دانشکده مهندسی برق

یادگیری تقویتی در کنترل

تمرین اول: مباحث مقدماتی و مسئله MAB

استاد: دکتر سعید شمقدری

دانشجو: سیده ستاره خسروی

پائیز ۱۴۰۳

چکیده

در این تمرین سوالاتی در خصوص مباحث مقدماتی یادگیری تقویتی و مفاهیم آن مطرح شده که در طول گزارش پاسخ آن‌ها ارائه شده است. در ادامه بیشتر به مسئله‌ی Multi Armed Bandit و مفاهیم آن در سوالات تمرین پرداخته می‌شود که پاسخ آن‌ها در فصول مربوطه آورده شده است.

واژه‌های کلیدی: یادگیری تقویتی، راهزن چند دست

فهرست مطالب

عنوان صفحه

فهرست تصاویر و نمودارها

عنوان صفحه

فصل ۱: مفاهیم مقدماتی

۱.۱ مقدمه

در این فصل به سوالات اول تا سوم پرداخته می‌شود، این سوالات درخصوص مفاهیم پایه‌ای یادگیری تقویتی هستند. در هرکدام از زیر فصل‌ها به پاسخ هر سوال می‌پردازیم.

۱.۲ سوال اول

صورت سوال: موارد زیر را به صورت خلاصه شرح دهید.

الف) یادگیری تقویتی در و وجه تمایز آن از سایر روش‌های یادگیری

پاسخ:

یک سوال اساسی که در ابتدای یادگیری روش یادگیری تقویتی مطرح می‌شود این است که چه چیزی یادگیری تقویتی را از سایر الگوریتم‌های یادگیری ماشین متمایز می‌کند؟ در اینجا یک جمله‌ی کلیدی مطرح می‌شود:

“The answer is, There is no supervisor! Only a signal called reward.”

درواقع مهم‌ترین تفاوت یادگیری تقویتی با روش‌های یادگیری دیگر مانند یادگیری نظارت شده این است که داده‌ها برچسب (لیبل) ندارند و درواقع معلمی وجود ندارد که بگوید پاسخ صحیح به ازای داده‌ی ورودی خاص چه خواهد بود. در یادگیری تقویتی نیازی به دیتاستی که شامل لیبل باشد نیست و به پاسخ صحیح نیز دسترسی وجود ندارد.

وجه شباهت یادگیری تقویتی با یادگیری غیرنظارتی در این است که هردو به دیتاست لیبل‌دار نیازی ندارند اما نقطه تمایز آن‌ها در اینجاست که در یادگیری غیرنظارتی جهت گروه‌بندی لازم است مدل ساختار داده را بفهمد که در یادگیری تقویتی نیازی به یافتن ساختار نیست. و همانطور که در جمله‌ی کلیدی که بیان شد نیاز به سیگنال پاداش است و بیشینه کردن آن.

از سایر تفاوت‌ها می‌توان به این موضوع اشاره کرد که در یادگیری تقویتی فیدبک و تاثیر اعمال در دراز مدت دریافت می‌شود یا به عبارتی به تاخیر می‌افتد و مانند سایر روش‌های دیگر سریعا بازخورد اعمال دریافت نمی‌شود. در یادگیری تقویتی زمان اهمیت پیدا می‌کند و عملی خوب است که در دراز

مدت بیشینه پاداش را به ارمغان بیاورد. در یادگیری تقویتی عمل فعلی می‌تواند برروی داده‌های دریافتی در لحظات بعد نیز تاثیر بگذارد.

ب) سه عنصر اساسی یادگیری تقویتی: عامل، محیط و تعامل میان آن‌ها

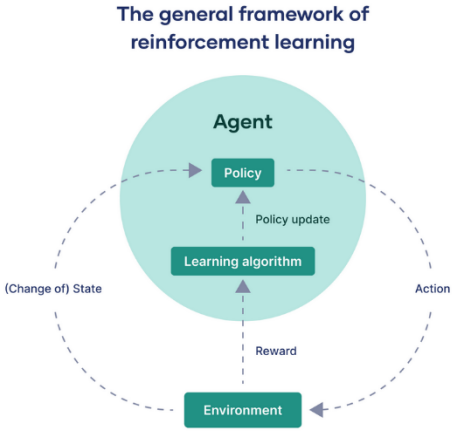
پاسخ:

در این بخش برای انتقال مفهوم عامل از یک مثال کنترلی استفاده می‌کنیم. فرض کنید یک سیستم موجود است که می‌خواهیم آن را کنترل کنیم که دینامیک مشخصی از آن در دسترس نیست. در اینجا عامل همان کنترل کننده‌ای است که باید در تعامل با محیط/فرایند یادبگیرد و بتواند به شکل مطلوبی سیستم را کنترل کند. در اینجا هرآنچه که خارج از عامل است محیط نامیده می‌شود. محیط می‌تواند واقعی و یا شبیه سازی باشد. هردو محیط و عامل با توجه به اهداف تعیین می‌شوند.

عامل همان هوش مصنوعی است که قرار است در تعامل با دنیای خارج از خودش که محیط نامیده می‌شود، یادبگیرد و به هدف که بیشینه کردن پاداش است برسد. عامل در هر گام زمانی عملی انجام می‌دهد، از محیط پاداش دریافت می‌کند، و محیط به حالت جدید می‌رود.

محیط در گام زمانی مذکور یک عمل از عامل دریافت می‌کند و به ازای آن به عامل پاداشی می‌دهد و وارد حالت بعدی خود می‌شود.

عامل باید با توجه به حالت موجود و پردازش پاداش دریافتی بهترین عمل را انتخاب کند، که یادگیری در این تعامل اتفاق می‌افتد.



شکل ۱: دیاگرام کلی یادگیری تقویتی

پ) مفهوم پاداش و تاثیر آن بر رفتار عامل.

پاسخ:

پاداش یک سیگنال بازخورد عددی است.

"Reward R_t is a scalar feedback signal"

پاداش بیان می‌دارد، که عامل در گام زمانی مشخص چقدر خوب عمل کرده است، حالت جدید چقدر به حالت مطلوب نزدیک است.

