دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر

گرایش هوش مصنوعی

بررسی خصوصیات همگرایی روش‌های ترکیبی یادگیری تقویتی با تخمین تابع

نگارش:

بابک به‌ساز

استاد راهنما:

دکتر رضا صفابخش

تیر ۱۳۸۶

بسمـه تعالـــی

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
معاونت پژوهشی

فرم اطلاعات پایان نامه
کارشناسی ارشد و دکترا

تاریخ: .....

پیوست: .....

نام و نام خانوادگی: بابک به‌ساز	**دانشجوی آزاد** ■	**بورسیه** □	**معادل** □
شماره دانشجوئی: ۸۳۱۳۱۱۶۶	**دانشکده:** مهندسی کامپیوتر	**رشته تحصیلی:** مهندسی کامپیوتر	
نام و نام خانوادگی استاد راهنما: رضا صفابخش			
عنوان پایان نامه به فارسی: بررسی خصوصیات همگرایی روش‌های ترکیبی یادگیری تقویتی با تخمین تابع			
عنوان پایان نامه به انگلیسی: Study of Convergence Characteristics of Reinforcement Learning with Function Approximation			
نوع پروژه: کارشناسی ارشد ■ / دکترا □	**کاربردی** □	**بنیادی** □	**توسعه‌ای** ■ **نظری** □
تاریخ شروع: مهرماه ۱۳۸۵	**تاریخ خاتمه:** تیرماه ۱۳۸۶	**تعداد واحد:** ۶	
سازمان تأمین کننده اعتبار: -			
واژه های کلیدی به فارسی: یادگیری تقویتی، روش‌های مبتنی بر ارزش، روش‌های مبتنی بر رویه، شبکه نیورو-فازی			
واژه های کلیدی به انگلیسی: Reinforcement Learning, Value-based Methods, Policy-based Methods, Neuro-Fuzzy Networks			
نظرها و پیشنهادها به منظور بهبود فعالیت های پژوهشی دانشگاه: **استاد راهنما:** **دانشجو:** **امضاء استاد راهنما:** **تاریخ:**			

نسخه ۱: معاونت پژوهشی

نسخه ۲: کتابخانه و به انضمام دو جلد پایان نامه به منظور تسویه حساب با کتابخانه و مرکز اسناد و مدارک علمی

چکیده

برای بسیاری سیستم‌ها، توانایی یادگیری یک مزیت مهم و حتی در بعضی موارد یک نیاز است. از ابتدا، برای ایجاد توانایی یادگیری دو ایده کلی بسیار مورد توجه بوده است. ایده اول که به یادگیری با نظارت منجر می‌شود، استفاده از زوج‌های آموزشی ورودی-خروجی است. در این نوع یادگیری، سعی بر آموزش عملکرد درست به سیستم، با تعدادی مثال است که هر مثال شامل خروجی مورد انتظار از سیستم برای یک ورودی معین است. ایده دیگر که به یادگیری بی‌نظارت منجر می‌شود، استفاده از قاعده‌مندی‌های موجود در ورودی است. در این نوع یادگیری، هیچ راهنمایی‌ای از خارج سیستم وجود ندارد و سعی بر کشف الگوها و قاعده‌مندی‌هایی در ورودی است که برای تولید خروجی مطلوب سیستم، مؤثر هستند.

از یک طرف، در یادگیری بانظارت تهیه زوج‌های آموزشی ورودی-خروجی در بعضی مسائل سخت و حتی گاهی ناممکن است. از طرف دیگر، بدلیل عدم وجود راهنمایی تعلیمی کافی در یادگیری بی‌نظارت، این روش در بسیاری مسائل کارایی مطلوب ندارد. بدلیل این مشکلات، در دو دهه اخیر، توجه به یک ایده کلی جدیدتر به نام یادگیری تقویتی جلب شده است که از نظر میزان راهنمایی تعلیمی، بین یادگیری بانظارت و بی‌نظارت قرار می‌گیرد. در این نوع یادگیری، سعی بر آموزش عملکرد مطلوب به سیستم، با دادن یک معیار عددی از کارایی فعلی آن است. از یک سو، تهیه معیاری عددی از کارایی سیستم، بسیار آسان‌تر از تهیه زوج‌های آموزشی ورودی-خروجی است و از سوی دیگر، میزان راهنمایی تعلیمی حاصل از این معیار عددی، می‌تواند برای راهنمایی سیستم به عملکرد مورد انتظار کافی باشد.

در گذشته، بیشترین توجه در یادگیری تقویتی بر روش‌های مبتنی بر جدول متمرکز بوده است. در این روش‌ها، برای هر وضعیت (یا وضعیت-عمل) سیستم یک خانه از حافظه برای نگه‌داری ارزش عددی آن وضعیت (یا وضعیت-عمل) اختصاص می‌یابد. به همین دلیل، استفاده از یادگیری تقویتی مبتنی بر جدول، در مسائلی با فضای بزرگ که وضعیت‌ها (یا وضعیت-عمل‌های) بسیاری دارند، تقریباً ناممکن می‌باشد. از این رو، روش‌هایی برای بکارگیری یادگیری تقویتی در این مسائل، از جمله روش‌های مبتنی بر ارزش و روش‌های مبتنی بر رویه که از تخمین تابع استفاده می‌کنند، طراحی شده‌اند. اما روش‌های مبتنی بر تخمین تابع، از خصوصیات همگرایی ضعیف‌تری نسبت به روش‌های مبتنی بر جدول برخوردارند که بررسی خصوصیات همگرایی آنها را برای استفاده درست از آنها پراهمیت می‌سازد.

در این پایان‌نامه، در ابتدا، به بررسی سه روش مبتنی بر ارزش براساس تجمیع وضعیت سخت، شبکه پرسپرون چند لایه، و شبکه سی‌مک، و همچنین، دو روش مبتنی بر رویه ری‌اینفورس و برخط باکستر و بارتلت پرداختیم. در آزمایشات تجربی این روش‌ها را بر روی سه مسأله یادگیری تقویتی ۱۰۰-راهزن مسلح، حفظ تعادل میله، و ربات ژیمناست، که به

ترتیب درجه پیچیدگی آسان، متوسط، و سخت دارند، اجرا کردیم. در این بررسی‌ها، تأثیر بعضی پارامترهای مهم هر روش در خصوصیات همگرایی آنها مورد مطالعه قرار گرفت. این بررسی‌ها نشان دهنده خصوصیات همگرایی بهتر روش‌های مبتنی بر رویه، هم از لحاظ نظری بدلیل وجود تضمین‌های همگرایی قوی‌تر و هم از لحاظ تجربی بدلیل نتایج بهتر، بود.

