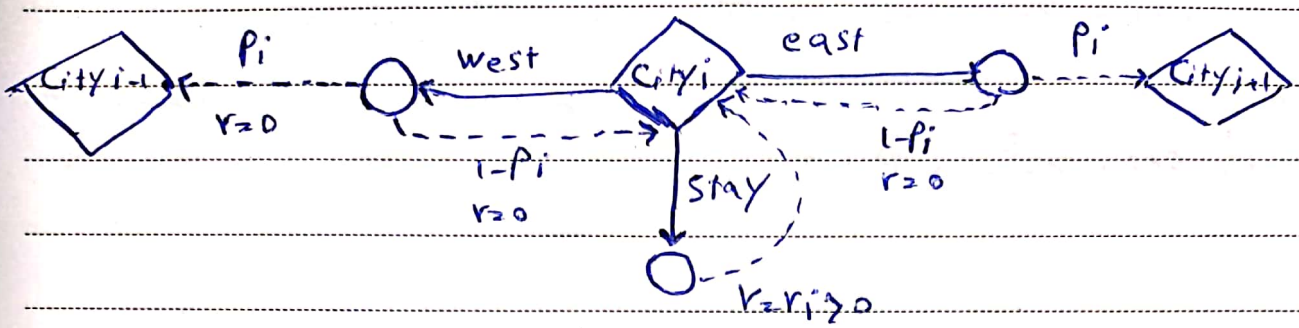
-1



الف)

$V_0^{stay}(i) = 0$ (ب

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

$$\Rightarrow V_1^{stay}(s) = \sum_{s'} T(i, stay, s') \left[R(i, stay, s') + \gamma V_k(s') \right]$$

$$= 1 \, T(i, stay, i) \left[R(i, stay, i) + \frac{1}{2} V_0(i) \right] = 1 \left[1 + 0 \right] = \boxed{1}$$

$$V_2^{stay}(i) = 1 \left[1 + \frac{1}{2} + 1 \right] = \boxed{1 + \frac{1}{2}}$$

$$V_3^{stay}(i) = 1 \left[1 + \frac{1}{2} \left(1 + \frac{1}{2} \right) \right] = \boxed{1 + \frac{1}{2} + \frac{1}{4}}$$

$$\vdots$$

$$V^{stay}(i) = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots \xrightarrow{\text{همگرا}} 2$$

می شود

پس از آنجایی که V^* همه استیت ها برابر 1 است و تمام state ها برابر خواهد بود زیرا وضعیت متقارنی دارند

$$V^*(1) = V^*(2) = \cdots = V^*(N)$$

ما در هر استیتی باشیم در هر حرکت دو انتخاب داریم که برای ما $r=1$ و stay که برای ما $r=0$ دارد و ما را به استیتی می برد که دقیقا از نقله r متناسب state فعلی است. بنابراین در هر حرکت بهتر است stay کنیم و سود $r=1$ را به $r=0$ ترجیح دهیم.

$$V^*(1) = \max_a Q(s, a, s') = \max\left(\underbrace{1*(1+\frac{1}{2}V^*(1)}_{\text{best action} = \text{stay}}, 1*(0+\frac{1}{2}V^*(2)), 1*(0+\frac{1}{2}V^*(0))\right)$$

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

پس می توانیم نتیجه بگیریم سیاست بهینه برای رسیدن به V^* در هر 6 state

stay کردن در آن State است. بنابراین در این مسئله ؛

$$V^*(5) = \underbrace{V^{stay}(5)}_{\text{محاسبه شده در قسمت ب) سؤال}} = 2 \quad \Rightarrow V^*(1) = 2$$

Subject: Year: Month: Day:

~~* فرض این است که سؤال MDP است و γ_i ها را می دانیم~~

ت) سیاست بهینه این است که ابتدا به state ای که بیشترین γ_i را دارد بروی

سپس برای همیشه در آن شهر stay کنی.

اثبات: در هر state ای که باشیم می توانیم به راست یا چپ حرکت کنیم.

فرض کنیم در S_i stat هستیم و استیت S_f بیشترین مقدار γ را برای stay دارد. فرض کنید در k اکشن می توانیم از S_i به S_f برویم.

برای شروع از S_i دو سیاست تعریف می کنیم:

1- از S_i با k اکشن به S_f می رویم و در آنجا stay می کنیم

2- هر سیاست دیگری غیر از سیاست ①

برای محاسبه utility بین از m مرحله داریم

① $utility = (m-k)\, R(S_f, stay, S_f)$

② در بهترین حالت در هر مرحله یک γ کسب کرده ایم که $\gamma \geq R(S_f, stay, S_f)$

① است زیرا اگر $\gamma = R(S_f, stay, S_f)$ باشد با این سیاست همان سیاست

$utility = m\, R(S_x, stay, S_x)$ ،

$R(S_x, stay, S_x) \geq R(S_f, stay, S_f)$ خواهد بود.

V^* ، حالی است که m به ∞ میل کند:

$$\frac{\text{utility ①}}{\text{utility ③}} = \lim_{m \to \infty} \frac{m R(S_f, stay, S_f) - K(R(S_f, stay, S_f))}{m R(S_n, stay, S_n)}$$

$$= \frac{R(S_f, stay, S_f)}{R(S_n, stay, S_n)} > 1 \Rightarrow \text{utility ①} > \text{utility ③}$$

(S, a, r, S')	$Q(1, stay)$	$Q(1, East)$	$Q(2, West)$	$Q(2, stay)$
Initial	0	0	0	0
$(1, stay, 4, 1)$	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