

### Exercise 3.4

1.

$s$	$a$	$s'$	$r$	$p(s', r   s, a)$
high	search	high	9 search	$\alpha$
high	wait	low	0	$1 - \alpha$
low	search	high	-3	$1 - \beta$
low	search	low	9 search	$\beta$
high	wait	high	1	9 wait
high	wait	low	0	0
low	wait	high	0	0
low	wait	low	1	9 wait
low	recharge	high	1	0
low	recharge	low	0	-

Exercise 3.15

$$\begin{aligned}
 3. G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \\
 G_{tc} &= (R_{t+1} + C) + \gamma(R_{t+2} + C) + \gamma^2(R_{t+3} + C) + \dots \\
 &= \left( R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \right) + C[1 + \gamma + \gamma^2 + \dots] \\
 &= G_t + C \cdot \frac{1}{1-\gamma}
 \end{aligned}$$

$$v_n(s) = E[G_t | S_t = s]$$

$$\begin{aligned}
 v_{\pi_c}(s) &= E[G_{tc} | S_t = s] \\
 &= E\left[G_t + \frac{C}{1-\gamma} | S_t = s\right]
 \end{aligned}$$

$$v_{\pi_c}(s) = E[G_t | S_t = s] + \frac{C}{1-\gamma}$$

$$v_{\pi_c}(s) = v_n(s) + \frac{C}{1-\gamma}$$

$$\frac{C}{1-\gamma} = v_c$$

Only the intervals between rewards are important and not the signs.

Since  $v_n(s) \forall s \in S$  gets an addition term of  $v_c$  which is a constant, the relative values of any state is not affected.

3.16

$$G_t = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{k-1} R_{t+k}$$

$$G_{t_c} = (R_{t+1} + c) + \gamma(R_{t+2} + c) + \dots + \gamma^{k-1}(R_{t+k} + c)$$

$$G_{t_c} = (R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots + \gamma^{k-1} R_{t+k}) \\ + c(1 + \gamma + \gamma^2 + \dots + \gamma^{k-1})$$

$$= G_t + c \frac{(1 - \gamma^k)}{1 - \gamma}$$

$$V_{\pi}(s) = E[G_t | S_t = s]$$

$$V_{\pi_c}(s) = E[G_{t_c} | S_t = s]$$

$$= E\left[G_t + c \frac{(1 - \gamma^k)}{1 - \gamma} \mid S_t = s\right]$$

$$= E[G_t | S_t = s] + c \frac{(1 - \gamma^k)}{1 - \gamma}$$

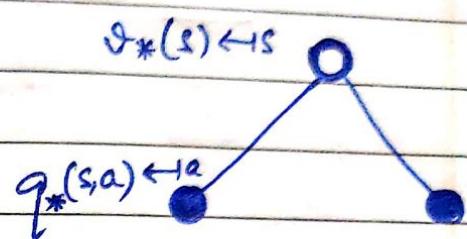
$$= V_{\pi}(s) + c \frac{(1 - \gamma^k)}{1 - \gamma}$$

The state value fraction of each state will be incremented by  $c \frac{(1 - \gamma^k)}{1 - \gamma}$ , the increment term is different for episodic task

But since the state value for each state gets incremented by the same constant term, there's no change in the relative values.

$$V_*(s) = \max_a q_*(s, a)$$

$$5 \quad V_*(s) = \max_{a \in A(s)} q_*(s, a)$$



Instead of taking the expected action value  $\mathbb{E}$ , for a state, we choose the maximum action value using  $q^*(s, a)$ .



$$8 \quad S_t = s, A_t = a, R_{t+1} = r$$

$$S_{t+1} = s', A_{t+1} = a', R_{t+2} = r'$$

$a'$  is an arbitrary action taken at time  $t+1$ .  
~~Arc is unaffected by  $s', a'$  and  $r'$~~