باید توجه کرد که کار عامل این است که پاداش درازمدت یا درواقع Cumulative reward را بیشینه کند. یادگیری تقویتی مبتنی بر فرضیه پاداش است:

"all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)."

در یادگیری تقویتی، یادگیری نوعی فرایند سیستماتیک است برای تنظیم پارامترهای سیاست به منظور رسیدن به سیاست بهینه، که این بروزرسانی سیاست از روی پاداش و جریمه انجام می‌شود.

درواقع پاداش برروی انتخاب عمل توسط عامل موثر است. پاداش باید به گونه‌ای انتخاب شود که بهبود سیاست و انتخاب عمل‌های بهینه‌تر را به خوبی نشان دهد و درواقع پاداش باید معنا دار تعریف شود.

پاداش می‌تواند برای هر عمل، برای انتقال به یک حالت خاص و یا پایان موفق بازی باشد.

ت) اهمیت تعریف هدف در وظایف یادگیری تقویتی.

پاسخ:

با توجه به اینکه در یادگیری تقویتی راه حل مسئله ارائه نمی‌شود و عامل لازم است برای رسیدن به هدف اعمالی را انجام دهد، بیان هدف بسیار کلیدی خواهد بود.

درواقع انتخاب هدف مناسب به عامل در انتخاب عمل کمک می‌کند، و ارزیابی عمل را ممکن می‌کند، هدف مشخص به عامل کمک می‌کند تا بین جستجو و بهره‌برداری تعادل برقرار کرده و با تنظیم

پاداش‌ها، به سمت رفتارهای مطلوب هدایت شود. همچنین، تعریف هدف انگیزه‌ای برای تلاش بیشتر در مواجهه با چالش‌ها و تمرکز بر جنبه‌های کلیدی در محیط‌های پیچیده ایجاد می‌کند.

در یادگیری تقویتی، هدف انتخاب عملی است که پاداش دراز مدت را بیشینه می‌کند. پس در اینجا درمی‌یابیم که انتخاب نوع پاداش نیز بسیار مهم است. در یادگیری تقویتی باید توجه داشت که ممکن است هر عملی تاثیرات طولانی مدت داشته باشد و بهتر است که پاداش‌های کوتاه مدت برای بیشینه شدن پاداش دراز مدت، قربانی شوند.

ث) تمایز بین وظایف اپیزودیک و مداوم

پاسخ:

در یادگیری تقویتی، وظایف به دو نوع اپیزودیک و مداوم تقسیم می‌شوند. در وظایف اپیزودیک، مسئله به مراحل یا اپیزودهایی تقسیم می‌شود که هر کدام دارای یک نقطه‌ی شروع هستند و با یک وضعیت پایانی به اتمام می‌رسند. در وظایف اپیزودیک تصمیمات گرفته شده در یک اپیزود تاثیری بر اپیزود دیگر نمی‌گذارد، درواقع اپیزود فعلی ربطی به اپیزود قبل و آنچه که در آن اتفاق افتاده ندارد، اما در وظایف مداوم، نقطه‌ی شروع وجود دارد ولی هیچ وضعیت پایانی مشخصی وجود ندارد و عامل به صورت پیوسته در حال تعامل با محیط است. به عنوان مثال، کنترل راه رفتن یک ربات در محیطی که دائماً در حال تغییر است و یا کنترل فشار مربوط به یک مخزن در یک پالایشگاه، یک وظیفه مداوم است. درواقع نوع وظیفه به محیط مرتبط می‌شود.

ج) تعادل بین Exploration و Exploitation و اهمیت آن در یادگیری تقویتی.

پاسخ:

یادگیری تقویتی مانند یادگیری از طریق آزمون و خطاست، و در آن عامل باید کشف کند که یک سیاست خوب براساس تجربه‌ای که در محیط بدست آورده، چه می‌تواند باشد بدون آنکه پاداش زیادی را از دست بدهد یا به عبارتی بیش از حد جریمه شود. جست و جو به عامل کمک می‌کند تا اطلاعات مفیدی درباره‌ی محیط کسب کند، و بهره برداری از اطلاعات شناخته شده برای به حداکثر رساندن

پاداش بهره می‌برد. با بهره برداری، براســـاس شـــناخت تا لحظه‌ی جاری به ســـمت حالت بهینه حرکت می‌کنیم و جســـت و جو نیز باعث شـــناخت بهتر در طول زمان می‌گردد و کمک می‌کند تا بتوانیم عمل‌ها بهتری را انتخاب کنیم. در یادگیری تقویتی چون هدف بیشـــینه کردن پاداش در طولانی مدت اســـت، تعادل میان جســـت و جو و بهره برداری بســـیار حســـاس می‌شـــود، اســـتفاده محض از بهره برداری منجر به سیاست زیر بهینه می‌شود و عامل نیاز دارد تا کل فضای محیط را جست و جو کند تا عمل‌های بهتری را بتواند انتخاب کند.

چ) تمایز میان اقدام و سیاست

پاسخ:

ســیاســت درواقع تابع رفتار عامل اســت، بیان می‌دارد که چه عملی(یا اقدامی) انتخاب شــود، ســیاســت بیان می‌دارد که در حالت‌های مختلف عمل مناسبی را انتخاب کنیم، به بیان بهتر تابعی که در یک حالت خاص یک عمل خاص را انتخاب می‌کند، سیاست نام دارد، و یادگیری تقویتی نیز در حقیقت یک نگاشت از حالت به عمل است.

ح) تمایز بین پاداش لحظه‌ای و پاداش بلند مدت

پاسخ:

پاداش لحظه‌ای مربوط به این اســـت که عامل بلافاصـــله پس از انجام یک عمل پاداشـــی را دریافت می‌کند، این پاداش مربوط به انجام آن عمل در آن گام زمانی اســـت و اطلاعاتی درخصـــوص نتیجه‌ی طولانی مدت عمل نمی‌دهد.

پاداش بلندمدت نیز مربوط به پاداش تجمعی‌ای اســت که عامل در طول زمان و با انجام یک ســری از اعمال دریافت می‌کند، این پاداش شـــامل پاداش‌های لحظه‌ای و نتیجه‌ی اعمال در طولانی مدت نیز می‌باشد.

