مدرسین: **دکتر فدایی و**
دکتر یعقوب‌زاده

بخش کتبی

MDP

سوال اول

یک جدول داریم که هر خانه در آن با شماره ردیف و ستون شناخته می‌شوند (ابتدا ردیف). Agent همیشه از حالت (1,1) که با حرف S مشخص شده شروع می‌کند. دو حالت هدف نهایی وجود دارد، (2,3) با پاداش 5+ و (1,3) با پاداش 5-. پاداش‌ها در حالت‌های غیر نهایی صفر می‌باشد. تابع Transition به گونه‌ای است که حرکت مورد نظر Agent (شمال، جنوب، غرب یا شرق) با احتمال 0.8 اتفاق می‌افتد. با احتمال 0.1 به هر یک از حالت‌های عمود بر جهت مورد نظر می‌رسد. اگر برخوردی با دیوار رخ دهد، Agent در همان حالت باقی می‌ماند.

		+5
S		-5

1) نتایج دو دور اول Value Iteration را با مقدار تخفیف[1] 0.9 محاسبه کنید. توجه کنید که خانه‌های (3, 1) و (3, 2) دارای Value ثابت می‌باشند. راهنمایی:

$$V_{i+1}(s) = \max_a \left(\sum_{s'} T(s,a,s')(R(s,a,s') + \gamma V_i(s')) \right)$$

[1] Discount Factor

S	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)
V_0	0	0	-5	0	0	5
V_1	0	0	-5	0	4	5
V_2	0	2.38	-5	2.88	4	5

2) Policy را با توجه به جدول ارزش‌های بالا محاسبه کنید (برای خانه‌هایی که دو یا چند Action با مقدار برابر وجود دارد می‌توانید هر کدام را به دلخواه انتخاب کنید).

S	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)
π*(S)	بالا	بالا	-	راست	راست	-

3) حال فرض کنید که تابع Transition را ندارید، اکنون برای اینکه بتوانید Policy بهینه را بدست آورید باید از روش‌هایی مانند Q-Learning و یا Monte Carlo استفاده کنید. Monte Carlo یک روش Model-Free می‌باشد، درباره این روش جستجو کنید و با کمک آن به سوال زیر پاسخ دهید. حال فرض کنید agent ما Policy-ای که همیشه به سمت راست برود را انتخاب می‌کند و سه آزمایش زیر را اجرا می‌کند، تخمین‌های Monte Carlo (کاربرد مستقیم) برای خانه‌های (1,1) و (2,2) با توجه به این مسیرها چیست؟

I) (1,1)–(1,2)–(1,3)
II) (1,1)–(1,2)–(2,2)–(2,3)
III) (1,1)–(2,1)–(2,2)–(2,3)

برای محاسبه تخمین‌ها، میانگین پاداش‌های دریافتی در مسیرهایی که از حالت‌های مشخص شده گذر کرده‌اند را می‌گیریم.

$V(1,1) = (-5 + 5 + 5)/3 = 5/3 = 1.666$
$V(2,2) = (5 + 5)/2 = 5$

4) اگر فرض کنیم Agent بر اساس TD-Learning یاد می‌گیرد، با استفاده از نرخ یادگیری 0.1 و با فرض مقادیر اولیه صفر (به جز خانه‌های نهایی)، بعد از 2 مرحله Iteration، برای هر خانه چه Value-ای داریم؟ (امتیازی)

$$V(s) = V(s) + \alpha(r + \gamma V(s') - V(s))$$

به روزرسانی ها بعد از آزمایش اول:

$$V((1, 1)) = 0 + 0.1(0 + 0.9 \times 0 - 0) = 0$$
$$V((1, 2)) = 0 + 0.1(-5 + 0.9 \times 0 - 0) = -0.5$$

به روزرسانی ها بعد از آزمایش دوم:

$$V((1, 1)) = 0 + 0.1(0 + 0.9 \times -0.5 - 0) = -0.045$$
$$V((1, 2)) = -0.5 + 0.1(0 + 0.9 \times 0 + 0.5) = -0.45$$
$$V((2, 2)) = 0 + 0.1(5 + 0.9 \times 0 - 0) = 0.5$$

به روزرسانی ها بعد از آزمایش سوم:

$$V((1, 1)) = -0.045 + 0.1(0 + 0.9 \times 0 + 0.045) = -0.0405$$
$$V((2, 1)) = 0 + 0.1(0 + 0.9 \times 0.5 - 0) = 0.045$$
$$V((2, 2)) = 0.5 + 0.1(5 + 0.9 \times 0 - 0.5) = 0.95$$

DQN

سوال اول

درباره الگوریتم و کاربردهای DQN[2] تحقیق کنید و مطالبی که متوجه شدید را به طور خلاصه توضیح دهید.

https://www.sciencedirect.com/topics/computer-science/deep-q-network

[2] Deep Q-Network