



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 مأموریت‌های بین مداری	۲
۲	۲-۲ بازی دیفرانسیلی	۲
۲	۳-۲ یادگیری تقویتی	۲
۲	۴-۲ ارتباط بین یادگیری تقویتی و بازی دیفرانسیلی	۲
۳	یادگیری تقویتی	۳
۳	۱-۳ مفاهیم اولیه	۳
۴	۱-۱-۳ حالت و مشاهدات	۴
۴	۲-۱-۳ فضای عمل	۴
۴	۳-۱-۳ سیاست	۴
۵	۴-۱-۳ مسیر	۵
۵	۵-۱-۳ تابع پاداش و بازگشت	۵

۶	۳-۱-۶ ارزش در یادگیری تقویتی
۷	۳-۲ عامل گرادیان سیاست عمیق قطعی
۸	۳-۲-۱ یادگیری Q در DDPG
۹	۳-۲-۲ سیاست در DDPG
۱۰	۳-۲-۳ اکتشاف و بهره‌برداری در DDPG
۱۰	۳-۲-۴ شبکه‌کد DDPG
۱۲	۳-۳ عامل گرادیان سیاست عمیق قطعی تاخیری دوگانه
۱۳	۳-۳-۱ اکتشاف و بهره‌برداری در TD3
۱۳	۳-۳-۲ شبکه‌کد TD3
۱۵	۳-۴ عامل عملگر نقاد نرم
۱۵	۳-۴-۱ یادگیری تقویتی تنظیم‌شده با آنتروپی
۱۵	۳-۴-۲ سیاست در SAC
۱۶	۳-۴-۳ تابع ارزش در SAC
۱۶	۳-۴-۴ تابع Q در SAC
۱۶	۳-۴-۵ معادله بلمن در SAC
۱۶	۳-۴-۶ یادگیری Q
۱۷	۳-۴-۷ سیاست در SAC
۱۸	۳-۴-۸ اکتشاف و بهره‌برداری در SAC
۱۸	۳-۴-۹ شبکه‌کد SAC
۲۰	۳-۵ عامل بهینه‌سازی سیاست مجاور
۲۱	۳-۵-۱ سیاست در الگوریتم PPO
۲۱	۳-۵-۲ اکتشاف و بهره‌برداری در PPO
۲۲	۳-۵-۳ شبکه‌کد PPO
۲۳	۴ مدل‌سازی محیط یادگیری سه جسمی

فهرست جداول

فهرست تصاویر

۳-۱ حلقه تعامل عامل و محیط	۴
----------------------------	---

فهرست الگوریتم‌ها

۱	گرایان سیاست عمیق قطعی	۱۱
۲	عامل گرایان سیاست عمیق قطعی تاخیری دوگانه	۱۴
۳	عامل عملگرد نقاد نرم	۱۹
۴	بهینه‌سازی سیاست مجاور (PPO-Clip)	۲۲

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 بین‌مداری

۲-۲ بازی دیفرانسیلی

۳-۲ یادگیری تقویتی

۴-۲ ارتباط بین یادگیری تقویتی و بازی دیفرانسیلی

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 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 شده‌است.

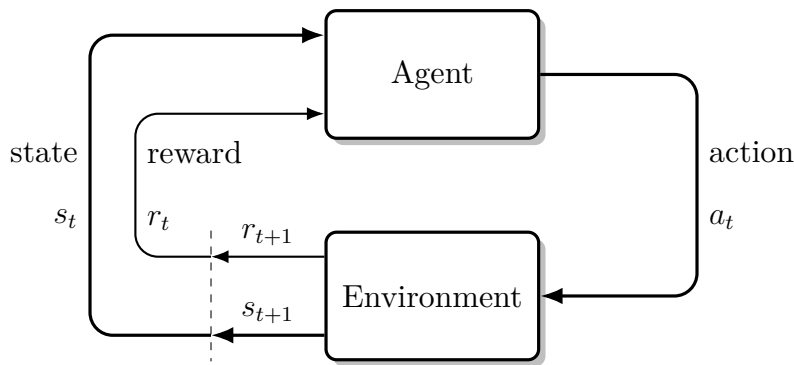
¹Reinforcement Learning (RL)

