**Midterm 2**

1. Short Answer.
   1. Value function is a predication of future reward whereas reward is the actual feedback value.
   2. Total discounted reward (summation of rewards with discount factor and reward function)
   3. Sarsa has conditions that it must follow. Q-Learning uses “off policy” to learn its action relative to the ε-greedy policy where as Sarsa uses conditionals to learn its action relative to the actual conditions. Sarsa looks up the next policy value and Q-Learning looks up the next maximum policy value.
   4. Q-Learning is “off policy” since it uses the ε-greedy algorithm to decide its next action maximizing value. It constantly goes for the optimal value action per the ε-greedy algorithm.
   5. A greedy algorithm can get stuck in a sub-optimal action whereas a ε-greedy algorithm will continue exploring. There will be minimum regret with the ε-greedy algorithm.
   6. No knowledge of transitions or rewards for Markov Decision Process. The algorithm cannot make predictions about the next state or reward.
2. You have communicated with it, but it has gotten stuck in a sub-optimal loop since you are using a greedy algorithm and it has found a way to achieve the reward.
3. G0 = -2 + (.5)(3) + (.5)2(5) + (.5)3(2) + (.5)4(-1) = .9375

G1 = 3 + (.5)(5) + (.5) 2(2) + (.5)3(-1) = 5.875

G2 = 5 + (.5)(2) + (.5) 2(-1) = 5.75

G3 = 2 + (.5)(-1) = 1.5

G4 = -1

G5 = discount = .5 (Episode has ended so no return here?)

This answer comes from slide 8: <https://mycourses.msstate.edu/bbcswebdav/pid-1293814-dt-content-rid-12450844_1/courses/ca861.201730.Group01/RL_Model_Free_Prediction_Handout.pdf> and slide 15: <https://mycourses.msstate.edu/bbcswebdav/pid-1290034-dt-content-rid-11804369_1/courses/ca861.201730.Group01/RL_MDP_Handout.pdf>

1. G0 = 3 + (.9)5 + (.9)2(5) + (.9)3(5) + (.9)4(5) + (.9)5(5)…

(ie. Summation of 3 + 5(.9)n) where n is the reward number

G1 = 5 + (.9)5 + (.9)2(5) + (.9)3(5) + (.9)4(5) + (.9)5(5)…

(ie Summation of 5(.9)n-1)

1. MDP
   1. v\*(2) = 4
   2. q\*(4,W) = 3.25
   3. vπ(4) = 6

See test for work. I apologize for some of the writing being so small, but please let me know if you have any questions and I will clarify.

1. Value Estimates
   1. Monte Carlo
      1. A = G0 = [1+(-4)(.9)+(0)+(5)(.9)3+(2)(.9)4 ]/1 = 2.3572
      2. C = G2 = [(0)+(5)(.9)+(2)(.9)2]/1 = 6.12
      3. B = (G1 + G3 / 2) = [(-4)+(0)+(5)(.9)2+(2)(.9)3]/2 = .754
      4. D = G4 = 2
   2. TD(0)
   3. TD(λ), λ = .5

See test for the above problem

1. See test for comment.