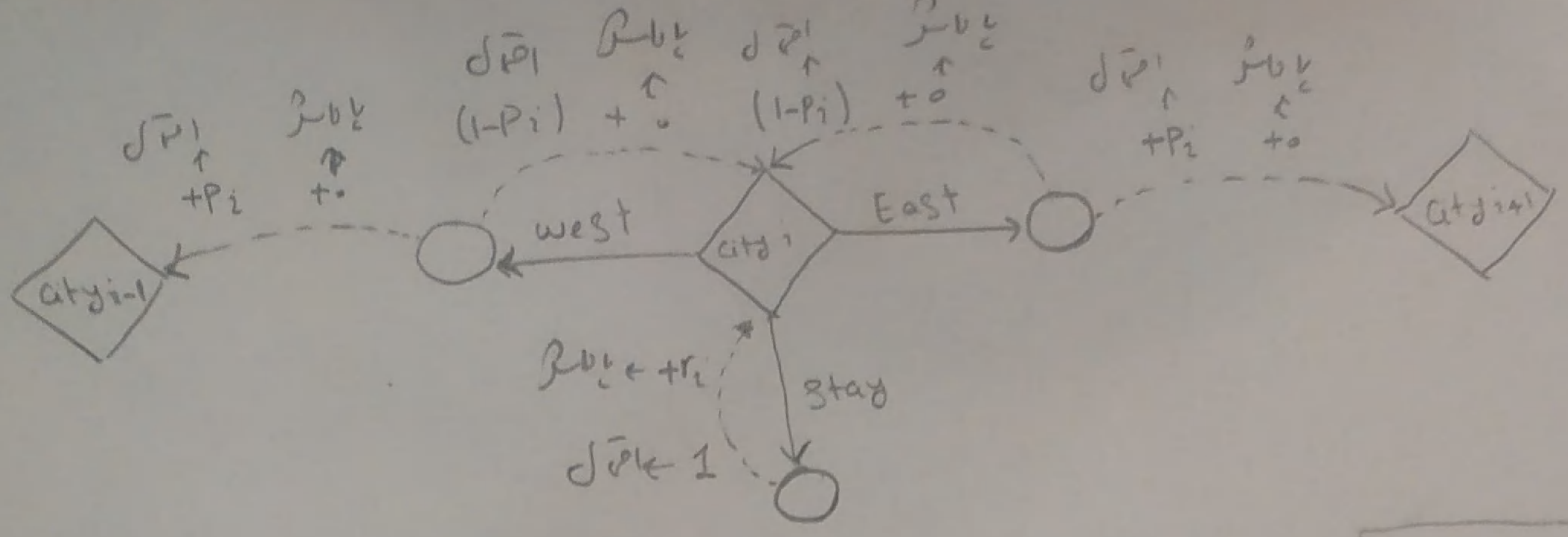
سوال ① الف)



ب) $\boxed{p_1 = 1,\ r_1 = 1,\ \gamma = 0.5}$

$$V^{stay}(1) = \sum T(s, stay, s') \left[R(s, stay, s') + \gamma V^{stay}(s') \right]$$

$$\Rightarrow V^{stay}(1) = 1 \left[1 + 0.5 V^{stay}(1) \right] \rightarrow V^{stay}(1) = 1 + 0.5 V^{stay}(1)$$

$$\Rightarrow 0.5 V^{stay}(1) = 1 \Rightarrow \boxed{V^{stay}(1) = 2}$$

پ) از الگوریتم value iteration استفاده می‌کنیم. نشان می‌دهیم $V_k(1)$ به ازای k به سمت بی‌نهایت به ۲ همگرا است

همه را برابر با صفر می‌ذاریم

$V_0(s) \longrightarrow$

$$V_{k+1}(s) = \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s') \right]$$

$$V_1(1) = \max \Big(1 \times (\underset{\downarrow \atop 1}{1 + 0.5 V_0(1)}) + 1 \times (0 \underset{\downarrow \atop 0}{+} 0.5 \times V_0(2)) \Big)$$

برای تمامی شهرها هم $V_1(1) = 1$, $V_1(2) = 1$, …, $V_1(n) = 1$ حال برای $\underline{V_2}$ حساب میکنیم

$$V_2(1) = \max \Big(\underbrace{(1 + 0.5 \times 1)}_{\boxed{1.5}},\ 1 \times (0 \underset{\downarrow \atop 0.5}{+} 0.5 \times 1) \Big)$$

