



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<sup>۱</sup> 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<sup>۲</sup> 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

---

<sup>1</sup>Game Theory

<sup>2</sup>Nash Equilibrium

# فهرست مطالب

۱	مقدمه	۱
۱	۱-۱ انگیزه پژوهش	۱
۱	۲-۱ تعریف مسئله	۱
۲	۳-۱ اهداف و نوآوری	۲
۲	۴-۱ یادگیری تقویتی	۲
۲	۵-۱ یادگیری تقویتی چند عاملی	۲
۲	۶-۱ محتوای گزارش	۲
۳	۲ پیشینه پژوهش	۳
۳	۱-۲ ماموریت‌های بین مداری	۳
۵	۲-۲ یادگیری تقویتی	۵
۵	۳-۲ یادگیری تقویتی چندعاملی	۵
۶	۳ یادگیری تقویتی	۶
۶	۱-۳ مفاهیم اولیه	۶
۷	۳-۱-۱ حالت و مشاهدات	۷
۷	۳-۱-۲ فضای عمل	۷
۷	۳-۱-۳ سیاست	۷
۸	۳-۱-۴ مسیر	۸

۸	۵-۱-۳ تابع پاداش و بازگشت
۹	۶-۱-۳ ارزش در یادگیری تقویتی
۱۰	۷-۱-۳ معادلات بلمن
۱۱	۸-۱-۳ تابع مزیت
۱۲	۲-۳ عامل گرادیان سیاست عمیق قطعی
۱۲	۱-۲-۳ یادگیری Q در DDPG
۱۴	۲-۲-۳ سیاست در DDPG
۱۴	۳-۲-۳ اکتشاف و بهره‌برداری در DDPG
۱۴	۴-۲-۳ شبکه‌د DDPG
۱۶	۳-۳ عامل گرادیان سیاست عمیق قطعی تاخیری دوگانه
۱۷	۱-۳-۳ اکتشاف و بهره‌برداری در TD3
۱۷	۲-۳-۳ شبکه‌د TD3
۱۹	۴-۳ عامل عملگر نقاد نرم
۱۹	۱-۴-۳ یادگیری تقویتی تنظیم‌شده با آنتروپی
۱۹	۲-۴-۳ سیاست در SAC
۲۰	۳-۴-۳ تابع ارزش در SAC
۲۰	۴-۴-۳ تابع Q در SAC
۲۰	۵-۴-۳ معادله بلمن در SAC
۲۱	۶-۴-۳ یادگیری Q
۲۱	۷-۴-۳ سیاست در SAC
۲۲	۸-۴-۳ اکتشاف و بهره‌برداری در SAC
۲۳	۹-۴-۳ شبکه‌د SAC
۲۴	۵-۳ عامل بهینه‌سازی سیاست مجاور
۲۵	۱-۵-۳ سیاست در الگوریتم PPO

۲۶	۲-۵-۳	اکتشاف و بهره‌برداری در PPO
۲۶	۳-۵-۳	شبکه‌د PPO
۲۸	۴	یادگیری تقویتی چند عاملی
۲۸	۱-۴	تعاریف و مفاهیم اساسی
۲۹	۲-۴	نظریه بازی‌ها
۲۹	۱-۲-۴	تعادل نش
۳۰	۲-۲-۴	بازی مجموع صفر
۳۱	۳-۴	الگوریتم‌های یادگیری تقویتی چندعاملی
۳۲	۵	مدل‌سازی محیط یادگیری سه جسمی
۳۳	۱-۵	مسئله سه جسمی محدود دایره‌ای
۳۴	۶	شبیه‌سازی عامل در محیط سه جسمی
۳۵	۷	سخت‌افزار در حلقه عملکرد عامل در محیط



# فهرست جداول

## فهرست تصاویر

۷	.....	۱-۳ حلقه تعامل عامل و محیط
۳۳	.....	۱-۵ هندسه مسئله سه بدنه محدود

# فهرست الگوریتم‌ها

۱	گرایان سیاست عمیق قطعی	۱۵
۲	عامل گرایان سیاست عمیق قطعی تاخیری دوگانه	۱۸
۳	عامل عملگرد نقاد نرم	۲۳
۴	بهینه‌سازی سیاست مجاور (PPO-Clip)	۲۷

