امیرحسین باقری 98105621

تمرین 7.1

سوال اول:

بله ممکن است

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

داریم:

فرض کنید که در مرحله i ام Action a را انجام دهیم ⟸ $\max \leftarrow \sum_{s_1'} T(s_1,a,s_1')\left[R(s_1,a,s_1')+\gamma V^{\pi_i}(s_1')\right]$

و مرحله i+1 نیز Action a را انجام دهیم: $\max \leftarrow \sum_{s_2'} T(s_2,a,s_2')\left[R(s_2,a,s_2')+\gamma V^{\pi_i}(s_2')\right]$

اما تضمینی وجود ندارد که در مرحله i+2 Action a انتخاب بهینه باشد.

دلیل اول ⟸ میزان reward $V^{\pi_i}(s_2')$ برای i+2 , i+1 متفاوت است.

مثال: $\gamma = 0.9$

Action{a, b, c}

$V[s_1] = 1$

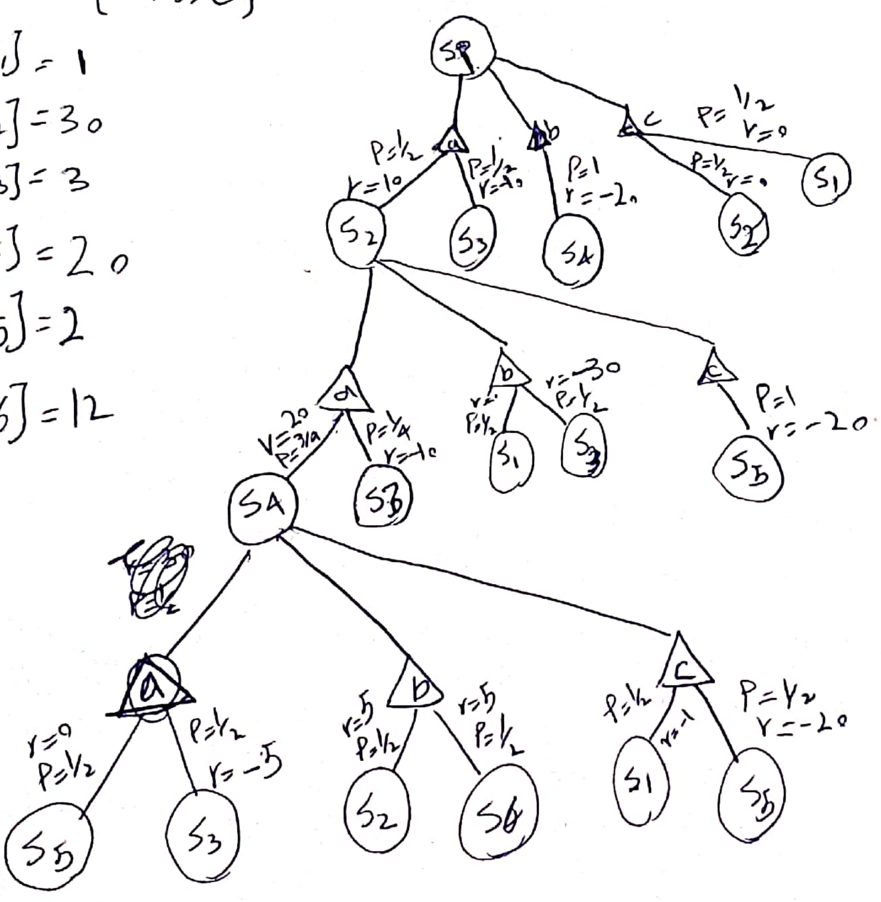
$V[s_2] = 30$

$V[s_3] = 3$

$V[s_4] = 20$

$V[s_5] = 2$

$V[s_6] = 12$



→ با اکشن a به s_2 می‌رویم. فرض کنید.

→ باز هم action a انتخاب و فرض کنید به S4 رفته‌ایم.

→ در انتها اکشن بهینه برابر b است.

بنابراین Policy تغییر کرده است.

