



به نام خدا  
دانشگاه تهران  
پردیس دانشکده‌های فنی  
دانشکده برق و کامپیوتر

درس سیستم‌های هوشمند

## تمرین شماره 6

دی ماه 1401

## فهرست سوالات

- سوال 1: یادگیری تقویتی مبتنی بر مدل (تحلیلی) : ..... 3
- سوال 2: یادگیری تقویتی مبتنی بر مدل (پیاده سازی) : ..... 4
- سوال 3: یادگیری تقویتی غیرمبتنی بر مدل (پیاده سازی) : ..... 5
- الف) حل محیط بازی بدون استفاده از روش Q-Learning و مبتنی بر پیمایش رندوم ..... 6
- ب) حل محیط بازی با استفاده از روش Q-Learning و مبتنی بر پیمایش هوشمندانه ..... 6
- نکات تحویل: ..... 8

## سوال 1: یادگیری تقویتی مبتنی بر مدل (تحلیلی):

در این سوال قصد داریم با الگوریتم تصمیم مبتنی بر تکرار<sup>1</sup> از روش های یادگیری تقویتی مبتنی بر مدل<sup>2</sup> در قالب یک سوال تحلیلی آشنا شویم و منطق الگوریتم را بصورت دقیق تر بررسی نماییم.

سه مرحله از الگوریتم *Policy-Iteration* را روی جدول زیر اجرا کنید.

فرض نمایید به دلیل وجود نامعینی، به احتمال  $0.6a$  به سمت جهت دلخواه و با احتمال  $0.2b$  به دو جهت مجاور خواهیم رفت. مقدار  $\alpha$  را در الگوریتم  $0.2$  فرض نمایید و همچنین اینکه عامل<sup>3</sup> برای حرکت کردن مجازات نخواهد شد.

|  |  |  |    |
|--|--|--|----|
|  |  |  | 3  |
|  |  |  | -2 |
|  |  |  |    |

شکل 1: جدول سوال یک

\* توجه نمایید که  $a$  رقم یکان و  $b$  رقم دهگان شماره دانشجویی شما است.

\* به عنوان نمونه اگر شماره دانشجویی شما به 54 ختم میشود عامل با احتمال  $0.64$  به سمت جهت دلخواه و با احتمال  $0.25$  به دو جهت مجاور خواهد رفت.

---

<sup>1</sup> Policy Iteration

<sup>2</sup> Model-Based Reinforcement Learning

<sup>3</sup> Agent

## سوال 2: یادگیری تقویتی مبتنی بر مدل (پیاده سازی) :

در این سوال قصد داریم که در قالب حل مساله زیر با الگوریتم *Value-Iteration* آشنا شویم.

تعریف مساله:

فردی تصمیم دارد که در یک فستیوال بخت آزمایی شرکت نماید. شرکت در این بخت آزمایی باین گونه است که وی میبایست بر روی توالی پرتاب یک سکه بگونه ای شرط بندی کند که اگر سکه شیر بیاید به همان اندازه ای که روی آن سمت سکه شرط بسته هست پول برنده شود و در صورتی که سکه خط بیاید به همان اندازه پول از دست بدهد. این روند مادامی ادامه می یابد که یا او به هدف خود که برنده شدن  $100\$$  است دست پیدا کند یا اینکه تمام پول خود را ببازد.

در مدل سازی مساله بالا لطفا به نکات زیر توجه نمایید:

- در این مساله احتمال شیر آمدن سکه را با  $p_h$  نمایش خواهیم داد.
- شرکت کننده در هر مرحله پرتاب تصمیم میگیرد که چه مقدار از سرمایه خود را شرط ببندد و این مقدار پول، یک مقدار صحیح است.
- در مدل سازی مساله، پاداش برای هر گذاری که منجر به پیروزی شود برابر  $+1$  خواهد بود و در بقیه موارد برابر  $0$  خواهد بود.
- از آنجا که در این مسئله، با یک زنجیره مارکوفی بدون تخفیف<sup>1</sup> روبه رو هستیم، مقدار گاما  $1$  لحاظ خواهد شد.

عاملی طراحی نمایید که با الگوریتم *Value-Iteration*، سیاست بهینه ای را برای شیوه شرط بندی شرکت کننده در هر مرحله ارائه دهد. (لطفا این پیاده سازی و ترسیم نمودارها را برای دو مقدار  $p_h = 0.25$  و  $p_h = 0.55$  انجام دهید.) همچنین اینکه دو منحنی سیاست بهینه برحسب سرمایه و همچنین منحنی مقدار ارزش های گذار<sup>2</sup> برحسب سرمایه را نیز بصورت گرافیکی ترسیم نمایید.

\* برای شهود بهتر در تفسیر نمودارها توصیه میشود منحنی سیاست بهینه برحسب سرمایه را با دستور `plt.bar()` ترسیم نمایید. (سعی نمایید که الگوی سیاست بهینه بدست آمده را بازای دو مقدار  $p_h$  مختصرا تفسیر نمایید.)