در بخش‌ها قبلی به این موضوع اشاره شد که در یادگیری تقویتی، هدف انتخاب عملی است که پاداش دراز مدت را بیشـــینه می‌کند. پس در اینجا درمی‌یابیم که انتخاب نوع پاداش نیز بســـیار مهم اســـت. در

یادگیری تقویتی باید توجه داشت که ممکن است هر عملی تاثیرات طولانی مدت داشته باشد و بهتر است که پاداش‌های کوتاه مدت برای بیشینه شدن پاداش دراز مدت، قربانی شوند.

خ) مفهوم تابع ارزش و نقش آن در یادگیری تقویتی

پاسخ:

دو نوع تابع ارزش در یادگیری تقویتی مطرح می‌شود، اولی تابع ارزش حالت است ($V(s)$) که بیان می‌دارد حالت s چقدر ارزشمند است و چقدر می‌تواند امتیاز بیشتری را تولید کند.

تابع بعدی تابع ارزش حالت- عمل است ($Q(s,a)$) که بیان می‌دارد اگر در حالت s باشیم و عمل a را انجام دهیم، چقدر ارزشمند خواهد بود و چقدر ما را به امتیاز بالاتر هدایت می‌کند.

در همه‌ی موارد باید طولانی مدت لحاظ شود، می‌خواهیم در طولانی مدت عملی انجام شود که به صورت میانگین بیشترین پاداش را تولید کند.

تابع ارزش درواقع به پاداش دراز مدت اشاره می‌کند، برای یک reward و یک state خاص تعریف می‌شود و به policy ربط دارد. معیار انتخاب عمل بهینه نیز تابع ارزش است.

۱.۳ سوال دوم

صورت سوال: نمونه‌ای از سناریو یا مشکل دنیای واقعی را ارائه داده و نحوه تعریف محیط و تعامل آن با عامل از طریق چرخه حالت-اقدام-پاداش را توضیح دهید.

پاسخ:

مثالی که در این قسمت ارائه می‌کنیم، در خصوص بحث Obstacle Avoidance در سیستم‌های خودمختار است، مانند خودروی خودران یا پرنده‌های عمودپرواز، تنها سنسوری هم که در این بخش به سیستم خودمختار کمک می‌کند دوربین آن است (Partial Observability)، در واقع ربات یا عامل فقط شامل یک دوربین و سیستم بینایی ماشین است و از محل دقیق خود اطلاعی ندارد.

دراینجا عامل برای یک خودروی خودران، مسیری تعریف می‌شود شامل موانع مختلف و pedestrians که شامل علائم راهنمایی، جدول، عابرین پیاده و سایر خودروها است.

برای پرنده‌ی عمود پرواز نیز می‌توان محیطی شامل ساختمان‌ها و دکل‌ها و کابل‌های برق تعریف کرد.

عامل نیز همان سیستم هوش مصنوعی پرنده و یا خودروی خودران است که از طریق تعامل با محیط باید یادبگیرد و از برخورد با موانع اجتناب کند.

در این بحث می‌توان محیط و عامل را به صورت شبیه سازی تعریف کرد.

داده‌ی دریافتی عامل از محیط نیز همان فریمی است که دوربین آن ضبط می‌کند و در کنار آن یک هوش مصنوعی نظارتی نیز صرفا می‌تواند برای کمک به عامل pedestrianها و موانع را آشکار کند (detection) اما اطلاعی از محل دقیق قرار گیری، فاصله با مانع و غیره نداریم.

فضای اعمال می‌تواند شامل کم کردن سرعت، تغییر مسیر، توقف، تغییر ارتفاع (پرنده) باشد.

پاداش‌های مثبت زمانی داده می‌شوند که عامل با موفقیت از موانع دوری کرده و مسیر ایمن را طی کند. پاداش منفی (جریمه) در صورت برخورد با مانع یا نزدیک شدن بیش از حد به آن تعلق می‌گیرد. هدف عامل این است که پاداش بلندمدت را با اجتناب مداوم از موانع و حرکت ایمن به حداکثر برساند.

۱.٤ سوال سوم

صورت سوال: **درمورد چالش‌ها و محدودیت‌های بالقوه یادگیری تقویتی شرح دهید. حداقل دو چالش را شناسایی و مورد بحث قرار دهید و آن‌ها را درکاربردهای واقعی یادگیری تقویتی توضیح دهید.**

پاسخ:

یکی از چالش‌ها طراحی و فراهم کردن محیط است، محیط هم می‌تواند شبیه سازی باشد و هم واقعی، در محیط واقعی بحث ایمنی و خسارت مطرح می‌شود درحالی که همزمان، در محیط واقعی می‌توان به دقت خوب رسید، و همچنین قرارگیری عامل در محیط واقعی ضروری نیز می‌باشد. در محیط شبیه‌سازی نیز سرعت یادگیری بالاتر است، می‌توان حالات مختلف را شبیه سازی کرد و همچنین به لحاظ ایمنی بهتر است اما شبیه سازی تفاوت بالقوه با واقعیت دارد.

چالش دیگر تعیین پاداش است، طراحی پاداش مناسب دشوار است؛ زیرا یک پاداش نامناسب می‌تواند عامل را به سمت رفتارهای نامطلوب هدایت کند و به همگرایی نرسیم. به علاوه اینکه وقتی پاداش را با

تاخیر بدهیم ممکن است فرایند یادگیری پیچیده شود، مثلا در مسیریابی یک ربات و یافتن بهترین و کوتاه‌ترین مسیر، فقط هنگامی پاداش بدهیم که به مقصد رسیدیم، یادگیری مشکل می‌شود و در این فرایند ممکن است حالاتی داشته باشیم که ربات اصلا به مقصد نرسد و راه اشتباه برود.

بحث دیگر ناپایداری است، که می‌تواند در محیط‌های داینامیک و پیچیده رخ دهد، تغییرات کوچک در پارامترهای عامل یا محیط می‌تولند منجر به ناپایداری و حتی واگرایی شود. برای مثال، در بازارهای مالی نوسانات کوچک در بازار می‌تواند به تصمیمات نادرست و زیان‌های بزرگ منجر شود، زیرا سیستم نتوانسته به درستی با شرایط جدید سازگار شود.