همچنین، یک سیستم نیورو-فازی جدید بر اساس روش‌های مبتنی بر رویه، طراحی کردیم. معماری این سیستم جدید با ایجاد تغییراتی در معماری یک سیستم موجود که آن را برای مسأله‌های یادگیری تقویتی اپیزودی مناسب می‌سازد، بدست آمده است. علاوه بر این، همگرایی الگوریتم یادگیری آن را به یک ماکزیمم محلی امیدریاضی میانگین پاداش اثبات کردیم. این سیستم نیورو-فازی، در حالیکه تمامی فواید معمول سیستم‌های نیورو-فازی را دارد، دارای این خصوصیت اضافه است که در چارچوب تقویتی عمل می‌کند و برای آموزش آن به جای زوج‌های آموزشی ورودی-خروجی تنها به یک سیگنال تقویتی نیاز است. در نهایت، مقایسه نتایج این سیستم جدید با پنج روش قبلی نشان‌دهنده برتری واضح کلی آن (با در نظر گرفتن نتیجه سه مسأله با هم) بر آنها بود. در مسأله ۱۰۰-راهزن مسلح، تمامی روش‌ها قابل مقایسه با هم بودند و به عملکرد مطلوب رسیدند. در مسأله حفظ تعادل میله، سیستم جدید بهترین عملکرد و در مسأله ربات ژیمناست، با اختلافی ناچیز دومین بهترین عملکرد را داشت. این نتایج در حالی بدست آمده است که از دانش قبلی در روش نیورو- فازی استفاده نشده است.

فهرست مطالب

فصل ۱

مقدمه

برای بسیاری سیستم‌ها، توانایی یادگیری یک مزیت و حتی بسیاری مواقع یک ضرورت است. حال سیستم مذکور می‌تواند، یک حیوان باشد، که احتیاج به پیدا کردن غذا و اجتناب از صیادان دارد؛ یا می‌تواند یک ربات تمیزکننده باشد، که مجبور به شارژ کردن باطری‌هایش در بازه‌های زمانی ثابت است. به این نوع سیستم‌های هوشمند، عامل[1] گفته می‌شود. یادگیری به عامل‌ها توانایی پیدا کردن راهی بهینه یا نزدیک به بهینه را، برای تأمین نیازهایشان یا رسیدن به اهداف‌شان، در محیط فعلی می‌دهد. حتی اگر محیط تغییر کند، عامل‌ها هنوز با یادگیری می‌توانند کارهاشان را به صورت بهینه یا نزدیک به بهینه انجام دهند.

کارهای زیادی در زمینه هوش مصنوعی و یادگیری ماشین برای حل این نوع مسئله‌ها انجام شده است. مثلاً، روش‌های حلی مانند درخت‌های جستجو، جستجوی مکاشفه‌ای، برنامه‌نویسی پویا، و سیستم‌های مبتنی بر قانون را می‌توان نام برد. ولی این روش‌ها توانایی تولید سیستمی جامع که بتواند مجموعه‌ای متنوع از مسأله‌ها را در محیط‌های بلادرنگ و واقعی حل کند، ندارند. دریک محیط واقعی، تعداد وضعیت‌های ممکن معمولاً بسیار زیاد یا نامتناهی است و اگر هر وضعیت جداگانه مورد توجه قرار گیرد، احتیاج به مقدار نامحدودی زمان است. حل یک مسأله به صورت بلادرنگ، بر مقدار زمانی که برای حل مسأله صرف می‌شود، محدودیت می‌گذارد. شاید بهای دست‌یابی به یک راه‌حل سریع، این باشد که به جواب بهینه نرسیم. ولی همین‌قدرهم که به جواب نزدیک به بهینه برسیم، در بیشتر مواقع به اندازه کافی خوب است. یک الگوریتم بلادرنگ خوب، باید بتواند به صورت افزایشی با اضافه شدن زمان اختصاص داده شده به حل مسأله، راه‌حل بهتری تولید کند.

[1] Agent

یک توسعه جالب در زمینه هوش مصنوعی و یادگیری ماشین، یادگیری تقویتی[۲] است. منظور از یادگیری تقویتی این است که عامل یک روند سعی و خطا در پیش می‌گیرد، اگر موفق باشد، پاداش می‌گیرد وگرنه جریمه می‌شود. راه‌حل موردنظر، معمولاً یک سری از اعمال است که عامل باید با یک ترتیب مشخص انجام دهد تا به هدف برسد. روش سعی و خطا از نظر تکنیکی از جهت داده‌های آموزشی بسیار مناسب است (زیرا در مسائل واقعی زوج‌های ورودی-خروجی معمولاً به ندرت در دسترسند). علاوه بر این، یادگیری تقویتی یک روش یادگیری بدون معلم است. فایده یک معلم این است که یادگیری می‌تواند بسیار سریع انجام شود. اما مشکل بزرگ آن این است که معلم باید تمام محیط مسأله را بشناسد و این بندرت در محیط‌های واقعی امکان‌پذیر است. اما یادگیری تقویتی لزوماً احتیاجی به دانش قبلی در مورد دینامیک محیط و مدلی از آن ندارد.

روش‌های پایه یادگیری تقویتی (که به روش‌های مبتنی بر جدول[۳] معروفند) نیاز به حافظه‌ای متناسب با تعداد وضعیت‌ها و اعمال دارند. با اینکه این دسته از روش‌ها معمولاً از تضمین همگرایی به جواب بهینه برخوردارند، اما بدلیل نیاز به منابع پردازشی و حافظه‌ای زیاد، استفاده از آنها برای بسیاری از مسائل واقعی که فضای وضعیت‌ها و اعمال آنها بزرگ است، با مشکل روبرو می‌شود. برای رفع این مشکل تلاش‌های زیادی صورت گرفته است و سه دسته روش کلی ارائه شده است: روش‌های سلسله مراتبی، روش‌های مبتنی بر ارزش[۴]، و روش‌های مبتنی بر رویه[۵]. مهمترین مشکل این سه دسته روش، عدم وجود تضمین‌های همگرایی مناسب برای آنهاست. از این رو، برای استفاده درست و معقول از آنها، نیاز به مطالعاتی، هم از لحاظ نظری و از لحاظ تجربی، بر روی خصوصیات همگرایی آنها می‌باشد. در این پایان‌نامه، به بررسی خصوصیات همگرایی روش‌های مبتنی بر ارزش و مبتنی بر رویه می‌پردازیم. در واقع، این دو دسته روش ترکیب یادگیری تقویتی با تخمین تابع می‌باشند.