# 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 یادگیری تقویتی

۵-۱ یادگیری تقویتی چند عاملی

۶-۱ محتوای گزارش

## فصل ۲

### پیشینه پژوهش

#### ۱-۲ مأموریت‌های 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 [۵، ۶]، 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 یادگیری تقویتی

۳-۲ یادگیری تقویتی چندعاملی



## فصل ۳

# یادگیری تقویتی

### ۱-۳ مفاهیم اولیه

دو بخش اصلی یادگیری تقویتی<sup>۱</sup> شامل عامل<sup>۲</sup> و محیط<sup>۳</sup> 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<sup>۴</sup> 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<sup>۵</sup> 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.

---

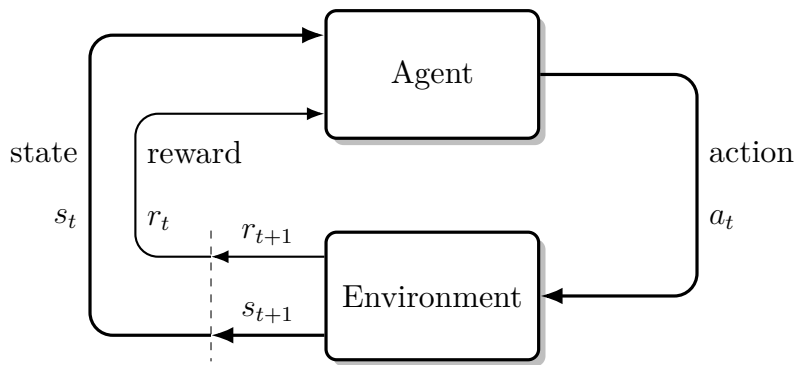
<sup>۱</sup>Reinforcement Learning (RL)

<sup>۲</sup>Agent

<sup>۳</sup>Environment

<sup>۴</sup>Reward

<sup>۵</sup>Return



شکل ۱-۳: حلقه تعامل عامل و محیط

### ۱-۱-۳ حالت و مشاهدات

حالت<sup>۶</sup> ( $s$ ) توصیف کاملی از وضعیت محیط است. همه‌ی اطلاعات محیط در حالت وجود دارد. مشاهده<sup>۷</sup> ( $o$ ) 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<sup>۸</sup> یا پیوسته<sup>۹</sup> باشد. در این پژوهش فضای عمل پیوسته و محدود به یک بازه مشخص است.

### ۳-۱-۳ سیاست

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

<sup>۶</sup>State

<sup>۷</sup>Observation

<sup>۸</sup>Discrete

<sup>۹</sup>Continuous

<sup>۱۰</sup>Policy

<sup>۱۱</sup>Deterministic

<sup>۱۲</sup>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  $\theta$  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<sup>۱۳</sup> توالی‌ای از حالت‌ها و عمل‌ها در محیط است.

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

گذار حالت<sup>۱۴</sup> 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<sup>۱۵</sup> به حالت فعلی محیط، آخرین عمل انجام شده و حالت بعدی محیط بستگی دارد. تابع پاداش را می‌توان به صورت زیر تعریف کرد.

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

<sup>13</sup>Trajectory

<sup>14</sup>State Transition

<sup>15</sup>Reward 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<sup>۱۶</sup> گفته می‌شود. یکی از انواع بازگشت، بازگشت بدون تنزیل<sup>۱۷</sup> با افق محدود<sup>۱۸</sup> است که مجموع پاداش‌های به‌دست‌آمده در یک بازه زمانی ثابت و از مسیر  $\tau$  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<sup>۱۹</sup> 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<sup>۲۰</sup> با نماد  $\gamma$  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<sup>۲۱</sup> یک حالت یا جفت حالت-عمل ضروری است. منظور از ارزش، بازگشت مورد انتظار<sup>۲۲</sup> 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<sup>۲۳</sup> ( $V^\pi(s)$ ): خروجی این تابع بازگشت مورد انتظار است در صورتی که از حالت  $s$  شروع شود و همیشه طبق سیاست  $\pi$  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<sup>۲۴</sup> ( $Q^\pi(s, a)$ ): خروجی این تابع بازگشت مورد انتظار است در صورتی که از حالت  $s$  شروع شود، یک اقدام دلخواه  $a$  (که ممکن است از سیاست  $\pi$  نباشد) انجام شود و سپس