بحث دیگر همان تعادل میان بهره برداری و جست و جو است، گفتیم که استفاده محض از بهره برداری منجر به سیاست زیر بهینه می‌شود و نمی‌تواند عمل‌های بهینه را بیاید، عامل نیاز دارد تا کل فضای محیط را جست و جو کند تا عمل‌های بهتری را بتواند انتخاب کند. همچنین اگر عامل بیش از حد جستجو کند، ممکن است زمان زیادی صرف یادگیری کند و نتایج زیربهینه بگیرد. فرض کنید دو نفر درحال بازی شطرنج هستند، عامل فقط ۳ استراتژی برای مات کردن یادگرفته است و با اتکا بر آن‌ها می‌خواهد حریف ماهر را شکست بدهد، در این صورت عامل همواره شکست خواهد خورد.

فصل ۲: مفاهیم راهزن چنددست

۲.۱ مقدمه

در این بخش به ارائه‌ی پاسخ برای مسائل مطرح شده در خصوص راهزن چند دست می‌پردازیم.

۲.۲ سوال چهارم

صورت سوال: **موارد زیر را درخصوص Multi-Armed Bandit توضیح دهید:**

الف) فرمول بندی مسائل Multi-Armed Bandit و ارتباط آن‌ها با سناریوهای تصمیم‌گیری متوالی به چه شکل است.

پاسخ)

در مسئله راهزن چند دست، فرض کنید یک عامل با مجموعه‌ای از بازوها روبرواست، که هر کدام از آنها توزیع احتمالی پاداش متفاوتی دارند. عامل می‌تواند در هر گام یا مرحله یکی از این بازوها را انتخاب کرده و بکشد (عمل) و سپس باتوجه به توزیع آن بازو، پاداشی دریافت می‌کند.

هربازو که نماینده یک عمل است می‌تواند توزیعی با امیدریاضی خاص و انحراف معیار مخصوص به خودش داشته باشد. وقتی عامل یک بازو را می‌کشد از توزیع آن یک پاداش تصادفی دریافت می۲کند که برای او ناشناخته است.

سیاست نیز درواقع این است که در هرمرحله کدام بازو کشیده شود تا به بیشینه پاداش طولانی مدت برسیم. باید توجه کرد که این بازی تک حالته است. عامل باید بین بهره‌برداری از بازوهایی که تاکنون کشیده است و عملکرد خوبی داشته‌اند و جستجو برای یافتن بازوهای بهتر تعادل برقرار کند.

هدف این است که پاداش تجمعی یا دراز مدت نیز بیشینه شود، ساده ترین حالت استفاده از میانگین پاداش‌ها است، روش‌های بهتری مانند sample average و weighted average هستند.

از روش‌های مختلف برای حل مسئله راهزن چند دست می‌توان به الگوریتم‌های e-greedy، UCB و Gradient Bandit اشاره کرد.

مسئله راهزن چنددست، مدل ساده‌ای از تصمیم‌گیری متوالی است که در آن عامل در هر مرحله یک تصمیم می‌گیرد و پاداش مربوط به آن تصمیم را مشاهده می‌کند و حالت و شرایط محیط نیز تفاوتی نمی‌کند. این مدل در واقع یک نسخه پایه از مسائل تصمیم‌گیری متوالی است که در آن‌ها با تغییر حالت و ... مواجه هستیم.

در سناریوهای تصمیم‌گیری متوالی، علاوه بر انتخاب عمل در هر لحظه، عامل با حالت‌های مختلف محیط مواجه است و تأثیر اعمالش می‌تواند به آینده نیز سرایت کند. بنابراین، مسئله راهزن چند دست را می‌توان به عنوان نسخه ساده‌ای از تصمیم‌گیری متوالی در نظر گرفت که در آن تنها یک تصمیم (کشیدن بازو) در هر مرحله وجود دارد و فقط یک حالت موجود است و پاداش‌ها با تاخیر داده نمی‌شوند.

کاربرد این مسئله در دنیای واقعی در آزمایشات بالینی، مدلسازی رفتارهای انسانی و طراحی سیستم‌های توصیه‌گر است.

ب) تمایز بین exploration و exploitation در این مسائل.

پاسخ)

در اینجا exploration یعنی کشیدن بازوهایی که تا به حال نکشیده‌ایم و ممکن است بهینه باشند یا نباشند، exploitation یعنی کشیدن بازویی که تا الان بیشترین پاداش را به ما داده‌است، درواقع در مورد دوم براساس اطلاعاتی که تا به حال از محیط کسب کرده‌ایم، عملی را انجام می‌دهیم که بیشترین پاداش را برای ما فراهم ساخته یا درواقع مقدار تابع ارزش را بیشینه کرده‌است، ولی با انجام جست و جو می‌خواهیم بیابیم آیا ممکن است بازویی باشد که وضعیت مارا بهتر کند یا خیر.

پ) مفهوم تخمین عمل-ارزش و اهمیت آن در یادگیری استراتژی بهینه.

باید بخاطر داشت که مقدار ارزش یک عمل برابر است با میانگین پاداش آن عمل وقتی که انتخاب می‌شود، یک راه تجربی برای استخراج آن و تخمین آن این است که میانگین پاداش‌ها را زمانی که از محیط دریافت می‌کنیم، برای هر عمل خاص محاسبه کنیم. (رابطه‌ی اول فصل دوم از کتاب ساتن)

وقتی که تعداد انجام یک عمل به بی‌نهایت میل می‌کند، مطابق قانون اعداد بزرگ ارزش آن عمل به سمت ارزش بهینه میل می‌کند. این روش یکی از روش‌های تخمین ارزش عمل است و طبیعتا بهترین نیست.

