به نام خداوند بخشنده مهربان

تمرین دوم:

Markov Decision Processes and Dynamic programming

سوال اول

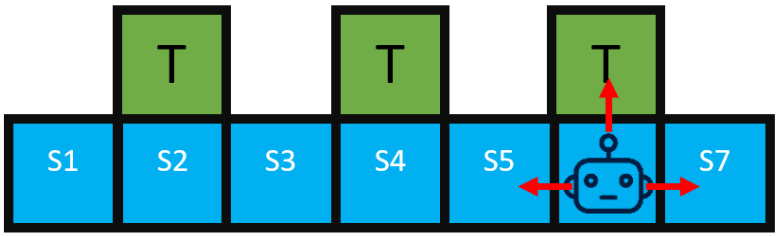
چه تفاوتی بین یادگیری تقویتی و یادگیری نظارتی (supervised learning) وجود دارد؟ توضیحات خود را برای دو مسئله شبیه به هم بیاورید که در یکی نیاز است تا از یادگیری تقویتی استفاده شود تا بتوانیم آن را حل کنیم و دیگری را با یادگیری نظارتی می‌توان حل نمود.

سوال دوم

در یک مسئله MDP اگر تابع ریوارد تحت یک تبدیل خطی (Linear transformation) تغییر یابد، آیا سیاست بهینه تغییر می‌کند؟(اثبات ریاضی و یا مثال نقض بیاورید و از حالت بدیهی ضرب کردن در صفر صرفنظر کنید.) آیا continuing task یا episodic task بودن در جواب این مسئله تغییری ایجاد می‌کند؟

سوال سوم

فرض کنید رباتی در یک محیط شطرنجی به صورت زیر فعالیت می‌کند. در هر اپیزود، ربات در یکی از خانه‌های ردیف پایین شروع به فعالیت می‌کند.(احتمال شروع در هر یک از خانه‌ها برابر دیگری است.) ربات می‌تواند به سمت چپ، راست و بالا حرکت کند. در صورتی که ربات تصمیم بگیرد به سمتی حرکت کند که دیوار وجود دارد، ربات در جای خود می‌ایستد. در صورتی که ربات وارد خانه‌های سبزرنگ گردد اپیزود تمام می‌شود.



شکل 1- ربات در محیط شطرنجی و انواع تصمیم‌های آن

سناریو اول:

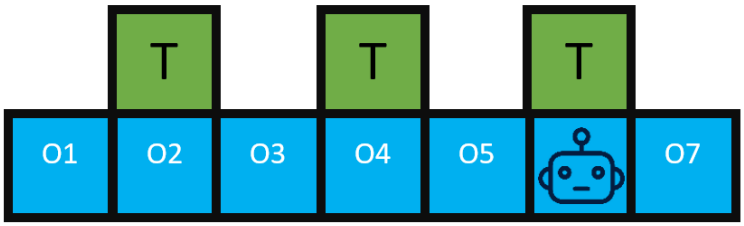
ربات محیط شطرنجی را به طور کامل می‌شناسد و در هرلحظه می‌داند که در کجای صفحه شطرنجی قرار دارد و براساس آن تصمیم گیری می‌کند. هدف آن است که ربات وارد ردیف دوم و یکی از خانه‌های سبزرنگ گردد. بدین اگر ربات وارد خانه‌های سبزرنگ گردد پاداش +1 دریافت می‌کند (و اپیزود تمام می‌شود) و اگر ربات از خانه‌ی آبی رنگی به خانه‌ی آبی رنگی دیگر برود پاداش 0 دریافت می‌کند. برای این تسک $\lambda = 0.9$ در نظرگرفته شده است.

الف) آیا تسکی که تعریف شده است را می‌توان با یک MDP نمایش داد؟ در صورت عدم امکان، توضیح دهید و در غیراین صورت MDP را به طور کامل مشخص کنید.

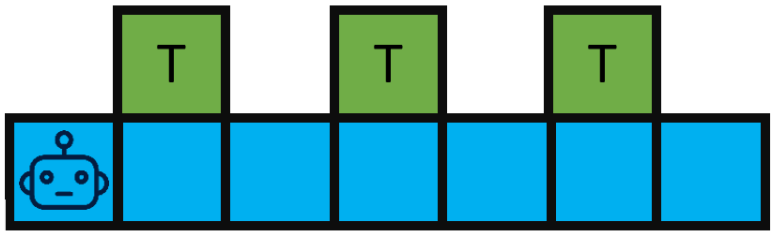
ب) برای حل این تسک چند سیاست deterministic بهینه وجود دارد؟ به روشی مناسب $\pi(s)$ را برای هریک نمایش دهید.