<sup>16</sup>Return

<sup>17</sup>Discount

<sup>18</sup>Finite-Horizon Undiscounted Return

<sup>19</sup>Infinite-Horizon Discounted Return

<sup>20</sup>Discount Factor

<sup>21</sup>Value

<sup>22</sup>Expected Return

<sup>23</sup>On-Policy Value Function

<sup>24</sup>On-Policy Action-Value Function

برای همیشه طبق سیاست  $\pi$  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<sup>۲۵</sup> ( $V^*(s)$ ): 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<sup>۲۶</sup> ( $Q^*(s, a)$ ): 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:

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

### 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:

$$V^\pi(s) = \mathbb{E}_{\substack{a \sim \pi \\ s' \sim P}} [r(s, a) + \gamma V^\pi(s')] \quad (۱۳-۳)$$

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

که در آن  $V^\pi(s)$  تابع ارزش حالت  $s$  تحت سیاست  $\pi$  است؛  $Q^\pi(s, a)$  تابع ارزش عمل  $a$  در حالت  $s$  تحت سیاست  $\pi$  است؛  $R(s, a)$  پاداش دریافتی پس از انجام عمل  $a$  در حالت  $s$  است؛  $\gamma$  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

<sup>25</sup>Optimal Value Function

<sup>26</sup>Optimal Action-Value Function

$\pi$  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<sup>۲۷</sup> توضیح داده می‌شود.

تابع مزیت  $A^\pi(s, a)$  که مربوط به سیاست  $\pi$  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  $\pi$  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  $\pi$  است و  $V^\pi(s)$  تابع ارزش حالت  $s$  تحت سیاست  $\pi$  است. این تابع مزیت نشان می‌دهد که انجام عمل  $a$  در حالت  $s$  نسبت به میانگین اعمال تحت سیاست  $\pi$  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.

<sup>27</sup> Advantage Function

## ۲-۳ عامل گرادیان سیاست عمیق قطعی

گرادیان سیاست عمیق قطعی<sup>۲۸</sup> الگوریتمی است که همزمان یک تابع  $Q$  و یک سیاست را یاد می‌گیرد. این الگوریتم برای یادگیری تابع  $Q$  از داده‌های غیرسیاست محور<sup>۲۹</sup> 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  $\phi$  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<sup>۳۰</sup> (MSBE) استفاده شده است که معیاری برای نزدیکی  $Q_\phi$  به حالت بهینه برای برآورده کردن معادله بلمن است.

<sup>28</sup>Deep Deterministic Policy Gradient (DDPG)

<sup>29</sup>Off-Policy

<sup>30</sup>Mean Squared Bellman Error

$$L(\phi, \mathcal{D}) = \mathbb{E}_{(s,a,r,s',d) \sim \mathcal{D}} \left[ \left( Q_{\phi}(s, a) - \left( r + \gamma(1-d) \max_{a'} Q_{\phi}(s', a') \right) \right)^2 \right] \quad (۱۹-۳)$$

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<sup>۳۱</sup> 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<sup>۳۲</sup> 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.

$$r + \gamma(1-d) \max_{a'} Q_{\phi}(s', a') \quad (۲۰-۳)$$

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  $\phi$  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  $\phi$  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  $\phi_{\text{targ}}$  نشان داده می‌شوند. در الگوریتم DDPG، شبکه هدف در هر به‌روزرسانی شبکه اصلی، با میانگین‌گیری پولیاک<sup>۳۳</sup> به صورت زیر به‌روزرسانی می‌شود.

$$\phi_{\text{targ}} \leftarrow \rho \phi_{\text{targ}} + (1 - \rho) \phi \quad (۲۱-۳)$$

در رابطه بالا  $\rho$  یک ابرپارامتر<sup>۳۴</sup> است که بین صفر و یک انتخاب می‌شود. در این پژوهش این مقدار نزدیک به یک در نظر گرفته شده است.

