \{\pi_1, \pi_2, \dots, \pi_n\}$ بتوانند عملکرد بهینه را حتی در حضور A_d حفظ کنند. به طور ریاضی، مقاومت به صورت زیر تعریف می‌شود:

باید همچنان به حداکثر رساندن پاداش خود ادامه دهد. π_i رفتار مخرب نشان دهد، اگر A_d $\forall A_i \in \mathcal{A}$,

تعریف ریاضی ایمنی ایمنی در MARL به معنای اطمینان از این است که سیستم در هیچ حالت خطرناکی وارد نمی‌شود. به طور ریاضی، ایمنی می‌تواند به صورت مجموعه‌ای از قیود تعریف شود که سیاست‌های یادگرفته شده باید آن‌ها را رعایت کنند:

$$\forall A_i \in \mathcal{A}, \quad u_i(s_i, s_{-i}) \geq \theta_i \quad \text{همه } s_i \in S_i, s_{-i} \in S_{-i}$$

که در آن θ_i آستانه‌ای است که برای هر عامل A_i تعیین شده و نشان‌دهنده حداقل پاداش قابل قبول است.

روش‌های افزایش مقاومت و ایمنی

برای افزایش مقاومت و ایمنی در MARL، روش‌های متعددی مورد استفاده قرار گرفته‌اند که در زیر به برخی از آن‌ها پرداخته می‌شود:

- الگوریتم‌های مقاوم در برابر اختلالات: این الگوریتم‌ها به گونه‌ای طراحی شده‌اند که بتوانند به سرعت

با تغییرات محیطی و حضور اختلالات سازگار شوند. به عنوان مثال، الگوریتم‌های مبتنی بر یادگیری تطبیقی که قادر به تغییر سیاست‌های خود در پاسخ به تغییرات محیط هستند.

- **فریم‌ورک‌های ایمنی:** چارچوب‌هایی برای تضمین ایمنی در تعاملات چندعاملی طراحی شده‌اند که شامل محدود کردن فضای عملیاتی و اعمال قیود بر سیاست‌های یادگیری است. این فریم‌ورک‌ها معمولاً شامل روش‌هایی برای نظارت و تنظیم رفتار عامل‌ها به منظور جلوگیری از وقوع رفتارهای خطرناک هستند.
- **آموزش در حضور اختلالات:** با آموزش الگوریتم‌ها در محیط‌هایی که شامل اختلالات هستند، می‌توان مقاومت الگوریتم‌ها را افزایش داد. این روش به الگوریتم اجازه می‌دهد تا در مواجهه با اختلالات غیرمنتظره، سیاست‌های مقاوم‌تری یاد بگیرد.
- **استفاده از اصول نظریه بازی‌ها:** با بهره‌گیری از تعادل‌های نظریه بازی‌ها مانند تعادل نش (Nash Equilibrium)، می‌توان سیاست‌هایی طراحی کرد که در مواجهه با استراتژی‌های متغیر سایر عوامل، پایداری و ایمنی سیستم حفظ شود.

نمونه‌های کاربردی

برای نشان دادن کاربردهای عملی MARL در افزایش ایمنی و مقاومت، به چند مثال اشاره می‌شود:

سامانه‌های خودران: در خودروهای خودران چندعاملی، ایمنی یکی از اولویت‌های اصلی است. با استفاده از MARL و فرض کردن سایر خودروها به عنوان عوامل یا اختلالات، می‌توان الگوریتم‌هایی توسعه داد که در مواجهه با رفتارهای غیرمنتظره سایر خودروها، ایمن باقی بمانند.

مدیریت انرژی در شبکه‌های هوشمند: در شبکه‌های انرژی هوشمند، MARL می‌تواند برای مدیریت بهینه انرژی در حضور اختلالات مانند خرابی‌ها یا حملات سایبری استفاده شود. الگوریتم‌های مقاوم می‌توانند با تغییرات ناگهانی در تقاضا یا عرضه انرژی سازگار شده و پایداری شبکه را حفظ کنند.

ربات‌های همکاری‌کننده: در سیستم‌های رباتیک همکاری‌کننده، اطمینان از ایمنی تعاملات میان ربات‌ها حیاتی است. MARL می‌تواند برای طراحی سیاست‌های ایمن که در مواجهه با رفتارهای مخرب یا اختلالات داخلی، سیستم را پایدار نگه دارند، به کار رود.