سناریو دوم:

ربات از محیطی که در آن است هیچ شناختی ندارد. در چهار طرف ربات سنسورهای فاصله سنجی تعبیه شده است که به ربات می‌گوید آیا دیواری در کنار خود دارد یا خیر. برای مثال به شکل‌های زیر توجه کنید.



شکل 2- ربات در این حالت یک آرایه به صورت {right=1,up=1,left=1,down=0} دریافت می‌کند.

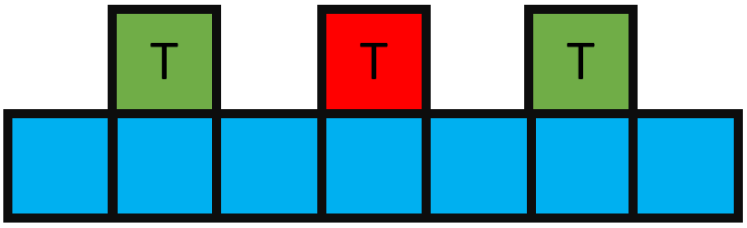


شکل 3- ربات در این حالت یک آرایه به صورت {right=1,up=0,left=0,down=0} دریافت می‌کند.

فرض کنید، هدف در اینجا مشابه سناریو اول باشد، از نظر المان‌های کلیدی یک MDP، در اینجا چه چیزی تغییر کرده است؟ آیا هنوز این مسئله با استفاده از MDP به گونه‌ای که در نهایت هدف ما براورده شود قابل نمایش است؟ (آیا سیاست بهینه همان چیزی هست که ما می‌خواهیم؟)

سناریو سوم:

فرض کنید همه چیز به مانند سناریو **دوم** است اما هدف ما تغییر کرده است و می‌خواهیم ربات تنها وارد دو خانه سبز رنگ چپ و راست گردد. زمانی که ربات وارد خانه قرمز رنگ در شکل زیر گردد ریوارد 1- دریافت خواهد کرد و اگر وارد خانه‌های سبزرنگ گردد ریوارد 1+ دریافت خواهد کرد. (بقیه ریواردها صفر هستند) برای این تسک $\lambda = 0.9$ در نظرگرفته شده است.



شکل 4- تغییر هدف

الف) آیا هنوز این مسئله با استفاده از MDP به گونه‌ای که در نهایت هدف ما براورده شود قابل نمایش است؟ (آیا سیاست بهینه همان چیزی هست که ما می‌خواهیم؟) در صورتی که می‌توانیم، سیاست بهینه مورد نظر را ارائه کنید و در صورتی که نمی‌توانیم، بگویید که با توجه به داده‌های ورودی ربات، چه کاری انجام دهیم تا ربات قادر باشد سیاستی بهینه‌ای را پیدا کند که هدف ما را نیز براورده کند.

ب) درصورتی که سناریو سوم را با استفاده از روش های dynamic programming حل کنید، پاسخ بهینه‌ای که یافت می‌شود مربوط به چه MDP است؟ حالت‌های این MDP را با $s_i \; i = 0,1, \dots$ نمایش دهید و مشخص کنید که خانه‌های شطرنجی مربوط به کدام یک از این حالت‌ها است و عمل‌ها را با $a_i \; i = 0,1,2, \dots$ نمایش دهید و مشخص کنید هر عمل مربوط به کدام عمل ربات است. ماتریس احتمال P را مشخص کنید.

قسمت پیاده سازی:

الف) MDP مشخص شده در قسمت سناریو سوم ب را در محیط پایتون پیاده سازی کنید و سیاست بهینه را با استفاده از الگوریتم‌های گفته شده در اول تمرین بدست آورید. سیاست بهینه بدست آمده را اگر برروی ربات قرار دهیم، در هر قسمت از محیط شطرنجی چه تصمیمی خواهد گرفت؟

ب) سناریو سوم را در حالتی در نظر بگیرید که ربات از مکان قرارگیری خود برروی نقشه مطلع است و دوباره با استفاده از الگوریتم‌ها سیاست بهینه را بدست بیاورید.

سوال چهارم

(برگرفته شده از تمرین بهار ۲۰۲۴ استنفورد) در این سوال به بررسی تاثیر طول اپیزود (Horizon) برروی سیاست عامل می‌پردازیم. یک ربات را درنظر بگیرید که وظیفه مدیریت سهام را برعهده دارد. (فرض کنید این مسئله به صورت MDP قابل نمایش است.)

