CSC 6260 – Advanced Topics in AI

Homework 1

Due: **Fri Feb 9th, 11:59pm**

**[35 total points] General Instructions**: Put your answers to the following problems into this word doc and submit a PDF document as an attachment under Assessments 🡪 Homework 1 for this course. Note that you may submit multiple times, but only the most recent entry submitted before the deadline will be graded. Only electronic submission will be considered, no handwriting. Please put your answer in the designated Answer area.

1. [2 points] At each time step the agent takes an \_\_\_\_\_\_\_.
2. Action
3. State
4. Reward

Answer: a

1. [2 points] What is the exploration/exploitation tradeoff?
2. The agent wants to explore the environment to learn as much about it as possible about the various actions. That way once it knows every arm’s true value it can choose the best one for the rest of the time.
3. The agent wants to explore to get more accurate estimates of its values. The agent also wants to exploit to get more reward. The agent cannot, however, choose to do both simultaneously.
4. The agent wants to maximize the amount of reward it receives over its lifetime. To do so it needs to avoid the action it believes is worst to exploit what it knows about the environment. However, to discover which arm is truly worst it needs to explore different actions which potentially will lead it to take the worst action at times.

Answer: b

1. [2 points] A policy is a function which maps \_\_\_\_ to \_\_\_\_.
2. States to values.
3. Actions to probability distributions over values.
4. States to probability distributions over actions.
5. States to actions.
6. Actions to probabilities.

Answer: d

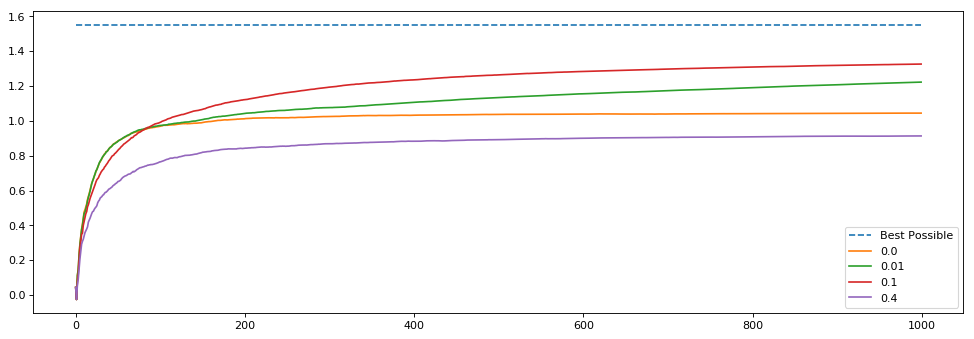
1. [2 points] Imagine the agent is learning in an episodic problem. Which of the following is true?
2. The agent takes the same action at each step during an episode.
3. The number of steps in an episode is always the same.
4. The number of steps in an episode is stochastic: each episode can have a different number of steps.

Answer: c

1. [2 points] If the reward is always +1 what is the sum of the discounted infinite return when
2. Infinity.

Answer: c

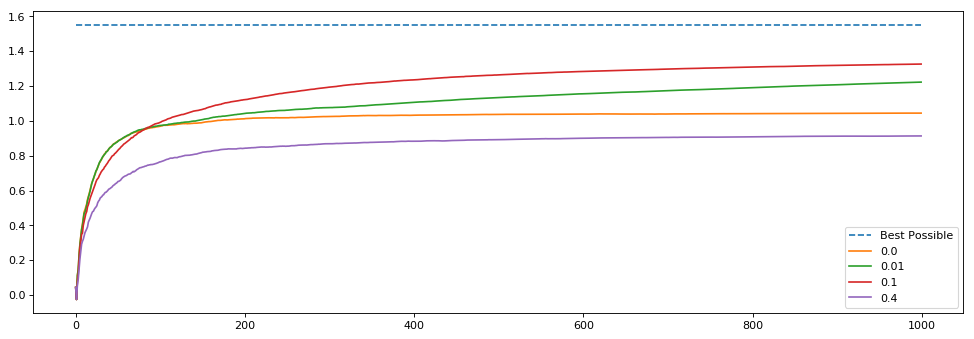
1. [4 points] Why do you think an agent with epsilon of 0.1 perform better over 1000 steps than an agent with epsilon of 0.01?



1. Epsilon of 0.1 is the optimal value for epsilon in general.
2. The 0.01 agent did not explore enough. Thus, it ended up selecting a suboptimal arm for longer.
3. The 0.01 agent explored too much causing the arm to choose a bad action too often.

Answer: b

1. [4 points] If exploration is so great, why do you think an agent with epsilon of 0.0 (a greedy agent) performs better than one with epsilon of 0.4?



1. Epsilon of 0.4 doesn’t explore often enough to find the optimal action.
2. Epsilon of 0.4 explores too often that it takes many sub-optimal actions causing it to do worse over the long term.
3. Epsilon of 0.0 is greedy, thus it will always choose the optimal arm.

Answer: b

1. [4 points] Imagine, an agent is in a maze-like grid world. You would like the agent to find the goal, as quickly as possible. You give the agent a reward of +1 when it reaches the goal and the discount rate is 1.0, because this is an episodic task. When you run the agent, it finds the goal but does not seem to care how long it takes to complete each episode. How could you fix this? (**Select all that apply**)
2. Give the agent -1 at each time step.
3. Give the agent a reward of 0 at every time step so it wants to leave.
4. Give the agent a reward of +1 at every time step.
5. Set a discount rate less than 1 and greater than 0, like 0.9.

Answer: b, d

1. [8 points] Suppose and we observe the following sequence of rewards: , and , with . Compute . Please show your work

Answer:

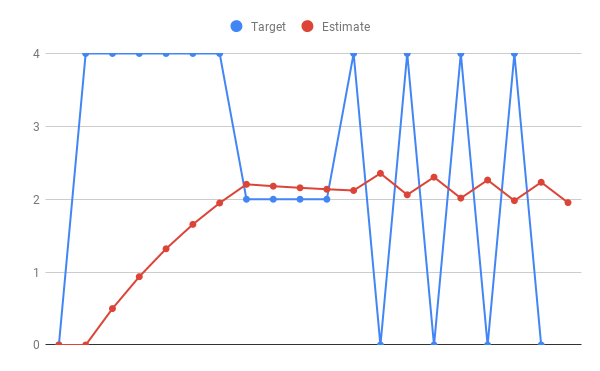
G0 = R1 + γ\*R2 + γ2\*R3 + γ3\*R4 + γ4\*R5

= -3 + 0.8 \* 5 + 0.82 \* 2 + 0.83 \* 7 + 0.84 \* 1

= -3 + 4 + 1.28 + 3.584 + 0.496 = 6.2736

1. [5 points] The equation used for incrementally updating the action values in the bandit problem is a key update rule we will use throughout the class. We discussed this equation extensively in class. The blue line is the target that we might estimate with this equation. The red line is our estimate plotted over time.

Given the estimate update in red, what do you think was the value of the step size parameter we used to update the estimate on each time step? Please explain why.



Answer: As we can see here, from Q0 to Q6 there is a steady increment, so all the rewards must be positive, while in Q6 to Q10, there is no change in the estimate value function, so the reward here must be equal to estimate value function itself.