\* همچنین لطفا توجه نمایید که درک مساله و مدل سازی درست فضای حالت<sup>3</sup> و فضای اقدام<sup>4</sup> در این مساله بسیار حائز اهمیت میباشد، بنابراین در گزارشکار خود بطور شفاف و دقیق جزییات مدل سازی و پیاده سازی الگوریتم را توضیح دهید.

---

<sup>1</sup> Undiscounted MDF

<sup>2</sup> Value Estimates

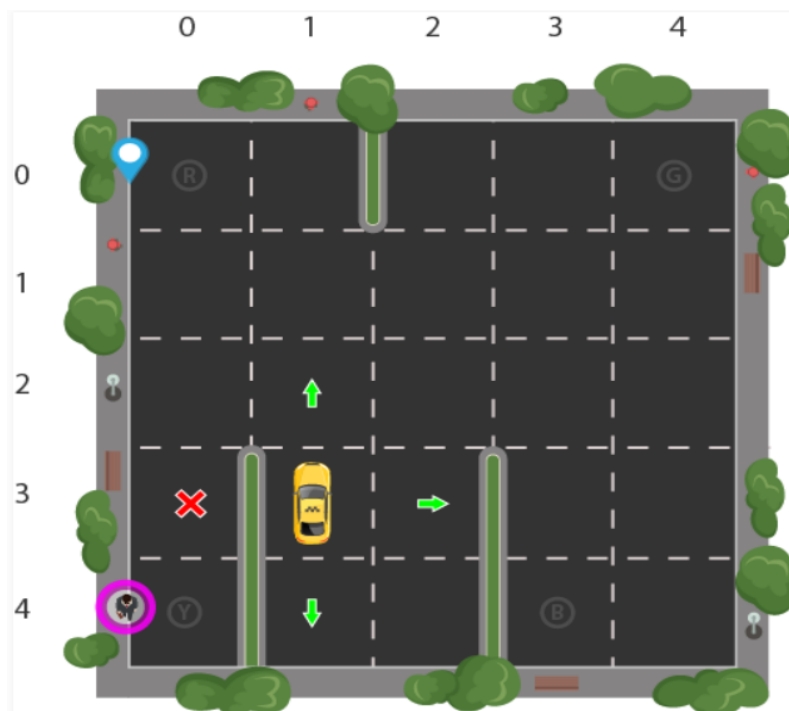
<sup>3</sup> State Space

<sup>4</sup> Action Space

### سوال 3: یادگیری تقویتی غیرمبتنی بر مدل (پیاده سازی) :

در این قسمت قصد داریم در قالب محیط یک بازی به بررسی الگوریتم های یادگیری تعاملی پردازیم و بررسی کنیم با حرکت هوشمندانه در محیط نسبت به حرکت رندوم تا چه اندازه میتوانیم بصورت بهینه تر به مقصد که در این مثال رساندن مسافر به مقصد مورد نظر میباشد دست یابیم.

محیط بازی مورد نظر بصورت زیر میباشد . هدف در این محیط این است که تاکسی زرد مسافر را از یک لوکیشن سوار کند و در یک لوکیشن دیگر پیاده نماید.



شکل 2: محیط بازی

تاکسی زرد در حرکت خود تابع یکسری قوانین است که میبایست رعایت شوند: (1) مورد اول اینکه راننده درمورد وقت مسافران اهمیت بخرج دهد و در کمترین زمان ممکن آنها را در مقصد مورد نظر پیاده کند. (2) مسافران را در مقصد درست پیاده نماید.

راننده تاکسی زرد به بازی هر مسافری که آن را با موفقیت به مقصد مورد نظرش برساند یک پاداش مثبت برابر 20+ دریافت خواهد کرد و در صورتی که مسافر را به مقصد اشتباه برساند جریمه خواهد شد و پاداشی منفی برابر 10- خواهد گرفت و در نهایت نیز از آنجایی که سوخت تاکسی محدود است و طولانی شدن مسیر منجر به تمام شدن بنزین خواهد شد لذا هر حرکت ای که منجر به رسیدن به مقصد نشود نیز راننده را متحمل جریمه خواهد کرد و در طی آن، عامل پاداشی منفی برابر 1- خواهد گرفت. همچنین اینکه اگر دقت کنید راننده در مسیر خود دیوارهایی را میبیند که در صورتی که به آنها برخورد کند جریمه خواهد شد و پاداشی منفی برابر 1- دریافت خواهد کرد و در مکان خود باقی خواهد ماند.

همچنین لطفا توجه نمایید که راننده تاکسی در مجموع شش حرکت {حرکت به بالا، حرکت به چپ، حرکت به راست، حرکت به پایین، سوار کردن مسافر و پیاده نمودن مسافر} را میتواند انجام دهد.

برای استفاده از محیط پیاده سازی شده، کافی است فایل قرار داده شده در پوشه تمرین را به کدهای خود اضافه نمایید. به علاوه فایلی به عنوان راهنما هم قرار داده شده که نشان میدهد چگونه میتوانید به فضای حالت و عمل دسترسی بیابید.

