**بسم الله الرحمن الرحیم**

دانشگاه علم و صنعت ایران

تابستان ۱۴۰۰

یادگیری عمیق | تمرین سری دوازدهم | تحویل: چهارشنبه ۲ تیر

---

**۱.** در این سوال با مطالعه مقاله زیر، با یک معماری شبکه عصبی جدید برای یادگیری تقویتی model-free آشنا می‌شوید. معماری Dueling بدون تخمین تأثیر عمل عامل، برای هر حالت تعیین می‌کند که کدام یک از حالات ارزشمند هستند. این معماری زمانی مناسب است که یک سری اعمال تأثیر مرتبطی بر روی محیط ندارند. در شکل دوم مقاله این امر را می‌توان مشاهده نمود.

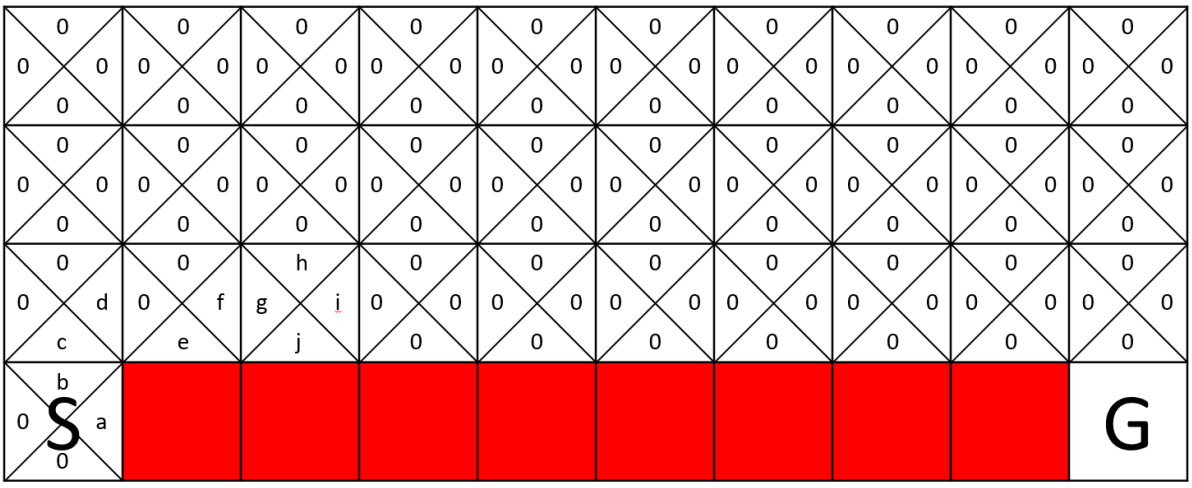
با مطالعه بخش سوم مقاله، جزییات این معماری، تابع Advantage و تابع value را شرح دهید. همچنین ذکر کنید که چرا از رابطه ۹ در مقاله استفاده شده است.

Wang, Ziyu, et al. "Dueling network architectures for deep reinforcement learning." *International conference on machine learning*. PMLR, 2016.

**۲.** فرض کنید در مسئله زیر قرار است عامل هوشمند یاد بگیرد که نقطه S به نقطه G برود (مقادیر نوشته شده مربوط به ارزش هر state و action هستند). سیستم پاداش مسئله به این صورت است که اگر به G برسد پاداش 100+ دریافت می‌کند و اگر در هر کدام از خانه‌های قرمز قرار بگیرد پاداش 100- دریافت کرده و در هر دو صورت یک episode به پایان می‌رسد. از آنجائیکه رسیدن سریعتر به G مهم است، هر گام دارای پاداش 1- است.

حال فرض کنید در یک episode به ترتیب عمل‌های زیر انجام شده است: بالا، راست، راست، پائین.

اگر مقادیر مشخص شده در جدول برابر باشند با: a=-20, b=+2, c=-1, d=+4, e=-30, f=+6, g=+2, h=+1, i=+10, j=-40، مقدار به روز شده این مقادیر در انتهای این episode را برای دو روش SARSA و Q-Learning محاسبه کنید.



**۳.** الگوریتم Monte Carlo دارای دو نسخه متفاوت every-visit و first-visit برای به روز رسانی stateهایی است که بیش از ۱ بار در یک episode مشاهده می‌شوند. از این لینک برای مطالعه تفاوت‌های این دو نسخه استفاده کرده و به سوال زیر پاسخ دهید.