<sup>31</sup>Replay Buffers

<sup>32</sup>Overfit

<sup>33</sup>Polyak Averaging

<sup>34</sup>Hyperparameter



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<sup>۳۵</sup> 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<sup>۳۶</sup> (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<sup>۳۷</sup> در سیاست‌های DDPG، در زمان آموزش نویز به عمل‌ها اضافه می‌شود. نویسندگان مقاله DDPG [۲۶] توصیه کرده‌اند که نویز OU<sup>۳۸</sup> با هم‌بندی زمانی<sup>۳۹</sup> اضافه شود. در زمان بهره‌برداری<sup>۴۰</sup> 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 [۲۷] پیاده‌سازی شده‌است.

<sup>35</sup>Stochastic Gradient Descent

<sup>36</sup>Gradient Ascent

<sup>37</sup>Exploration

<sup>38</sup>Ornstein-Uhlenbeck

<sup>39</sup>Time-Correlated

<sup>40</sup>Exploitation

ورودی: پارامترهای اولیه سیاست  $(\theta)$ ، 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<sup>۴۱</sup> 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<sup>۴۲</sup>: الگوریتم TD3 به جای یک تابع  $Q$ ، دو تابع  $Q_{\phi_1}$  و  $Q_{\phi_2}$  را یاد می‌گیرد (از این رو دوگانه<sup>۴۳</sup> نامیده می‌شود) و از کوچک‌ترین مقدار این دو  $Q_{\phi_1}$  و  $Q_{\phi_2}$  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  $Q_{\phi_1}$  و  $Q_{\phi_2}$  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<sup>۴۴</sup>: الگوریتم TD3 سیاست را با تاخیر بیشتری نسبت به تابع  $Q$  به‌روزرسانی می‌کند. در مرجع [۲۸] توصیه شده‌است که برای هر دو به‌روزرسانی تابع  $Q$ ، یک به‌روزرسانی سیاست انجام شود.

<sup>41</sup>Twin Delayed Deep Deterministic Policy Gradient (TD3)

<sup>42</sup>Clipped Double-Q Learning

<sup>43</sup>twin

<sup>44</sup>Delayed Policy Updates

این دو ترفند منجر به بهبود قابل توجه عملکرد TD3 نسبت به DDPG پایه می‌شوند. در نهایت سیاست با به حداکثر رساندن  $Q_{\phi_1}$  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 [۲۹] پیاده‌سازی شده‌است.

---

## الگوریتم ۲ عامل گرادیان سیاست عمیق قطعی تاخیری دوگانه

---

ورودی: پارامترهای اولیه سیاست  $(\theta)$ ، پارامترهای تابع  $Q$   $(\phi_1, \phi_2)$ ، 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<sup>۴۵</sup> 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<sup>۴۶</sup> 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<sup>۴۷</sup> 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.

<sup>۴۵</sup>Soft Actor Critic (SAC)

<sup>۴۶</sup>Target Policy Smoothing

<sup>۴۷</sup>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<sup>۴۸</sup> است.

### ۳-۴-۳ تابع ارزش در SAC

اکنون می‌توان تابع ارزش کمی متفاوت را بر اساس این مفهوم تعریف کرد.  $V^\pi$  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  $Q^\pi$  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  $V^\pi$  و  $Q^\pi$  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.

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

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

---

<sup>48</sup>Trade-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<sup>۴۹</sup> 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 [۳۰]، از یک سیاست گاوسی<sup>۵۰</sup> 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.

<sup>49</sup>Reparameterization

<sup>50</sup>Squashed 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<sup>۵۱</sup> 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  $Q^{\pi_\theta}$  را با یکی از تخمین‌زننده‌های تابع خود جایگزین کنیم. برخلاف TD3 که از  $Q_{\phi_1}$  (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  $\alpha$  به طور صریح تعادل بین اکتشاف و بهره‌برداری را کنترل می‌کند، به‌طوری‌که مقادیر بالاتر  $\alpha$  به اکتشاف بیشتر و مقادیر پایین‌تر  $\alpha$  به بهره‌برداری بیشتر منجر می‌شود. مقدار بهینه  $\alpha$  (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.

