A11) We can write

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

which can be simplified as :

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

Hence R_{t+2} is dependent on S_t, A_t

~~Although we can also write R_{t+2} as~~

A12) $E[R_{t+2}|s,a] =$

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

A13) $v_\pi(s) = E[G_t \mid S_t = s]$

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

$$= \sum_a E[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] \cdot \pi(a|s)$$

$$= \sum_{a,r,s'} E[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a, S_{t+1} = s', R_{t+1} = r] \cdot \pi(a|s)\, p(r, s' \mid s, a)$$

$$= \sum_{a,r,s'} \left(E[R_{t+1} \mid S_t = s, A_t = a, S_{t+1} = s', R_{t+1} = r] + \gamma E[G_{t+1} \mid S_t = s, A_t = a, S_{t+1} = s', R_{t+1} = r] \right) \cdot \pi(a|s)\, p(r, s' \mid s, a)$$

$$v_\pi(s) = \sum_{a,r,s'} \pi(a|s)\, p(r, s' \mid s, a) \cdot \left\{ r + \gamma v_\pi(s') \right\}$$

A 14)

t	1	2	3	4
R_t	2	-1	10	-3
G_t	3.625	3.25	8.5	-3

We know that

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

Now $R_t = c \quad \forall \; t$

$$G_t = c \sum_{T=0}^{\infty} \gamma^T$$

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

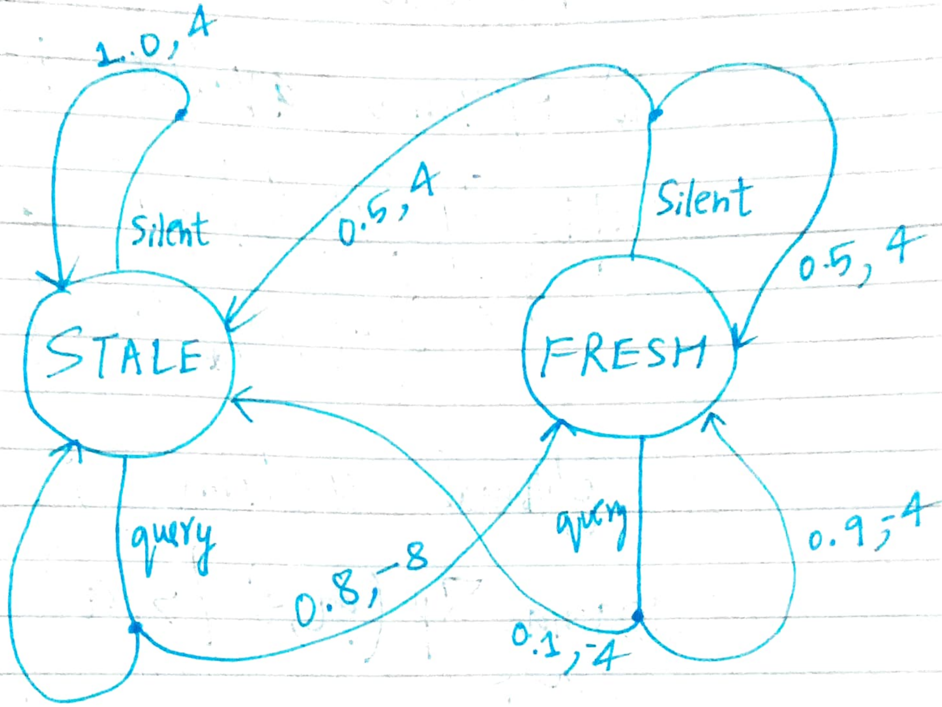
A15) We will chose the action that maximizes the expected return.

$$\pi_*(s) = \underset{a}{\operatorname{argmax}}\, E[G_t \mid S_t = s, A_t = a]$$

$$= \underset{a}{\operatorname{argmax}} \sum_{r,s'} p(r, s' \mid s, a) \cdot \{ r + V_*(s')\, \gamma \}$$

A1b)

1



2

$\gamma = 0.5$

reward_stale = -10.0

reward_fresh = +10.0

$T = 3$

value(a^*)	$t=0$	$t=1$	$t=2$	$t=3$
stale	5.75(s)	3.5 (s)	-1.0(s)	-10.0
fresh	6.3 (s)	4.75(s)	4.0(s)	10.0

Hence optimal policy is to stay silent always and never query.

Steps: i) Use dynamic programming, start calculation with t=3 backwards

ii) Optimal value is

$$\max_a \sum_{r,s'} p(r,s'|s,a) \cdot \{r + V_*(s')\}$$

Optimal action is

$$\operatorname{argmax}_a \sum_{r,s'} p(r,s'|s,a) \cdot \{r + V_*(s')\}$$

(3) Value iteration :

k	v(stale)	v(fresh)
0	0.0	0.0
1	4.0	4.0
2	6.0	6.0
3	7.0	7.0
4	7.5	7.5

Using $V_{k+1}(s) = \max_a \sum_{r,s'} p(r,s'|s,a) \cdot \{r + \gamma V_k(s')\}$

Policy iteration

Initially, $v(\text{stale}) = v(\text{fresh}) = 0$

$\pi(\text{stale}) = \text{silent}$

$\pi(\text{fresh}) = \text{silent}$

iteration 1

evaluation:

$v(\text{stale}) = 4 + 0.5\, v(\text{stale})$

$\Rightarrow \boxed{v(\text{stale}) = 8}$

$$v(\text{fresh}) = 0.5\{4 + 0.5\, v(\text{stale})\} + 0.5\{4 + 0.5\, v(\text{fresh})\}$$

$$v(\text{fresh}) = 8$$

policy ~~iter~~ improvement:

$\pi(\text{fresh}) = \text{silent}$

$\pi(\text{stale}) = \text{silent}$

policy is stable so all iterations ahead are same.

Used Bellman Equations for evaluation and $\pi(s) = \arg\max_a \sum_{r,s'} p(r, s' | s, a) \{r + \gamma v(s')\}$.

for improvement