Question 1

s	a	s'	r	$r(s,a,s')$	$p(s',r \mid s,a)$	$p(s',r \mid s,a)$
high	search	high	10	r_s	α	$+0.1 \alpha \cdot r_s$
high	search	low	-10	r_s	$1-\alpha$	$-0.1(1-\alpha) r_s$
low	search	high	0	-3	$1-\beta$	0
low	search	low	-10	r_s	β	$-0.1 \beta r_s$
high	wait	high	10	r_w	1	$0.1 r_w$
high	wait	low	10	r_w	0	–
low	wait	high	0	–	0	–
low	wait	low	10	r_{wait}	1	$0.1 r_w$
low	rech.	high	0	0	1	–
low	rech.	low	0	–	0	–

Since $r(s,a,s') = \dfrac{\sum r \, p(s',r \mid s,a)}{p(s' \mid s,a)}$

① $r_s = \dfrac{10 \cdot p(s',r \mid s,a)}{p(\text{high} \mid \text{high}, \text{search})} \rightarrow \alpha$

$p(s',r \mid s,a) = \alpha \cdot 0.1 \cdot r_s$

② $r_s = \dfrac{-10}{(1-\alpha)} \, p(s',r \mid s,a)$

$p(\text{low}, -10 \mid \text{high}, \text{search}) = -0.1 \cdot (1-\alpha) r_s.$

Question 2 - Given the gridworld environment,

Solving the Bellman equation, $v_\pi(s)$:

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

Solving for state 1 for illustration -

$$\begin{aligned} v_\pi(1) = {} & 0.25\times 1\,[-1 + \gamma v_\pi(1)] + \\ & 0.25\times 1\,[-1 + \gamma v_\pi(1)] + \\ & 0.25\times 1\,[0 + \gamma v_\pi(2)] + \\ & 0.25\times 1\,[0 + \gamma v_\pi(4)] \end{aligned}$$

Similarly

$$\begin{aligned} v_\pi(2) = {} & 0.25\times 1\,[-1 + \gamma v_\pi(2)] + \\ & 0.25\times 1\,[0 + \gamma v_\pi(1)] + \\ & 0.25\times 1\,[0 + \gamma v_\pi(3)] + \\ & 0.25\times 1\,[0 + \gamma v_\pi(4)] \end{aligned}$$

Thus

$$\underbrace{\begin{bmatrix} v_\pi(1) \\ v_\pi(2) \\ \vdots \\ v_\pi(25) \end{bmatrix}}_{25\times 1} = \underbrace{\begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_{25} \end{bmatrix}}_{25\times 1} + \gamma \underbrace{\underbrace{P_\pi}_{25\times 25}\;\underbrace{v_\pi}_{25\times 1}}_{25\times 1}$$

25 → Total no. of states

$$\therefore\; v_\pi = R_\pi + \gamma P_\pi v_\pi$$

$$v_\pi - \gamma P_\pi v_\pi = R_\pi$$

$$\boxed{v_\pi = (I - \gamma P_\pi)^{-1} R_\pi} \quad \forall\, s \in S.$$

Question 3

Exercise 3.15

$$v_\pi(s) = E_\pi[R_t \mid S_t = s]$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid S_t = s\right]$$

Adding a constant to each reward -

Let $\hat{r}_{t+k+1} = r_{t+k+1} + c$

$$\hat{v}_\pi(s) = E_\pi[\gamma^k \hat{r}_{t+k+1} \mid S_t = s]$$

$$= E_\pi[\gamma^k r_{t+k+1} \mid S_t = s] + E_\pi[\gamma^k c \mid S_t = s]$$

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

Thus we see the best policy doesn't change with addition of constant (c) at each reward.

Exercise 3.16

Adding a constants at episodic tasks:

$$v_\pi(s) = E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots + \gamma^{k-1} R_{t+k}]$$

$$\hat{v}_\pi(s) = E[(R_{t+1} + c) + \gamma(R_{t+2} + c) + \dots + \gamma^{k-1}(R_{t+k} + c)]$$

$$= c[1 + \gamma + \gamma^2 + \dots + \gamma^{k-1}] + v_\pi(s)$$

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

$k \to$ small value, (A) (A) is closer to terminal state, thus adding a constant only ~~Adds~~ Prolongs / reduces the steps to terminal states.

Notes	Appointment

FRIDAY 3 MAY 2019

Question 4 Solving the Bellman optimality eq^n;

$$v^*(s) = \max_a \sum_r \sum_{s'} (r + \gamma v_*(s'))\, p(s', r | s, a)$$

Thus,

$$v^*(s) \geq R(s) + \gamma \max_{a \in A} \sum p(s'|s,a)\, v(s')$$

Thus taking $|A|$ as the linear constraints;

$$v(s) \geq R(s) + \gamma \sum_{s' \in S} p(s'|s,a)\, v(s') \quad \forall a \in A$$

Thus, $\underset{v}{\text{minimise}} \sum_s v(s)$

subject to $v(s) \geq R(s) + \gamma \sum_{s' \in S} p(s'|s,a)\, v(s') \; \forall a \in A, s \in A$

Solving for optimal policy,

if we optimise,

$$\min_v \sum_s d(s)\, v(s)$$

subject to $v(s) \geq R(s) + \gamma \sum_{s' \in S} p(s'|s,a)\, v(s') \quad \forall a \in A, s \in S$

Here $d(s)$ is the distribution over states.

Adding $\mu(s,a)$ for each constraint,

$$\underset{\mu(s,a)}{\text{maximise}} \sum_{s \in S} R(s) \sum_{a \in A} \mu(s,a)$$

Subject to $\sum_{a \in A} \mu(s', a) = d(s') + \sum_{s \in S} \sum_{a \in A} p(s'|s,a)\, \mu(s,a) \quad \forall s' \in S.$

where $\mu(s,a) = \sum_{t=0}^{\infty} \gamma^t P(S_t = s, A_t = a)$

Thus $\pi^*(s) = \max_{a \in A} \mu(s,a)$

Notes	Appointment

Question 5

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

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

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

$$= \max_a \sum_r \sum_{s'} (r + \gamma v_*(s'))\, p(s', r \mid s, a)$$

Thus,

$$\boxed{v_*(s) = \max_a q_*(s,a)}$$

value of each state when taken an action a.

Question 6

The bug described in question 4.4 to find a way such that policy doesn't keep on switching in case of finding multiple policies

In this case,

pseudo code

if $v_\pi(s) << \theta$:

don't update the $v_\pi(s)$

θ increase by a small value to break 'continuous update of states.