مطالب این گزارش به ترتیب زیر خواهد آمد. در فصل دوم، مطالبی برای آشنایی با یادگیری تقویتی و روش‌های مبتنی برجدول ارائه می‌شود، و به اختصار مفاهیم روش‌های سلسله‌مراتبی، مبتنی بر ارزش و مبتنی بر رویه بیان می‌شود. در فصل سوم، روش‌های مبتنی بر ارزش مورد بررسی قرار می‌گیرند و نتایج پیاده‌سازی سه نمونه از آنها بحث می‌شود. معرفی روش‌های مبتنی بر رویه در فصل چهارم صورت می‌گیرد و نتایج پیاده‌سازی دو روش از این دسته ارائه می‌شود. در فصل پنجم، یک سیستم نیورو-فازی جدید بر پایه یک روش مبتنی بر رویه ارائه و نتایج پیاده‌سازی آن بررسی می‌شود. مقایسه نتایج تجربی شش روش پیاده‌سازی‌شده (شامل روش جدید) در فصل ششم انجام می‌شود و در نهایت، نتیجه‌گیری نهایی و پیشنهادات در فصل هفتم می‌آید.

[2] Reinforcement Learning
[3] Table-Based
[4] Value-Based
[5] Policy-Based

فصل ۲

یادگیری تقویتی

هنگامی که به طبیعت پدیده یادگیری می‌اندیشیم، ایده یادگیری به وسیله تعامل با محیط به احتمال زیاد، اولین ایده‌ای است که به ذهن ما خطور می‌کند. هنگامی که نوزادی به اطراف خود نگاه می‌کند، دست خود را در هوا تکان می‌دهد و یا بازی می‌کند. هیچ معلمی بالای سر وی وجود ندارد و رابط‌های حسی و حرکتی تنها وسایل ارتباطی او با محیط هستند. اما ادامه تعامل با محیط و تمرین با این رابط‌ها تولیدکننده میزان زیادی اطلاعات در مورد علت‌ها و معلول‌ها، نتایج اعمال، و روش رسیدن به اهداف خود است. چنین منابع اطلاعاتی، در طول زندگی ما، بدون شک یکی از منابع اخذ دانش در مورد محیط اطراف و حتی خود ما محسوب می‌شوند.

یادگیری تقویتی، در حقیقت نگرشی محاسباتی و ریاضی به پدیده و فرآیند یادگیری از طریق تعامل با محیط است و در یادگیری تقویتی هدف کشف روند فرایند واقعی یادگیری در انسان یا حیوانات نیست. به عبارت دیگر، یادگیری تقویتی، از طریق تحلیل‌های ریاضی مدل‌هایی را عرضه و ارزیابی می‌کند که در مسائل علمی و اقتصادی می‌تواند بعنوان ماشین‌های یادگیرنده مورد استفاده قرار گیرند. در این فصل، هدف ایجاد دیدی تحلیلی و محاسباتی نسبت به مساله یادگیری تقویتی است. در اولین قدم، یادگیری تقویتی را با یادگیری با نظارت[6] مقایسه خواهیم کرد و به بیان تفاوت‌های این دو خواهیم پرداخت. سپس، به بیان تعریفی دقیق از صورت و اجزای مساله یادگیری تقویتی و روش‌های آن خواهیم پرداخت [۱].

۲-۱ یادیگری تقویتی در مقابل یادگیری با نظارت

از دید تحلیلی و بطور خلاصه، یادگیری رویه‌ای از انجام اعمال است بطوریکه در دراز مدت سیگنال پاداش عددی‌ای که در ازای انجام هر عمل به یادگیرنده داده می‌شود را بیشنیه نماید. در یادگیری تقویتی، برخلاف بسیاری از اشکال یادگیری

[6] Supervised Learning

ماشین، به یادگیرنده گفته نمی‌شود که چه عملی را انجام دهد و این وظیفه یادگیرنده است که دریابد که با انجام چه اعمالی پاداش بیشتری را بدست می‌آورد. در حالات خاصی از یادگیری تقویتی، انجام یک عمل، نه تنها باعث دریافت پاداش شده بلکه محیط را در وضعیتی قرار می‌دهد که انجام اعمال بعدی پاداش‌های دیگری را به دنبال خواهد داشت. این دو خصوصیت، یعنی جستجوی مبتنی بر سعی و خطا و پاداش همراه با تاخیر[۷]، در حقیقت دو خصوصیت متمایزکننده یادگیری تقویتی از دیگر اشکال یادگیری ماشین هستند.

یادگیری تقویتی. از جهات بنیادی دیگری نیز با یادگیری با نظارت متفارت است. یادگیری با نظارت، در حقیقت یادگیری با استفاده از مثال‌هایی است که توسط یک مربی یا ناظر خارجی در اختیار یادگیرنده قرار می‌گیرد. با وجود اهمیت بسیار زیاد، این روش به تنهایی جهت یادگیری به وسیله تعامل با محیط مناسب نیست. در هنگام تعامل با محیط، معمولاً بدست آوردن مثال‌هایی از رفتار بهینه و درست که مشخص‌کننده تمام شرایطی که عامل یادگیرنده باید در آنها تصمیم‌گیری کند ممکن نیست. در حقیقت از این لحاظ نیز، در چنین محیط‌هایی، از عامل یادگیرنده انتظار می‌رود که از تجارب خود، تعامل بهینه با محیط را فرا بگیرد.

