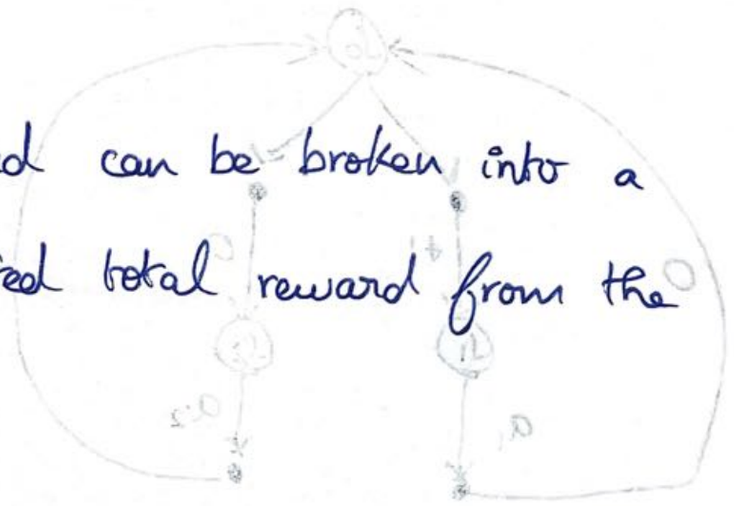
Name: Hithesh Shanmugam

Assignment 3

CSC 580 – Artificial Intelligence II

Problem 1: Exercise 3.17



The expected value of the total reward can be broken into a sum of immediate reward and expected total reward from the next state.

Using the definition of $q_\pi(s,a)$ with E_π we can write the equation as

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

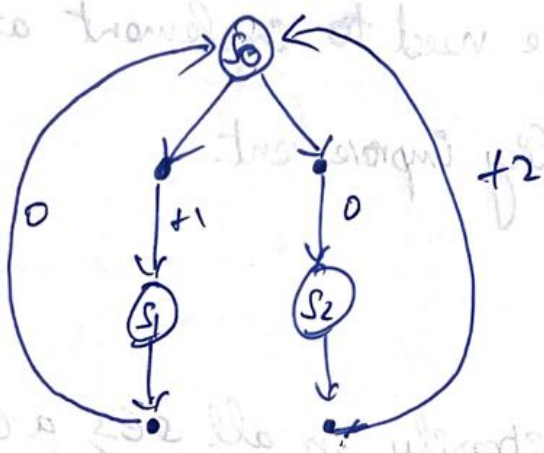
$$q_\pi(s,a) = E_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a]$$

The second equation is after substituting $G_t = R_{t+1} + \gamma G_{t+1}$

Replacing the E_π and we get

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

Problem 2: Exercise 3.22



We can derive the equations of $G_{\pi left}$ and $G_{\pi right}$ by the following equations

$$G_{\pi left} = \sum_{i=0}^{\infty} \gamma^{2i} \quad , \quad G_{\pi right} = \sum_{i=0}^{\infty} 2\gamma^{1+2i}$$

$$G_{\pi left} = \frac{1}{1-\gamma^2} \quad , \quad G_{\pi right} = \frac{2\gamma}{1-\gamma^2}$$

Now for $\gamma = 0.9, 0.5, 0$ we need to find which is optimal,

$\gamma = 0.9$ (Case 1)

$$G_{\pi left} = \frac{1}{1-(0.9)^2} \qquad G_{\pi right} = \frac{2(0.9)}{1-(0.9)^2}$$

$$= \frac{1}{1-0.81} \qquad = \frac{1.82}{1-0.81}$$

$$= \frac{1}{0.19} \qquad = \frac{1.82}{0.19}$$

$$G_{\pi left} \simeq 5.26 \qquad G_{\pi right} \simeq 9.58$$

$G_{\pi right} > G_{\pi left}$

So In this case $G_{\pi right}$ is the optimal

$\gamma = 0.5$ (Case 2)

