• درس یادگیری تقویتی   تمرین شماره ۲

• ۸۱۰۱۰۱۲۰۴

محمدطاها حاجیلی

• سوال شماره ۱ :

• MDP — markove Decision Process

در این سوال یک جدول داریم که مختصات آن با شکل‌های درون دستور نوشته شده است.

Agent در حالت (۱,۱) قرار گرفته است که هدف آن رسیدن به خانه‌هایی با بیشترین مقدار در هدف نهایی را دارد. همچنین خانه‌های هدف نهایی مشخص شده‌اند.

حال تابع Transition را این گونه تعریف می‌کنیم که Agent باید با احتمال $\frac{8}{10}$ به سمت هدفی که می‌خواهد برود برود و با احتمال $\frac{1}{10}$ می‌تواند به دیگر جهت‌ها برود. به این صورت محیط non deterministic داریم. حال اگر به دیوار برخورد کند Agent باید در همان حالت بماند.

هدف ۱ : یافتن نتایج ۲ دور اول Value Iteration را محاسبه کرده با مقدار تخفیف $\frac{9}{10}$ می‌دانیم که مقادیر خانه‌های (۱,۳) و (۲,۳) تغییر نخواهند کرد زیرا نهایی هستند.

• حال مقادیر را این گونه قرار می‌دهیم:

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

که با به جمع مقادیر از Value Iteration استفاده کنیم.

• این که Value Iteration

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

تابع Transition ، مقدار پاداشی که می‌گیریم ، مقدار

• Value در اصل مقداری است که مقدار است که به آن ببریم ← که آیا باید برای آن اینجا برویم اگر جابجا شویم.

(۲,۱)	(۲,۲)	
(۱,۱)	(۱,۲)	

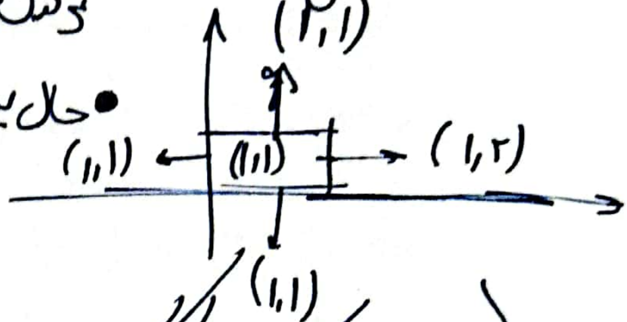
• حال برای هرکدام از استیت ها داریم.

$$V(1,2) = \max_a \left(\sum_{s'} T(s,a,s') \left(R(s,a,s') + \gamma V_i(s') \right) \right)$$

s' → استیت هایی که از استیت (1,1) می‌توان به آن حالت رفت که (2,1) و (1,2) و (1,1) جابجا شوند / نرفت

action ها را یکی یکی می‌روم

• حال برای هرکدام از action محاسبه کنیم.



• در صورت سوال گفته شده اگر به دیوار برخورد کند به خودش بازمی‌گردد.

① up ↑

$$\sum_{s'} T((1,2), up, s') \left(R(s,a,s') + \gamma V(s') \right) \qquad \gamma = \frac{9}{10}$$

Reward = 0 → • مقدار پاداش که برای جای‌گذاری به این ها صفر بود

$$= \left(\frac{8}{10}\right)\left(0 + \frac{9}{10} \cdot 0\right)_{(2,1)} + \left(\frac{1}{10}\right)\left(0 + \frac{9}{10} \cdot 0\right)_{(1,2)} + \left(\frac{1}{10}\right)\left(0 + \frac{9}{10} \cdot 0\right)_{(1,1)} = 0$$

② down ↓

$$\sum_{s'} T((1,1), down, s') \left(R((1,1), down, s') + \gamma V(s') \right)$$

• در این حالت هم همانند قبل به ازای تمام استیت هایی که می‌توان به آن رفت یعنی (1,2) و (1,1) و (1,1) هم مقدار Reward برابر با صفر است و هم مقدار نهایی خود آن ها هم صفر هستند پس لذا مجموع آن ها هم برابر با صفر است.

