(1) از فرمول های زیر برای Evaluation و Policy Improvement در الگوریتم policy iteration استفاده می کنیم.

Evaluation:

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

Improvement:

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

~~برای استفاده از این سیاست π_i ابتدا باید مقادیر همگرا شود لذا ابتدا هم $V(s)$ ها صفر می کنیم و چندین تکرار انجام می دهیم~~

ابتدا $V(s)$ ابتدا $V(s)=0$ می گذاریم برای هم حالت ها و سپس چندین تکرار انجام می دهیم تا به π_i برسیم. در ابتدا سیاست همه حالت ها را اتمام بازی می گذاریم.

init:
$\forall s: V(s)=0$
$\pi_0(s)=$ finish

iterate:

حالت	s_1	s_2	s_3	s_4	s_5	s_6
$V_0(s)$	0	0	0	0	0	0
π_0	اتمام	اتمام	اتمام	اتمام	اتمام	اتمام
$V_1^{\pi_0}(s)$	1	2	3	4	5	6
π_1	تاس	تاس	اتمام	اتمام	اتمام	اتمام
$V_2^{\pi_1}(s)$	2.5	2.5	3	4	5	6
π_2	تاس	تاس	اتمام	اتمام	اتمام	اتمام
$V_3^{\pi_2}(s)$	2.83	2.83	3	4	5	6
π_3	تاس	تاس	اتمام	اتمام	اتمام	اتمام
$V_4^{\pi_3}(s)$	2.94	2.94	3	4	5	6
π_4	تاس	تاس	اتمام	اتمام	اتمام	اتمام
$V_5^{\pi_4}(s)$	2.98	2.98	3	4	5	6
π_5	تاس	تاس	اتمام	اتمام	اتمام	اتمام
$V_6^{\pi_5}(s)$	2.99	2.99	3	4	5	6

$V(s_1), V(s_2)$ بعد از چندین مرحله به 3 همگرا می شوند.

۱.

حالت	s_1	s_2	s_3	s_4	s_5	s_6
π_i	تاس ریختن	تاس ریختن	اتمام بازی	اتمام بازی	اتمام بازی	اتمام بازی
$V^{\pi_i}(s)$	3	3	3	4	5	6

$$V^{\pi_i}(s_1) = \frac{1}{6}\left((-1+1\times3)+(-1+1\times3)+(-1+1\times3)+(-1+1\times4)+(-1+1\times5)+(-1+1\times6)\right)$$

$$V^{\pi_i}(s_1) = \frac{1}{6}\times18 = 3 \qquad V^{\pi_i}(s_2) = \underset{\text{محاسبه می‌شود}}{\text{مشابه } s_1} = 3$$

$$V^{\pi_i}(s_3) = 1(0+1\times3) = 3 \qquad V^{\pi_i}(s_4) = 1(0+1\times4) = 4 \qquad V_{\pi_i}(s_5) = 1(0+1\times5) = 5$$

$$V^{\pi_i}(s_6) = 1(0+1\times6) = 6$$

۲.

حالت	s_1	s_2	s_3	s_4	s_5	s_6
π_i	تاس ریختن	تاس ریختن	اتمام بازی	اتمام بازی	اتمام بازی	اتمام بازی
π_{i+1}	تاس ریختن	تاس ریختن	اتمام بازی یا تاس ریختن	اتمام بازی	اتمام بازی	اتمام بازی

s_1 — actions → تاس: $Q=3$ → argmax = تاس
s_1 — actions → اتمام: $Q=1$ ($Q(s_1, finish)$)

s_2 — actions → تاس: $Q=\frac{1}{6}(18)=3$ → argmax = تاس
s_2 — actions → اتمام: $Q=2=1(0+1\times2)$

s_3 → تاس: $Q=\frac{1}{6}(18)=3$
s_3 → اتمام: $Q=1(0+1\times3)=3$
→ argmax = اتمام / تاس
هردو تصمیم optimal اند.

s_4 → تاس: $Q=3$
s_4 → اتمام: $Q=4$ → argmax = اتمام

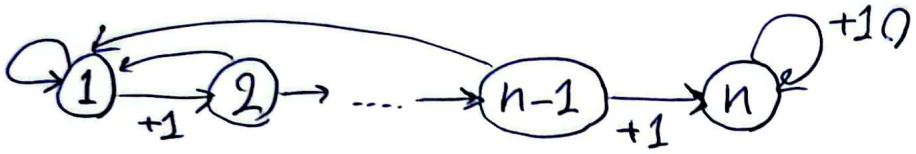
s_5 → تاس: $Q=\frac{1}{6}(18)=3$
s_5 → اتمام: $Q=5$ → argmax = اتمام

s_6 → تاس: $Q(s_6, dice)=3$
s_6 → اتمام: $Q(s_6, finish)=6$ → argmax = finish
اتمام

بله؛ مقادیر به دست آمده بهینه هستند و policy همگرا (converge) شده است.
و دیگر نیاز به بروز رسانی ندارد. چراکه policy از حالت π_i به π_{i+1} تغییر خاصی نداشته است. چراکه در حالت S_3 هردو تصمیم optimal (بهینه) هستند و نمی توان از آن به عنوان تغییر سیاست یاد کرد و همچنین S_1, S_2, S_4, S_5, S_6 هیچ تغییری نکرده اند.
اما باید توجه کنیم که ممکن است value ها دیرتر از policy ها همگرا شوند و وقتی policy ها همگرا هستند value ها همگرا نباشند.
The policy often converges long before the values.

