RL Homework 2

Submitted By:
Snehal Gupta
2016201

Q1

Q1
given:

$S = \{low, high\}$

$A(high) = \{search, wait\}$

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

search at high
- leaves high with probability α
- reduces it to low with probability $1-\alpha$

search at low
- leaves low with probability β
- depletes the battery with $1-\beta$ [Robot is rescued and battery is recharged to high]

Reward
- 1 for each can collected
- -3 when got rescued

r_{search} - Expected no. of cans the robot will collect after searching.

r_{wait} - Expected no. of cans the robot will collect while waiting.

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

This has been obtained using

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

Q2

According to Bellman equation for $V\pi$,

$$V\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \varphi V\pi(s')\right]$$

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

$$\sum_a \pi(a|s) \sum_{s',r} \varphi\, p(s',r|s,a)\, V\pi(s')$$

$$V\pi(s) - \sum_a \pi(a|s) \sum_{s',r} \varphi\, p(s',r|s,a)\, V\pi(s')$$

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

For $s = s'$

$$\text{Coeff of } V\pi(s) = \left(1 - \sum_a \pi(a|s) \sum_{s',r} \varphi\, p(s',r|s,a)\right)$$

For $s \neq s'$

$$\text{Coeff of } V\pi(s) = \cancel{\sum_{s',r}} \left(\sum_a \pi(a|s) \sum_{s',r} \varphi\, p(s',r|s,a)\right)$$

$$\text{Coeff of } V\pi(s) \times V\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\, r$$

Q3

Q3

Ex 3.15

Signs of the rewards are not important and only the intervals between them are important. This is because, rewards can be made of the same sign by adding/subtracting a large positive constant to/from all the rewards. This leads to increase/decrease in the value function by a constant, which does not affect the algorithm.

To Prove: Adding a constant c to all the rewards adds a constant v_c to the value of the states, and, thus does not affect the relative values of any states under any policies.

Proof:

We know that,

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad \text{and} \quad v_\pi(s) = E_\pi[G_t \mid S_t = s, \cancel{A_t=a}]$$

After adding constant c to all the rewards,

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

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

$$= E\left[\sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) \,\middle|\, S_t = s\right]$$

$$= E\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=0}^{\infty} \gamma^k c \,\middle|\, S_t = s\right]$$

$$= E\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right] + E\left[\underbrace{\sum_{k=0}^{\infty} \gamma^k c}_{\text{constant term}} \,\middle|\, S_t = s\right]$$

$$= v_\pi(s) + \sum_{k=0}^{\infty} \gamma^k c$$

$$v_\pi'(s) = v_\pi(s) + \underbrace{\frac{c}{1-\gamma}}_{\text{constant term } v_c} \quad [\text{since } 0 \le \gamma < 1]$$

We can observe that the value function increases only by a constant v_c and hence, does not affect the relative values of any states under any policies.

Ex 3.16

In case of episodic task,
Let terminal time be T.

Solving the equation,

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

$$= E\left[\sum_{k=0}^{T} \gamma^k (R_{t+k+1} + c) \,\middle|\, S_t = s\right]$$

$$= E\left[\sum_{k=0}^{T} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right] + E\left[\sum_{k=0}^{T} \gamma^k c \,\middle|\, S_t = s\right]$$

$$= v_\pi(s) + \underbrace{c\left(\frac{\gamma^{T+1} - 1}{\gamma - 1}\right)}_{v_c}$$

Here, v_c is a function of T. So, Value function will differ in different episodes, but same for one episode. Since episodes are independent, the algorithm won't be affected.

Q5

Q5

$$V_*(s) = \max_{a \in A(s)} q_{\pi*}(s, a)$$

$$= \max_{a} \mathbb{E}_{\pi*}[G_t \mid S_t = s, A_t = a]$$

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

$$= \max_{a} \mathbb{E}[R_{t+1} + \gamma V_*(S_{t+1}) \mid S_t = s, A_t = a]$$

$$= \max_{a} \sum_{s', r} p(s', r \mid s, a)[r + \gamma V_*(s')] \quad \text{—} \; (1)$$

Also,

$$V_*(s') = \max_{a' \in A(s')} q_*(s', a') \quad \text{—} \; (2)$$

Substituting (2) into (1)

$$V_*(s) = \max_{a} \sum_{s', r} p(s', r \mid s, a)\left[r + \gamma \max_{a' \in A(s')} q_*(s', a')\right]$$