Michael Owen

CSC580

Assignment 1

Part 1

Exercise 4.1

$$q_\pi(11, down) = -1 + v_\pi(Terminal) = -1 + 0 = \mathbf{-1}$$

$$q_\pi(7, down) = -1 + v_\pi(11) = -1 - 14 = \mathbf{-15}$$

Exercise 4.2

$$v_\pi(15) = 0.25[(-1-20) + (-1-22) + (-1-14) + (-1 + v_\pi(15))$$

$$= 0.25\left(-60 + v_\pi(15)\right)$$

$$= -15 + 0.25\,(v_\pi(15))$$

Since $v_\pi(15) == v_\pi(13) = -20$ we get:

$$\boldsymbol{v_\pi(15) = -15 + 0.25(-20) = -20}$$

If the dynamics are changed so that we can transition between 13 and 15, nothing is really changed since the two states both hold the same value. Therefore, we still have $\boldsymbol{v_\pi(15) = -20}$

4.5 on Next page

Exercise 4.5

1. **Initialization** Q(s, a) ∈ R and π(s) ∈ A(s) arbitrarily for all s ∈ S, a ∈ A

2**. Policy Evaluation**

Loop:

Δ ← 0

Loop for each s ∈ S and a ∈ A:

q = Q(s, a)

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

Δ← $\max(\Delta, |q - Q(s, a)$

Until $\Delta < \theta$ $(a\ small\ positive\ number\ deteriming\ the\ accuracy\ of\ estimation)$

4. **Policy Improvement**

$policy - stable \leftarrow true$

$For\ each\ s \in$ S and a ∈ A:

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

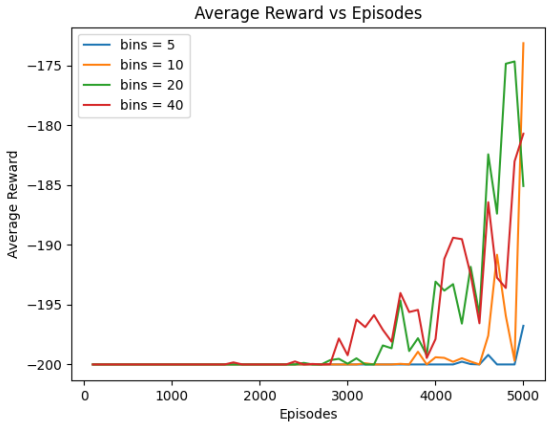
$\pi(s) \leftarrow argmax_a\ Q(s, a)$

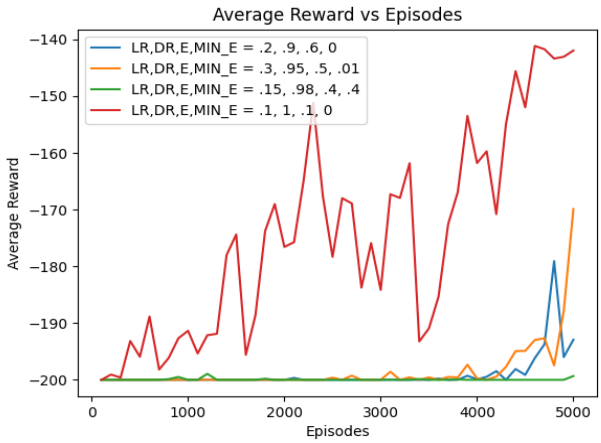
$If\ old - action$ not ∈ $\{a_i\}$ $, then\ policy - stable \leftarrow false$

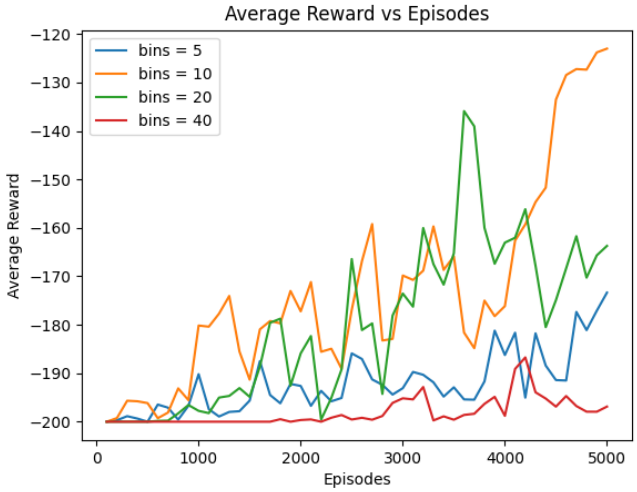
$If\ policy - stable, then\ stop\ and\ return\ Q \approx q_*\ and\ \pi \approx \pi_*$; else go to 2

Part 2 - Simulation writeup on next page

Part 2: Write-up for Mountain car simulation







Note:

LR = Learning Rate, DR = Discount Rate or Gamma, E = Epsilon,

(Figure 1 – top left – default parameter values)

(Figure 2 – top right – bins = 10 for all)

(Figure 3 – bottom – parameter values: LR, Gamma/DR, Eps, Min_Eps = .1, 1, .1, 0)

-Max Horizontal position and velocity at that position: [0.5373, 0.0412]

Summary of results

After conducting several experiments I've found that certain parameter values provide better results (i.e. higher average rewards per 100 episodes). The parameter with the largest effect

seemed to be Epsilon. I found that as the value for Epsilon was decreased, the agent was able to successfully reach the goal state earlier (relative to episode) and at a higher frequency. I also found that using a Min_Epsilon close or equal to 0 was better – which makes sense because as the q table updates over time and approaches near optimal values, it's probably better to exploit more versus experiment. I experimented with different number of bins but found that the rewards with 10 bins were at least as good, or better, as other bin sizes I've experimented with – such as 5, 20, and 40 bins. I noticed (with the parameter combination used in Figure 3) that with bins > 10, it took longer (more episodes) for successes. Conversely, as shown in the first figure with the original parameters, successes were achieved in earlier episodes with larger bin sizes – which shows that with different parameters, bin sizes effect the results differently. Also, when using the same value for Epsilon and Min_Epsilon (see green line in chart), there were very few successful episodes. Thus, decaying Epsilon to a value near or equal to 0 was much better. Finally, regarding learning rate and gamma/discount rate, I found that a lower value for learning rate (0.1) along with a higher value for gamma (1.0) produced the lowest average reward. As shown in Figure 2 (red-line) and Figure 3, the 'best' (highest average reward) parameter combination I found was: learning rate = .1, gamma=1, epsilon = .1, min_epsilon = 0, and bins = 10.