یکی دیگر از مسائلی که تنها در یادگیری تقویتی مطرح بوده و در سایر اشکال یادگیری مطرح نیست، موازنه میان کشف تجارب جدید و بکارگیری تجارب قبلی[۸] است. به عبارت دیگر، این گزینه، که عامل یادگیرنده جهت انجام اعمال بعدی، اعمالی را انتخاب کند که در گذشته انجام داده و پاداش‌های زیادی را از آنها کسب کرده است، بسیار منطقی به نظر می‌رسد. اما از طرف دیگر، این حقیقت هم، که جهت یافتن چنین اعمالی، عامل باید دست به تجربه و انجام اعمالی که تاکنون انجام نداده است بزند، قابل تامل است. به عبارت دیگر، در عمل، اتخاذ هیچ کدام از این روش‌ها به تنهایی یادگیرنده را به مقصود خود نمی‌رساند و راه‌حل اصلی ترکیبی از بکارگیری تجارب قبلی و دست زدن به تجارب جدید است. این اصل و روش‌های پیاده‌سازی و بکارگیری آن در مسائل یادگیری تقویتی، یکی از مسائلی است که دانشمندان و محققین در چند دهه گذشته حجم وسیعی از تحقیقات خود را بر روی آن متمرکز کرده‌اند، و این در حالی است که در یادگیری تحت نظارت چنین بحثی به هیچ وجه مطرح نیست.

۲-۲ مساله یادگیری تقویتی

در این بخش به بررسی اجزای مسأله یادگیری تقویتی می‌پردازیم. سپس با حالت خاصی از مسأله یادگیری تقویتی که خصوصیات مناسبی برای تحلیل‌های نظری دارد و بیشتر تحقیقات بر روی آن متمرکز شده است، آشنا می‌شویم.

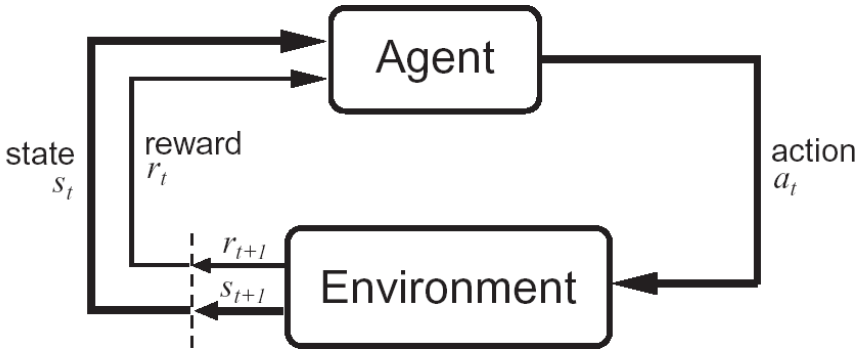
۲-۲-۱ رابط محیط - عامل

همانطور که گفته شد، مسأله یادگیری تقویتی، در حقیقت، مساله یادگیری از طریق تعامل با محیط جهت دستیابی به هدفی تعیین شده است. در این مساله، به عنصری که تقش یادگیرنده و تصمیم‌گیرنده را ایفا می‌کند، عامل[۹] گفته می‌شود و موجودیتی که عامل با آن تعامل می‌کند، محیط[۱۰] نام دارد. عامل و محیط، پیوسته در طول زمان به تعامل با یکدیگر ادامه

[7] Delayed Reward
[8] Exploration vs. Exploitation Trade-off
[9] Agent
[10] Environment

می‌دهند. بدین صورت که عامل به انتخاب و انجام اعمال پرداخته و محیط به اعمال عامل پاسخ داده و وضعیت جدیدی را ایجاد می‌کند. علاوه بر این، محیط به عامل پاداش[11] می‌دهد که در حقیقت یک سیگنال عددی است که هدف عامل بیشینه کردن آن در طول زمان است. مجموعه مشخصات کامل یک محیط، در حقیقت یک مساله یادگیری تقویتی را تعریف می‌کند، که به آن اصطلاحاً تکلیف[12] گفته می‌شود.

به عبارت دقیق‌تر، عامل و محیط در دنباله‌ای از گام‌های گسسته زمانی $t=1,2,3,\ldots$ به تعامل با یکدیگر پرداخته، بدین صورت که در هر گام زمانی t، عامل تصویری از وضعیت محیط یعنی $s_t \in S$، که S مجموعه وضعیت‌ها محیط می‌باشد، را دریافت کرده و براساس آن اقدام به انتخاب یک عمل مانند $a \in A(s_t)$، که $A(s_t)$ مجموعه اعمال قابل انجام در وضعیت s_t است، نماید. در گام زمانی بعدی، عامل به عنوان نتیجه عمل خود، پاداشی عددی مانند $r_{t+1} \in \Re$ دریافت می‌کند و خود را در وضعیت جدیدی مانند s_{t+1} می‌بیند. شکل ۲-۱ این روند را به صورت کاملاً واضح نشان می‌دهد. در هر گام زمانی، عامل نگاشتی از وضعیت‌ها محیط به احتمال انتخاب هر یک از اعمال قابل انجام، تولید می‌کند. این نگاشت، اصطلاحاً، رویه[13] عامل نامیده شده و با π_t نشان داده می‌شود، که $\pi_t(s,a)$ احتمال $a_t = a$ به شرط $s_t = s$ است. روش‌های یادگیری تقویتی، در حقیقت، چگونگی تغییر رویه با توجه به تجربه صورت گرفته توسط عامل را بیان می‌کنند، و هدف عامل بیشینه نمودن مجموع پاداش دریافتی در درازمدت است.



شکل ۲-۱ تعامل محیط- عامل در یادگیری تقویتی [۱].