²Agent

³Environment

⁴Reward

⁵Return



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 شامل 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

^۶State

^۷Observation

^۸Discrete

^۹Continuous

^{۱۰}Policy

^{۱۱}Deterministic

^{۱۲}Stochastic

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:

$$Q^\pi(s, a) = \mathbb{E}_{\tau \sim \pi} [R(\tau) | s_0 = s, a_0 = a] \quad (۱۰-۳)$$

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:

$$V^*(s) = \max_{\pi} (V^\pi(s)) \quad (۱۱-۳)$$

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:

$$Q^*(s, a) = \max_{\pi} (Q^\pi(s, a)) \quad (۱۲-۳)$$

۲-۳ عامل گرادیان سیاست عمیق قطعی

گرادیان سیاست عمیق قطعی^{۲۷} الگوریتمی است که همزمان یک تابع Q و یک سیاست را یاد می گیرد. این الگوریتم برای یادگیری تابع Q از داده های غیرسیاست محور^{۲۸} و معادله بلمن استفاده می کند. این الگوریتم برای یادگیری سیاست نیز از تابع Q استفاده می کند.

این رویکرد وابستگی نزدیکی به یادگیری Q دارد. اگر تابع ارزش-عمل بهینه مشخص باشد، در هر حالت داده شده عمل بهینه را می توان با حل معادله (۱۳-۳) به دست آورد.

$$a^*(s) = \arg \max_a Q^*(s, a) \quad (۱۳-۳)$$

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 محیط هایی با فضا های عمل پیوسته مناسب باشد. آنچه این الگوریتم را برای فضای عمل پیوسته مناسب می کند، روش محاسبه $a^*(s)$ است. فرض می شود که تابع $Q^*(s, a)$ نسبت به آرگومان عمل مشتق پذیر است. مشتق پذیری این امکان را می دهد که یک روش یادگیری مبتنی بر گرادیان برای سیاست $\mu(s)$ استفاده شود. سپس، به جای اجرای یک بهینه سازی زمان بر در هر بار محاسبه $\max_a Q(s, a)$ ، می توان آن را با رابطه $\max_a Q(s, a) \approx Q(s, \mu(s))$ تقریب زد.

²⁵Optimal Value Function

²⁶Optimal Action-Value Function

²⁷Deep Deterministic Policy Gradient (DDPG)

²⁸Off-Policy

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.

$$Q^*(s, a) = \mathbb{E}_{s' \sim P} \left[r(s, a) + \gamma \max_{a'} Q^*(s', a') \right] \quad (۱۴-۳)$$

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^{۲۹} (MSBE) استفاده شده است که معیاری برای نزدیکی Q_ϕ به حالت بهینه برای برآورده کردن معادله بلمن است.

$$L(\phi, \mathcal{D}) = \mathbb{E}_{(s, a, r, s', d) \sim \mathcal{D}} \left[\left(Q_\phi(s, a) - \left(r + \gamma(1 - d) \max_{a'} Q_\phi(s', a') \right) \right)^2 \right] \quad (۱۵-۳)$$

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^{۳۰} تجربه‌شده 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^{۳۱} 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.