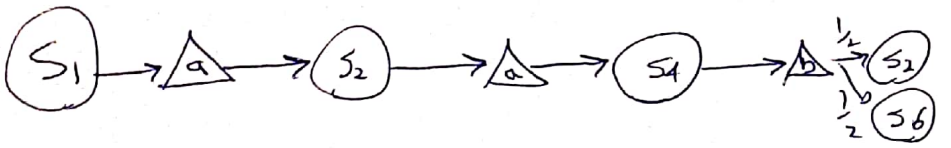
علت:

از آنجاکه با انتخاب هر اکشن با یک احتمالی به یک state دیگر می‌رویم،

حفظی است ⇐ ممکن است جدید بالا transition prob متفاوت است.

⇐ Policy می‌تواند تغییر کند

از مثال صفحه قبل داریم:



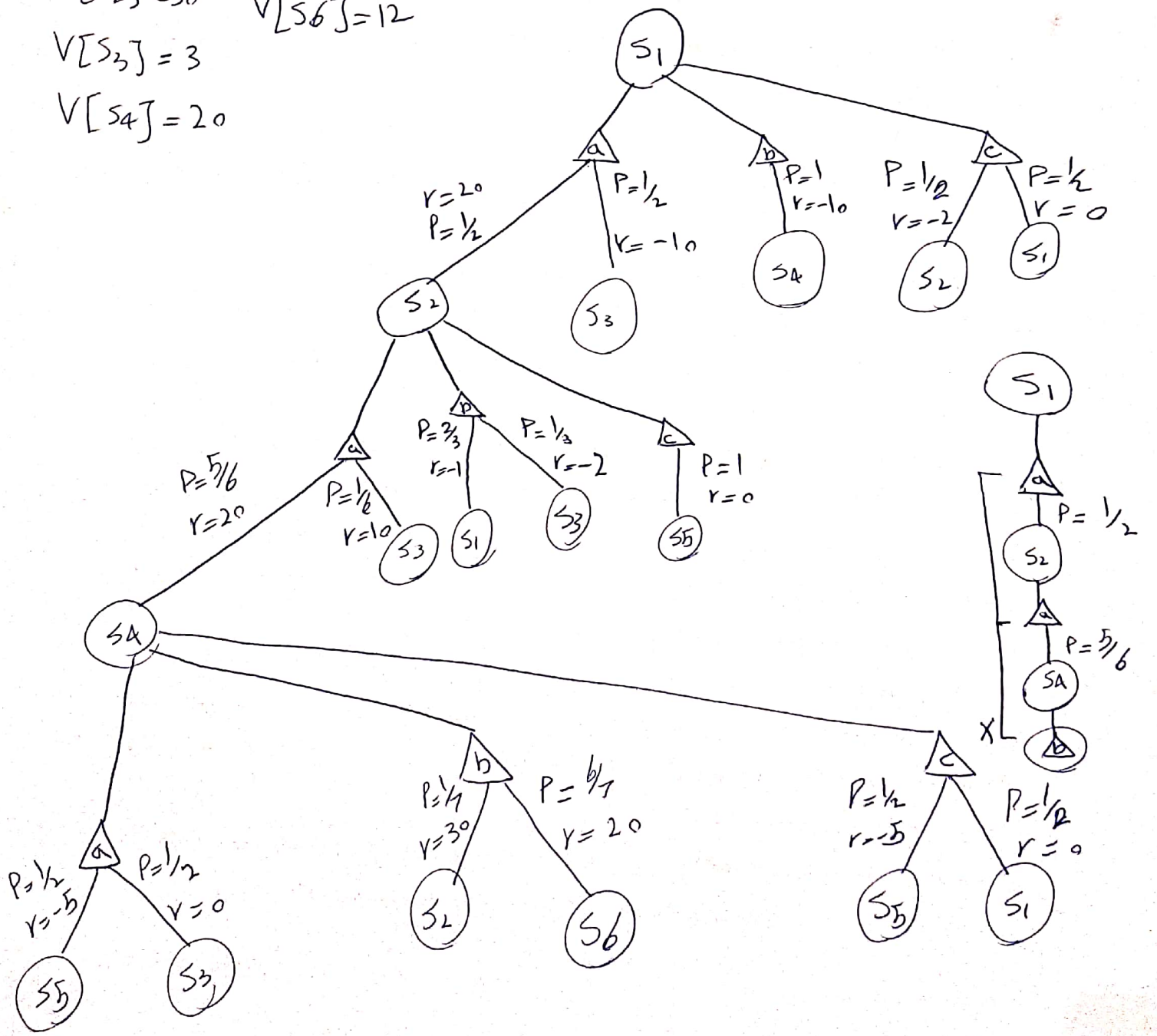
در MDP زیر Actions={a,b,c}

$V[S_1]=1$ $V[S_5]=2$

$V[S_2]=30$ $V[S_6]=12$

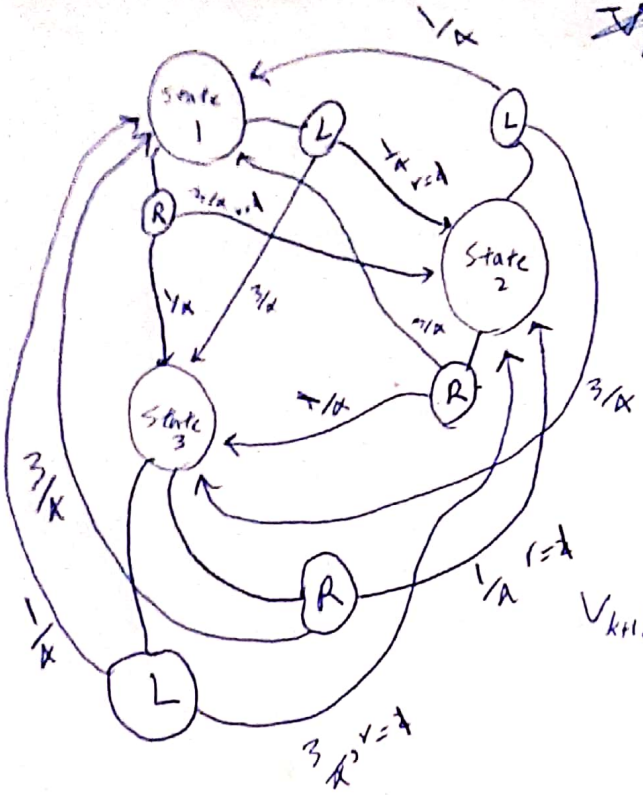
$V[S_3]=3$

$V[S_4]=20$



به طور کلی انتخاب هر Policy در هر حالت بستگی به expect هر اکشن دارد expect هر اکشن
تابعی از Transition prob و مقدار V^* هر state است. بنابراین از آنجا که
مقدار expect در Policy کسانی در حالت i و i+1 مقدار ماکسیمم گذاشت.
می توان گفت که در حالت i+2 نیز همان اکشن مقدار expect ماکسیمم را
دارا است. [حالت خنثی احاکردیم] بنابراین ممکن است Policy
تغییر کند.

سوال ۲



$V_{k+1}[s] = \max_{a}^{Arg} \left(\sum_{s'} T(s,s',a)\,(R(s,s') + \right.$

V_k	k=0	1	2
State 1	0	3/4	27/32
State 2	0	0	12/32
State 3	0	3/4	27/32

دقت کنید که فرض کردیم که $V_{s_{i+2}}$ reward $(s_i \to s_2) = 4$

$$V_2[s_1] = \max\left(\underbrace{\tfrac{1}{4} \times (1 + \tfrac{1}{2} \times 0) + \tfrac{3}{4} \times (0 + \tfrac{1}{2} \times \tfrac{3}{4})}_{L},\ \underbrace{\tfrac{3}{4}(1+0) + \tfrac{1}{4} \times (\tfrac{1}{2} \times \tfrac{3}{4})}_{R} \right) = \frac{3+24}{32} = \frac{27}{32}$$