حال برای $V_2(1) = V_2(2) = \dots = V_2(n) = \underline{1.5}$ ، V_3 را حساب میکنیم

$$V_3(1) = \max \Big(\underset{\downarrow \atop \boxed{1.75}}{(1 + 0.5 \times 1.5)},\ 1 \times (0 \underset{\downarrow \atop 0.75}{+} 1.5 \times 0.5) \Big)$$

ادامه قسمت ب: اگر دقت کنید دنباله ما دارد به 2 همگرا می‌شود پس نتیجه میگیریم

$$V^*(1) = 1 + \underbrace{\left(\left(\left(\left(1\times\tfrac{1}{2}\right)+1\right)\times\tfrac{1}{2}\right)+1\right)\times\tfrac{1}{2}\cdots}_{\downarrow}$$

سری جمله درون پرانتز برابر با ① می‌شود پس

$$\boxed{V^*(1)=2}$$

با الگوریتم Policy iteration در ابتدا سیاست اولیه π_0 را برابر با stay میگذاریم

	city1	city2	...	city n
$\pi_0 \to$	stay	stay	...	stay

$$\to V^{\pi_0}(1) = \sum T(s,\pi_0,s')\left[R(s,\pi_0,s') + \gamma V^{\pi_0}(s')\right]$$

$$\to V^{\pi_0}(1) = 1\times\left(1+0.5\,V^{\pi_0}(1)\right)$$

$$V^{\pi_0}(2) = 1\times\left(1+0.5\,V^{\pi_0}(2)\right)$$

$$\vdots$$

$$V^{\pi_0}(n) = 1\times\left(1+0.5\,V^{\pi_0}(n)\right)$$

$$\to \boxed{V^{\pi_0}(1) = V^{\pi_0}(2) = V^{\pi_0}(3) = \cdots = V^{\pi_0}(n) = 2}$$

حالا چون Policy ثابت است

$$\pi_1(1) = \max_a\left(\underset{\text{stay}}{1\left(1+\overset{1.25}{0.5\times 2}\right)},\ \underset{\text{eat}}{1\left(\overset{0.25}{0+0.5\times 2}\right)}\right) \to \underline{\pi_1(1)\to \text{stay}}$$

برای بقیه حالات هم $\pi_1(k)$ به ازای $n\geq k\geq 1$ برابر با stay می‌شود پس سیاست بهینه در State است

$\boxed{V^*(1)=2}$ می‌شود

قسمت ت در صفحه بعد.

قسمت ت)

$$\pi^*(s) = \arg_a \max \underset{s' : \text{stay}}{\sum} T(s,a,s')\left[R(s,a,s') + \gamma v^*(s')\right]$$

$$\rightarrow \pi^*(\text{city } i) = \max\Big(1\times(r_i + 1\times v^*(\text{city } i)),\ P_i\left(0+1\times v^*(\text{city}(i+1))\right) + (1-P_i)\left(0+1\times v^*(\text{city } i)\right)$$

$$,\ P_i\left(0+1\times v^*(\text{city}(i-1))\right) + (1-P_i)\left(0+1\times v^*(\text{city } i)\right)\Big) = \max\Big(r_i + v^*(\text{city } i),\ v^*(\text{city } i) +$$

$$P_i v^*(\text{city}(i+1)) - P_i v^*(\text{city } i),\ v^*(\text{city } i) + P_i v^*(\text{city}(i-1)) - P_i v^*(\text{city}(i))\Big)$$

$$= \max\Big(r_i + v^*(\text{city}(i)),\ v^*(\text{city}(i)) + P_i\left(v^*(\text{city}(i+1)) - v^*(\text{city}(i))\right),$$

$$v^*(\text{city}(i)) + P_i\left(v^*(\text{city}(i-1)) - v^*(\text{city}(i))\right)\Big)$$

همان طور که مشاهده می‌کنید $v^*(\text{city}(i))$ در هر ۳ عبارت هست پس ماکزیمم بین 3 مقدار

$$\left\{r_i,\ P_i\left(v^*(\text{city}(i+1)) - v^*(\text{city}(i))\right),\ P_i\left(v^*(\text{city}(i-1)) - v^*(\text{city}(i))\right)\right\}$$