$$r + \gamma(1 - d) \max_{a'} Q_\phi(s', a') \quad (۱۶-۳)$$

²⁹Mean Squared Bellman Error

³⁰Replay Buffers

³¹Overfit

در هنگام کمینه کردن تابع خطای میانگین مربعات بلمن، سعی شده است تا تابع Q شبیه تر به هدف یعنی رابطه (۱۶-۳) 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 می شود. راه حل آن استفاده از یک مجموعه پارامترهایی است که با تأخیر زمانی به ϕ نزدیک می شوند. به عبارت دیگر، یک شبکه دوم ایجاد می شود که به آن شبکه هدف گفته می شود. شبکه هدف با تأخیر پارامترهای شبکه اول را دنبال می کند. پارامترهای شبکه هدف با نشان ϕ_{targ} نشان داده می شوند. در الگوریتم DDPG، شبکه هدف در هر به روزرسانی شبکه اصلی، با میانگین گیری پولیاک^{۳۲} به صورت زیر به روزرسانی می شود.

$$\phi_{\text{targ}} \leftarrow \rho \phi_{\text{targ}} + (1 - \rho) \phi \quad (۱۷-۳)$$

در رابطه بالا ρ یک ابرپارامتر^{۳۳} است که بین صفر و یک انتخاب می شود. در این پژوهش این مقدار نزدیک به یک در نظر گرفته شده است.

الگوریتم DDPG نیاز به یک شبکه سیاست هدف ($\mu_{\theta_{\text{targ}}}$) برای محاسبه عمل هایی که به طور 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 می آید یعنی با میانگین گیری پولیاک از پارامترهای سیاست در طول زمان آموزش استفاده می شود.

با در نظر گرفتن موارد اشاره شده، یادگیری Q در DDPG با کمینه کردن تابع خطای میانگین مربعات بلمن (MSBE) یعنی معادله (۱۸-۳) با استفاده از کاهش گرادیان تصادفی^{۳۴} انجام می شود.

$$L(\phi, \mathcal{D}) = \mathbb{E}_{(s,a,r,s',d) \sim \mathcal{D}} \left[\left(Q_{\phi}(s, a) - (r + \gamma(1-d)Q_{\phi_{\text{targ}}}(s', \mu_{\theta_{\text{targ}}}(s'))) \right)^2 \right] \quad (۱۸-۳)$$

۲-۲-۳ سیاست در DDPG

در این بخش یک سیاست تعیین شده $\mu_{\theta}(s)$ یاد گرفته می شود تا عملی را انجام می دهد که بیشینه $Q_{\phi}(s, a)$ رخ دهد. از آنجا که فضای عمل پیوسته است و فرض شده است که تابع Q نسبت به عمل مشتق پذیر است، رابطه زیر با استفاده از صعود گرادیان^{۳۵} (تنها نسبت به پارامترهای سیاست) بیشینه می شود.

$$\max_{\theta} \mathbb{E}_{s \sim \mathcal{D}} [Q_{\phi}(s, \mu_{\theta}(s))] \quad (۱۹-۳)$$

^{۳۲}Polyak Averaging

^{۳۳}Hyperparameter

^{۳۴}Stochastic Gradient Descent

^{۳۵}Gradient Ascent

۳-۲-۳ اکتشاف و بهره‌برداری در DDPG

برای بهبود اکتشاف^{۳۶} در سیاست‌های DDPG، در زمان آموزش نویز به عمل‌ها اضافه می‌شود. نویسندگان مقاله DDPG [۱] توصیه کرده‌اند که نویز OU^{۳۷} با هم‌بندی زمانی^{۳۸} اضافه شود. در زمان بهره‌برداری^{۳۹} 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 شده است. در این پژوهش الگوریتم ۱ در محیط پایتون با استفاده از کتابخانه TensorFlow [۲] پیاده‌سازی شده است.