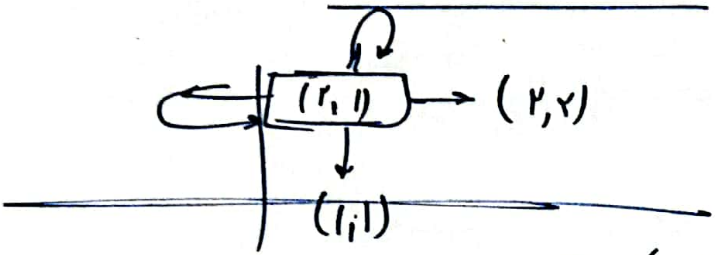
$$\longrightarrow \sum_{s'} T((1,1), down, s') \left(R((1,1), down, s') + \gamma V(s') \right) = 0$$

③ Right
④ Left

برای این دو حالت هم دقیقاً مانند حالت ۱ و ۲ مقدار برابر با خود خانه گذشته می‌شود.

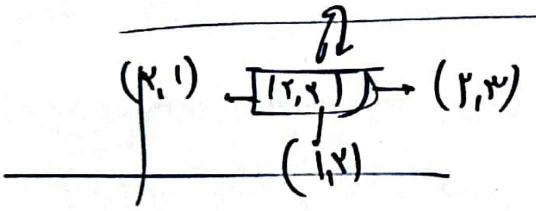
$$\Rightarrow V(1,4) = \max_a (0) = 0 \rightarrow \boxed{V(1,4) = 0}$$

• حال (۲,۱) را بررسی می‌کنیم.



در این حالت هم مانند قبل به استیت های (۲,۲) و (۱,۱) یعنی خودش یعنی (۲,۱) می‌رویم. و در این حالت هم تمام پاداش ها برابر با صفر است و هم چنین مقدار نهایی کل استیت ها برابر با صفر است.

$$\Rightarrow V(2,1) = 0$$



• حال (۲,۲) را بررسی می‌کنیم. در این حالت مقدار (۲,۳) ۵ است. حال باید تک تک محاسبه کنیم.

$$\max_a \sum_{s'} T((2,2), a, s')\left(R((2,2), a, s') + \gamma V(s')\right)$$

برای مختلف a بررسی می‌کنیم.

① a → up ↑

$$\rightarrow \sum_{s'} = \frac{8}{10}\left(0+\frac{9}{10}(0)\right) + \frac{1}{10}\left(0+\frac{9}{10}(0)\right) + \frac{1}{10}\left(0+\frac{9}{10}(5)\right) = 0.45$$

$$\sum_{s'} = \frac{8}{10}\left(5+\frac{9}{10}(5)\right) + \frac{1}{10}\left(0+0\left(\frac{9}{10}\right)\right) + \frac{1}{10}\left(0+0\left(\frac{9}{10}\right)\right) = 7.6$$

$$\sum_{s'} = \frac{8}{10}\left(0+\frac{9}{10}(0)\right) + \frac{1}{10}\left(0+0\left(\frac{9}{10}\right)\right) + \frac{1}{10}\left(0+5\left(\frac{9}{10}\right)\right) = 0.45$$

④ $\sum_{s'} = 0 \quad \rightarrow \max = \boxed{7.6}$

• این مقادیر را با این نحوه در نظر گرفته‌ام که مقدار Reward در رفتن به استیت‌های پایانی باید لحاظ شده باشد و همچنین مقادیر V_0 هم در استیت‌های آن هم باید برابر با لحاظ و شده باشد همان طور که در صورت سوال بیان شده بود.

حل برای (1,2) به این شکل:

$$\to \max_a \left[\sum_{s'} T(s,a,s')\left(R(s,a,s') + \gamma V(s')\right)\right]$$

$$① \text{ up } \to \sum\left(\frac{8}{10}\left(0+0\left(\frac{9}{10}\right)\right)\right) + \left(\frac{1}{10}\left(0+5\left(\frac{9}{10}\right)\right)\right) + \left(\frac{1}{10}\left(-5+5\left(\frac{9}{10}\right)\right)\right) = -0.95$$

$$② \downarrow \sum\left(\frac{8}{10}\left(0+0\left(\frac{9}{10}\right)\right)\right) + \left(\frac{1}{10}\left(0+\frac{9}{10}(0)\right)\right) + \left(\frac{1}{10}\left(-5+5\left(\frac{9}{10}\right)\right)\right) = -0.95$$