$P[R_{t+2} = r | S_t = s, A_t = a]$  is unaffected by  $a'$  &  $s'$   
 $\Rightarrow P[R_{t+2} = r | S_t = s, A_t = a]$  is calculated over all possible  $a', s'$

$$\Rightarrow P[R_{t+2} = r' | S_t = s, A_t = a]$$

$$= \sum_{s', a'} \pi(a' | s') \cdot P[S_{t+1} = s' | S_t = s, A_t = a] \cdot P[R_{t+2} = r' | S_{t+1} = s', A_t = a]$$

$$= \sum_{s', a'} \pi(a' | s') p(s' | s, a) p(r' | s', a')$$

$$= \sum_{s', a'} \pi(a' | s') \left[ \sum_a [p(r, s' | s, a)] \right] \left[ \sum_{s''} p(r', s'' | s', a') \right]$$

where  $s'' = s_{t+2}$

$$9 \quad E[R_{t+2} | S_t = s, A_t = a]$$

$$= \sum r' P[R_{t+2} = r' | S_t = s, A_t = a]$$

$$= \sum_{r' \in R} r' P[$$

$$= \sum \left[ r' \sum_{s', a'} \pi(a' | s') \left[ \sum_r p(r, s' | s, a) \right] \left[ \sum_{s''} p(r', s'' | s', a') \right] \right]$$

using the expression in previous question.

$$10 \quad v_{\pi}(s) = E[G_t | S_t = s]$$

$$= E[R_{t+1} + \gamma G_{t+1} | S_t = s]$$

$$= E[R_{t+1} | S_t = s] + \gamma E[G_{t+1} | S_t = s]$$

$$= E[R_{t+1} | S_t = s] + \gamma E[E[G_{t+1} | S_t = s] | S_t = s]$$

$$= E[R_{t+1} | S_t = s] + \gamma E[v_{\pi}(s_{t+1}) | S_t = s]$$

$$= E[R_{t+1} + \gamma v_{\pi}(s_{t+1}) | S_t = s]$$

$$= \sum_{s', r} (r + \gamma v_{\pi}(s')) p(s', r | s)$$

$$= \sum_a \pi(a | s) \sum_{s', r} (r + \gamma v_{\pi}(s')) p(s', r | s, a)$$

$$11. R_1 = 2, R_2 = -1, R_3 = 10, R_4 = -3$$

$$G_3 = R_4 = -3$$

$$G_2 = R_3 + \gamma R_4 = 10 + 0.5 \times (-3) = 8.5$$

$$G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 = -1 + \gamma G_2 = -1 + 0.5 \times 8.5 \\ = 3.25$$

$$G_0 = R_1 + \gamma G_1 = 2 + 0.5 \times 3.25 \\ = 3.625$$

Q If agent receives constant reward  $c$ , at every time step,  $R_{t+k} = c$ ,  $k \geq 0$

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \\ &= c + \gamma c + \gamma^2 c + \dots \\ &= c(1 + \gamma + \gamma^2 + \dots) \\ &= \frac{c}{1-\gamma} \end{aligned}$$

12 We know,  $\pi^*(s) = \operatorname{argmax}_a q^*(s, a)$

Now consider  $q^*(s, a)$

$$q^*(s, a) = \max_{\pi} q_{\pi}(s, a)$$

$$q_{\pi}(s, a) = E_{\pi}[G_t \mid S_t = s, A_t = a]$$

$$= E_{\pi}[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a]$$

$$= E_{\pi}[R_{t+1} \mid S_t = s, A_t = a] + \gamma E_{\pi}[G_{t+1} \mid S_t = s, A_t = a]$$

$$= E_{\pi}[R_{t+1} \mid S_t = s, A_t = a] + \gamma E_{\pi}[E[G_{t+1} \mid S_t = s] \mid S_t = s, A_t = a]$$

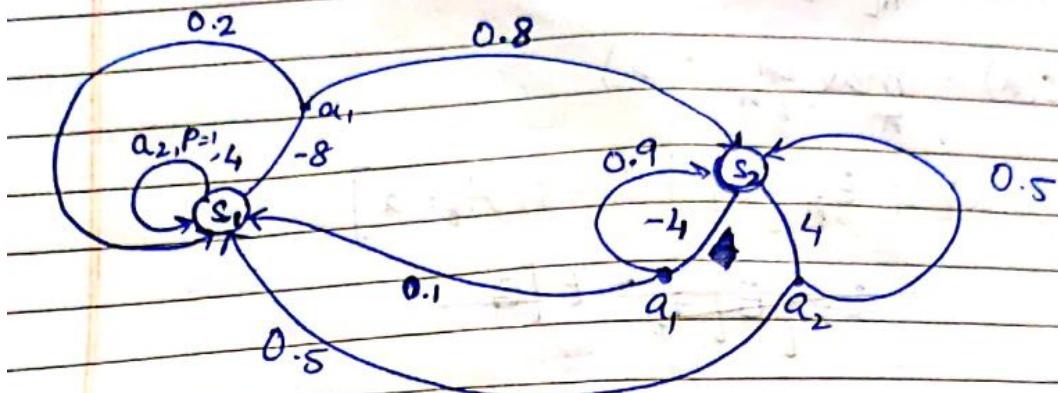
$$= E_{\pi}[R_{t+1} \mid S_t = s, A_t = a] + \gamma E_{\pi}[v_{\pi}(s_{t+1}) \mid S_t = s, A_t = a]$$