³⁶Exploration

³⁷Ornstein–Uhlenbeck

³⁸Time-Correlated

³⁹Exploitation

الگوریتم ۱ گرادیان سیاست عمیق قطعی

ورودی: پارامترهای اولیه سیاست (θ) ، پارامترهای تابع $Q(\phi)$ ، بافر تکرار بازی خالی (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 به روزرسانی فرا رسیده است:

۸: به ازای هر تعداد به روزرسانی:

۹: یک دسته تصادفی گذر از یک حالت به حالت دیگر، $B = \{(s, a, r, s', d)\}$ ، از 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 به روزرسانی 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 به روزرسانی 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 به روزرسانی 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^{۴۰} یکی از الگوریتم‌های یادگیری تقویتی است که برای حل مسائل کنترل در محیط‌های پیوسته طراحی شده است. 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^{۴۱}: الگوریتم TD3 به جای یک تابع Q ، دو تابع Q_{ϕ_1} و Q_{ϕ_2} را یاد می‌گیرد (از این رو دوگانه^{۴۲} 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 (۲۰-۳) آورده شده است.

$$y(r, s', d) = r + \gamma(1 - d) \min_{i=1,2} Q_{\phi_i, \text{targ}}(s', a'(s')) \quad (20-3)$$

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

$$L(\phi_1, \mathcal{D}) = \mathbb{E}_{(s,a,r,s',d) \sim \mathcal{D}} \left(Q_{\phi_1}(s, a) - y(r, s', d) \right)^2 \quad (21-3)$$

$$L(\phi_2, \mathcal{D}) = \mathbb{E}_{(s,a,r,s',d) \sim \mathcal{D}} \left(Q_{\phi_2}(s, a) - y(r, s', d) \right)^2 \quad (22-3)$$

- به‌روزرسانی‌های تاخیری سیاست^{۴۳}: الگوریتم TD3 سیاست را با تاخیر بیشتری نسبت به تابع Q به‌روزرسانی می‌کند. در مرجع [۳] توصیه شده است که برای هر دو به‌روزرسانی تابع Q ، یک به‌روزرسانی سیاست انجام شود.

⁴⁰Twin Delayed Deep Deterministic Policy Gradient (TD3)

⁴¹Clipped Double-Q Learning

⁴²twin

⁴³Delayed 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:

$$\max_{\theta} E_{s \sim \mathcal{D}} [Q_{\phi_1}(s, \mu_{\theta}(s))] \quad (23-3)$$

۱-۳-۳ اکتشاف و بهره‌برداری در TD3

الگوریتم TD3 یک سیاست قطعی را به صورت غیر سیاست محور آموزش را 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 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 [۴] 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^{۴۴} 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^{۴۵} 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^{۴۶} 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 گونه ای 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:

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

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.

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

⁴⁴Soft Actor Critic (SAC)

⁴⁵Target Policy Smoothing

⁴⁶Entropy-Regularized Reinforcement Learning

که در آن ($\alpha > 0$) ضریب مبادله^{۴۷} 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.

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

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.

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

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

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

۵-۴-۳ معادله بلمن در SAC

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

$$Q^\pi(s, a) = \mathbb{E}_{\substack{s' \sim P \\ a' \sim \pi}} [R(s, a, s') + \gamma (Q^\pi(s', a') + \alpha H(\pi(\cdot|s')))] \quad (۲۴-۳)$$

$$= \mathbb{E}_{s' \sim P} [R(s, a, s') + \gamma V^\pi(s')] \quad (۲۵-۳)$$

۶-۴-۳ یادگیری Q

با در نظر گرفتن موارد اشاره‌شده، یادگیری Q در SAC با کمینه‌کردن تابع خطای میانگین مربعات بلمن (MSBE) یعنی معادله (۶-۴-۳) با استفاده از کاهش گرادیان انجام می‌شود.

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

⁴⁷Trade-Off

در معادله (۶-۴-۳) تابع هدف برای روش یادگیری تقویتی SAC به صورت زیر تعریف می‌شود.