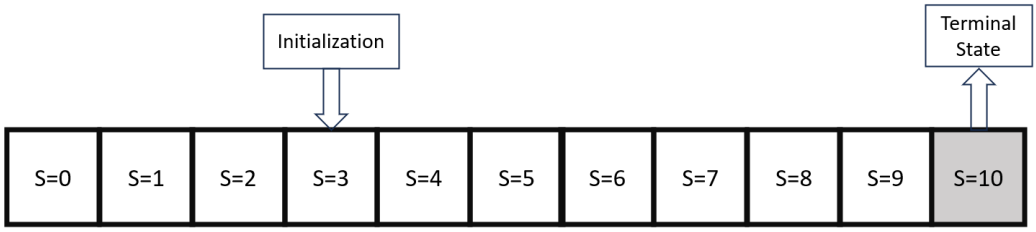
s به تعداد سهامی گفته می‌شود که در حال حاضر ربات آن را دارد (که همواره عددی صحیح بین [0,10] است) در هر لحظه، ربات دو انتخاب دارد: بفروشد (در صورت امکان s به میزان یک واحد کم می‌شود) و یا بخرد (در صورت امکان s به میزان یک واحد زیاد می‌شود).

- اگر $s>0$ باشد و عامل فروش انجام دهد، برای فروش پاداش $+1$ دریافت می‌کند و سطح سهام به $s-1$ تغییر می‌یابد. اگر $s=0$ باشد، هیچ اتفاقی نمی‌افتد.
- اگر $s<9$ باشد و عامل خرید انجام دهد، پاداشی دریافت نمی‌کند و سطح سهام به $s+1$ تغییر می‌یابد.
- مالک سهام دوست دارد در پایان روز موجودی به طور کامل تامین شود، بنابراین اگر سطح سهام به حداکثر مقدار $s=10$ برسد، عامل پاداش $+100$ دریافت می‌کند.
- حالت $s=10$ همچنین یک وضعیت نهایی است و اگر به آن برسد، مسئله پایان می‌یابد.

تابع پاداش که به صورت $r(s, a, s')$ نمایش داده می‌شود، به طور خلاصه به شکل زیر است:

- $r(s,sell,s-1)=1$ for $s>0$
- $r(0,sell,0)=0$
- $r(s,buy,s-1)=0$ for $s<9$
- $r(9,buy,10)=100$ این شرط نشان می‌دهد که تغییر از $s=9$ به $s=10$ ریوارد $+100$ می‌دهد و سطح سهام به حداکثر مقدار خود می‌رسد.

فرض می‌شود که سطح سهام همیشه در ابتدای روز از $s=3$ شروع می‌شود. ما بررسی خواهیم کرد که چگونه سیاست بهینه عامل با تنظیم افق محدود H مسئله تغییر می‌کند. به یاد داشته باشید که افق H به محدودیتی در تعداد گام‌های زمانی اشاره دارد که عامل می‌تواند قبل از پایان اپیزود با MDP تعامل داشته باشد، صرف نظر از اینکه به یک حالت نهایی رسیده باشد یا خیر. ما به بررسی ویژگی‌های سیاست بهینه (سیاستی که بیشترین پاداش اپیزود را کسب می‌کند) با تغییر افق H خواهیم پرداخت. (در حالت زمانی محدود discount factor=1 است)



شکل 5- تصویر مسئله

به عنوان مثال، فرض کنید H=4 باشد. عامل می‌تواند برای سه گام فروش انجام دهد و از s=3 به s=2، سپس s=1 و در نهایت s=0 انتقال یابد و برای هر عمل فروش پاداش‌های 1+، 1+ و 1+ دریافت کند. در گام چهارم، موجودی خالی است، بنابراین می‌تواند فروش یا خرید انجام دهد که در هر صورت هیچ پاداشی دریافت نمی‌کند. سپس مسئله به دلیل اتمام زمان خاتمه می‌یابد.

الف) آیا با شروع از حالت اولیه s=3 می‌توان مقداری از H انتخاب کرد که در نتیجه آن، سیاست بهینه هم گام‌های خرید و هم گام‌های فروش را در طول اجرا انجام دهد؟ پاسخ خود را توضیح دهید.

ب) با شروع از حالت اولیه s=3، برای چه مقادیری از H، سیاست بهینه به موجودی کاملاً پر می‌رسد؟ به عبارت دیگر، یک بازه برای H ارائه دهید.