$$= E_{\pi}[R_{t+1} + \gamma v_{\pi}(s_{t+1}) \mid S_t = s, A_t = a]$$

$$q^*(s, a) = \max_{\pi} E[R_{t+1} + \gamma v_{\pi}(s_{t+1}) \mid S_t = s, A_t = a]$$

$$\pi^*(s) = \operatorname{argmax}_a E[R_{t+1} + \gamma v_{\pi}(s_{t+1}) \mid S_t = s, A_t = a]$$

13

State $s_1 \rightarrow \text{stale}$  $s_2 \rightarrow \text{fresh}$ Action $a_1 \rightarrow \text{query}$  $a_2 \rightarrow \text{remain silent}$ 

Initially,  $\pi(a_1 | s_1) = \pi(a_2 | s_1) = \frac{1}{2}$   
 $\pi(a_1 | s_2) = \pi(a_2 | s_2) = \frac{1}{2}$

$$v_\pi(s_1) = 0, v_\pi(s_2) = 0$$

$$\gamma = \frac{1}{2}$$

Iteration - 1

$$v_\pi(s_1) = \max(a, b)$$

$$a = \sum_{s', r} p(s', r | s, a_1) [r + \gamma v_\pi(s')]$$

$$= 0.2 \times (-8 + 0) + 0.8 \times (-8 + 0) \\ = -8$$

$$b = \sum_{s', r} p(s', r | s, a_2) [r + \gamma v_\pi(s')]$$

~~$$= 1 \times \left( 4 + \frac{1}{2} \times (-8) \right) = 0 - \frac{1 \times 4}{2} + 1$$~~

$$v_\pi(s_1) = 0$$

$$b = \frac{1}{2}$$

~~v<sub>t</sub>~~

$$b = 1 \times \left( 4 + \frac{1}{2} \times 0 \right) = 4$$

$$\Rightarrow v(s_1) = 4$$

$s_2$

$$v(s_2) = \max(a, b)$$

action  $a_1$       action  $a_2$

$$a = \sum_{s', r} p(s', r | s, a) [r + \gamma v(s')]$$

$$= 0.9 [-4 + 0.5 \times 0] + 0.1 [-4 + 0.5 \times 4]$$

$$= -0.9 - 3.6 + 0.1 [-2]$$

$$= -3.8$$

$$b = 0.5 [4 + 0.5 \times 0] + 0.5 [4 + 0.5 \times 4]$$

$$= 2 + 3 = 5$$

$$v(s_2) = 5$$

Iteration - 2

$s_1$

$$v(s_1) = (a, b)$$

$$a = \sum_{s', r} p(s', r | s, a) [r + \gamma v(s')]$$

$$= 0.2 [-8 + 0.5 \times 4] + 0.8 [-8 + 0.5 \times 5]$$

$$= 0.2 [-6] + 0.8 [-8 + 2.5]$$

$$= -1.2 + 0.8 \times (-5.5) = -5.6$$

$$b = 1 \times [4 + 0.5 \times 4] = 6$$

$$v(s_1) = 6$$

$$v(s_2) = \max(a, b)$$

$$\begin{aligned}a &= 0.9 \times [-4 + 0.5 \times 5] + 0.1 \times [-4 + 0.5 \times 6] \\&= 0.9 \times (-1.5) + 0.1 \times (-1) \\&= -1.35 - 0.1 \\&= -1.45\end{aligned}$$

$$\begin{aligned}b &= 0.5 \times (4 + 0.5 \times 6) + 0.5 \times (4 + 0.5 \times 5) \\&= 0.5(7) + 0.5(6.5) \\&= 6.75 \\v(s_2) &= 6.75\end{aligned}$$