$$③ \to \sum\left(\frac{8}{10}\left(-5+5\left(\frac{9}{10}\right)\right)\right) + \left(\frac{1}{10}\left(0+\frac{9}{10}(0)\right)\right) + \left(\frac{1}{10}\left(0+\frac{9}{10}(0)\right)\right) = -4.4$$

$$④ \leftarrow \sum(0+0+0) = 0 \implies \max = 0$$

$$\to \max(-0.95,\ -0.95,\ 0,\ -4.4) = 0$$

پس مقادیر V_1 این گونه خواهد شد بعد از یک Iteration

0	4,4	+5
0	0	-5

• حل دور دوم را به این شکل:

ابتدا استیت (1,1) به این شکل:

$$\max_a \sum T(s,a,s')\left(R(s,a,s') + \gamma V_1(s')\right)$$

$$① \to \sum \frac{8}{10}\left(0+\frac{9}{10}(0)\right) + \frac{1}{10}\left(0+\frac{9}{10}(0)\right) + \frac{1}{10}\left(0+\frac{9}{10}(0)\right) = 0$$

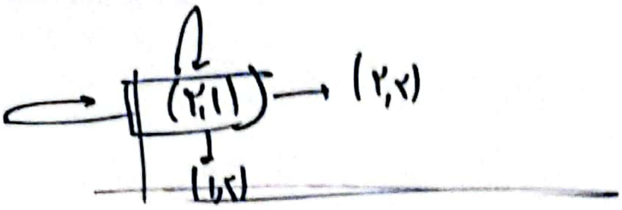
$$② \to \sum \frac{8}{10}\left(0+\frac{9}{10}(0)\right) + \frac{1}{10}\left(0+\frac{9}{10}(0)\right) + \frac{1}{10}\left(0+\frac{9}{10}(0)\right) = 0$$

$$③ \to 0$$

$$④ \to 0$$

$$\to \max(0, \dots, 0) = 0$$

• این استیت هیچ تغییری از لحاظ مقادیر نکرده $V_2(1,1) = 0$



حالت (۲,۱)

↓ ① Σ ۸/۱۰ (0+0(۹/۱۰)) + ۱/۱۰ (0+0(۹/۱۰)) + ۱/۱۰ (0+ ۷,۶(۹/۱۰)) = 0,۶۸٤

↑ ② Σ ۸/۱۰ (0+0(۹/۱۰)) + ۱/۱۰ (0+0(۹/۱۰)) + ۱/۱۰ (0+۷,۶(۹/۱۰)) = 0,۶۸٤

→ ③ Σ ۸/۱۰ (0+ ۹/۱۰(۷,۶)) + ۱/۱۰ (0+0(۹/۱۰)) + ۱/۱۰ (0+0(۹/۱۰)) = ۵,٤۷۲

← ④ Σ ۸/۱۰ (0+۹/۱۰) + 0 + 0 = 0

• → max = ۵,٤۷۲ → V(۲,۱) = ۵,٤۷۲

• حالت (۲,۲) :

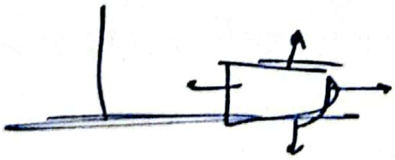
↓ ① Σ ۸/۱۰ (0+0(۹/۱۰)) + ۱/۱۰ (0+۹/۱۰ (0)) + ۱/۱۰ (0+۹/۱۰(۵)) = 0,۴۵

↑ ② Σ ۸/۱۰ (۷,۶(۹/۱۰)+0) + ۱/۱۰ (0+۹/۱۰(0)) + ۱/۱۰ (۵+۹/۱۰(۵)) = ۶,٤۲۲

→ ③ Σ ۸/۱۰ (۵(۹/۱۰)+۵) + ۱/۱۰ (0+۹/۱۰) + ۱/۱۰ (0+۹/۱۰(۷,۶)) = ۸,۲۸٤

← ④ Σ ۸/۱۰ (0+0(۹/۱۰)) + ۱/۱۰ (0+0(۹/۱۰)) + ۱/۱۰ (0+۹/۱۰(۷,۶)) = 0,۶۸٤

→ max = ۸,۲۸٤ → V۲(۲,۲) = ۸,۲۸٤



حالت (۱,۲)