ساده‌ترین راه برای انتخاب عمل این است که عملی را انتخاب کنیم که بیشترین ارزش تخمین زده شده را داراست. این روش درواقع نوعی انتخاب حریصانه و یا مبتنی بر ایده‌ی بهره برداری است. (رابطه‌ی ۲ در فصل دوم کتاب ساتن)

$$A_t \doteq \underset{a}{\operatorname{argmax}}\, Q_t(a)$$

این روش از دانشی که تا الان کسب کردیم بهره می‌برد و به سایر اعمال نگاه نمی‌کند تا شاید عمل بهتری را بیابد. بهترین راه این است که در بیشتر موارد براساس بهره برداری عمل کنیم، و در زمان‌هایی

با یک احتمال کم در حد ε به صورت رندوم از میان سایر اعمال نیز عملی را انتخاب کنیم (با احتمال مساوی) به این روش همان $\varepsilon - greedy$ می‌گوییم. این موضوع تضمین می‌دهد که تمامی اعمال به تعدا زیاد انتخاب می‌شوند و زمانی که گام‌های زمانی به سمت بی‌نهایت می‌روند، ارزش تمام اعمال به مقدار بهینه خودشان میل می‌کند. در این قسمت احتمال انتخاب عمل بهینه به بیشتر از مقدار $1 - \varepsilon$ نیز میل خواهد کرد.

ت) مفهوم استراتژی greedy

پاسخ:

در پاسخ قسمت قبل به این موضوع اشاره کردیم مجددا تکرار می‌کنیم:

ساده‌ترین راه برای انتخاب عمل این است که عملی را انتخاب کنیم که بیشترین ارزش تخمین زده شده را داراست. این روش درواقع نوعی انتخاب حریصانه و یا مبتنی بر ایده‌ی بهره برداری است. (رابطه‌ی ۲ در فصل دوم کتاب ساتن)

$$A_t \doteq \underset{a}{\operatorname{argmax}}\, Q_t(a)$$

این روش از دانشی که تا الان کسب کردیم بهره می‌برد و به سایر اعمال نگاه نمی‌کند تا شاید عمل بهتری را بیابد.

ث) استراتژی‌های e-greedy و UCB به عنوان روش‌های متعادل‌ساز اکتشاف و بهره برداری.

پاسخ:

گفتیم که در روش e-greedy، با احتمال ε از میان سایر اعمال به صورت تصادفی (که احتمال انتخاب هرکدام با هم برابر است) یک عملی را انتخاب می‌کنیم و با احتمال $1 - \varepsilon$ نیز عمل بهینه را انتخاب خواهیم کرد، این موضوع باعث می‌شود علاوه بر بهره برداری، سایر اعمال را نیز تست کنیم، و اگر عملی یافتیم که بهینه‌تر از قبلی است در ادامه در حالت حریصانه آن را انتخاب کنیم. در این روش انتخاب از میان سایر اعمال به نوعی کورکورانه است و فرقی بین عمل‌ها وجود ندارد. در روش UCB ما با آگاهی بیشتری جست و جو را انجام می‌دهیم، در این روش عمل‌ها را با توجه به پتانسیل آن‌ها برای بهینه بودن

و میزان عدم قطعیت آن انتخاب می‌کنیم، هر عمل که بهینه‌تر است و عدم قطعیت آن کمتر است در سیاست مذکور انتخاب می‌شود.

$$A_t \doteq \underset{a}{\text{argmax}}[\, Q_t(a) + c\sqrt{\frac{lnt}{N_t(a)}}]$$

ترم زمانی موجود در رابطه‌ی بالا باعث می‌شود که در طول زمان همه‌ی عمل‌ها انتخاب بشوند حتی اگر به Q^* هم رسیده باشیم. وقتی زمان می‌گذرد، ترم زمانی بزرگتر می‌شود، این باعث می‌شود در طول زمان باز هم بتوانیم همه‌ی عمل‌ها را انتخاب کنیم. به علاوه مقدار N که به تعداد دفعه‌ی انتخاب شدن یک عمل اشاره دارد، باعث می‌شود که عمل‌هایی که یک مدت زمان طولانی انتخاب نشده‌اند نیز دوباره انتخاب بشوند، عمل بهینه نیز با توجه به مقدار ارزشی که دارد نیز انتخاب می‌گردد. در ابتدا که هیچکدام از اعمال انتخاب نشده‌اند نیز چون N صفر است، کسر بیشینه مقدار است و این باعث می‌شود که در گام‌های ابتدایی تمامی عمل‌ها انتخاب شوند. این موضوع مانند شرایط اولیه‌ی خوش‌بینانه عمل می‌کند با این تفاوت که در شرایط اولیه خوش‌بینانه فقط در زمان‌های ابتدایی جست و جو خواهیم داشت اما در UCB این جست و جو همواره ادامه دارد، و مانند روش اپسیلون-حریصانه، این جست و جو به صورت یکنواخت نیست و به عدم قطعیت نیز مرتبط است.

ج) ارزیابی و مقایسه‌ی استراتژی‌های مختلف انتخاب عمل از منظر همگرایی

پاسخ:

در روش greedy با توجه به اینکه صرفا حریصانه عمل می‌کنیم و جست و جویی موجود نیست به سمت نقطه‌ی بهینه همگرا نمی‌شویم، لازم است جست و جو نیز انجام شود، به همین علت روش e-greedy بهتر است زیرا جست و جو را فراهم می‌سازد، هر چند که این جست و جو نیز به صورت یکنواخت است. در روش مذکور، انتخاب مقدار اپسیلون نیز بسیار مهم است، اپسیلون بزرگ باعث می‌شود احتمال انتخاب عمل بهینه در مجموع بالا نباشد و پاداش میانگین که دریافت می‌شود زیر حالت بهینه باشد، انتخاب مقدار خیلی کوچک برای اپسیلون فرایند یادگیری را نیز کند می‌کند اما در مجموع و در زمان طولانی می‌توان به پاداش میانگین بیشتری رسید و در موارد بیشتری عمل بهینه انتخاب خواهد

شـد. روش UCB نیز تضـمین می‌دهد که تمامی اعمال در ابتدا تسـت می‌شـوند، و در ادامه همچنان این جست و جو با توجه به میزان عدم قطعیت ادامه می‌یابد. در همگرایی این روش نیز مقدار پارامتر c بسیار موثر است زیرا که نوعی بهره است که در عدم قطعیت موثر است. روش gradient bandit نیز روشی دیگر است که از طریق پاداش به سیاست می‌رسیم بدون آنکه ارزشی تخمین بزنیم.

در همه‌ی این روش‌ها با فرض انتخاب پارامترهای مناسـب و در مدت زمان یکسـان UCB مقدار Average Reward بیشتری را به ارمغان می‌آورد. به علاوه در انتخاب روش لازم است به حساسیت روش نسبت به تغییر پارامترهای آن توجه داشت که از این لحاظ مجددا UCB وضعیت بهتری دارد.