محیط‌های صنعتی: در محیط‌های صنعتی که شامل چندین عامل نظیر ربات‌ها و ماشین‌آلات است، ایمنی و مقاومت سیستم‌ها از اهمیت بالایی برخوردار است. با استفاده از MARL، می‌توان الگوریتم‌هایی توسعه داد که در مواجهه با اختلالات مانند خرابی تجهیزات یا خطاهای انسانی، عملکرد سیستم را حفظ کنند.

چالش‌ها و فرصت‌ها

با وجود مزایای متعدد MARL در افزایش ایمنی و مقاومت سیستم‌ها، چالش‌هایی نیز در این زمینه وجود دارد:

- **پیچیدگی مدل‌سازی اختلالات:** مدل‌سازی دقیق اختلالات و رفتارهای مخرب می‌تواند پیچیده و زمان‌بر باشد، به ویژه در محیط‌های پیچیده و پویا.
- **هماهنگی میان عوامل:** هماهنگی موثر میان عوامل در حضور اختلالات نیازمند الگوریتم‌های پیچیده و کارآمد است که بتوانند به سرعت و بهینه با تغییرات محیطی سازگار شوند.
- **تعادل میان کارایی و ایمنی:** حفظ تعادل میان بهینه‌سازی کارایی سیستم و تضمین ایمنی در مواجهه با اختلالات یک چالش بزرگ است که نیازمند طراحی دقیق الگوریتم‌ها است.
- **نیاز به داده‌های بزرگ:** برای آموزش الگوریتم‌های مقاوم و ایمن، نیاز به مجموعه‌های داده بزرگ و متنوعی است که شامل انواع اختلالات و رفتارهای مخرب باشند.

با این حال، فرصت‌های بزرگی نیز در این زمینه وجود دارد. توسعه الگوریتم‌های پیشرفته MARL که بتوانند به طور موثر با اختلالات مواجه شوند، می‌تواند منجر به سیستم‌های هوشمند و خودکار با کارایی و ایمنی بالاتر شود. همچنین، پژوهش‌های ادامه‌دار در زمینه نظریه بازی‌ها و روش‌های مقاوم‌سازی الگوریتم‌های یادگیری تقویتی، پتانسیل بالایی برای بهبود MARL و کاربردهای آن فراهم می‌آورد.

نتیجه‌گیری

در نهایت، ایمنی و مقاومت در یادگیری تقویتی چندعاملی به عنوان عوامل کلیدی در توسعه سیستم‌های هوشمند و خودکار مطرح می‌گردند. با فرض کردن اختلالات و توسعه الگوریتم‌های مقاوم، می‌توان عملکرد سیستم‌های MARL را در مواجهه با چالش‌های مختلف بهبود بخشید. با توجه به اهمیت و کاربردهای گسترده MARL در حوزه‌های علمی و صنعتی، تحقیقات در زمینه افزایش ایمنی و مقاومت این الگوریتم‌ها همچنان ادامه خواهد یافت تا به سیستم‌هایی با قابلیت‌های بیشتر و عملکردی پایدارتر دست یابند.

۴-۲-۴ الگوریتم‌های یادگیری تقویتی چندعاملی

فصل ۵

مدل سازی محیط یادگیری سه جسمی

مسیرهای فضایی سنتی تحت تأثیر گرانش یک جسم مرکزی (خورشید، زمین یا سیاره‌ای دیگر) شکل می‌گیرند و توسط اجسام سوم تحت تأثیر قرار می‌گیرند. تحلیل‌های مخروطی پیوسته نشان می‌دهند که چگونه می‌توان طراحی را از یک جسم مرکزی به جسم دیگر منتقل کرد، هنگامی که فضاپیما از حوزه تأثیر یک جسم عبور می‌کند. در برخی موارد، مأموریت فضاپیما آن را در ناحیه‌ای از فضا قرار می‌دهد که به‌طور همزمان تحت تأثیر دو جسم بزرگ است. این مسیرها نمی‌توانند از تحلیل دو جسم با اختلالات جسم سوم استفاده کنند، بلکه باید تأثیرات هر دو جسم به‌طور همزمان در نظر گرفته شوند. برای درک این مسئله، ابتدا به مطالعه مسئله عمومی سه جسمی خواهیم پرداخت و چگونگی اعمال آن به سیستم‌های واقعی را بررسی خواهیم کرد. سپس، برخی از مسیرها و تحلیل‌های جالبی که توسط فضاپیماهای مدرن استفاده می‌شوند، ارائه خواهیم داد.

