(1) Probability of picking an even arm = $2x$

" " " " odd arm = x

No. of even arms = 5

" " odd arms = 5

$\Rightarrow$

$5(2x) + 5(x) = 1$

$\Rightarrow x = \frac{1}{15}$

R_a: Reward drawn upon choosing arm a.

(RV)

$E[R_a | a=i] = i$

$E[R_a] = \sum_i R_{a=i} \cdot P(a=i)$

$= \frac{2}{15} \times \frac{1}{15} + \frac{2}{15} \times 2 \times \frac{1}{15}$

$= \frac{1}{15}[1+3+5+7+9] + \frac{2}{15}[2+4+6+8+10]$

$= \frac{1}{15}[1+3+5+7+9+4+8+12+16+20]$

$= \frac{1}{15}[85] = \frac{85}{15} = \frac{17}{3}$

As each picking of arm is different from one another and the distribution from which reward is drawn is same. $\therefore$ Rewards at different instants are iid Random variables.

$\Rightarrow E\left[\sum_{t=1}^{10} R_a\right] = 10 \sum E[R_a]$

$= \frac{170}{3}$

(2) $S_1 = \{1, 2, 4, 5, 7, 9, 10\}$

$S_2 = \{3, 6, 8\}$.

R_a : Reward drawn upon choosing arm a.

$E_1 = E[R_a | a \in S_1] = P(0) \times 0 + P(1) \times 1$

$= 0.5 \times 0 + 0.5 \times 1$

$= 0.5$

$E_2 = E[R_a | a \in S_2] = P(0) \times 0 + P(0.2) \times 0.2 + P(1) \times 1$

$= 0.2 \times 0.3 + 1 \times 0.4$

$= 0.06 + 0.4 = 0.46$

as $E_1 > E_2$

optimal action must be set S_1.

→ Policy 1.

$$\pi_1(a) = \begin{cases} 0.2 & a=1 \\ 0.8 & a=2 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi_2(a) = \begin{cases} 0.3 & a=4 \\ 0.7 & a=5 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi_3(a) = \begin{cases} 0.5 & a=9 \\ 0.5 & a=2 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi_4(a) = \begin{cases} 0.2 & a=4 \\ 0.2 & a=5 \\ 0.6 & a=10 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi_5(a) = \begin{cases} 0.8 & a=9 \\ 0.1 & a=4 \\ 0.1 & a=1 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi_6(a) = \begin{cases} 0.1 & a=7 \\ 0.9 & a=5 \\ 0 & \text{otherwise} \end{cases}$$

(3) Initial estimates

$Q(1) = 1$

$Q(2) = 2$

$Q(3) = 3$

$\varepsilon = 0.5$

$\{0, 1\}$.

t=1 Explore

$a = 2$

$R(a=2) = 0$

$Q(2) = 1$

$Q(1) = 1$
$Q(2) = 1$
$Q(3) = 3$

t=2 Exploit

$a = 3$

~~$Q(3) = 0$~~ $R(3) = 0$

$Q(3) = 1.5$

$Q(1) = 1$
$Q(2) = 1$
$Q(3) = 1.5$

t=3 Explore

$a = 1$

$R(1) = 1$

$Q(1) = 1$

$Q(1) = 1$
$Q(2) = 1$
$Q(3) = 1.5$

t=4 Exploit

$a = 3$

$R(3) = 0$

$Q(3) = 1$

$Q(1) = 1$
$Q(2) = 1$
$Q(3) = 1$

t=5 Explore

$a = 2$

$R(2) = 1$

$Q(2) = 1$

$Q(1) = 1$
$Q(2) = 1$
$Q(3) = 1$

t=6 Exploit

$a = 1$

$R(a=1) = 0$

$Q(1) = 0.667$

$Q(1) = 0.667$

$Q(2) = 1$

$Q(3) = 1$

(4) Current state : S

Next state : S'

action taken : a

reward : r

S	a	S'	r	$p(s', r \mid s, a)$
h	s	h	r_s	α
h	s	L	r_s	$1-\alpha$
h	w	h	r_w	1
h	w	L	-	0
L	s	H	r_p	$1-\beta$
L	s	L	r_s	β
L	w	H	-	0
L	w	L	r_w	1
L	re	H	0	1
L	re	L	-	0

States : {high, low}

A(high) = {search, wait}

A(low) = {search, wait, recharge}

ACRONYMS:

high → h

low → L

search → s

wait → w

recharge → re

reward of search → r_s

" " wait → r_w

reward penalty → r_p

DATE
PAGE

3.15

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

$$G_t' = \sum_{k=0}^{\infty}$$

Adding a constant c.

$$G_t' = \sum_{k=0}^{\infty} \gamma^k (R_{k+t+1} + c)$$

$$G_t' = \sum_{k=0}^{\infty} \gamma^k R_{k+t+1} + \sum_{k=0}^{\infty} \gamma^k c \quad ; \gamma < 1$$

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

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

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

$$= \mathop{E}_{\pi}\left[G_t + \frac{c}{1-\gamma} | S_t = s\right]$$

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

Value function will change but new policy will stay same when a constant gets added to all the rewards.