چ) چند نمونه از کاربردهای عملی مسئله Multi-Armed Bandit

پاسخ:

در موارد قبل به موضوعات آزمایشات بالینی، مدلسازی رفتارهای انسانی و طراحی سیستم‌های توصیه گر اشـاره کردیم. یک مثال را با توضـیح اگر بخواهیم بیان کنیم، انتخاب بهترین بنر تبلیغاتی اسـت. فرض کنید برای کمپین یک محصول خاص می‌خواهیم مردم را دعوت به خرید بلیط کنیم، اینکه تبلیغات ما به چه صـورت باشـد مهم اسـت فرض کنید ۱۰ بنر تبلیغاتی داریم از میان آن‌ها باید موردی را برای تبلیغات گسـترده انتخاب کنیم که بیشـترین سـود را به ارمغان می‌آورد. یا یک سـایت خبری چه خبری را به یک مشاهده کننده سایت نشان دهد یا تلاش برای مسیریابی تطبیقی در یک شبکه برای کاهش میزان تاخیر.

۲.۳ سوال پنجم

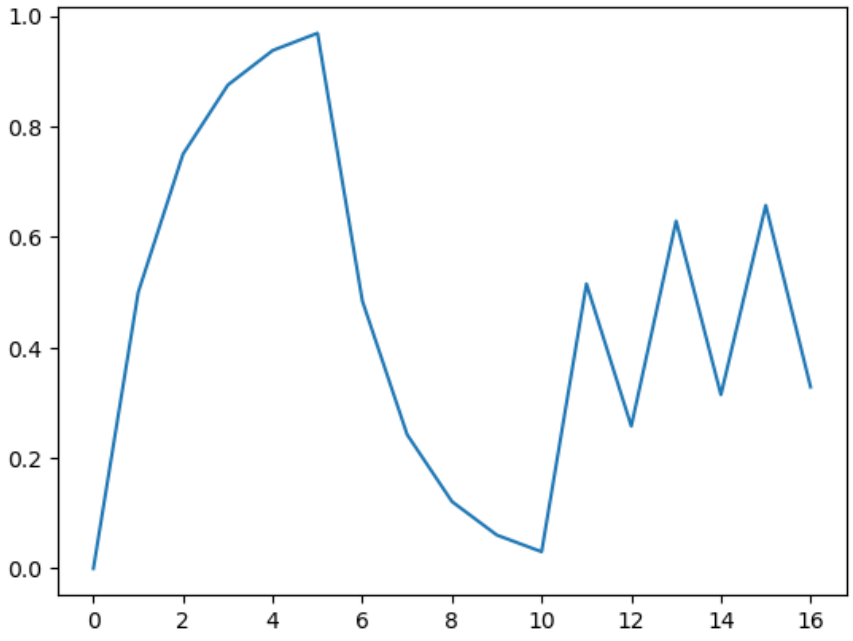
صورت سوال:

۵) با در نظر گرفتن رابطه به‌روزرسانی (۲.۵) و $Q_0 = 0$، تخمینی از Q_t را به ازای زمان‌های آتی به‌دست آورید.
در تمام قسمت‌های این سوال، سیگنال هدف را از $t = 0$ تا $t = 15$ به‌صورت زیر در نظر می‌گیریم:
[۱، ۱، ۱، ۱، ۱، ۰، ۰، ۰، ۰، ۰، ۰، ۱، ۰، ۱، ۰، ۱، ۰]

الف)

الف- به ازای اندازه گام $\alpha = 0.5$ ، تخمین‌های Q_t مربوط به مراحل زمانی ۱۵-۱ را به‌دست آورده، آنها را در نمودار ترسیم نموده و نقاط برآورد را با یک خط به هم متصل کنید. این تخمین در $t = 4$ چقدر به ۱ نزدیک است؟ برای یک لحظه فرض کنید که سیگنال هدف تا انتها ۱ باقی می‌ماند. بدون ترسیم، تخمین Q_t در $t = 10$ و $t = 20$ چقدر به ۱ نزدیک خواهد بود؟

پاسخ:



شکل ۲: نمودار ارزش بر حسب زمان

با استفاده از کد نوشته شده برای خطا داریم:

Error in t4: 0.0625

اگر با پاداش برابر ۱ ادامه پیدا کند:

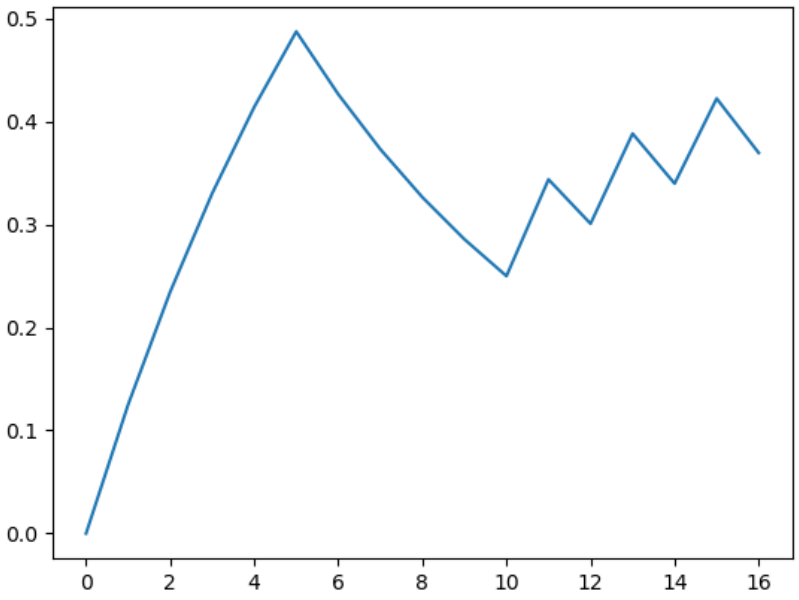
Error in t10: 0.9697265625

Error in t20: 0.041962623596191406

ب)

ب- با یک نمودار جدید، بخش نموداری قسمت الف را این بار به ازای اندازه گام $\alpha = \frac{1}{8}$ تکرار نمایید.