چارچوبی که هم‌اکنون تعریف شد، چارچوبی خلاصه و انعطاف‌پذیر بوده و برای روش‌های مختلف، جهت حل مسائل گوناگون قابل استفاده است. به عنوان مثال، منظور از گام‌های زمانی، الزاماً، بازه‌های ثابت و با طول یکسان از زمان واقعی نیست. به عبارت بهتر، گام‌های زمانی می‌توانند نشان‌دهنده مراحل پشت سر هم از فرآیند تصمیم‌گیری و انجام اعمال باشند. همچنین اعمال نیز می‌توانند در سطوح مختلف تعریف شوند. به همین صورت، حالات و وضعیت‌های محیط نیز می‌توانند به صورت‌های مختلف تعریف شوند. در بعضی موارد بخشی از داده‌هایی که یک وضعیت را توصیف می‌کنند ممکن است براساس حافظه عامل از اطلاعات دریافتی قدیمی‌تر از محیط و یا براساس دانشی کاملاً درونی شکل گرفته باشند و همچنین در مورد اعمال نیز، بعضی از اعمال می‌توانند اعمالی کاملاً درونی محسوب شوند. به عنوان مثال، عمل انتخاب موضوعی که عامل به آن

[11] Reward
[12] Task
[13] Policy

فکر کند و یا مکانی که به آن توجه کند، نوعی عمل به حساب می‌آید. به طور کلی، اعمال تمام تصمیم‌گیری‌هایی را شامل می‌شود که عامل می‌خواهد چگونگی انجام آنها را فرا بگیرد و وضعیت‌ها، دانش در مورد هر چیزی می‌باشد که جهت تصمیم‌گیری مفید است.

اصولاً، مرز بین عامل و محیط معمولاً همانند مرز فیزیکی یک ربات و یا بدن یک حیوان نیست و اکثراً، این مرز بیش از این مقدار به عامل نزدیک می‌شود. به عنوان مثال، موتورها و اتصالات و همچنین سخت‌افزار حسگر معمولاً بخشی از محیط فرض می‌شوند. همچنین، پاداش‌ها نیز به طور معمول در درون سیستم‌های یادگیرنده محاسبه می‌شوند ولی موجودیتی خارج از عامل فرض می‌شوند. این مساله را می‌توان با یک قانون کلی مدل‌سازی نمود. بدین صورت که هر چیزی که با دلخواه عامل مستقیماً قابل تغییر نیست، به عنوان موجودیتی خارج از عامل و جزئی از محیط فرض می‌شود. همچنین این فرض که عامل هیچ دانشی از محیط ندارد نیز در بعضی مواقع نادرست است. به عنوان مثال، عامل معمولاً تا حدی از طریقه محاسبه پاداش‌ها به عنوان تابعی از وضعیت محیط و عمل انجام شده مطلع است، اما ما همیشه فرآیند محاسبه پاداش را فرآیندی خارج از عامل فرض می‌کنیم، زیرا در حقیقت این فرآیند بوده که تکلیف عامل را تعریف کرده و بنابراین باید به گونه‌ای باشد که مستقیماً و با دلخواه عامل قابل تغییر نباشد. اما در عمل، مسائلی نیز وجود دارد که عامل با وجودیکه از تمامی جزئیات و فرآیندهای محیط مطلع است، باز هم با یک مساله بسیار پیچیده یادگیری تقویتی روبرو است.

چارچوب یادگیری تقویتی، خلاصه‌سازی قابل توجهی از مساله یادگیری هدف-گرا[14] به وسیله تعامل با محیط است. این چارچوب به گونه‌ای طراحی شده که هر نوع مساله یادگیری تقویتی، هر نوع سیستم حسی، کنترلی و هدف قابل کاهش به سه سیگنال رد و بدل‌شونده میان محیط و عامل است. این چارچوب شاید جهت توصیف تمام مسائل یادگیری رویه تصمیم گیری کافی نباشد، ولی در عمل ثابت شده که چارچوبی بسیار مفید و قابل استفاده می‌باشد. این نکته نیز قابل ذکر است که شیوه توصیف وضعیت‌ها و اعمال تأثیر قابل توجهی در کارآیی روش‌های یادگیری تقویتی دارد. انتخاب شیوه‌های توصیف، در این گونه مسائل، در حال حاضر بیشتر هنر محسوب می‌شود تا علم و با تجربیات طراح مدل یادگیرنده ارتباط مستقیم دارد.

۲-۲-۲ اهداف، پاداش‌ها و سود[15]

در یادگیری تقویتی، هدف و مقصود نهایی عامل بر مبنای سیگنال پاداشی که از محیط به عامل فرستاده می‌شود، تعریف می‌شود و در هر گام زمانی، پاداش عددی حقیقی مانند $r_t \in R$ است. هدف عامل، بیشینه نمودن مجموع پاداش دریافتی است و این نه به مفهوم بیشینه نمودن پاداش فعلی بلکه بیشینه نمودن پاداش کلی درمدت زمان طولانی است.

استفاده از سیگنال پاداش جهت تبیین و توصیف هدف عامل یکی از اساسی‌ترین خصوصیات متمایزکننده یادگیری تقویتی از دیگر روش‌های موجود است. با وجودیکه ممکن است چنین روشی جهت قاعده‌مند کردن اهداف در نگاه اول محدودکننده به نظر برسد، اما در عمل کاملاً انعطاف‌پذیر و قابل استفاده نشان داده است. به عنوان مثال، جهت یادگیری طریقه راه رفتن به یک ربات، محققان در هر گام زمانی، پاداشی متناسب با میزان حرکت ربات به جلو در نظر گرفته‌اند و یا اینکه جهت یادگیری یک رابط برای خروج از یک مارپیچ، در هر گام زمانی که ربات در مارپیچ است، پاداش ۱- برای آن در

[14] Goal-based
[15] Return

نظر گرفته می‌شود و با خروج ربات از مارپیچ به آن پاداشی برابر ۱+ می‌دهند. پاداش ۱- و یا به عبارت دیگر جریمه ۱- ربات را به خروج از مارپیچ در زمان کوتاه‌تر تشویق می‌کند.

همانطور که گفته شد در مثال‌های بالا، عامل پس از یادگیری موفق، قادر است مجموع پاداش‌های دریافتی خود را بیشینه نماید. پس در صورتیکه از عامل انتظار رفتار خاصی را داریم، باید پاداش‌ها را به گونه‌ای تنظیم کنیم که بیشینه شدن آن معادل با برآورده شدن انتظار و هدف مورد نظر ما باشد. بنابراین، پاداش‌ها باید واقعاً با مقصود ما مرتبط باشند. همچنین، سیگنال پاداش محلی جهت وارد کردن دانش عامل راجع به مساله و روش رسیدن به هدف آن نیست. به عنوان مثال، یک بازیکن شطرنج تنها باید در حالت بردن بازی پاداش بگیرد.نه در حالاتی که اهداف فرعی و زیر هدف‌های بازی مانند زدن مهره حریف و بدست گرفتن کنترل بخشی استراتژیک از صفحه برآورده می‌شوند. در صورتیکه دست یافتن به این اهداف فرعی با پاداش همراه شود، عامل ممکن است راه‌هایی برای برآورده کردن آنها بیابد که هدف اصلی را برآورده نکنند. به عنوان مثال، بازیکن شطرنج ممکن است مهره حریف را حتی به قیمت باختن بازی بزند. به طور خلاصه و مختصر، سیگنال پاداش در حقیقت، وسیله جهت انتقال هدف ما به عامل است، نه وسیله جهت انتقال روشی که با آن می‌خواهیم به هدف برسیم.