3.16

An episodic will end at some finite Time T.

$$G_t = \sum_{k=0}^{T} \gamma^k R_{k+t+1}$$

assuming $\gamma = 1$

$$G_t = \sum_{k=0}^{T} R_{k+t+1}$$

$$G'_t = \sum_{k=0}^{T} (R_{k+t+1} + C)$$

$$= \sum_{k=0}^{T} R_{k+t+1} + \sum_{k=0}^{T} C$$

$$G_t' = G_t + cT$$

$$V_\pi(s) = E_\pi[G_t | S_t = s]$$

$$V'_\pi(s) = E_\pi[G'_t | S_t = s]$$

$$= E_\pi[G_t + cT | S_t = s]$$

$$= E_\pi[G_t | S_t = s] + C E[T | S_t = s]$$

Addition of a constant in Rewards of an episodic task would results in drastic changes in value function.

As T is Random variable, its expectation would change from problem to problem & episode to episode.

This might even result in change policy.

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

(11) R_{t+2} is conditioned on S_t, A_t if S_{t+1} is not taken into consideration.

$$P(R_{t+2}=r'' \mid S_t=s, A_t=a)$$

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

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

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

(MDP Assumption)

$$P(R_{t+2}=r'' \mid S_t=s, A_t=a)$$

$$= \sum_{s''} P(R_{t+2}=r'', S_{t+2}=s'' \mid S_t=s, A_t=a)$$

$$P(R_{t+2}=r'', S_{t+2}=s'' \mid S_t=s, A_t=a) =$$

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

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

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

$$P(R_{t+2}=r'' \mid S_t=s, A_t=a) = \sum_{s''}\sum_{s'}\sum_{a'} p(r'',s'' \mid s',a')\,\pi(a' \mid s')\,p(s' \mid s,a)$$

(12)

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

$$\sum_{s''}\sum_{s}\sum_{a} p(r'',s'' \mid s',a')\,\pi(a' \mid s')\,p(s' \mid s,a)\,[R_{t+2}]$$

(13)

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

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

$$= \sum_{s}\sum_{a}\sum_{g} [R_{t+1} + \gamma G_{t+1}]\, P(s \mid s)\, P(G_{t+1}=g \mid S_t=s)$$

$$= \sum_{s}\sum_{g} [s + \gamma g]\, P(s,g \mid s)$$

$$= \sum_{s',s,a,g} [s+\gamma g]\, P(s,g,s',a \mid s)$$

$$= \sum_{s',s,a,g} [s + \gamma g]\, P(s,g,s' \mid s,a)\,\pi(a \mid s)$$

$$= \sum_{a}\pi(a \mid s) \sum_{s',r} p(s,s' \mid s,a)\Big[s + \gamma \sum_{g} g\, P(g \mid s',s,a)\Big]$$

$$= \sum_{a}\pi(a \mid s) \sum_{s',r} P(s,s' \mid s,a)\Big[s + \gamma \sum_{g} g\, P(g \mid s')\Big]$$

$$v_k(s) = \sum_{a}\pi(a \mid s) \sum_{s',r} P(s,s' \mid s,a)\,[s + \gamma\, v_k(s')]$$

(14) $R_1 = 2,\ R_2 = -1,\ R_3 = 10,\ R_4 = -3$

$\gamma = 0.5$

$G_4 = 0$

$G_3 = -3 + 0.5 \times 0 = -3$

$G_2 = 10 + 0.5(-3) = 8.5$

$G_1 = -1 + 0.5 \times 8.5 = 3.25$

$G_0 = 2 + 0.5 \times 3.25 = 3.625$

$R_{t+k+1} = C$ (Constant Reward)

$$G_t = \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1})$$

$$= \sum_{k=0}^{\infty} \gamma^k (C) = C \sum_{k=0}^{\infty} \gamma^k$$

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

(15) $V_*(s)$ ← given $\forall s \in S$

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_a \sum_{s',r} p(s', r|s, a)(r + \gamma V_*(s')) \\ 0 & \text{otherwise} \end{cases}$$

$$\pi_*(a|s) = \begin{cases} 1, & \text{if } a = \arg\max_a q^*(s, a) \\ 0, & \text{otherwise} \end{cases}$$

(16) Current State : s
Next State : s'
Reward : r
Action : a

States = {fresh, stale}
action = {query, silent}

s	a	s'	r	P(s', r | s, a)
Fresh	query	fresh	-4	0.9
Fresh	query	stale	-4	0.1
Fresh	silent	Fresh	4	0.5
Fresh	silent	Stale	4	0.5
Stale	query	Fresh	-8	0.8
Stale	query	Stale	-8	0.2
Stale	silent	Fresh	4	0
Stale	silent	stale	4	1

