Exercise 3.4

1.

s	a	s'	r	$p(s', r \mid s, a)$
high	search	high	r_{search}	α
high	search	low	r_{search}	$1-\alpha$
low	search	high	-3	$1-\beta$
low	search	low	r_{search}	β
high	wait	high	1	r_{wait}
high	wait	low	0	0
low	wait	high	0	0
low	wait	low	1	r_{wait}
low	recharge	high	1	0
low	recharge	low	0	-

Exercise 3.15

3. $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

$G_{tc} = (R_{t+1} + c) + \gamma(R_{t+2} + c) + \gamma^2(R_{t+3} + c) + \cdots$

$= \left(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots\right) + c\left[1 + \gamma + \gamma^2 + \cdots\right]$

$= G_t + c \cdot \frac{1}{1-\gamma}$

$v_\pi(s) = E\left[G_t \mid S_t = s\right]$

$v_{\pi_c}(s) = E\left[G_{tc} \mid S_t = s\right]$

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

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

$v_{\pi_c}(s) = v_\pi(s) + \frac{c}{1-\gamma}$

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

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

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

3.16

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

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

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

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

$$v_\pi(s) = E\left[G_t \mid S_t = s\right]$$

$$v_{\pi_c}(s) = E\left[G_{t_c} \mid S_t = s\right]$$

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

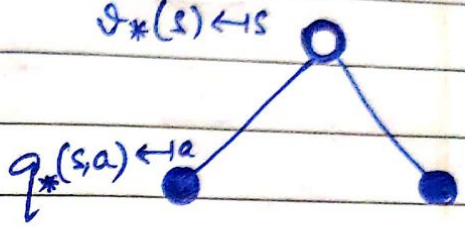
$$= E\left[G_t \mid S_t = s\right] + c\frac{(1-\gamma^k)}{(1-\gamma)}$$

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

The state value 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

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



Instead of taking the expected action value for a state, we choose the maximum action value using $q_*(s,a)$

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

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

a' is an arbitrary action taken at time $t+1$.

~~r is unaffected by s', a' and r'~~

$P[R_{t+2} = r \mid S_t = s, A_t = a]$ is unaffected by a' & s'

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

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

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

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

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

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

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

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

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

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

using the expression in previous question.

10 $v_\pi(s) = E[G_t \mid S_t = s]$

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

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

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

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

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

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

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

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

$G_3 = R_4 = -3$

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

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

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

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

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$$

$$= c + \gamma c + \gamma^2 c + \ldots$$

$$= c(1 + \gamma + \gamma^2 + \ldots)$$

$$= \frac{c}{1 - \gamma}$$

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

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

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

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

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

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

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

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

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

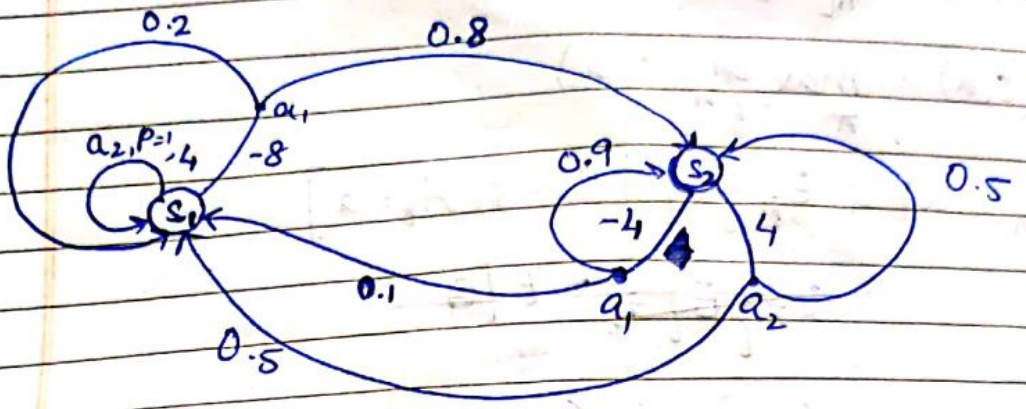
$$q^*(s,a) = \max_a E[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a]$$