تاکنون، گفته شده که هدف عامل یادگیرنده بیشینه کردن پاداش دریافتی در مدت زمان طولانی است. اما به توصیف قاعده‌مند و دقیق این عبارت نپرداخته‌ایم. فرض کنید که دنباله پاداش‌های دریافتی از زمان t به بعد به صورت $r_{t+1}, r_{t+2}, r_{t+3}, \ldots$ بوده است. در این صورت چه چیز از این دنباله را باید بیشینه کنیم؟ در حالت کلی، هدف ما بیشینه کردن سود یا R_t است که به صورت تابعی از دنباله پاداش‌ها تعریف می‌شود. در ساده‌ترین حالت، سود مجموعه پاداش‌هاست:

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \ldots + r_T, \qquad (۱-۲)$$

که T آخرین گام زمانی می‌باشد. این تعریف، در کاربردهایی با معنی است که گام زمانی پایانی در آنها دارای مفهوم باشد. به عبارت دیگر روند تعامل عامل با محیط به زیر دوره‌هایی، که اپیزود[۱۶] نامیده می‌شوند، تقسیم می‌گردد. به عنوان مثال، هر بار انجام یک بازی و یا تمام کردن یک مارپیچ نمونه‌هایی از اپیزودها هستند. هر اپیزود با وضعیت‌هایی خاص که به آن وضعیت پایانی[۱۷] گفته می‌شود، تمام می‌شود و بلافاصله پس از آن سیستم وارد یک وضعیت از وضعیت‌های شروع اپیزود می‌گردد. به مسائلی که بدین صورت هستند، مسائل اپیزودیک[۱۸] گفته می‌شود. در مسائل اپیزودیک، در بعضی مواقع، نیاز داریم که مجموعه وضعیت‌های غیر پایانی را، که S نامیده می‌شود، از مجموعه تمام وضعیت‌ها، از جمله وضعیت‌های پایانی، که S^+ نامیده می‌شود، جدا کنیم.

از طرف دیگر، در بسیاری از مسائل، تعامل میان عامل و محیط، به صورت طبیعی، اپیزودیک نیست و بدون حد و مرز ادامه پیدا می‌کند. به چنین مسائلی، مسائل ادامه‌دار[۱۹] گفته می‌شود. فرمول (۲-۱) برای این گونه مسائل، کاربردی ندارد، زیرا گام زمانی نهایی $T = \infty$ بوده و بنابراین سود، که سعی در بیشینه کردن آن داریم، می‌تواند به راحتی به سمت بینهایت

[16] Episode
[17] Terminal State
[18] Episodic Tasks
[19] Continuing Tasks

میل کند. بنابراین، بهتر است که جهت سود از تعریفی استفاده کنیم که از لحاظ مفهومی پیچیده‌تر ولی از لحاظ ریاضی ساده‌تر باشد.

مفهوم جدیدی که ما به آن نیاز داریم، مفهوم کاهش[۲۰] است. بر طبق این مفهوم، عامل سعی در انتخاب اعمالی دارد که مجموع پاداش‌های دریافتی کاهش یافته[۲۱] آنها بیشینه باشد. به زبان دقیق‌تر، عامل a_t را طوری انتخاب می‌کند که سود کاهش یافته[۲۲] مورد انتظار آن یعنی:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \qquad (۲-۲)$$

بیشینه گردد. در فرمول ۲-۲، $0 \le \gamma \le 1$ عددی بین صفر و یک بوده و نرخ کاهش[۲۳] نامیده می‌شود.

نرخ کاهش، در حقیقت تعیین‌کننده ارزش فعلی پاداش‌های آتی است. به عبارت دیگر، پاداشی که k گام زمانی بعد دریافت شود، ارزشی معادل γ^{k-1} برابر ارزشی را دارد که در صورتیکه در زمان حال دریافت می‌شد، داشت. در صورتیکه $\gamma < 1$ باشد و مجموعه $\{r_k\}$ دارای حد پایین و بالای معلوم باشد، سود کاهش یافته به مقداری محدود و غیر از بینهایت همگرا خواهد شد. در صورتیکه $\gamma = 0$ باشد، عامل تنها به بیشینه کردن پاداش فعلی علاقه دارد و اصطلاحاً عاملی نزدیک‌بین خوانده می‌شود. با افزایش γ به سمت ۱، پاداش‌های آتی اهمیت بیشتری پیدا می‌کنند و به عبارتی عامل دوراندیش می‌شود.

۲-۲-۳ خاصیت مارکف[۲۴]

همانطور که در قسمت‌های گذشته گفته شد، در چارچوب یادگیری تقویتی، عامل تصمیمات خود را براساس سیگنالی دریافتی از محیط تحت عنوان وضعیت محیط، می‌گیرد. در این قسمت، در مورد سیگنال وضعیت و اطلاعاتی که در درون خود دارد، بحث خواهد شد و در ادامه خصوصیتی خاص از دسته‌ای از محیط‌ها و سیگنال‌های وضعیت آنها به نام خاصیت مارکف معرفی خواهد شد که دارای اهمیتی ویژه است.