<sup>51</sup>Likelihood 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 [۲۹] پیاده‌سازی شده است.

الگوریتم ۳ عامل عملگرد نقاد نرم

ورودی: پارامترهای اولیه سیاست  $(\theta)$ ، پارامترهای تابع  $Q$   $(\phi_1, \phi_2)$ ، 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<sup>۵۲</sup> یک الگوریتم بهینه‌سازی سیاست مبتنی بر گرادینان است که برای حل مسائل کنترل مسئله‌های یادگیری تقویتی استفاده می‌شود. این الگوریتم از الگوریتم TRPO<sup>۵۳</sup> 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<sup>۵۴</sup> است، مشابه روشی که در الگوریتم TRPO استفاده شده است. با این حال، به جای اعمال یک محدودیت سخت<sup>۵۵</sup>، 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<sup>۵۶</sup> 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.

<sup>52</sup>Proximal Policy Optimization (PPO)

<sup>53</sup>Trust Region Policy Optimization

<sup>54</sup>Kullback-Leibler (KL) Divergence

<sup>55</sup>Hard Constraint

<sup>56</sup>Clipping

### 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<sup>۵۷</sup> 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  $\epsilon$  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**<sup>۵۸</sup> 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.

<sup>57</sup>Stochastic Gradient Descent (SGD)

<sup>58</sup>Early 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<sup>۵۹</sup> شود.

### ۳-۵-۳ شبه‌کد PPO

در این بخش الگوریتم PPO پیاده‌سازی شده آورده شده است. در این پژوهش الگوریتم<sup>۴</sup> در محیط پایتون با استفاده از کتابخانه PyTorch [۲۹] پیاده‌سازی شده است.

---

<sup>59</sup>Local Optima

---

## الگوریتم ۴ بهینه‌سازی سیاست مجاور (PPO-Clip)

---

ورودی: پارامترهای اولیه سیاست  $(\theta_0)$ ، پارامترهای تابع ارزش  $(\phi_0)$

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  $V_{\phi_k}$ .

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<sup>۱</sup> 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<sup>۲</sup> معرفی شده تا رفتارهای توزیعی چندعاملی را تنظیم کند. رابطه رقابت میان عامل‌ها در سناریوهای مختلف تحلیل شده و آن‌ها با الگوریتم‌های معمول یادگیری تقویتی چندعاملی ترکیب شده‌اند. 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 شده است.

## ۴-۱ تعاریف و مفاهیم اساسی

یادگیری تقویتی چندعاملی<sup>۳</sup> به بررسی چگونگی یادگیری و تصمیم‌گیری چندین عامل مستقل در یک محیط مشترک پرداخته می‌شود. برای تحلیل دقیق و درک بهتر این حوزه، اجزای اصلی آن شامل عامل، سیاست و مطلوبیت<sup>۴</sup> در نظر گرفته می‌شوند که در ادامه به صورت مختصر و منسجم تشریح می‌گردند.

- عامل: یک موجودیت مستقل به عنوان عامل تعریف می‌شود که به صورت خودمختار با محیط تعامل کرده و بر اساس مشاهدات رفتار سایر عامل‌ها، سیاست‌هایش انتخاب می‌گردند تا سود حداکثر یا ضرر حداقل حاصل شود. در سناریوهای مورد بررسی، چندین عامل به صورت مستقل عمل می‌کنند؛ اما اگر

---

<sup>1</sup>Multi-Agent

<sup>2</sup>Equilibrium

<sup>3</sup>Multi-Agent Reinforcement Learning (MARL)

<sup>4</sup>Utility

تعداد عامل‌ها به یک کاهش یابد، MARL به یادگیری تقویتی معمولی تبدیل می‌شود.

- سیاست: برای هر عامل در MARL، سیاستی خاص در نظر گرفته می‌شود که به عنوان روشی برای انتخاب اقدامات بر اساس وضعیت محیط و رفتار سایر عامل‌ها تعریف می‌گردد. این سیاست‌ها با هدف به حداکثر رساندن سود و به حداقل رساندن هزینه طراحی شده و تحت تأثیر محیط و سیاست‌های دیگر عامل‌ها قرار می‌گیرند.