$$G_{\pi left} = \frac{1}{1-(0.5)^2} = \frac{1}{1-0.25} = \frac{1}{0.75}$$

$$G_{\pi left} \simeq 1.34$$

$$G_{\pi right} = \frac{2(0.5)}{1-(0.5)^2} = \frac{1}{1-0.25} = \frac{1}{0.75}$$

$$G_{\pi right} \simeq 1.34$$

$G_{\pi right} = G_{\pi left}$ so in this case both are optimal

$\gamma = 0$ (Case 3)

$$G_{\pi left} = \frac{1}{1-(0)^2} = \frac{1}{1}$$

$$G_{\pi left} = 1$$

$$G_{\pi right} = \frac{2(0)}{1-(0)^2} = \frac{0}{1}$$

$$G_{\pi right} = 0$$

$G_{\pi left} > G_{\pi right}$

In this case the $G_{\pi left}$ is the optimal solution.

Problem 3: Exercise 4.1

$\gamma = 1$

0.0	-14	-20	-22
-14	-18	-20	-20
-20	-20	-18	-14
-22	-20	-14	0.0

→ 7

→ 11

→ 7 down

→ 11 down

The policy given here is equiprobable random policy

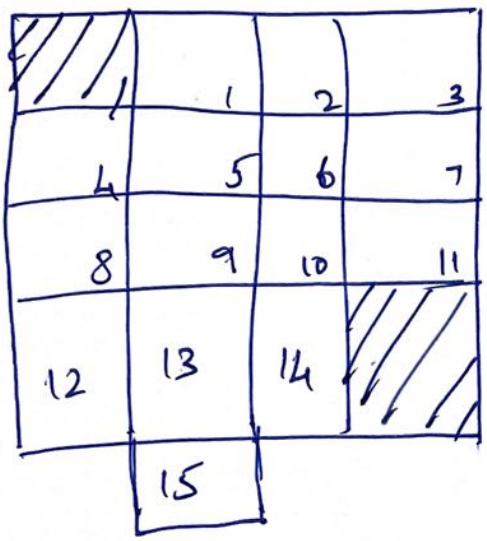
If we apply that

$$q_\pi(11, \text{down}) = -1 + v_\pi(T) = -1 + 0 \quad \text{(from the table)}$$
$$= -1$$

$$q_\pi(7, \text{down}) = -1 + v_\pi(11) = -1 + (-14) \quad \text{(from the table)}$$
$$= -15$$

$q_\pi(11, \text{down}) = -1$, $q_\pi(7, \text{down}) = -15$

Problem 4 Exercise 4.2



Adding the state 15 to the bottom of the state 13 will give the result as

$$V_{\pi}(15) = -1 + 0.25(-20 - 22 - 14 + V_{\pi}(15))$$

$$= -15 + 0.25\, V_{\pi}(15)$$

$$V_{\pi}(15) = -15/0.75$$

$$V_{\pi}(15) = -20$$

We don't need to recalculate the whole for changing the dynamics. The states 15 and 13 is exactly the same thus they must share the same state value as -20.

Problem 5: Exercise 4.5

In this the total no of process we need to implement are Initialization, Policy Evaluation, Policy improvement.

Step 1 : Initialization

$Q(s,a) \in R$ and $\pi(s) \in A(s)$ arbitrarily for all $s \in S$, $a \in A$

Step 2: Policy Evaluation

Loop :

$\Delta \leftarrow 0$

Loop for each $s \in S$ and $a \in A$:

$q := Q(s,a)$

$Q(s,a) \leftarrow \sum_{s',r} p(s',r|s,a)[r + \gamma \sum_{a'} \pi(a'|s') Q(s',a')]$

$\Delta \leftarrow \max(\Delta, |q - Q(s,a)|)$

until $\Delta <$ (a small positive number determining the accuracy of estimation

Step 3: Policy Improvement

policy-stable $\leftarrow$ True

for each $s \in S$ and $a \in A$:

old-action $\leftarrow \pi(s)$

$\pi(s) \leftarrow \arg\max_a Q(s,a)$

If old-action $\notin$ {a:}, which is the set of equi-best solutions from $\pi(s)$ then policy-stable $\leftarrow$ false

If policy-stable, then stop and return $Q \simeq q_*$ and $\pi \simeq \pi_*$; else goto 2nd step