در فصل اول معادلات عمومی حرکت اجسام متعدد را معرفی کردیم. علی‌رغم وجود ده ثابت حرکت، هیچ‌کس نتوانسته است مسئله عمومی سه جسمی را به‌صورت تحلیلی بسته حل کند—ممکن است که این کار غیرممکن باشد. بنابراین، تحقیقات بر روی ساده‌سازی مسئله عمومی متمرکز شده است. ساندمن در سال ۱۹۱۲ یک راه‌حل سری توانی یافت. زمانی که این راه‌حل با شرایط اولیه ترکیب شود، ارزیابی‌های عددی از مسیرها را در یک بازه زمانی محدود ارائه می‌دهد. برای مطالعه کامل مسئله، به کتاب Szebehely (۱۹۶۷) مراجعه کنید. یکی از راه‌حل‌های تحلیلی خاص—مسئله سه جسمی محدود—از زمان اوایل و لاگرانژ شناخته شده است. اخیراً، از تکنیک‌های عددی برای تولید راه‌حل‌ها استفاده شده است.

دانشمندان برای صدها سال مسئله سه جسمی را مطالعه کرده‌اند، اگرچه بیشتر تحقیقات از دهه ۱۹۶۰، با استفاده از فناوری محاسباتی مدرن، انجام شده است. تحلیل‌های آنها انواع مختلفی از مدارهای سه جسمی را کشف کرده است که بسیاری از آنها برای فضاپیماها بسیار مفید هستند. در این بخش، چندین گزینه مدار سه جسمی را بررسی خواهیم کرد.

۱-۵ مسئله سه جسمی محدود دایره‌ای

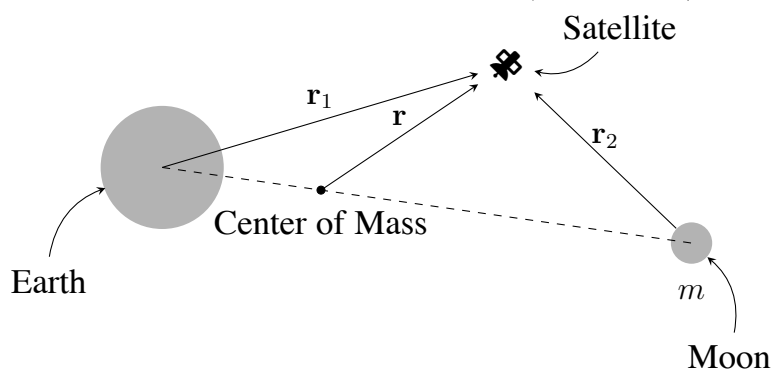
مسئله عمومی سه جسمی مسیره‌ای سه جسم پرجرم دلخواه را بررسی می‌کند که تحت تأثیر جاذبه متقابل قرار دارند. با این حال، این مسئله به مراتب عمومی‌تر از آن چیزی است که طراحان مأموریت‌های عملیاتی نیاز دارند. دو ساده‌سازی رایج وجود دارد که برای قابل دسترس‌تر و کاربردی‌تر کردن این مسئله در مسیره‌ای واقعی معمولاً انجام می‌شود. این ساده‌سازی‌ها عبارتند از:

۱. جرم جسم سوم (ماهواره) در مقایسه با اجسام اصلی ناچیز است.

۲. اجسام اصلی و ثانویه در مدارهای دایره‌ای حول مرکز جرم که بین دو جسم قرار دارد حرکت می‌کنند.

مسئله به دست آمده معمولاً به نام مسئله سه جسمی محدود دایره‌ای (CRTBP) شناخته می‌شود. گاهی اوقات، فرض دیگری نیاز به حرکت فقط در صفحه مداری اجسام اصلی و ثانویه دارد. برای این مسئله سه جسمی محدود دایره‌ای در صفحه، مؤلفه Z معادلات صفر می‌شود. پیدا کردن معادلات ساده برای بیان راه‌حل‌ها در مسئله سه جسمی دشوار است. کتاب‌های کامل در این زمینه وجود دارد (Szebehely, ۱۹۶۷). هدف ما در اینجا توصیف حرکات کیفی و برجسته کردن چندین راه‌حل کلاسیکی است که شناخته شده هستند.