پاسخ:

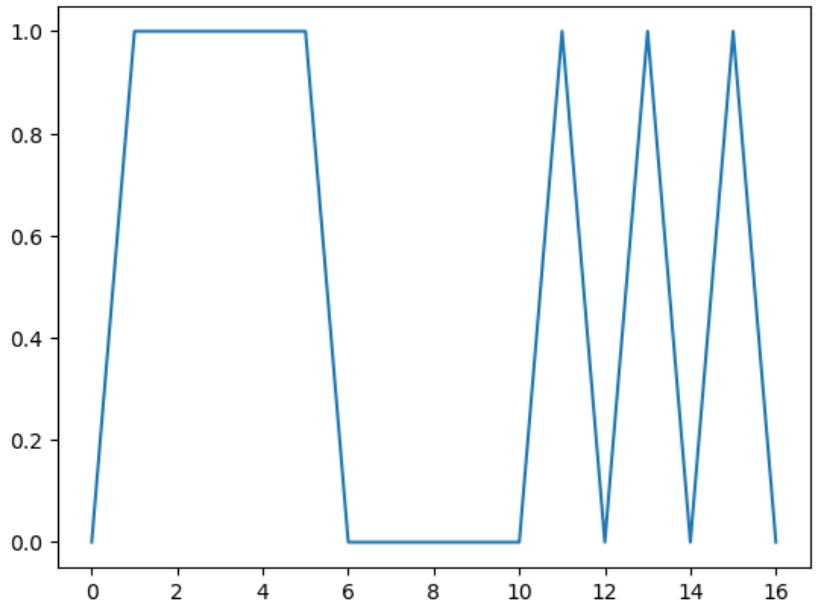


شکل ۳: نمودار ارزش برحسب زمان در آلفا یک هشتم

پ)

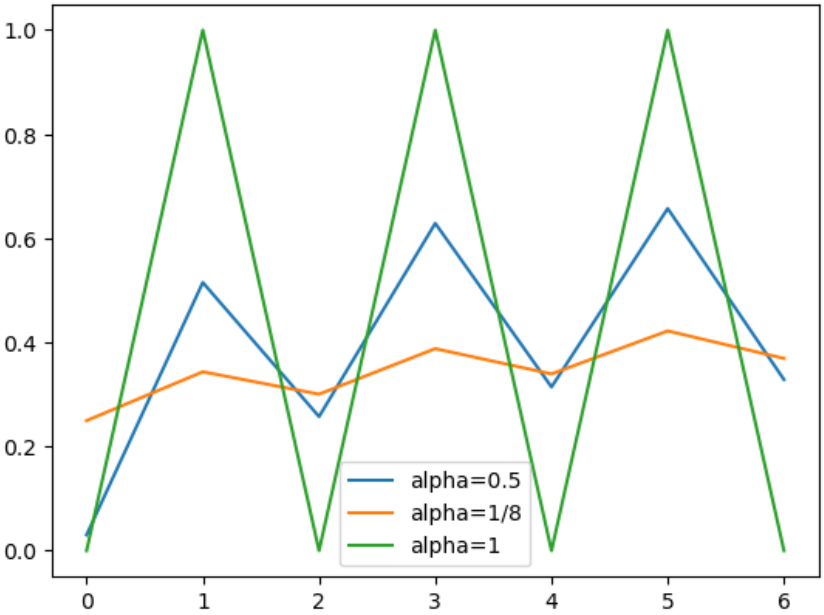
پ- نمودار سوم را به ازای اندازه گام $\alpha = 1$ تهیه کنید. وقتی هدف متناوب است (یعنی $t = 10$ تا $t = 15$) کدام اندازه گام تخمین‌هایی با خطای کوچک‌تری ایجاد می‌کند؟

پاسخ:



شکل ۴: نمودار ارزش برحسب زمان در آلفا برابر ۱

حال برای هر ۳ گام در زمانی که نوسان می‌کنند، هر ۳ را در یک نمودار رسم می‌کنیم.



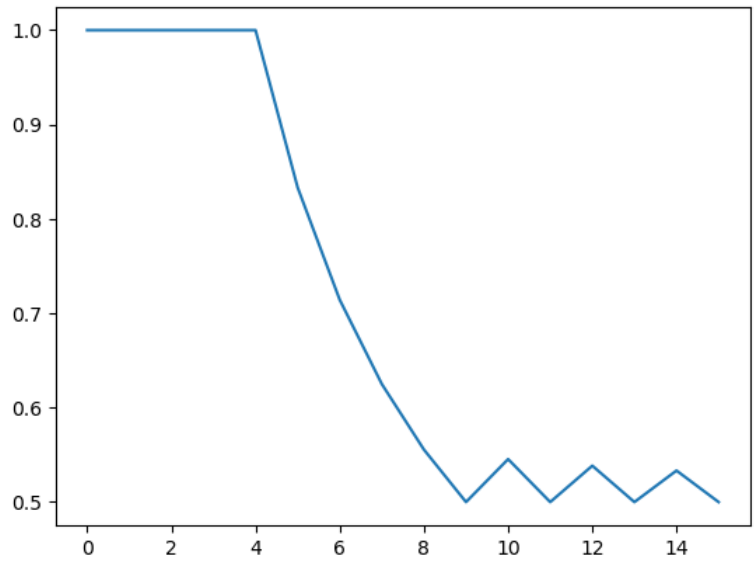
شکل ۵: نمودار ارزش بر حسب زمان در ۳ گام

همانطور که مشاهده می‌گردد با انتخاب اندازه گام بیشتر، میزان نوسان بیشتر می‌شود، به صورت میانگین خطای آلفا برابر ۱ کمتر از آلفا برابر ۰.۵ و آلفا برابر ۰.۱۲۵ است اما در مقاطعی آلفا برابر ۱ خطای بسیار زیادی حدود ۱ ایجاد می‌کند، که این موضوع با مقدار کمتری برای حالت آلفا برابر ۰.۵ نیز برقرار است.