اصولاً، سیگنال وضعیت باید حتماً شامل داده‌هایی باشد که مستقیماً از حسگرها دریافت می‌شوند، اما علاوه بر این، می‌تواند شامل داده‌های دیگری هم باشد. به عبارت دیگر، وضعیت می‌تواند نسخه‌ای پردازش شده و بسیار سطح بالا از داده‌های دریافتی از حسگرها و یا متشکل از ساختارهای داده‌ای پیچیده‌ای باشد که در طول زمان و براساس دنباله‌ای از داده‌های دریافتی از حسگرها ساخته شده باشد. از طرف دیگر، از سیگنال وضعیت انتظار نمی‌رود که همه چیز در مورد محیط و جزئیات آن را به عامل منتقل کند. به عنوان مثال از عامل امدادگری که وظیفه آن رسیدگی به حال تعدادی مصدوم یک حادثه است، انتظار نمی‌رود که قبل از رسیدن به محل حادثه و معاینه مصدومین از تمام آسیب‌های داخلی یک مصدوم

---

[20] Discounting
[21] Discounted Reward
[22] Discounted Return
[23] Discount Rate
[24] Markov Property

بیهوش مطلع باشد. در چنین حالاتی، قسمتی از اطلاعات وضعیت محیط از دید عامل مخفی است که این اطلاعات می‌توانست برای عامل جهت تصمیم‌گیری مفید باشد. اما عامل نمی‌تواند آنها را در اختیار داشته باشد.

سیگنال وضعیتی که درحالت ایده‌آل مورد نظر ماست، سیگنالی است که اطلاعات دریافت شده قبلی از محیط را در خود خلاصه کرده بطوریکه تمام اطلاعات مفید و قابل استخراج در مورد محیط، در آن موجود باشد. سیگنال وضعیتی که چنین خاصیتی را داشته باشد، سیگنال مارکف و یا دارای خاصیت مارکف گفته می‌شود. به عنوان مثال وضعیت صفحه شطرنج در یک بازی شطرنج و معدودی پارامتر مانند تعداد حرکات انجام شده، وضعیتی دارای خاصیت مارکف است، زیرا شامل تمام اطلاعاتی است که بتوان براساس آن اتفاقات رخ داده در بازی تا این زمان را دریافت و بازی را ادامه داد.

حال به تعریفی تحلیلی و قاعده‌مند خاصیت مارکوف در مسائل یادگیری تقویتی می‌پردازیم. جهت ساده‌سازی فرمول‌ها، فرض می‌کنیم که تعداد وضعیت‌ها و پاداش‌ها متناهی است. این فرض بدون از دست رفتن کلیت تعریف، به ما این امکان را می‌دهد که به جای انتگرال‌ها و توابع چگالی احتمال از مجموع و توابع جرم احتمال استفاده کنیم. برای شروع پاسخ محیط در زمان $t+1$ نسبت به عمل انجام شده در زمان t را در نظر می‌گیریم. در فراگیرترین حالت علت و معلولی، این پاسخ ممکن است به تمام رویدادهایی که قبلاً رخ داده بستگی داشته باشد. در این حالت، دینامیک محیط تنها با استفاده از توابع احتمالی کامل قابل توصیف است:

$$\Pr\{s_{t+1}=s', r_{t+1}=r \mid s_t, a_t, r_t, s_{t-1}, a_{t-1}, \ldots, r_1, s_0, a_0\} \qquad (۲-۳)$$

که این رابطه به ازای تمامی مقادیر s' و r و رخدادهای قبلی، یعنی $a_0, s_0, r_1, \ldots, r_t, a_t, s_t$ برقرار است.

در صورتیکه سیگنال وضعیت دارای خاصیت مارکف باشد، آنگاه پاسخ محیط در زمان $t+1$ تنها به وضعیت موجود و عمل انجام شده در زمان t وابسته است، که در آن حالت، دینامیک محیط تنها به وسیله معادله:

$$\Pr\{s_{t+1}=s', r_{t+1}=r \mid s_t, a_t\} \qquad (۲-۴)$$

قابل توصیف است. به عبارت دیگر یک سیگنال وضعیت دارای خاصیت مارکف است، اگر و تنها اگر رابطه ۲-۳ با رابطه ۲-۴ به ازای تمام مقادیر s' و r و $a_0, s_0, r_1, \ldots, r_t, a_t, s_t$ معادل باشد. در این حالت محیط و تکلیف مورد نظر نیز دارای خاصیت مارکف خواهند بود.

در صورتیکه محیطی دارای خاصیت مارکوف باشد، آنگاه دینامیک تک گام آن ما را قادر می‌سازد که با داشتن وضعیت و عمل فعلی به پیش‌بینی وضعیت بعدی و پاداش آن بپردازیم. می‌توان نشان داد که، با تکرار این کار، قادریم تنها با داشتن اطلاعات وضعیت فعلی، به خوبی زمانی که تمام اطلاعات قبلی را داریم، به پیش‌بینی تمام رخدادهای آینده بپردازیم.

حتی در مواردی که سیگنال وضعیت محیط مارکوف نیست، در نظر گرفتن وضعیت فعلی به عنوان تخمینی از یک وضعیت مارکوف، فرض مناسبی است. به عبارت دقیق‌تر، کافی یا مناسب بودن دانش در مورد وضعیت فعلی جهت تخمین پاداش‌ها و انتخاب اعمال مد نظر ما است. همچنین، در حالاتی که مدلی از محیط نیز فرا گرفته می‌شود، قابل پیش‌بینی بودن وضعیت‌های بعدی با توجه به وضعیت فعلی نیز مدنظر است. بنابراین، با نزدیکتر کردن وضعیت مساله به وضعیت مارکوف، فرآیند یادگیری تقویتی کارآیی بهتری خواهد داشت. به تمام دلایل ذکر شده، در نظر گرفتن وضعیت فعلی به عنوان تخمینی از یک وضعیت مارکوف بسیار مفید خواهد بود و البته نباید فراموش کرد که سیگنال وضعیت ممکن است به طور کامل خاصیت مارکوف را بر آورده نکند.

۲-۲-٤ فرایند تصمیم‌گیری مارکف[۲۵]

در یک مساله یادگیری تقویتی، تکلیفی که دارای خاصیت مارکف باشد، فرآیند تصمیم‌گیری مارکف نامیده می‌شود و در صورتیکه فضای وضعیت‌ها و اعمال آن محدود و متناهی باشند. آنگاه فرآیند تصمیم‌گیری مارکف متناهی[۲۶] نامیده می‌شود. فرآیندهای تصمیم‌گیری مارکفی متناهی اهمیت ویژه‌ای در شکل‌دهی نظریه یادگیری تقویتی دارند و پایه بخشی اساسی از نظریات اساسی آن به شمار می‌روند.