- مطلوبیت: مطلوبیت هر عامل بر اساس نیازها و وابستگی‌هایش به محیط و سایر عامل‌ها تعریف شده و به صورت سود منهای هزینه، با توجه به اهداف مختلف محاسبه می‌شود. در سناریوهای چندعاملی، از طریق یادگیری از محیط و تعامل با دیگران، مطلوبیت هر عامل بهینه می‌گردد.

در این چارچوب، برای هر عامل در MARL تابع مطلوبیت خاصی در نظر گرفته شده و بر اساس مشاهدات و تجربیات حاصل از تعاملات، یادگیری سیاست به صورت مستقل انجام می‌شود تا ارزش مطلوبیت به حداکثر برسد، بدون اینکه مستقیماً به مطلوبیت سایر عامل‌ها توجه شود. این فرآیند ممکن است به رقابت یا همکاری میان عامل‌ها منجر گردد. با توجه به پیچیدگی تعاملات میان چندین عامل، تحلیل نظریه بازی‌ها به عنوان ابزاری مؤثر برای تصمیم‌گیری در این حوزه به کار گرفته می‌شود. بسته به سناریوهای مختلف، این بازی‌ها در دسته‌بندی‌های متفاوتی قرار داده شده که در بخش‌های بعدی بررسی خواهند شد.

## ۲-۴ نظریه بازی‌ها

نظریه بازی‌ها شاخه‌ای از ریاضیات است که به مطالعه تصمیم‌گیری در موقعیت‌هایی می‌پردازد که نتیجه انتخاب‌های هر فرد به تصمیمات دیگران وابسته است. این نظریه چارچوبی برای تحلیل تعاملات میان بازیکنان ارائه می‌دهد و در حوزه‌های مختلفی مانند اقتصاد، علوم سیاسی، زیست‌شناسی و علوم کامپیوتر کاربرد دارد. در این فصل، دو مفهوم کلیدی نظریه بازی‌ها یعنی تعادل نش و بازی‌های مجموع صفر بررسی شده است.

### ۱-۲-۴ تعادل نش

تعادل نش<sup>۵</sup> یکی از بنیادی‌ترین مفاهیم در نظریه بازی‌ها است که توسط جان نش در سال ۱۹۵۰ معرفی شد. این مفهوم به مجموعه‌ای ازها اشاره دارد که در آن هیچ بازیکنی نمی‌تواند با تغییر یک‌جانبه سیاست خود، سود بیشتری به دست آورد، به شرطی که سیاست‌های سایر بازیکنان ثابت بماند.

<sup>5</sup>Nash Equilibrium



- تعریف تعادل نش: فرض کنید یک بازی با  $n$  بازیکن داریم. هر بازیکن  $i$  دارای مجموعه سیاست‌های  $\Pi_i$  و تابع مطلوبیت  $u_i : \Pi_1 \times \Pi_2 \times \dots \times \Pi_n \rightarrow \mathbb{R}$  است. یک مجموعه سیاست  $\pi^* = (\pi_1^*, \pi_2^*, \dots, \pi_n^*)$  تعادل نش نامیده می‌شود اگر برای هر بازیکن  $i$  و هر سیاست  $\pi_i \in \Pi_i$  در وضعیت  $s$  داشته باشیم:

$$u_i(\pi_i^*, \pi_{-i}^*, s) \geq u_i(\pi_i, \pi_{-i}^*, s) \quad (۱-۴)$$

در اینجا،  $\pi_{-i}^*$  نشان‌دهنده سیاست‌های همه بازیکنان به جز بازیکن  $i$  است. در ادامه پژوهش جهت استفاده از چارچوب نظریه بازی در یادگیری تقویتی تابع مطلوبیت به‌گونه‌ای تعریف شده است که برابر با تابع ارزش  $u_i(\pi_i, \pi_{-i}, s) = V_i^{\pi_i, \pi_{-i}}(s)$  باشد.