### ~~\*3~~ Iteration 3

$$v(s_1) = \max(a, b)$$

$$\begin{aligned} a &= 0.8 \times (-8 + 0.5 \times 6.75) + 0.2 \times (-8 + 0.5 \times 6) \\ &= -4.7 \end{aligned}$$

$$b = 1 \times (4 + 0.5 \times 6)$$

$$= 7$$

$$\boxed{v(s_1) = 7} \quad \rightarrow$$

$$v(s_2) = \max(a, b)$$

$$\begin{aligned} a &= 0.9 \times (-4 + 0.5 \times 6.75) \\ &\quad + 0.1 \times (-4 + 0.5 \times 7) \\ &= -4 + 0.5 \times 6.775 \\ &= -0.6125 \end{aligned}$$

$$\begin{aligned} b &= 0.5(4 + 0.5 \times 7) + 0.5(4 + 0.5 \times 6.75) \\ &= 0.5[7.5 + 7.375] \\ &= 7.4375 \end{aligned}$$

$$\boxed{v(s_2) = 7.4375}$$

## Iteration 4

$$v(s_1) = \max(a, b)$$

$$\begin{aligned} a &= 0.8(-8 + 0.5 \times 7.44) + 0.2(-8 + 0.5 \times 7) \\ &= -8 + 0.5[7.44 \times 0.8 + 0.2 \times 7] \\ &= -4.324 \end{aligned}$$

$$\begin{aligned} b &= 1 \times (4 + 0.5 \times 7) \\ &= 7.5 \end{aligned}$$

$$v(s_1) = 7.5$$

$$v(s_2) = \max(a, b)$$

$$\begin{aligned} a &= 0.9(-4 + 0.5 \times 7.44) + 0.1(-4 + 0.5 \times 7.5) \\ &= -4 + 0.5[0.9 \times 7.44 + 0.1 \times 7.5] \\ &= -0.277 \end{aligned}$$

$$\begin{aligned} b &= 0.5(4 + 0.5 \times 7.5) + 0.5(4 + 0.5 \times 7.44) \\ &= 4 + 0.5 \times 0.5 \times [15] \\ &= 7.735 \end{aligned}$$

$$v(s_2) = 7.735$$



$\gamma = 0.5$     $\delta = 0.2$   
Policy Iteration    $v(s_1) = 0$ ,  $v(s_2) = 0$ ,  $\pi(a_1|s_1) = 0.5$ ,  $\pi(a_2|s_1) = 0.5$   
Policy evaluation    $\pi(a_2|s_2) = 0.5$ ,  $\pi(a_1|s_2) = 0.5$

$$\begin{aligned}
 v_1(s_1) &= 0.5 \times [0.2(-8 + 0) + 0.8(-8 + 0)] \\
 &\quad + 0.5[1 \times (4 + 0)] \\
 &= 0.5[-8 + 2] = -3
 \end{aligned}$$

$$\begin{aligned}
 v_1(s_2) &= 0.5 \left[ 0.5 \times (4 + 0.5 \times (-3)) + 0.5 \times (4 + 0.5 \times (0)) \right] \\
 &\quad + 0.5[0.9 \times (-4 + 0.5 \times (0)) + 0.1(-4 + 0.5 \times (-3))] \\
 &= 0.5[1.25 + 2 - 3.6 - 0.55] \\
 &\quad - 0.5 = -0.45
 \end{aligned}$$

$$\begin{aligned}
 v_2(s_1) &= 0.5 \left[ 0.2 \times (-8 + 0.5 \times (-3)) + 0.8(-8 + 0.5 \times (-0.45)) \right] \\
 &= 0.5[-1.9 + 6.58] \\
 &\quad + 0.5[2.5] \\
 &= -2.99
 \end{aligned}$$

$$\begin{aligned}
 v_2(s_2) &= 0.5 \times [0.5 \times (4 + 0.5 \times (-2.99)) + 0.5(4 + 0.5 \times (-0.45))] \\
 &\quad + 0.5[0.1(-4 + 0.5 \times (-2.99)) + 0.9(-4 + 0.5 \times (-0.45))] \\
 &= 0.5[1.2525 + 1.8875 - 0.549 - 3.802] \\
 &= -0.60
 \end{aligned}$$

## Policy Improvement

$$\pi(s_i) = \operatorname{argmax}(a, b)$$

$$a = 0.2 \times (-8 + (-2.99)_{0.5}) + 0.8 (-8 + 0.5 \times -0.6)$$

$$b = 4 + 0.5 \times (-2.99)$$

$$\pi(s_1) = a_2$$

$$\pi(s_2) = \operatorname{max}(a_1, a_2)$$

$$\pi(s_2) = 0.5 \times (4 + 0.5 \times (-2.99)) + 0.5 (4 + 0.5 \times (-0.6))$$

$$\pi(s_2) = \operatorname{argmax}(a_1, a_2)$$

$$a = 0.5 (4 + 0.5 \times (-2.99)) + 0.5 (4 + 0.5 \times (-0.6))$$

$$b = 0.5 (-4 + 0.5 (-0.6)) + 0.1 (-4 + 0.5 (-2.99))$$

$$\pi(s_2) = a_2$$

We again go to the policy evaluation step, followed by the policy improvement step and conclude,

$$\left. \begin{array}{l} \pi(s_1) = a_2 \\ \pi(s_2) = a_2 \end{array} \right\} \text{Stays silent.}$$

