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,

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

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

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

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

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

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

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

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,

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

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

$$= E\left[\sum_{k=0}^{T} \gamma_k R_{t+k+1} \mid S_t = s\right] + E\left[\sum_{k=0}^{T} \gamma_k c \mid S_t = s\right]$$

$$= v_\pi(s) + E\left[\sum_{k=0}^{T} \gamma_k c \mid S_t = s\right]$$

Here, $v_c = E\left[\sum_{k=0}^{T} \gamma_k c \mid S_t = s\right]$ is a function of T and T is a random variable that normally varies from episode to episode. Different episodes will have different value-functions.

New G_t is $G_t + c\left(\frac{1-\gamma^T}{1-\gamma}\right)$

$\Rightarrow$ It will increase v_π when T increases.

Example

Consider an episodic task with one state S and two actions A_1 and A_2.

$A_1 \rightarrow$ Agent goes to terminal state with reward 1

$A_2 \rightarrow$ Agent goes back to S with reward 0.

On adding 1 to each reward, when A_2 is performed forever, ~~Gt =~~ return is $\frac{1}{1-\gamma}$ which can be bigger than 2 if $\gamma < \frac{1}{2}$.

Q5

Q5 $$v_*(s) = \max_{a \in A(s)} q_{\pi^*}(s,a)$$

$$= \max_{a} E_{\pi^*}\left[G_t \mid S_t = s, A_t = a\right]$$

$$= \max_{a} E_{\pi^*}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right]$$

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

Also,

$$q_*(s,a) = E\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right] \quad (2)$$

Substituting (2) into (1)

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