XCS234 Assigment 1

Ammar Husain

July 2022

1 Q1

For the purpose of this question lets abbreviate the actions as follows:
Right and down = RD = 0
Right and up = RU = 1

1.1 a

The optimal actions with $\gamma = 0.9$ are as follows:

$r_s = -4 : 1$

Optimal policy is unique and gamma does matter. A lower gamma makes the policy more greedy and it ends up reaching the greenblock rather than give up right away since $r_s < r_r$

$r_s = -1 : [0, 0, 1, 0], [0, 0, 0, 1]$

The optimal policy here is not unique, there are 2 of them. Gamma does not matter in this case since the optimal path length is smaller than r_r so even if the agent accrues the negative reward it still comes out ahead reaching the green block.

$r_s = 0 : [0, 0, 1, 0], [0, 0, 0, 1]$

The optimal policy here is not unique, there are 2 of them. Gamma does not matter in this case since the agent is not getting any reward at every step so is not incentivized to go longer.

$r_s = 0 : [0, 0, 1, 1, 0, 0]$

The optimal policy here is unique and this is the longest route it can take to the green block. Gamma matters here because a low gamma value will greedly try to get to the goal rather than take the longer way thereby accumulating more rewards.

1.2 b

Given any possible configuration of the grid, the value of r_s that will always cause the optimal policy to return the shortest path to the green target is 0. This is because the discount factor pushes the agent towards the goal while not wasting time.

The optimal action from square 27 in this case is right down = 0.

1.3 c

$r_s = -4 : [1, 1], [1, 0]$ Optimal policy is not unique and gamma does matter. A lower gamma makes the policy more greedy and it ends up reaching the greenblock rather than give up right away

$r_s = -1 : [1, 1], [1, 0]$ Optimal policy is not unique and gamma does matter. A lower gamma makes the policy more greedy and it ends up reaching the greenblock rather than give up right away

$r_s = 0 : [0, 0, 1, 1, 0, 1, 0, 1, 0]$ Optimal policy is not unique. Gamma does not matter in this case since the agent is not getting any reward at every step so is not incentivized to go longer

$r_s = 1$: Agent does not terminate

Agent keeps getting a reward at every step so has no incentive to terminate.

Value of r_s that would return the shortest path is 0. Optimal action from square 27 using this value is $[1, 1, 1, 0, 1, 0]$

1.4 d

$$G_{old,t} = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{1}$$

Now if we substitute $r = a(c + r)$. We get

$$G_{new,t} = a(c+r_{t+1})+\gamma a(c+r_{t+2})+\gamma^2 a(c+r_{t+3})+... = ac\sum_{k=0}^{\infty}\gamma^k + a\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}$$

$\mathrm{G}_{new,t} = ac\sum_{k=0}^{\infty}\gamma^k + aG_{old,t}$
(2)

Therefore

$$V^{\pi}_{new}(s) = E_{\pi}[G_{new,t}|x_t = s]$$

$\mathrm{V}^{\pi}_{new}(s) = E_{\pi}[ac\sum_{k=0}^{\infty}\gamma^k + aG_{old,t}|x_t = s]$
$\mathrm{V}^{\pi}_{new}(s) = ac\sum_{k=0}^{\infty}\gamma^k + aE_{\pi}[G_{old,t}|x_t = s]$
$\mathrm{V}^{\pi}_{new}(s) = ac\sum_{k=0}^{\infty}\gamma^k + aV^{\pi}_{old}(s)$
(3)

2 Q2

We need to prove that for any arbitrary initial state s_{start}:

$V_1^{\pi}(s_{start}) \geq V_1^{\pi+}(s_{start})$

Since we know that $\pi^+(s) = \pi(s)$

if $s \notin S^+$

let's assume that $s_{start} \in S^+$

Now we can compute Q value for the starting state where we take the first action given policy π^+ and then follow policy π moving forward:

$Q_t^\pi(s_{start}, \pi^+(s_{start}, t)) = r(s_{start}, a^+) + \gamma V_\pi(s_{start})$

This is because in every state $s \in S^+$ action a^+ leads to the same state with probability 1. And we already know that $r(s_{start}, a^+) = 1$. Therefore:

$Q_t^\pi(s_{start}, \pi^+(s_{start}, t)) = 1 + \gamma \sum_{s' \in S} P(s'|s_{start}, \pi(s_{start}))V_\pi(s_{start})$

Conversely if we now take the first action given policy π and then follow policy π moving forward we get:

$Q_t^\pi(s_{start}, \pi(s_{start}, t)) = r(s_{start}, a_t) + \gamma \sum_{s' \in S} P(s'|s_{start}, \pi(s_{start}))V_\pi(s_{start})$

which equals

$Q_t^\pi(s_{start}, \pi(s_{start}, t)) = H + \gamma \sum_{s' \in S} P(s'|s_{start}, \pi(s_{start}))V_\pi(s_{start})$

Now lets take the difference of both Q values we have so far:

$Q_t^\pi(s_{start}, \pi^+(s_{start}, t)) - Q_t^\pi(s_{start}, \pi(s_{start}, t)) = 1 + \gamma \sum_{s' \in S} P(s'|s_{start}, \pi(s_{start}))V_\pi(s_{start}) - H - \gamma \sum_{s' \in S} P(s'|s_{start}, \pi(s_{start}))V_\pi(s_{start})$

Since after the first step both policies follow π the second terms cancel out when calculated in expectation of $x \sim \pi$

$E_{x\sim\pi}(Q_t^\pi(s_{start}, \pi^+(s_{start}, t)) - Q_t^\pi(s_{start}, \pi(s_{start}, t))) = 1 - H$

Summing over the finite time horizon we get

$\sum_{t=1}^{H} E_{x\sim\pi}(Q_t^\pi(s_{start}, \pi^+(s_{start}, t)) - Q_t^\pi(s_{start}, \pi(s_{start}, t))) = H(1 - H)$

And since H is a positive number this expression turns out to be less than zero

$\sum_{t=1}^{H} E_{x\sim\pi}(Q_t^\pi(s_{start}, \pi^+(s_{start}, t)) - Q_t^\pi(s_{start}, \pi(s_{start}, t))) = H(1 - H) \leq 0$

Given the performance difference lemma we know that

$V_1^\pi(s_{start}) - V_1^{\pi+}(s_{start}) = \sum_{t=1}^{H} E_{x\sim\pi}(Q_t^\pi(s_{start}, \pi^+(s_{start}, t)) - Q_t^\pi(s_{start}, \pi(s_{start}, t)))$

Therefore:

$V_1^\pi(s_{start}) - V_1^{\pi+}(s_{start}) \leq 0$

which proves that

$V_1^\pi(s_{start}) \geq V_1^{\pi+}(s_{start})$

3 Q3

3.1 a

In class we have already seen that $||BV_k - BV_j|| \leq \gamma||V_k - V_j||$ for $\gamma < 1$

We can use the same proof to state that the Bellman backup will be a contraction operator for non stationary discount factor as long as $\gamma_k < 1$ for all timesteps. This can be easily shown for $\gamma_k = 1 - \frac{a}{k+1}$ where all $k \geq 1$. Therefore $\frac{a}{k+1} > 0$ and $\gamma_k < 1$.

3.2 c