14 To Prove → The Policy improvement step either improves the current policy or the current policy is optimal.

We prove it <sup>in</sup> 2 parts:

1) Policy improvement either improves the policy or leaves it unchanged

2) If the policy is unchanged, it is optimal.

Proof

1) We say  $\pi'$  is better than  $\pi$  if  $v_{\pi'} \geq v_{\pi} \forall s$   
we define  $T_{\pi}$  &  $T$  as

$$T_{\pi}(f(s)) = E_{\pi}[R_{t+1} + \gamma f(s_{t+1}) | s_t]$$

Let  $\pi' \neq \pi$  be

$$T(f(s)) = \max (R_{t+1} + \gamma f(s_{t+1}) | s_t, A_t)$$

$$T(f(s)) = \max_{a \in A(s)} E[R_{t+1} + \gamma f(s_{t+1}) | s_t = s, A_t = a]$$

Corollary

We know, for the policy improvement step

$$TV_{\pi_k}(s) = T_{\pi_{k+1}}(v_{\pi_k}) \quad \text{--- ①}$$

Since we pick the greedy action in  $TV_{\pi_k}$ , this would clearly be a better choice ~~as~~ than any other action since the ~~as~~ state value will become suboptimal.

$$\text{① } TV_{\pi_k} \Rightarrow T_{\pi_{k+1}} v_{\pi_k} \geq T_{\pi_k} v_{\pi_k} \quad \text{--- ②}$$

$$T_{\pi_k} v_{\pi_k} = v_{\pi_k} \quad \text{--- ③}$$

since  $v_{\pi_k}$  is the fixed point  
of  $T_{\pi_k} v_{\pi_k}$

From ①, ② & ③

$$T_{n_{k+1}} v_{n_k} \geq v_{n_k} \quad \text{--- } ④$$

We know that

$$T_{n_{k+1}} (T_{n_{k+1}} v_{n_k}(s)) \geq T_{n_{k+1}} v_{n_k} \quad \text{--- } ⑤$$

~~$T_{n_{k+1}} v_{n_k}$~~

From ④ & ⑤

$$T_{n_{k+1}} (T_{n_{k+1}} v_{n_k}(s)) \geq v_{n_k}(s)$$

Similarly repeating, we get

$$T_{n_{k+1}}^N (v_{n_k}(s)) \geq v_{n_k}(s) \quad \text{--- } ⑥$$

Here,  $v_{n_{k+1}}(s)$  is the fixed point for  $T_{n_{k+1}}$

$T_{n_{k+1}}^{N+1} (v_{n_k}(s))$  converges to  $v_{n_{k+1}}(s)$

$$\Rightarrow \lim_{N \rightarrow \infty} T_{n_{k+1}}^N (v_{n_k}(s)) = v_{n_{k+1}}(s) \quad \text{--- } ⑦$$

From ⑥ & ⑦

$$\Rightarrow v_{n_{k+1}}(s) \geq v_{n_k}(s)$$

Hence proved.

2) we need to prove,

if,  $T_{\pi_{k+1}} V_{\pi_k}(s) = V_{\pi_k}(s)$  — ⑧

then the policy is optimal.

$T_{\pi_{k+1}}$  has a unique fixed point  $V_{\pi_{k+1}}$

$$V_{\pi_k}(s) = V_{\pi_{k+1}}(s) \quad \forall s$$

Also,  $T_{\pi_{k+1}} V_{\pi_k}(s) = T_{\pi_{k+1}}(s)$  — ⑨

Thus from ⑧ & ⑨

$$V_{\pi_k}(s) = T_{\pi_{k+1}} V_{\pi_k}(s)$$

Then

$$\Rightarrow V_{\pi_k}(s) = \max_a E [R_{t+1} + \gamma V_{\pi_k}(s_{t+1}) | s_t = s, A_t = a]$$

Thus,  $V_{\pi_k}(s)$  is the optimal Policy.