در طول کار با کتابخانه OpenAi Gym برای این محیط بازی به نکات زیر توجه نمایید:

- R,G,Y,B چهار لوکیشنی هستند که هم مسافران در آنجا سوار میشوند و هم میتوانند مقصد نهایی مسافران باشند.
- در این محیط بصورت پیشفرض لوکیشن با رنگ آبی نشان دهنده مکان مسافر فعلی و لوکیشن با رنگ بنفش نشان دهنده مقصد مسافر فعلی است.
- همچنین اینکه مادامی که تاکسی خالی از مسافر باشد با رنگ زرد در نقشه به نمایش درخواهد آمد و مادامی که حاوی مسافر باشد با رنگ سبز به نمایش درخواهد آمد.

## الف) حل محیط بازی بدون استفاده از روش Q-Learning و مبتنی بر پیمایش رندوم

در ابتدا قصد داریم تا با استفاده از یک loop بینهایت کاملا بصورت رندوم در محیط گام برداریم تا منجر به رسیدن یک مسافر به مقصد درست شود. این پیاده سازی را بر مبنای پیمایش کاملا رندوم انجام دهید و تعداد گام ها و مقدار کل جریمه را در طول این پیاده سازی گزارش کنید.

در پیاده سازی این قسمت برای گام نهادن در محیط میتوانید از دستور `env.action_space.sample()` استفاده نمایید.

\* در طول این پیاده سازی برای نمایش گرافیکی حرکت تاکسی میبایست در طول loop اطلاعات هر بار `render` شدن محیط را در یک دیکشنری ذخیره کنید تا بتوانید پس از پایان حلقه بصورت گرافیکی حرکت تاکسی را به تصویر بکشید.

## ب) حل محیط بازی با استفاده از روش Q-Learning و مبتنی بر پیمایش هوشمندانه

برای محیط بیان شده، الگوریتم یادگیری بر اساس معیار Q را شبیه سازی کنید. (نتایج را با حالتی که از الگوریتم Q استفاده نمیکردیم مقایسه نمایید).

\* پارامترهایی مانند میانگین پاداش در هر جابه جایی<sup>۱</sup>، میانگین تعداد گام ها در هر سفر<sup>۲</sup> و میانگین کل مقدار جریمه ها در طول هر اپیزود<sup>۳</sup> میتواند در مقایسه نتایج مفید باشد.

با استفاده از پیاده سازی بالا باین سوال پاسخ دهید. دقت کنید که برای پاسخ خود دلیل ارائه کنید. آیا میتوانید با تغییر دادن پاداشها به مقادیر معقول به همگرایی سریعتری برسید؟ منطق خود را برای این تغییرات ذکر کرده و آن را شبیه سازی کنید.

### نکات سوال سوم:

- برای کدهایی که ضمیمه میکنید، حتما گزارشکار دقیق با ذکر جزئیات پیاده سازی بنویسید. ( این گزارش ها معیار تفاوت کد شما با مدل های مشابه موجود در اینترنت خواهد بود. )
- از پاسخ های روشن در گزارشکار خود استفاده نمایید و تمام فرضیات خود را به طور شفاف بیان نمایید.
- همچنین توجه نمایید که ایده های خلاقانه در *tune* کردن هایپر پارامترها و بکارگیری بهینه تر الگوریتم -Q Learning در این سوال میتواند نمره امتیازی در بر داشته باشد و یا اینکه کاستی نمرات شما در این بخش را جبران نماید.
- بطور ویژه در این تمرین درک و مدلسازی درست مسئله ها اهمیت بیشتری نسبت به پیاده سازی صرف خواهد داشت که در پی آن گزارشکار دقیقتری را نیز میطلبید پس لطفا سعی کنید در گزارشکار خود بصورت شفاف جزئیات مدلسازی و پیاده سازی خود را ذکر نمایید.

---

<sup>1</sup> Average Rewards per move

<sup>2</sup> Average number of timesteps per trip

<sup>3</sup> Average number of penalties per episode

## نکات تحویل:

- مهلت تحویل این تمرین **10 بهمن** میباشد.
- انجام این تمرین به صورت یک نفره است.
- برای انجام این تمرین تنها مجاز به استفاده از زبان برنامه نویسی پایتون هستید.
- در صورت وجود تقلب نمره تمامی افراد شرکت کننده در آن **-100** لحاظ میشود.
- لطفا پاسخ تمرین خود را (به همراه کد/گزارش سوال کامپیوتری) به صورت زیر در صفحه درس آپلود نمایید:

HW [HW number] \_ [Last name] \_ [Student number].zip

- در صورت وجود هر گونه ابهام یا مشکل لطفاً به مسئولان حل تمرین ایمیل بزنید.

Question #1 : [Erfan Hajhashemi](#) + [Oveys Delafrooz](#)

Question #2 : [Oveys Delafrooz](#) + [Mohammad Heydari](#)

Question #3 : [Mohammad Heydari](#) + [Erfan Hajhashemi](#)