$\tfrac{1}{4} + \tfrac{9}{32} = \tfrac{3}{8}$ ، $\tfrac{3}{4} + \tfrac{3}{32} =$

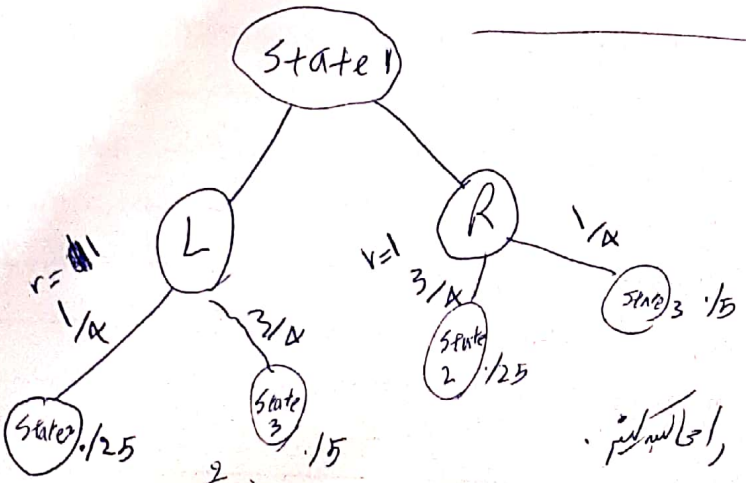
$$V_2[s_2] = \max\left(\tfrac{1}{4} \times (\tfrac{3}{4} \times \tfrac{1}{2}) + \tfrac{3}{4} \times \tfrac{1}{2} \times \tfrac{3}{4},\ \tfrac{3}{4} \times (\tfrac{1}{2} \times \tfrac{3}{4}) + \tfrac{1}{4} \times \tfrac{1}{2} \times \tfrac{3}{4} \right) = \frac{12}{32}$$

$\tfrac{9}{32} + \tfrac{3}{32}$

$$V_2[s_3] = \max\left(\tfrac{1}{4} \times \tfrac{1}{2} \times \tfrac{3}{4} + \tfrac{3}{4}(1 + \tfrac{1}{2} \times 0),\ \tfrac{1}{4} \times (1 + \tfrac{1}{2} \times 0) + \tfrac{3}{4} \times \tfrac{1}{2} \times \tfrac{3}{4} \right) = \frac{27}{32}$$

$\tfrac{3}{32} + \tfrac{3}{4}$ ، $\tfrac{1}{4} + \tfrac{9}{32}$

ب



باید $E(L)$ و $E(R)$ را محاسبه کنیم.

$$E(L) = \tfrac{3}{4} \times (\tfrac{1}{2} \times \tfrac{1}{2}) + \tfrac{1}{4} \times (1 + \tfrac{1}{2} \times \tfrac{1}{4}) = \frac{3}{16} + \frac{9}{32} = \frac{15}{32}$$

$\tfrac{2}{8}$ ، $1 + \tfrac{1}{8}$

$$E(R) = \tfrac{1}{4} \times (\tfrac{1}{2} \times \tfrac{1}{2}) + \tfrac{3}{4} \times (1 + \tfrac{1}{2} \times \tfrac{1}{4}) = \frac{1}{16} + \frac{24}{32} = \frac{26}{32}$$

$1 + \tfrac{1}{8}\tfrac{2}{8}$

بنابراین اکشن بهینه حتما R است.

با انتخاب R با احتمال $\frac{3}{4}$ به state ۲ و با احتمال $\frac{1}{4}$ reward کاهش می‌یابد.

به ۳ state کاهش می‌یابد. بنابراین با توجه به همسایه‌ی expectimax و policy بهینه انتخاب action R است.

V \ k	0	1	2
state1	0	$\frac{3}{4}$	$\frac{27}{32}$
state2	0	0	$\frac{3}{8}$
state3	0	$\frac{3}{4}$	$\frac{27}{32}$

خلاصه:

الف ←

ب ← Action A زیرا مقدار expectimax آن بیشتر است.