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

نماد عمل بعدی را به جای a' به \tilde{a}' تغییر داده شده تا مشخص شود که عمل‌های بعدی باید آخرین سیاست نمونه‌برداری 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.

$$V^{\pi}(s) = \mathbb{E}_{a \sim \pi} [Q^{\pi}(s, a)] + \alpha H(\pi(\cdot|s)) \quad (۲۶-۳)$$

$$= \mathbb{E}_{a \sim \pi} [Q^{\pi}(s, a) - \alpha \log \pi(a|s)] \quad (۲۷-۳)$$

روش بهینه‌سازی سیاست از ترفند پارامتری‌سازی مجدد^{۴۸} استفاده می‌کند، که در آن نمونه ای از $\pi_{\theta}(\cdot|s)$ با محاسبه یک تابع قطعی از وضعیت، پارامترهای سیاست و نویز مستقل استخراج می‌شود. در این پژوهش مانند نویسندگان مقاله SAC [۵]، از یک سیاست گاوسی^{۴۹} فشرده استفاده شده است. بر اساس این روش نمونه‌ها مطابق با رابطه زیر بدست می‌آیند:

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

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

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^{۵۰} است که معمولاً باعث افزایش واریانس و ناپایداری در آموزش می‌شود.

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

⁴⁸Reparameterization

⁴⁹Squashed Gaussian Policy

⁵⁰Likelihood Ratio Trick

برای به دست 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:

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

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، به طوری که 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.

۹-۴-۳ شبه‌کد SAC

در این بخش الگوریتم SAC پیاده‌سازی شده آورده شده است. در این پژوهش الگوریتم ۳ در محیط پایتون با استفاده از کتابخانه PyTorch [۴] پیاده‌سازی شده است.

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^{۵۱} یک الگوریتم بهینه‌سازی سیاست مبتنی بر گرادین است که برای حل مسائل کنترل مسئله‌های یادگیری تقویتی استفاده می‌شود. این الگوریتم از الگوریتم TRPO^{۵۲} 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^{۵۳} است، مشابه روشی که در الگوریتم TRPO استفاده شده است. با این حال، به جای اعمال یک محدودیت سخت^{۵۴}، 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^{۵۵} 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.

⁵¹Proximal Policy Optimization (PPO)

⁵²Trust Region Policy Optimization

⁵³Kullback-Leibler (KL) Divergence

⁵⁴Hard Constraint

⁵⁵Clipping

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:

$$\theta_{k+1} = \arg \max_{\theta} \mathbb{E}_{s,a \sim \pi_{\theta_k}} [L(s, a, \theta_k, \theta)] \quad (۲۸-۳)$$

در این پژوهش برای به حداکثر رساندن تابع هدف، چندین گام بهینه‌سازی گرادیان کاهشی تصادفی^{۵۶} اجرا شده است. در معادله بالا L به صورت زیر تعریف شده است:

$$L(s, a, \theta_k, \theta) = \min \left(\frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s, a), \text{clip} \left(\frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)}, 1 - \epsilon, 1 + \epsilon \right) A^{\pi_{\theta_k}}(s, a) \right) \quad (۲۹-۳)$$

که در آن ϵ یک فرایمتر است که مقدار آن معمولاً کوچک است. این فرایمتر مشخص می‌کند که چقدر اندازه گام بهینه‌سازی باید محدود شود. در این پژوهش مقدار $\epsilon = 0.2$ انتخاب شده است.

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^{۵۷} 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.

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

^{۵۶}Stochastic Gradient Descent (SGD)

^{۵۷}Early Stopping

می‌کند تا از پاداش‌هایی که قبلاً پیدا کرده است، بهره‌برداری کند. البته این موضوع می‌تواند منجر به گیر افتادن خط‌مشی در بهینه‌های محلی^{۵۸} 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 [۴] پیاده‌سازی شده است.

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 کاهش گرادیان.

⁵⁸Local Optima

فصل ۴

مدل سازی محیط یادگیری سه جسمی

فصل ۵

شبیه‌سازی عامل در محیط سه جسمی

Bibliography

- [1] 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.
- [2] 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.
- [3] S. Fujimoto, H. van Hoof, and D. Meger. Addressing function approximation error in actor-critic methods, 2018.
- [4] 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.
- [5] 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