ابتدا، بیایید از یک چارچوب مختصات همزمان برای مسئله سه جسمی استفاده کنیم، همانطور که در شکل ۱۲-۱۳ نشان داده شده است.* معادلاتی برای یک چارچوب باریک‌ناح (ثابت) به زودی معرفی خواهیم کرد. این چارچوب از نمادهای کوچک x ، y و z برای اجزای فردی استفاده می‌کند. زیرنویس S با محورهای نشان می‌دهد که مبدأ در مرکز جرم (مرکز سیستم) است و چارچوب با سرعت زاویه‌ای qS می‌چرخد.



شکل ۱-۵: هندسه مسئله سه بدنه محدود

فصل ۶

شبیه‌سازی عامل در محیط سه جسمی

۱-۶ آموزش عامل در یادگیری توقیتی

۱-۱-۶ تعریف عامل و فرآیند آموزش

در یادگیری توقیتی، عامل (Agent) موجودیتی است که در یک محیط قرار دارد و با آن تعامل می‌کند تا به هدف خاصی برسد. هدف عامل در این مدل، یادگیری استراتژی‌هایی است که باعث افزایش پاداش (reward) یا کاهش هزینه (cost) در درازمدت می‌شود. این فرآیند شامل دریافت اطلاعات از محیط، انتخاب اقدام‌ها (actions) و به‌روزرسانی سیاست‌ها یا مدل‌های تصمیم‌گیری است. فرآیند یادگیری عامل به‌طور کلی شامل سه مرحله اصلی است:

- انتخاب استراتژی‌ها (Policy Selection): انتخاب بهترین استراتژی برای انجام اقدامات در هر وضعیت.
- دریافت پاداش (Reward Reception): دریافت پاداش از محیط برای هر اقدام انجام‌شده.
- بهبود سیاست‌ها (Policy Improvement): با توجه به پاداش‌ها و تجربیات، به‌روزرسانی سیاست‌ها برای انتخاب بهتر اقدامات در آینده.

در این بخش، به معرفی مهم‌ترین الگوریتم‌های یادگیری توقیتی پرداخته می‌شود که برای آموزش عامل‌ها در این زمینه استفاده می‌شوند:

الگوریتم‌ها

- **DDPG (Deep Deterministic Policy Gradient):** این الگوریتم یک روش یادگیری تقویتی مبتنی بر سیاست است که از یک شبکه عصبی برای تخمین سیاست استفاده می‌کند. این الگوریتم برای حل مسائل با فضای حالت و اقدام پیوسته کاربرد دارد.
 - **TD³ (Twin Delayed Deep Deterministic Policy Gradient):** الگوریتم TD³ نسخه بهبود یافته‌ای از DDPG است که با استفاده از استراتژی‌هایی مانند تاخیر در به‌روزرسانی سیاست و استفاده از دو شبکه عمل می‌کند تا از پدیده‌های نوسانی در یادگیری جلوگیری کند.
 - **SAC (Soft Actor-Critic):** الگوریتم SAC یک روش مبتنی بر سیاست است که به‌طور همزمان به بهینه‌سازی سیاست و تابع ارزش پرداخته و از مفهوم انتروپی نرم برای بهبود بهره‌وری یادگیری استفاده می‌کند. این الگوریتم به‌ویژه در مسائل با فضای حالت و اقدام پیوسته موثر است.
 - **PPO (Proximal Policy Optimization):** الگوریتم PPO یکی از الگوریتم‌های یادگیری تقویتی مبتنی بر سیاست است که از روش‌های بهینه‌سازی موقتی برای بهبود سیاست‌های عامل استفاده می‌کند. این الگوریتم به‌ویژه برای مسائل با فضای حالت و اقدام گسسته و پیوسته مناسب است.
- این الگوریتم‌ها از طریق بهینه‌سازی تدریجی سیاست عامل‌ها، قادر به یادگیری و بهبود استراتژی‌های خود برای رسیدن به پاداش بیشتر در طول زمان هستند. انتخاب هرکدام از این الگوریتم‌ها به ویژگی‌های خاص محیط و چالش‌های یادگیری بستگی دارد.

۶-۱-۲ زمان و نوع پردازنده در آموزش عامل

آموزش عامل‌ها معمولاً به شدت به زمان و منابع پردازشی نیاز دارد. بسته به نوع الگوریتم و پیچیدگی محیط، زمان آموزش می‌تواند متفاوت باشد. در این بخش، دو موضوع کلیدی بررسی می‌شود: تأثیر زمان بر الگوریتم‌ها و تأثیر نوع پردازنده‌ها.