فرآیند تصمیم‌گیری مارکف متناهی، اصولاً به وسیله مجموعه وضعیت‌ها و اعمال و همچنین به وسیله دینامیک تک‌گام محیط تعریف می‌گردد. با داشتن هر وضعیت و عمل، مانند s و a، احتمال ایجاد هر یک از وضعیت‌ها مانند s' به عنوان وضعیت بعدی برابر:

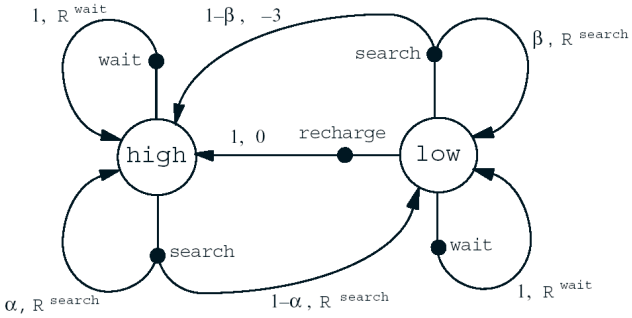
$$\mathrm{P}_{ss'}^{a} = \Pr\{s_{t+1} = s' \mid s_t = s, a_t = a\} \qquad (۵-۲)$$

خواهد بود. به چنین مقادیری، احتمالات انتقال[۲۷] گفته می‌شود. به همین صورت با داشتن s و a و وضعیت بعدی s'، مقدار مورد انتظار پاداش بعدی برابر است با:

$$R_{ss'}^{a} = E\{r_{t+1} \mid s_t = s, a_t = a, s_{t+1} = s'\} \qquad (۶-۲)$$

مقادیر $\mathrm{P}_{ss'}^{a}$ و $R_{ss'}^{a}$، به طور کامل قادر به توصیف مهمترین جنبه‌های دینامیکی فرآیندهای تصمیم‌گیری مارکف متناهی هستند.

یکی از ابزارهای مفید مورد استفاده جهت خلاصه‌سازی و نشان دادن دینامیک فرآیندهای تصمیم‌گیری مارکف متناهی، گراف انتقال است. در گراف انتقال، دو نوع گره وجود دارد که عبارتند از گره‌های وضعیت و گره‌های عمل. در یک گراف انتقال به ازای هر وضعیت یک گره وضعیت وجود دارد و به ازای هر دوتایی وضعیت- عمل، یک گره عمل موجود است. شروع از وضعیت s و انجام عمل a ما را از طریق یک خط به گره عمل (s,a) متصل می‌کند و پاسخ محیط به آن عمل نیز با پیکان‌های خروجی از این گره عمل نشان داده می‌شود که به عنوان مثال یکی از این پیکان‌های خروجی به گره وضعیت s' رفته و s' را به عنوان وضعیت بعدی معین می‌کند. هر پیکان خروجی معادل سه تایی (s,s',a) است و مقادیر $\mathrm{P}_{ss'}^{a}$ و $R_{ss'}^{a}$ بر روی آن قرار می‌گیرد. شکل ۲-۲ نمونه‌ای از یک گراف انتقال را نشان می‌دهد.



شکل ۲-۲ نمونه‌ای از یک گرف انتقال [۱].

---

[25] Markov Decision Process

[26] Finite Markov Decision Process

[27] Transition Probabilities

۲-۳ توابع ارزش

اکثریت قریب به اتفاق الگوریتم‌های یادگیری تقویتی براساس تخمین توابع ارزش طراحی شده‌اند. این روش‌ها که در ادامه به آنها اشاره می‌شود، روش‌های پایه یادگیری تقویتی را تشکیل می‌دهند. توابع ارزش، توابعی از وضعیت‌ها (و یا دوتایی‌های وضعیت- عمل) می‌باشند که حاوی تخمینی از ارزش بودن در این وضعیت (و یا حاوی تخمینی از ارزش انجام یک عمل در یک وضعیت خاص) برای عامل هستند. منظور از ارزش، در حقیقت، تخمینی از مجموع پاداش‌های آتی و یا به صورت دقیق‌تر سود مورد انتظار است. البته پاداش‌هایی که عامل در آینده دریافت خواهد کرد، به اعمالی که انجام خواهد داد وابسته است. به همین جهت، توابع ارزش، برای هر رویه، به طور جداگانه تعریف می‌گردند.

بدین ترتیب ارزش یک وضعیت تحت رویه π، اصطلاحاً $V^{\pi}(s)$، برابر سود مورد انتظار با شروع از وضعیت s و دنبال کردن رویه π است. برای هر فرآیند تصمیم‌گیری مارکف، میتوانیم $V^{\pi}(s)$ را به صورت قاعده‌مند تعریف کنیم:

$$V^{\pi}(s) = E_{\pi}\{R_t \mid s_t = s\} = E_{\pi}\{\sum_{k=1}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\}, \qquad (۷-۲)$$

در معادله بالا، منظور از $E_{\pi}\{\}$ مقدار مورد انتظار با دنباله کردن رویه π توسط عامل است. مقدار این تابع ارزش برای وضعیت‌های پایانی برابر صفر است. تابع ارزش V^{π}، اصطلاحاً تابع ارزش وضعیت[۲۸] نامیده می‌شود.

به همین صورت، ارزش انجام عمل a، در وضعیت s، تحت رویه π، اصطلاحاً $Q^{\pi}(s,a)$، برابر با سود مورد انتظار با شروع از وضعیت s، انجام عمل a، و پیگیری رویه π بوده و به صورت:

$$\begin{aligned} Q^{\pi}(s,a) &= E_{\pi}\{R_t \mid s_t = s, a_t = a\} \\ &= E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a\} \end{aligned} \qquad (۸-۲)$$

تعریف می‌شود که آن را اصطلاحاً تابع ارزش-عمل[۲۹] برای رویه π می‌نامیم.

خصوصیت بنیادی توابع ارزش که در نظریه یادگیری تقویتی و برنامه‌نویسی پویا مورد استفاده قرار می‌گیرد، قابلیت توصیف ارتباط مقادیر آن در وضعیتهای مختلف، در معادلات خاص بازگشتی است. به عبارت بهتر به ازای هر رویه مانند π و هر وضعیت مانند s، شرط زیر بین ارزش s و ارزش وضعیت‌های ممکن پس از آن صادق است:

---

[28] State Value Function

[29] Action-Value Function