← بنابراین مقدار ماکسیمم ، محاسبات در صفحه 5 آمده است

ث)

$$\rightarrow \text{sample} = R(s,a,s') + \gamma \max_{a'} (s',a')$$

در ابتدا همه $Q(s,a)$ برابر با صفر است

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\,[\text{sample}]$$

$\gamma = 0.5$ ، $\alpha = 0.5$

sample 1: $(1, \text{stay}, 4, 1)$

$$\text{sample 1} = 4 + \cancel{\gamma \max_{a'} Q(1,a')}^{\,0} = 4 \quad , \quad Q(1,\text{stay}) \leftarrow (1-0.5)\cancel{Q(1,\text{stay})} + \overset{0.5}{\alpha}(4)$$

$$\rightarrow \boxed{Q(1,\text{stay}) = 2}\ \checkmark,\ Q(1,\text{east}) = 0,\ Q(2,\text{west}) = 0,\ Q(2,\text{stay}) = 0$$

Sample 2: $(1, \text{East}, 0, 2)$

$$\text{sample 2} = 0 + \cancel{\gamma \max_{a'} (2,a')}^{\,0} = 0 \qquad Q(1,\text{east}) = (1-\alpha)\cancel{Q(1,\text{east})} + \alpha\,\cancel{\text{sample 2}} = 0$$

$$\rightarrow Q(1,\text{stay}) = 2,\ Q(1,\text{east}) = 0,\ Q(2,\text{west}) = 0,\ Q(2,\text{stay}) = 0$$

Sample 3: (2, stay, 6, 2)

$\rightarrow$ Sample 3 $= 6 + \cancel{\gamma \max_{a'} \underset{0}{Q(2, a')}} = \boxed{6}$

$Q(2, stay) \leftarrow (1-\alpha)\,\cancel{\underset{0}{Q(2, stay)}} + \overset{0.5}{\alpha}\,\overset{6}{\text{Sample 3}} = \boxed{3}$

$Q(1, stay) = 2,\ Q(1, east) = 0,\ Q(2, west) = 0,\ Q(2, stay) = 3$

Sample 4: (2, west, 0, 1)

$\text{Sample 4} = R(s, a, s') + \gamma \max_{a'} Q(s', a') = 0 + \overset{0.5}{\gamma} \max_{a'} \overset{2}{Q(1, a')} = \boxed{1}$

$a = stay$

$Q(2, west) = (1-\alpha)\, Q(2, west) + \alpha\ \text{Sample 4} = 0.5 \times 0 + \alpha \times 1 = 0.5$

$Q(1, stay) = 2,\ Q(1, east) = 0,\ Q(2, west) = 0.5,\ Q(2, stay) = 3$

Sample 5: (1, stay, 4, 1)

$\text{Sample 5} = 4 + \cancel{\overset{0.5}{\gamma} \max_{a} \overset{1}{Q(1, a')}} = \boxed{5}$

$Q(1, stay) \leftarrow (1-\alpha)\, Q(1, stay) + \alpha\ \text{Sample 5} = 0.5 \times 2 + \underset{2.5}{\cancel{0.5 \times 5}} = \boxed{3.5}$

پس مرحله جستجوی [illegible]

	Q(1, stay)	Q(1, East)	Q(2, west)	Q(2, stay)
(1, stay, 4, 1) ←	2	0	0	0
(1, East, 0, 2) ←	2	0	0	0
(2, stay, 6, 2) ←	2	0	0	3
(2, west, 0, 1) ←	2	0	0.5	3
(1, stay, 4, 1) ←	3.5	0	0.5	3

ادامه قسمت ت) باید مقدار Max

$$\max\{r_i, p_i(v^*(city(i+1)) - v^*(city(i))), p_2(v^*(city(i-1)) - v^*(city(i)))\}$$

اسم آن کشوری که بین این سه مقدار max کدام بیشتر به مقدار r_i , p_2 دارد

اگر r_iها برابر باشند آنگاه حاصل $v^*(city(i-1)) - v^*(city(i))$ برابر با صفر می‌شود دلیل اینکه همه

این شهرها با هم برابر است در قسمت پ که $r_i=1$ یعنی برابر همه یکی بود گفته شد در آن قسمت سیاست

Stay انتخاب شود اگر حاصل $v^*(city(i))$ از بقیه شهرهای مجاور بیشتر باشد در آن صورت باز سیاست

Stay انتخاب شود