حالت	S_1	S_2	S_3	S_4	S_5	S_6
$\pi(i)$	ریختن تاس	ریختن تاس	اتمام بازی	اتمام بازی	اتمام بازی	اتمام بازی
$V(\pi(i))$	۳	۳	۳	۴	۵	۶

حالت	S_1	S_2	S_3	S_4	S_5	S_6
$\pi(i)$	ریختن تاس	ریختن تاس	اتمام بازی	اتمام بازی	اتمام بازی	اتمام بازی
$\pi(i+1)$	تاس ریختن	تاس ریختن	تاس ریختن / اتمام	اتمام	اتمام	اتمام

همان طور که در جدول مشخص می باشد policy ها در حالت i و حالت i+1 تغییر پیدا کرده است.
در حالت ۳ هم با هر دو تصمیم بهینه اند در نتیجه
نمی توان از آن به عنوان بهترین policy عنوان کرد. در نتیجه policy همگرا یا Converge نشده است.
البته امکان دارد policy ثابت باشد ولی value ها بعد از policy همگرا شوند. اصولاً در مسائل کاربردی policy زودتر از value ها همگرا می شود.

توضیحات عملی برای سؤال ۱

از این فرمول ها برای policy evaluation، policy improvement و policy iteration استفاده می‌کنیم.

evaluation:

$$V^{\pi_i}_{k+1}(s) = \sum_{s'} T(s,\pi_i(s),s')[R(s,\pi_i(s),s') + \gamma V_k^{\pi_i}(s')]$$

Improvement:

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

$V(s)=0$ را مقداردهی اولیه می‌کنیم تا این الگوریتم بار دیگر به π_i برسیم. policy ها برای همه حالت‌ها در ابتدا f می‌باشند. مراحل همگرایی.

finish: f

out: T

حالت‌ها	S_1	S_2	S_3	S_4	S_5	S_6
$V_0(S)$	۰	۰	۰	۰	۰	۰
π_0 / $V_1^{\pi_0}(s)$	f ۱	f ۲	f ۳	f ۴	f ۵	f ۶
π_1 / $V_2^{\pi_1}(s)$	T ۲/۵	T ۲/۵	f ۳	f ۴	f ۵	f ۶
π_2 / $V_3^{\pi_2}(s)$	T ۲/۸۳	T ۲/۸۳	f ۳	f ۴	f ۵	f ۶
π_3 / $V_4^{\pi_3}(s)$	T ۲/۹۴	T ۲/۹۴	f ۳	f ۴	f ۵	f ۶
π_4 / $V_5^{\pi_4}$	T ۲/۹۸	T ۲/۹۸	f ۳	f ۴	f ۵	f ۶
π_5 / $V_6^{\pi_5}$	T ۲/۹۹	T ۲/۹۸	f ۳	f ۴	f ۵	f ۶
π_i / $V^{\pi_i}(s)$	T ۳	T ۳	f ۳	f ۴	f ۵	f ۶

حل کار ادامه: $V(S)=0$ $\pi_0(S)=f$

$V(S_1)$ و $V(S_2)$ بعد از چندین دور ۳ همگرا می‌شوند.

$V^{\pi_1}(S_1)=\frac{1}{4}\left[(-1+1\alpha 3)+(-1+1\alpha 3)+(-1+1\alpha 3)+(-1+1\alpha 4)+(-1+1\alpha 5)+(-1+1\alpha 6)\right]$

برای حالت ۲:

$V^{\pi_i}(S_1)=\frac{1}{4}\alpha 18=3$	S_1: آتش $Q=3$ max ؛ fold $Q=1$ $Q(S,f)$
$V^{\pi_i}(S_2)=3$	S_2 { $T: Q=\frac{1}{6}(18)=3$ max ; $f: 2=1(0+1\alpha 2)$
$V^{\pi_i}(S_3)=3$	S_3 { $T: Q=\frac{1}{6}(18)=3$; $f: Q=1(0+1\alpha 3)=3$ — هر دو optimal هستند
$V^{\pi_i}(S_4)=1(0+1\alpha 4)$	S_4 { T $Q=3$; f $Q=4 \to$ max
$V^{\pi_i}(S_3)=1(0+1\alpha 3)=3$	S_5 { $T: Q=\frac{1}{6}(18)=3$; $f: Q=5 \to$ max
$V^{\pi_i}(S_6)=1(0+1\alpha 6)=6$	S_6 { T $Q=3$; f $Q(S_6,f)=6 \to$ max
$V^{\pi_i}(S_5)=1(0+1\alpha 5)=5$	

② ۱- policy بهینه: به ازای همه State ها راست می‌رویم در نهایت، در loop می‌مانیم

۱	۲	۳	...	n-۲	n-۱	n
R	R	R	R ...	R	R	Loop

۲- مقدار بهینه $V^*(n)$

$V_0(n) = 0$ Initial State(0) مقداردهی اولیه

$V_1(n) = 10 + \frac{1}{2} \times 0 = 10$ $V_3(n) = 10 + \frac{10}{2} + \frac{10}{2^2}$

$V_2(n) = 10 + \frac{1}{2} \times 10 = 15$ $V_4(n) = 10 + \frac{10}{2} + \frac{10}{2^2} + \frac{10}{2^3}$

$\vdots$

$$V_n(n) = 10 \times \left(1 + \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^{n-1}}\right) \xrightarrow{n \to \infty} 10 \times 2 = 20$$

جمع هندسی: ۲ میل می‌کند

۲۰ همگرا می‌شود

۳- مقدار بهینه برای $V^*(k)$ به ازای $k = 1, \dots, n-1$ ها:

$$V^*(1) = V^*(n-(n-1)) = 1 + \frac{1}{2} + \cdots + \left(\frac{1}{2}\right)^{n-2} + \frac{20}{2^{n-1}} = 2$$

$$V^*(2) = V^*(n(n-2)) = 1 + \frac{1}{2} + \cdots + \left(\frac{1}{2}\right)^{n-3} + \frac{20}{2^{n-2}} = 2$$

$\vdots$

$$V^*(n-1) = V^*[n-1] = \left(\frac{1}{2}\right)^0 + \frac{20}{2} = 11$$

$V(n) = 20$

$V(n-1) = 1 + \frac{20}{2}$

$V(n-2) = 1 + \frac{1}{2} + \frac{20}{4}$

$V(n-3) = 1 + \frac{1}{2} + \frac{1}{2^2} + \frac{20}{2^3}$

$V(n-k) = 1 + \frac{1}{2} + \cdots + \left(\frac{1}{2}\right)^{k-1} + \frac{20}{2^k} \simeq 2$

به ازای k های بزرگ: همگرا به ۲ — به ازای k های بزرگ صفر همگرا

۴- به ازای iteration های ۱ و ۲ هر مقداری می‌توانند همگرا شوند.

	۱	۲	۳	...	n-۲	n-۱	n
V_0	۱۰.	۰	۰	...	۰	۰	۰
V_1	۱.	۱	۱	...	۱	۱	۱۰
V_2	۱/۵	۱/۵	۱/۵	...	۱/۵	۶	۱۵

در ستون‌های ۱ و ۲ تمام state ها می‌توانند همگرا شوند چرا که با حرکت به راست، پاداش می‌گیرند و همچنین state n هم پاداش ۲۰ می‌گیرد.

ع.م.س: ۱ تا n-۱

مقدار state ها به سمت بالا ترجیحی در state سمت راست می‌آید.