تأثیر زمان بر الگوریتم‌ها

زمان یادگیری به پیچیدگی محیط و الگوریتم‌های انتخابی بستگی دارد. برای مثال:

- در DDPG، که یک الگوریتم مبتنی بر سیاست است، زمان یادگیری معمولاً به تعداد وضعیت‌ها و اقدامات پیوسته بستگی دارد.

- در TD3، که نسخه بهبود یافته‌ای از DDPG است، زمان یادگیری ممکن است بیشتر باشد زیرا دو شبکه برای به‌روزرسانی استفاده می‌شود.
- SAC زمان یادگیری بیشتری نیاز دارد زیرا همزمان از دو فرآیند یادگیری سیاست و تابع ارزش استفاده می‌کند.
- در PPO، زمان یادگیری به‌طور معمول بستگی به تعداد تکرارها و حجم داده‌های ورودی دارد.

تأثیر نوع پردازنده‌ها

- پردازنده‌های مرکزی (CPU): برای الگوریتم‌های ساده‌تر مانند DDPG و TD3 که به پردازش موازی کمتری نیاز دارند، استفاده از CPU کافی است. - پردازنده‌های گرافیکی (GPU): الگوریتم‌های پیچیده‌تری مانند SAC و PPO که نیاز به پردازش موازی دارند، می‌توانند از GPU بهره‌برداری بیشتری کنند. این پردازنده‌ها می‌توانند زمان یادگیری را به‌طور چشمگیری کاهش دهند. - پردازنده‌های مخصوص یادگیری (TPU): در برخی شبیه‌سازی‌های پیچیده که به محاسبات بیشتری نیاز دارند، TPU می‌تواند زمان آموزش را به‌طور قابل توجهی کاهش دهد.

۳-۱-۶ نتیجه‌گیری

انتخاب الگوریتم‌های یادگیری توقیتی مناسب و پردازنده‌های بهینه، می‌تواند تأثیر زیادی بر کارایی آموزش عامل‌ها در محیط‌های پیچیده مانند محیط سه جسمی داشته باشد. الگوریتم‌هایی مانند DDPG، TD3، SAC و PPO هرکدام بسته به نوع چالش‌های محیط، مزایا و معایب خاص خود را دارند و انتخاب پردازنده مناسب می‌تواند به‌طور قابل توجهی زمان یادگیری و عملکرد عامل‌ها را بهبود بخشد.

فصل ۷

سخت افزار در حلقه عملکرد عامل در محیط

Bibliography

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] A. Rubinsztein, R. Sood, and F. E. Laipert. Neural network optimal control in astrodynamics: Application to the missed thrust problem. *Acta Astronautica*, 176:192–203, 2020.
- [8] 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.

- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] S. D. Smet and D. J. Scheeres. Identifying heteroclinic connections using artificial neural networks. *Acta Astronautica*, 161:192–199, 2019.
- [15] N. L. O. Parrish. *Low Thrust Trajectory Optimization in Cislunar and Translunar Space*. PhD thesis, University of Colorado Boulder, 2018.
- [16] 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.
- [17] 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.

- [18] 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.
- [19] 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.
- [20] B. Gaudet, R. Furfaro, and R. Linares. Reinforcement learning for angle-only intercept guidance of maneuvering targets. *Aerospace Science and Technology*, 99, 2020.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] D. Miller and R. Linares. Low-thrust optimal control via reinforcement learning. In *29th AAS/AIAA Space Flight Mechanics Meeting*, Ka’anapali, Hawaii, 2019.
- [25] 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.
- [26] nobelprize.org. Jean tirole, 2021. [Online; accessed October 17, 2024], Available at <https://www.nobelprize.org/prizes/economic-sciences/2014/tirole/facts/>.
- [27] 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.
- [28] 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.
- [29] S. Fujimoto, H. van Hoof, and D. Meger. Addressing function approximation error in actor-critic methods, 2018.
- [30] 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.
- [31] 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

In this study, a quadcopter stand with three degrees of freedom was controlled using game theory-based control. The first player tracks a desired input, and the second player creates a disturbance in the tracking of the first player to cause an error in the tracking. The move is chosen using the Nash equilibrium, which presupposes that the other player made the worst move.. In addition to being resistant to input interruptions, this method may also be resilient to modeling system uncertainty. This method evaluated the performance through simulation in the Simulink environment and implementation on a three-degree-of-freedom stand.

Keywords: Quadcopter, Differential Game, Game Theory, Nash Equilibrium, Three Degree of Freedom Stand, Model Base Design, Linear Quadratic Regulator



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