- اهمیت تعادل نش: تعادل نش نقطه‌ای را در بازی مشخص می‌کند که هر بازیکن بهترین پاسخ را نسبت به انتخاب‌های دیگران ارائه داده است. این مفهوم به‌ویژه در بازی‌های غیرهمکارانه، به‌عنوان پیش‌بینی رفتار منطقی بازیکنان استفاده می‌شود و در زمینه‌هایی مانند یادگیری تقویتی چند عامله کاربرد گسترده‌ای دارد.

## ۲-۲-۴ بازی مجموع صفر

بازی‌های مجموع صفر<sup>۶</sup> دسته‌ای از بازی‌ها هستند که در آن‌ها تابع ارزش یک بازیکن دقیقاً برابر با ضرر بازیکن دیگر است. به عبارت دیگر، مجموع ارزشهای همه بازیکنان در هر مرحله صفر است.

- تعریف بازی مجموع صفر: در یک بازی دو نفره، اگر تابع ارزش بازیکن اول  $(V_1^{(\pi_1, \pi_2)}(s))$  و بازیکن دوم  $(V_2^{(\pi_1, \pi_2)}(s))$  به‌گونه‌ای باشد که برای هر مجموعه سیاست  $(\pi_1, \pi_2)$  به صورت زیر باشد را یک بازی مجموع صفر نامیده می‌شود.

$$V_1^{(\pi_1, \pi_2)}(s) + V_2^{(\pi_1, \pi_2)}(s) = 0 \rightarrow V_1^{(\pi_1, \pi_2)}(s) = -V_2^{(\pi_1, \pi_2)}(s) \quad (۲-۴)$$

- سیاست بهینه در بازی مجموع صفر: در بازی‌های مجموع صفر، سیاست بهینه هر بازیکن، انتخابی است که تابع ارزش خود را در برابر بهترین پاسخ حریف به حداکثر برساند. این سیاست اغلب به تعادل نش منجر می‌شود. سیاست بهینه دو بازیکن در بازی مجموع صفر با تابع ارزش معادله  $(۲-۴)$  به صورت زیر است.

<sup>۶</sup>Zero-Sum Games

$$V_1^*(s) = \max_{\pi_1} \min_{\pi_2} V_1^{(\pi_1, \pi_2)}(s) \tag{۳-۴}$$

$$V_2^*(s) = \max_{\pi_2} \min_{\pi_1} V_2^{(\pi_1, \pi_2)}(s) \tag{۴-۴}$$

## 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

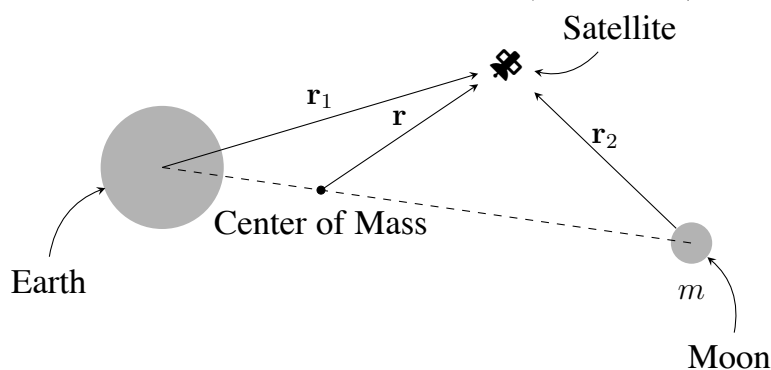
مسئله عمومی سه جسمی مسیره‌ای سه جسم پرجرم دلخواه را بررسی می‌کند که تحت تأثیر جاذبه متقابل قرار دارند. با این حال، این مسئله به مراتب عمومی‌تر از آن چیزی است که طراحان مأموریت‌های عملیاتی نیاز دارند. دو ساده‌سازی رایج وجود دارد که برای قابل دسترس‌تر و کاربردی‌تر کردن این مسئله در مسیره‌ای واقعی معمولاً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

## 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] 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.
- [27] 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.

- [28] S. Fujimoto, H. van Hoof, and D. Meger. Addressing function approximation error in actor-critic methods, 2018.
- [29] 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.
- [30] 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