عاملی را در فضای زیر در نظر بگیرید که از نقطه شروع به سمت نقطه هدف حرکت می‌کند. ارزش اولیه تمام حالت‌ها برابر صفر است ($V(s) = 0 \; for \; all \; s \in S$). در اپیزود اول عامل حالات مقابل را به ترتیب ملاقات می کند: ۱, ۲, ۳, ۸, ۷, ۱۲, ۱۶, ۲۰, ۲۱, ۲۲, ۲۱, ۲۲, ۱۷, ۱۸, ۲۳. اگر عامل از نسخه first -visit الگوریتم Monte Carlo برای تخمین تابع state-value استفاده کند، تابع ارزش هر حالت بعد از پایان اپیزود اول را محاسبه کنید (مقدار ارزش هر حالت را در جدولی مشابه وارد کنید). (خانه های سیاه شده به این معناست که عامل نمیتواند در آن‌ها حضور داشته باشد و مقادیر پاداش همه خانه‌ها به جز خانه‌های مشخص شده برابر صفر است. همچنین مقدار پارامتر های مسئله را به این شکل در نظر بگیرید. $\alpha = 0.5, \; \gamma = 1$) مسئله را برای زمانی که عامل از نسخه every-visit استفاده کند، مجددا حل نمایید.

| | | | | |
|---|---|---|---|---|
| Start<br>s=1 | s=2 | s=3 | s=4 | s =5 |
| s=6 | s=7 | s=8 | s=9 | s=10 |
| s=11 | s=12 | | s=13 | s=14 |
| s=15 | s=16 | | s=17 | s=18 |
| s=19 | s=20 | R=-5<br>s=21 | s=22 | R=+10<br>s=23<br>Goal |

**۴.** هدف از این سوال، پیاده سازی مفاهیمی است که در رابطه با مدل Deep Q-Learning در کلاس درس فرا گرفته اید. برای این منظور، لازم است که ابتدا با محیط gym آشنا شوید. gym یک ابزار برای توسعه و مقایسه الگوریتم‌های یادگیری تقویتی است. این ابزار شامل تعدادی محیط‌های شبیه‌سازی شده برای یادگیری و ارزیابی عامل است. برای مطالعه و آشنایی بیشتر با این ابزار به این لینک مراجعه نمایید. در این سوال قصد داریم با پیاده سازی الگوریتم Deep Q-Learning یک عامل را در محیط SpaceInvaders-v0 آموزش دهیم و عملکرد آن را با رسم نمودار میانگین پاداش‌های دریافتی ارزیابی نماییم. Discount factor و learning rate را به انتخاب خود با ذکر دلیل تعیین کنید. همچنین برای epsilon نیز از epsilon decay استفاده نمایید. در این سوال نیاز است که برای حذف همبستگی میان تجربیات، replay memory را نیز پیاده‌سازی نمایید. برای پیاده‌سازی از keras استفاده نمایید. لطفا کدهای خود را در فرمت ipynb با خروجی ارسال نمایید.

## نکات تکمیلی

۱) لطفاً پاسخ سوالات (تئوری و توضیحات پیاده‌سازی) را به طور گویا و به زبان فارسی و در صورت امکان تایپ همراه با سورس کدهای نوشته شده، در یک فایل فشرده شده به شکل HW12_YourStudentID.zip قرار داده و بارگذاری نمایید.

۲) منابع استفاده شده را به طور دقیق ذکر کنید.

۳) برای سهولت در پیاده‌سازی‌ها و منابع بیشتر، زبان پایتون پیشنهاد می‌شود. لطفا کدهای مربوطه را به طور جداگانه در فرمت py. یا ipynb. ارسال نمایید.

۴) ارزیابی تمرین‌ها براساس صحیح بودن راه حل‌ها، گزارش مناسب، بهینه بودن کدها و کپی نبودن می‌باشد.

۵) در مجموع تمام تمرین‌ها، تنها ۷۲ ساعت تاخیر در ارسال پاسخ‌ها مجاز است اما پس از آن به صورت خطی از نمره کسر خواهد شد (معادل با روزی ۵۰ درصد).

۶) در رابطه با پرسش و پاسخ در رابطه با تمرین‌ها می‌توانید در گروه مربوطه مطرح کنید.

موفق باشید.