نکته ۱: ما زمانی موجودی را کاملاً پر در نظر می‌گیریم که عمل خرید در حالت s=9 انتخاب شود و باعث انتقال به s=10 گردد. این شامل آخرین گام زمانی در افق نیز می‌شود.

نکته ۲: با انجام فقط عمل‌های خرید، عامل می‌تواند از s=3 به s=10 در H=7 گام برسد.

ج) اکنون به تنظیمات افق نامتناهی با وجود λ توجه کنید. به عبارت دیگر، هیچ محدودیت زمانی وجود ندارد و مسئله تنها زمانی خاتمه می‌یابد که به یک حالت نهایی برسیم. فرض کنید $\lambda = 0$ ، آنگاه سیاست بهینه در هنگام s=3 چه عملی انجام می‌دهد؟ سیاست بهینه در هنگام s=9چه عملی انجام می‌دهد؟

د) آیا در تنظیمات افق نامتناهی با وجود λ ، امکان انتخاب یک مقدار ثابت $\lambda \in [0,1)$ وجود دارد به طوری که سیاست بهینه با شروع از s=3 هرگز موجودی را به طور کامل پر نکند؟ در صورت وجود، بازه‌ای از λ که این شرط را براورده می‌کند پیدا کنید.

قسمت پیاده سازی:

مسئله را به صورت finite horizon و infinite horizon پیاده سازی کنید.

الف) با توجه به پاسخ خود، در سوال الف قسمت بالا، مقدارهایی از H انتخاب کنید که در آن پس از به دست آوردن سیاست بهینه با استفاده از الگوریتم های VI و PI ، سیاست بهینه ویژگی های زیر را داشته باشند:

- سیاست بهینه تنها خرید کند.
- سیاست بهینه تنها فروش کند.
- سیاست بهینه هم خرید و هم فروش انجام دهد.

ب) در حالت تنظیمات افق نامتناهی و به ازای $\lambda = 0$ سیاست بهینه را بدست بیاورید و با جواب خود در سوال ج قسمت بالا، مقایسه کنید.

ج) در حالت تنظیمات افق نامتناهی و به ازای $\lambda = 0.1$ و $\lambda = 0.9$ سیاست بهینه را جداگانه بدست بیاورید و با یکدیگر مقایسه کنید.

نکات تمرین

- استفاده از LLM ها در این تمرین مشکلی ندارد. اما در صورت استفاده لطفاً منبع و prompt خود را ذکر نمایید تا تقلب محسوب نشود.
- مهلت ارسال این تمرین تا پایان روز جمعه ۲۵ آبان ماه خواهد بود.
- انجام این تمرین به‌صورت یک‌نفره است. اما بحث و گفت‌وگو در پیامرسان درس مانعی ندارد.
- لطفاً گزارش و کد تمرین را در قالب یک فایل zip در سامانه ایلرن بارگذاری کنید.
- در صورت وجود سؤال و یا ابهام می‌توانید از طریق پیام‌رسان درس با دستیار آموزشی مصطفی حمیدی فرد در ارتباط باشید.

برای قسمت‌های پیاده سازی به موارد زیر توجه کنید:

ابتدا مسئله را به صورتی که گفته شده است در محیط پایتون پیاده سازی کنید. موارد زیر باید رعایت گردد:

- برای پیاده سازی MDP شما نیاز به توابع p(s_next,r|s,a)، r(s,a,s_next) برای گرفتن احتمال انتقال از یک حالت به حالت دیگر و محاسبه ریوارد دارید.
- توجه کنید که برای مسائل finite horizon باید راه حلی پیدا کنید تا بتوانید آن را با استفاده از روش‌های dynamic programming حل کنید.
- الگوریتم های policy iteration و value iteration باید به صورت توابعی نوشته شوند که اطلاعات MDP را گرفته و سیاست بهینه را خروجی دهد.
- برای هریک از قسمت های زیر در صورت نیاز به محاسبه سیاست بهینه، باید هر دو الگوریتم value iteration و policy iteration را برای آن اجرا کنید. تفاوت بین پاسخ این دو الگوریتم و میزان زمانی که طول می کشند تا به سیاست بهینه برسند را با یکدیگر مقایسه کنید.
- سیاست های بهینه را در هرقسمت به صورتی مناسب برروی یک شکل یا جدول نمایش دهید و در گزارش بیاورید.
- برای کدی که نوشته اید به صورت مناسب کامنت بگذارید در صورتی که کد خوانایی نداشته باشد **20 درصد نمره کسر** خواهد شد.