$V_\pi(\text{fresh}) \equiv V_F$

$V_\pi(\text{stale}) \equiv V_S$

$$V_F = \frac{1}{2}[0.9(-4 + 0.5V_F) + 0.1(-4 + 0.5V_S)] + \frac{1}{2}[0.5(4 + 0.5V_F) + 0.5(4 + 0.5V_S)]$$

$$= \frac{1}{2}[-4 + 0.45V_F + 0.05V_S + 4 + 0.25V_F + 0.25V_S]$$

$$= \frac{1}{2}[0.7V_F + 0.3V_S] \Rightarrow 2V_F = 0.7V_F + 0.3V_S$$

$$\Rightarrow \boxed{1.3V_F = 0.3V_S}$$

$$V_S = 0.5[0.8(-8 + 0.5V_F) + 0.2(-8 + 0.5V_S) + 1(4 + 0.5V_S)]$$

$$2V_S = -8 + 0.4V_F + 0.1V_S + 0.5V_S$$

$$2V_S = -8 + 0.4V_F + 0.6V_S$$

$$1.4V_S = -8 + 0.4V_F = -8 + 0.092V_S$$

$$V_S = (-8)/(1.307) = -6.1176, \quad V_F = -1.4117$$

(17) Policy improvement theorem.

$\pi'(s)$ is the improved policy over $\pi(s)$.

$\pi'(s)$ is same as $\pi(s)$, except $\pi'(s) = a \neq \pi(s)$

$$V_\pi(s) \leq q_\pi(s, \pi'(s))$$

$$= \mathbb{E}[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, A_t = \pi'(s)]$$

$$= \mathbb{E}_{\pi'}[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s]$$

$$\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s]$$

$$= \mathbb{E}_{\pi'}[R_{t+1} + \gamma \mathbb{E}_{\pi'}[R_{t+2} + \gamma V_\pi(S_{t+2}) \mid S_{t+1}, A_{t+1} = \pi'(S_{t+1})] \mid S_t = s]$$

$$= \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 V_\pi(S_{t+3}) \mid S_t = s]$$

$$\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \mid S_t = s]$$

$$= V_{\pi'}(s)$$

$$\pi'(s) = \arg\max_a q_\pi(s, a)$$

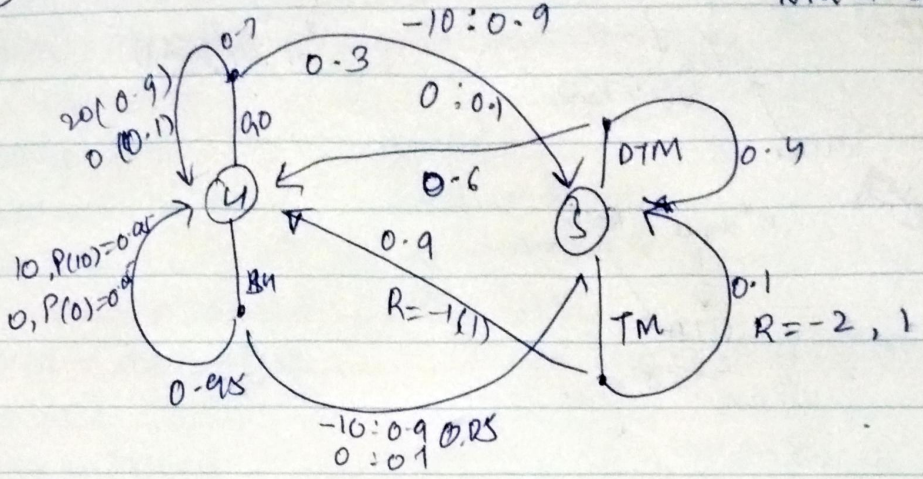
$$V_{\pi'}(s) = \max_a \mathbb{E}[R_{t+1} + \gamma V_{\pi'}(S_{t+1}) \mid S_t = s, A_t = a]$$

$$= \max_a \sum_{s', r} p(s', s \mid s, a)[r + \gamma V_{\pi'}(s')]$$

if $\pi'(s)$ is not better than $\pi(s)$, then $\pi'(s)$ must be optimal as it satisfies bellman optimality equation, so $\pi(s)$ will be optimal.

(4)

M ← ~~Mbook~~ Medicine
NM ← didn't took Medicine



S	a	s'	$\hat{s}$	P(s', $\hat{s}$ | s, a)
healthy	Stay home	healthy	10	0.9025
healthy	stay home	healthy	0	0.0475
healthy	stay home	sick	-10	0.045
healthy	stay home	sick	0	0.05
healthy	Go out	healthy	20	0.63
healthy	Go out	healthy	0	0.07
healthy	Go out	sick	-10	0.27
healthy	Go out	sick	0	0.03
sick	~~take~~ M	healthy	-1	0.9
sick	M	sick	-2	0.1
sick	NM	healthy	0	0.6
sick	NM	~~healthy~~ sick	0	0.4