ت)

ت- نمودار چهارم را به ازای اندازه گام $\alpha = \frac{1}{t}$ رسم نمایید. بر اساس این نمودارها، این اندازه گام برای چه مسائلی مناسب است؟ چرا همیشه انتخاب درستی نیست؟

پاسخ:



شکل ۶: نمودار ارزش برحسب زمان برای آلفای 1/t

این گام تضمین می‌کند که الگوریتم به‌تدریج همگرا می‌شود و به یک مقدار پایدار نزدیک می‌شود. زیرا با گذشت زمان، تأثیر نمونه‌های جدید کمتر می‌شود و الگوریتم به مقدار قبلی خود پایبند می‌ماند. زمانی که پاداش‌ها نوسانی هستند این مقدار می‌تواند مناسب باشد.

برخی مسائل، محیط یا پاداش‌ها ممکن است به مرور زمان تغییر کنند. در چنین مواقعی، استفاده از این آلفا مناسب نیست، چون باعث می‌شود الگوریتم به تغییرات جدید کندتر واکنش نشان دهد و با افزایش

زمان نیز پاداش‌های جدید کمتر اثر خواهند کرد زیرا که آلفا خیلی کوچک شده است و یادگیری بسیار کند می‌شود.

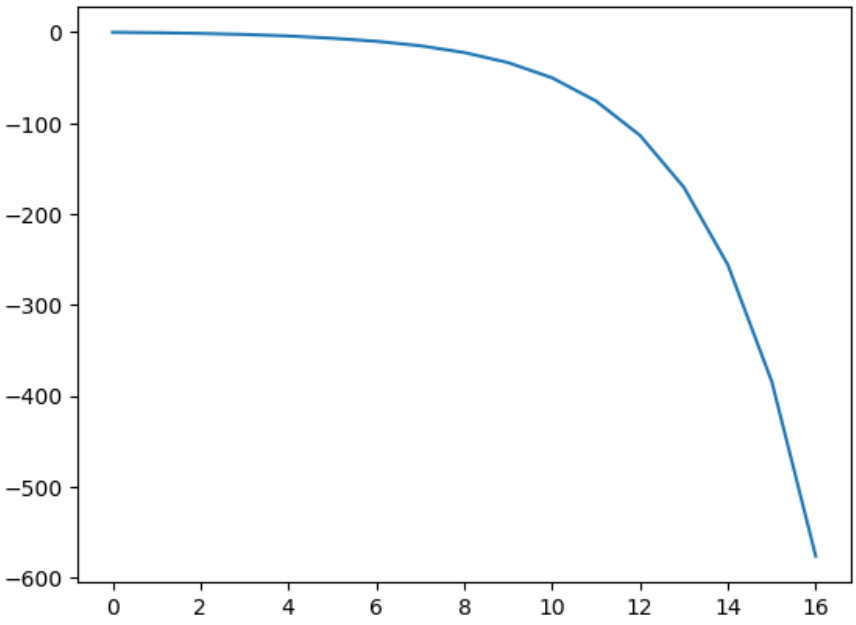
ث)

ث- اگر اندازه گام $\alpha = -0.5$ باشد چه اتفاقی می افتد؟ $\alpha = 1.5$؟ محدوده ایمن برای اندازه گام چقدر است؟ انتخاب درستی نیست؟

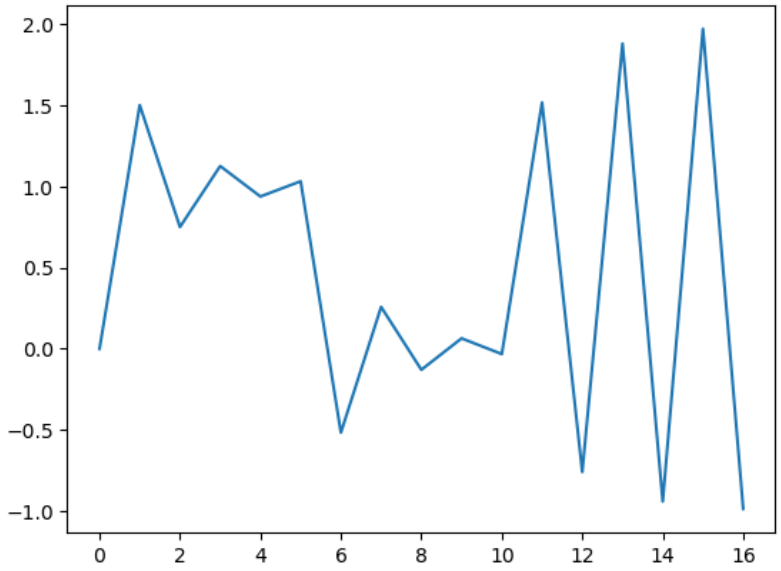
پاسخ:

با انتخاب مقدار منفی سیستم واگرا می‌شود، مقدار ۱.۵ نیز مناسب نیست، شرطی که برای انتخاب آلفا داریم این است که این مقدار بزرگتر از ۰ باشد و درون بازه‌ی بین ۰ تا ۱ قرار بگیرد، ضمنا جمع ضرایط در تخمین ارزش باید ۱ شود، که این مقادیر شروط را برای داشتن یک تخمین نا اریب نقض می‌کنند.

نمودارها نیز در ادامه موجود است:



شکل ۷: آلفای منفی ۰.۵



شکل ۸: آلفای بزرگتر از ۱ (۱.۵)

ج)

ج- حال فرض کنید پاداش $R_{t+1} = R_t + N(0,1)$ به طور تصادفی حرکت کند، به طوری‌که در آن $N(0,1)$ یک متغیر تصادفی توزیع شده نرمال با میانگین ۰ و واریانس ۱ باشد. در این مورد، کدام یک از این اندازه گام‌ها تخمینی با کمترین خطای مطلق را ایجاد می‌کند؟

پاسخ:

در این حالت باتوجه به اینکه محیط پویاست و قبلا نیز اشاره کردیم، بهتر است از آلفای ثابت استفاده کنیم. در این حالت برای داشتن تخمین ناریب لازم است به شروط آلفا توجه شود و مقدار آلفا را بین ۰ و ۱ انتخاب کنیم، در این حالت هرچه الفا کمتر باشد میزان نوسانات کمتر است، بهتر است آلفای برابر با ۰.۱۲۵ را انتخاب کنیم.