↓ ① Σ ۸/۱۰ (0+0(۹/۱۰)) + ۸/۱۰ (0+0(۹/۱۰)) + (۱/۱۰)(0+۹/۱۰(−۵)) = −0,۴۵

↑ ② Σ ۸/۱۰ (0+ ۷,۶(۹/۱۰)) + ۱/۱۰ (۹/۱۰(−۵)−۵) + ۱/۱۰ (0+0(۹/۱۰)) = ٤,۵۲۲

→ ③ Σ ۸/۱۰ (−۵+۹/۱۰(−۵)) + ۱/۱۰ (0+۹/۱۰(۷,۶)) + ۱/۱۰ (0+۹/۱۰(0)) = −۶,۹۱۶

← ④ Σ ۸/۱۰ (0+0(۹/۱۰)) + ۱/۱۰ (0+۹/۱۰(0)) + ۱/۱۰ (0+۷,۶(۹/۱۰)) = 0,۶۸٤

⇒ max = ٤,۵۲۲ → V(۱,۲) = ٤,۵۲۲

• سپس مقادیر نهایی به این شکل خواهد شد: در دور دوم از Value Iteration

5,472	8,284	+5
0	4,532	-5

که البته با این فرض بود که ما فقط مقدار Reward را در رفتن به خانه‌های پایانی بگیریم و هر بار مقدار خانه‌های پایانی ثابت بوده است:

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	7,6	5
V_2	0	4,532	-5	5,472	8,284	5

حال مقدار Policy را به این صورت حساب می‌کنیم

$\pi^*(s) =$

چون که برای ما مقادیر Q ها را نداریم پس مقدار Policy را این گونه حساب می‌کنیم:

$$\pi_{i+1}(s) = \arg\max_a \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V^{\pi}_i(s')\right]$$

از روی اسلایدهای درس

$\pi_{i+1}(1,1) = \arg\max:$

$\uparrow \sum \left(\frac{8}{10}\left(0+\frac{9}{10}(5,472)\right) + \frac{1}{10}\left(0+\frac{9}{10}(0)\right) + \frac{1}{10}\left(0+\frac{9}{10}(4,532)\right)\right) \simeq 4,34$

$\rightarrow \sum \left(\frac{8}{10}\left(0+\frac{9}{10}(4,532)\right) + \frac{1}{10}\left(0+\frac{9}{10}(5,472)\right) + \frac{1}{10}\left(0+\frac{9}{10}(0)\right)\right) \simeq 3,74$

$\leftarrow \sum \left(\frac{8}{10}\left(0+\frac{9}{10}(0)\right) + \frac{1}{10}\left(0+\frac{9}{10}(0)\right) + \frac{1}{10}\left(0+\frac{9}{10}(5,472)\right)\right) \simeq 0,492$

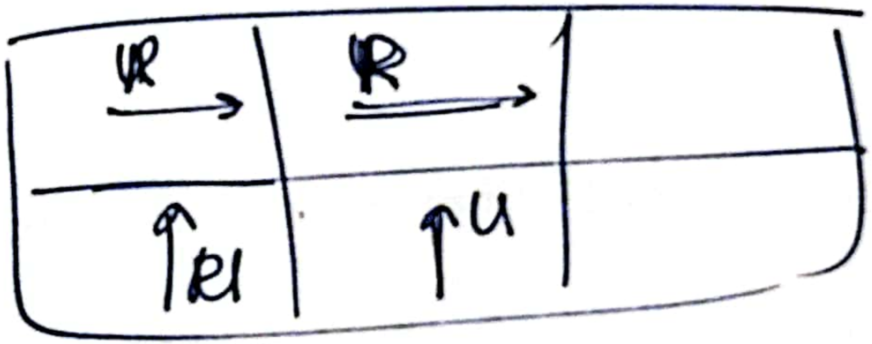
$\downarrow \sum \left(\frac{8}{10}\left(0+\frac{9}{10}(0)\right) + \frac{1}{10}\left(0+\frac{9}{10}(0)\right) + \frac{1}{10}\left(0+\frac{9}{10}(4,532)\right)\right) \simeq 0,409$