---

۲)



۱.

حالت	1	2	3	...	n-2	n-1	n
سیاست (π)	حرکت آزادانه	حرکت آزادانه	حرکت آزادانه	حرکت آزادانه	حرکت آزادانه	حرکت آزادانه	loop (خارجی شویم)

۲.

$$V^{\pi}_{k+1}(n) = \sum_{s'} \underbrace{T(s,\pi_i(s),s')}_{1}\left[\underbrace{R(s,\pi(s),s')}_{+10} + \underbrace{\frac{1}{2}}_{\gamma} V^{\pi}_{k}(n)\right] = Q(n, loop)$$

$$V^{\pi}_{k+1}(n) = 1 \times \left[10 + \frac{1}{2} V^{\pi}_{k}(n)\right]$$

$$V^{\pi}_{0}(n) = 0 \quad (init)$$

$$V^{\pi}_{1}(n) = 1 \times \left[10 + \frac{1}{2} 0\right] = 10$$

$$V^{\pi}_{2}(n) = 1 \times \left[10 + \frac{1}{2} 10\right] = 15$$

$$V^{\pi}_{3}(n) = 1 \times \left[10 + \frac{1}{2} 15\right] = 17.5$$

$$\vdots \qquad \left(\frac{1}{2}\right)10 + \left(\frac{1}{2}\right)^2 10$$

$$\vdots$$

$$V^{\pi}_{i}(n) = 1 \times \left[10 + \frac{10}{2} + \frac{10}{2^2} + \cdots + \frac{10}{2^{n-1}}\right] = 10 \times \frac{1-\left(\frac{1}{2}\right)^n}{1-\frac{1}{2}}$$

$n \to \infty$

$$V^{\pi}_{i}(n) = 20$$

$$V^{\pi}_{i}(n) = 10 \times \frac{1}{\frac{1}{2}}$$

پس $V^*(n)$ به 20 همگرا می‌شود. $V^*(n) = 20$

۳.

$$V^*(k) = \underbrace{1 + (\frac{1}{2})^1 + (\frac{1}{2})^2 + \dots + (\frac{1}{2})^i}_{1 + \frac{1}{2} + (\frac{1}{2})^2 + \dots + (\frac{1}{2})^{(n-1)-k}} + 20 \cdot (\frac{1}{2})^{i+1} = \sum_{i=0}^{i=n-1-k} (\frac{1}{2})^i + 20 \times (\frac{1}{2})^{n-k}$$

$$= \frac{1 - (\frac{1}{2})^{n-k}}{1 - \frac{1}{2}} + 20 \times (\frac{1}{2})^{n-k}$$

$n-k$ بزرگ شود به صفر همگرا می‌شود.

$n-k$ بزرگ شود به دو همگرا می‌شود.

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

k کوچک، n بزرگ $\longrightarrow V^*(k) = 2(1 - \underbrace{(\frac{1}{2})^\infty}_{=0}) + 20(0) = 2 \xrightarrow{\text{مثلاً}} V^*(1) = 2 \longrightarrow V^*(1) = 2$

هرچه k بزرگتر شود $V^*(k)$ بیشتری شود. $\longrightarrow$ مثلاً $V^*(n-1) = 2(1 - \frac{1}{2}) + 20 \cdot (\frac{1}{2}) = 11 \longrightarrow V^*(n-1) = 11$

هرچه k بزرگتر باشد (state) ارزش بیشتر دارد.

~~$V^*(1) = 2 \leftarrow k = 1$~~

~~$V^*(n-1) \leftarrow k = n-1$~~

۴.

حالت	1	2	3	4	...	n-3	n-2	n-1	n
$V_0^\pi(s)$	0	0	0	0	0	0	0	0	0
$V_1^\pi(s)$	1	1	1	1	1	1	1	1	10
$V_2^\pi(s)$	1.5	1.5	1.5	1.5	1.5	1.5	1.5	6	15

در iteration های اول و دوم تمام حالت‌ها نمی‌توانند صفر باشند چراکه با حرکت آزادانه حالت‌های اتا $n-1$ پاداش می‌گیرند و همچنین حالت n نیز پاداش می‌گیرد. $+10$