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

13

State
s_1 → stale
s_2 → fresh

Action
a_1 → query
a_2 → remains silent



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

$v_0(s_1) = 0$, $v(s_2) = 0$
$\gamma = \frac{1}{2}$

Iteration-1

s_1

$v_1(s_1) = \max(a, b)$

$a = \sum_{s',r} p(s', r|s, a_1)[r + \gamma v_0(s')]$

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

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

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

~~$v(s_1) = 0$~~ ~~$b = \frac{1}{2}$~~

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

$\Rightarrow v(s_1) = 4$

s_2

$v(s_2) = \max(a, b)$ → action a_1, → action a_2

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

$$= 0.9 \times \left[-4 + 0.5 \times 0\right] + 0.1\left[-4 + 0.5 \times (4)\right]$$

$$= -3.6 + 0.1\left[-2\right]$$

$$= -3.8$$

$$b = 0.5\left[4 + 0.5 \times 0\right] + 0.5\left[4 + 0.5 \times (4)\right]$$

$$= 2 + 3 = 5$$

$v(s_2) = 5$

Iteration - 2

$v(s_1) = (a, b)$ → a_1, → a_2

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

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

$$= 0.2\left[-6\right] + 0.8\left[-8 + 2.5\right]$$

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

$$b = 1 \times \left[4 + 0.5(4)\right] = 6$$

$v(s_1) = 6$

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

$a = 0.9 \times [-4 + 0.5 \times 5] + 0.1[-4 + 0.5 \times (6)]$

$= 0.9 \times (-1.5) + 0.1 \times (-1)$

$= -1.35 - 0.1$

$= -1.45$

$b = 0.5 \times (4 + 0.5 \times 6) + 0.5(4 + 0.5 \times 5)$

$= 0.5(7) + 0.5(6.5)$

$= 6.75$

$v(s_2) = 6.75$

~~t=3~~ Iteration 3

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

$a = 0.8 \times (-8 + 0.5 \times 6.75) + 0.2 \times (-8 + 0.5 \times 6)$

$= -4.7$

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

$= 7$

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

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

$a = 0.9 \times (-4 + 0.5 \times 6.75) + 0.1 \times (-4 + 0.5 \times 7)$

$= -4 + 0.5 \times 6.775$

$= -0.6125$

$b = 0.5(4 + 0.5 \times 7) + 0.5(4 + 0.5 \times 6.75)$

$= 0.5[7.5 + 7.375]$

$= 7.4375$

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

Iteration 4

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

$a = 0.8 \times (-8 + 0.5 \times 7.44) + 0.2(-8 + 0.5 \times 7)$

$= -8 + 0.5[7.44 \times 0.8 + 0.2 \times 7]$

$= -4.324$

$b = 1 \times (4 + 0.5 \times 7)$

$= 7.5$

$v(s_1) = 7.5$

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

$a = 0.9(-4 + 0.5 \times 7.44) + 0.1(-4 + 0.5 \times 7.5)$

$= -4 + 0.5[0.9 \times 7.44 + 0.1 \times 7.5]$

$= -0.277$

$b = 0.5(4 + 0.5 \times 7.5) + 0.5(4 + 0.5 \times 7.44)$

$= 7.735$

$v(s_2) = 7.735$

Policy Iteration

$\gamma = 0.5 \quad \theta = 0.2$

$v(s_1) = 0, \; v(s_2) = 0, \; \pi(a_1|s_1) = 0.5, \; \pi(a_2|s_1) = 0.5$

$\pi(a_2|s_2) = 0.5, \; \pi(a_1|s_2) = 0.5$

Policy evaluation

$$v_1(s_1) = 0.5 \times [0.2(-8+0) + 0.8(-8+0)]$$
$$+ 0.5 [1 \times (4+0)]$$

$$= 0.5 [-8+2] = -3$$

$$v_1(s_2) = 0.5 [0.5 \times (4 + 0.5 \times (-3)) + 0.5 \times (4 + 0.5 \times (0))]$$
$$+ 0.5 [0.9 \times (-4 + 0.5 \times (0)) + 0.1 (-4 + 0.5 \times (-3))]$$