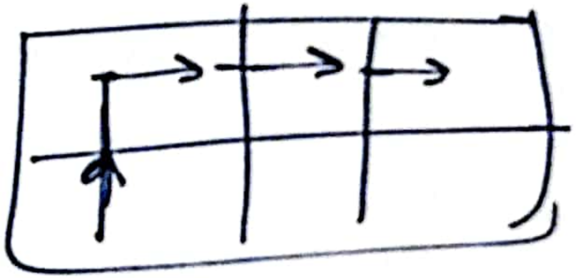
$\rightarrow$ max = ٤٫٣٤ $\rightarrow$ argmax = ↑ $\rightarrow$ up

پس باید up انتخاب شود

$$\pi_{i+1}(s) = \arg\max_{a} \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^{\pi}(s') \right]$$

که بهترین یعنی اکشن‌ها را محاسبه شود با این می‌گیریم

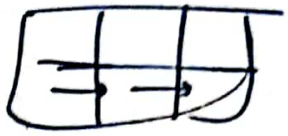




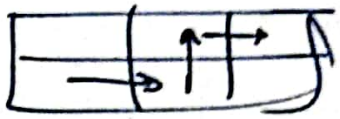
و مسیر این گونه خواهد شد

سوال ③ :

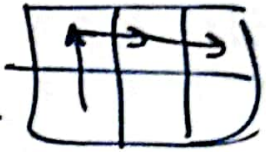
$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots$



Path ① → (1,1) → (1,2) → (1,3) → Rewards



Path ② → (1,1) → (1,2) → (2,2) → (2,3)



Path ③ → (1,1) → (2,1) → (2,2) → (2,3)

• Monte Carlo Estimations :

Path ① (1,1) → −5 (1,2) → −5

Path ② (1,1) → 5 (1,2) → 5 (2,2) → 5

Path ③ (1,1) → 5 (2,1) → 5 (2,2) → 5

: متوسط‌گیری •

$(1,1) \rightarrow \frac{-5+5+5}{3} = 1.67$

$(2,2) \rightarrow \frac{5+5}{2} = 5$

$V(1,1) = 1.67$

$V(2,2) = 5$

• سوال ④ : اگر Agent بر اساس مقادیر TD-learning یاد بگیرد با استفاده از نرخ یادگیری $\frac{1}{10}$
و با فرض مقادیر اولیه صفر (به جز خانه های نهایی که هستند) بعد از ② iteration مقادیر را
بخواهیم.

• در الگوریتم TD learning این گونه کار کنیم:

Temporal Difference Learning

$$\text{Sample} = R(s,\pi(s),s') + \gamma V^{\pi}(s')$$

• $V^{\pi}(s) = (1-\alpha)\, V^{\pi}(s) + (\alpha)\, \text{Sample}$

• به طوری که داریم میانگین وزن دار می‌گیریم. به این $\alpha \leftarrow$ Learning Rate

$$V^{\pi}(s) = (1-\alpha) V^{\pi}(s) + (\alpha)\left(R(s,\pi(s),s') + \gamma V^{\pi}(s')\right)$$

$T \Rightarrow (1,1)$

$$\rightarrow V(1,1) = \left(1-\frac{1}{10}\right)(0) + \frac{1}{10}\left(0 + \frac{9}{10}(0)\right) = 0$$

$$V(1,2) = \left(\frac{9}{10}\right) V^{\pi}(1,2) + \frac{1}{10}\left(0 + \frac{9}{10}(0)\right) = 0$$

$$V(1,1) = 0 + \frac{1}{10}\left(\frac{9}{10} + \frac{45}{100}\right) = -0{,}045$$

$$V(1,2) = -0{,}45 + 0{,}1(0{,}9 \times 0 + 0{,}45) = -0{,}405$$

$$V(2,2) = 0 + 0{,}1\left(0 + \frac{9}{10} \times 500\right) = \frac{45}{100}$$

S	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)
V_0	0	0	-5	0	0	5
V_1		-0,45	-5	0	0	5
V_2	-0,045	-0,405	-5	0	0,45	5