$$= 0.5 [1.25 + 2 - 3.6 - 0.55]$$
$$= -0.45$$

$$v_2(s_1) = 0.5 [0.2 \times (-8 + 0.5(-3)) + 0.8 (-8 + 0.5 \times (-0.45))]$$
$$= 0.5 [-1.9 - 6.58]$$
$$+ 0.5 [2.5]$$
$$= -2.99$$

$$v_2(s_2) = 0.5 \times [0.5 \times (4 + 0.5 \times (-2.99)) + 0.5 (4 + 0.5 \times (-0.4$$

$$+ 0.5 [0.1 (-4 + 0.5 (-2.99)) + 0.9 (-4 + 0.5 \times (-0.45$$

$$= 0.5 [1.2525 + 1.8875 - 0.549 - 3.802]$$

$$= -0.60$$

Policy Improvement

$\pi(s_1) = \text{argmax}(a, b)$ ($a \to a_1$, $b \to a_2$)

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

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

$\pi(s_1) = a_2$

~~$\pi(s_2) = \max(a, a$~~

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

$\Rightarrow$

$\pi(s_2) = \text{argmax}(a, b)$ ($a \to a_2$, $b \to a_1$)

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

$b = 0.9(-4 + 0.5(-0.6)) + 0.1(-4 + 0.5(-2.99))$

$\Rightarrow$

$\pi(s_2) = a_2$

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

$\pi(s_1) = a_2$

$\pi(s_2) = a_2$

stays silent.

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

We prove it in 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 Π' is better than π if $v_{\pi'} \geq v_{\pi}$ $\forall s$

We define T_π & T as

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

~~Let $\pi' \in \Pi$ be~~

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

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

We know, for the policy improvement step

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

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

$$\Rightarrow T v_{\pi_k} \geq T_{\pi_k} v_{\pi_k} \quad \text{——} (2)$$

$$T_{\pi_k} v_{\pi_k} = v_{\pi_k} \quad \text{——} (3)$$

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

From ①, ② & ③

$$T_{\pi_{k+1}} v_{\pi_k} \geq v_{\pi_k} \quad \text{——— ④}$$

We know that

$$T_{\pi_{k+1}} \left(T_{\pi_{k+1}} v_{\pi_k}(s) \right) \geq T_{\pi_{k+1}} v_{\pi_k} \quad \text{——— ⑤}$$

From ④ & ⑤

$$T_{\pi_{k+1}} \left(T_{\pi_{k+1}} v_{\pi_k}(s) \right) \geq v_{\pi_k}(s)$$

Similarly repeating, we get

$$T^{N}_{\pi_{k+1}} \left(v_{\pi_k}(s) \right) \geq v_{\pi_k}(s) \quad \text{——— ⑥}$$

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

$T^{N}_{\pi_{k+1}} (v_{\pi_k}(s))$ converges to $v_{\pi_{k+1}}(s)$

$$\Rightarrow \lim_{N \to \infty} T^{N}_{\pi_{k+1}} \left(v_{\pi_k}(s) \right) = v_{\pi_{k+1}}(s) \quad \text{——— ⑦}$$

From ⑥ & ⑦

$$\Rightarrow v_{\pi_{k+1}}(s) \geq v_{\pi_k}(s)$$

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

Thus from (8) & (9)

$$v_{\pi_k}(s) = T v_{\pi_k}(s)$$

$$\Rightarrow v_{\pi_k}(s) = \max E\left[R_{t+1} + \gamma v_{\pi_k}(s_{t+1}) \mid s_t = s, A_t = a\right]$$

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