Autonomous Networking A.A. 20-21
Written exam – 09/02/2021

(For homework integration answer only to E1, E4, E5)

Each True/False question is worth 1 points. Leaving a question blank is worth 0 points. Answering incorrectly is worth −1 points.

Q1. A small discount (close to 0) encourages greedy behavior. [true or false]
A discount close to zero will place extremely small values on rewards more than one step away, leading to greedy behavior that looks for immediate rewards.

Q2. A reward is a scalar feedback. [true or false]

Q3. The agent state is the information used to pick the next action. [true or false]

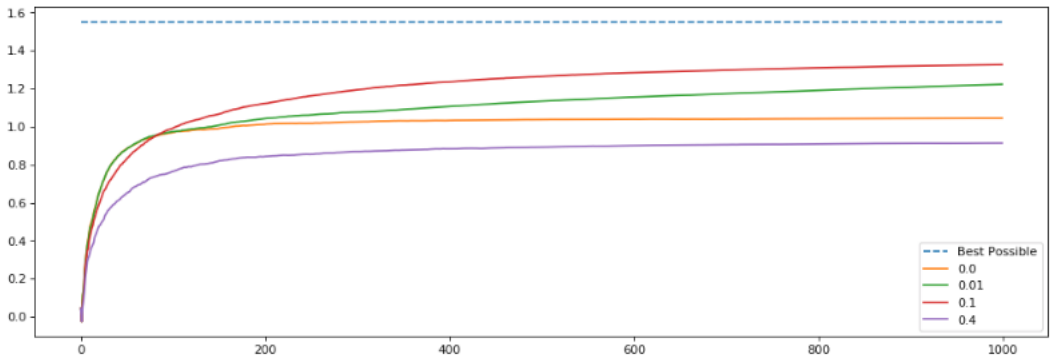
Q4. An MDP is a Markov Decision Parameter. [true or false]

Q5. In a K-armed bandit there are K states. [true or false] There are K possible actions and just 1 state

Q6. A policy is a function which maps states to values. [true or false]

Select all correct answers

Q7. Why did epsilon of 0.1 perform better over 1000 steps than epsilon of 0.01?



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

Q8. What is the exploration/exploitation tradeoff?

1. 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.

2. 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.

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.

Q9. What is the incremental rule (sample average) for action values?

1. $Q_{n+1} = Q_n + 1/n\ [R_n - Q_n]$
2. $Q_{n+1} = Q_n + 1/n\ [Q_n]$
3. $Q_{n+1} = Q_n + 1/n\ [R_n + Q_n]$
4. $Q_{n+1} = Q_n - 1/n\ [R_n - Q_n]$

Q10. What is the difference between a small gamma (discount factor) and a large gamma?

1. With a smaller discount factor the agent is more far-sighted and considers rewards farther into the future.
2. With a larger discount factor the agent is more far-sighted and considers rewards farther into the future.
3. The size of the discount factor has no effect on the agent.

Exercises and theoretical questions

E1.

Given a set of three actions {Transmit, Wait, Receive} and the the following Q-values for the state S, what is the probability (in terms of ε) that we will take each action on our next move when we follow an ε-greedy exploration policy (assuming any random operations are chosen uniformly from all actions)?

Q(S, Transmit) = 0.75

Q(S, Wait) = 0.25

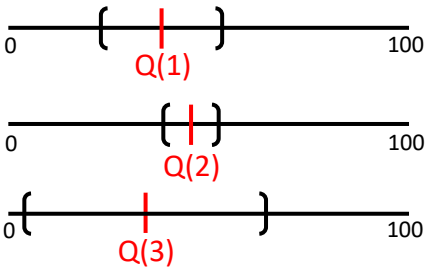
Q(S, Receive) = 0.5

Transmit $(1-\varepsilon)+ \varepsilon/3 = 1-(2\ \varepsilon)/3$

Wait ε/3

Receive ε/3

E2.
Suppose an agent can take 3 possible actions, *a=1,2,3*. Each action has its estimated *Q(a)* that is represented with uncertainty intervals on axis with increasing values as follows:



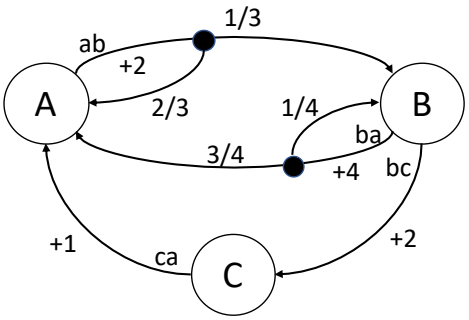
a) What action would the agent pick next if it is following the Upper Confidence Bound approach? Motivate your answer. Q3
b) What is the advantage, if any, of applying UCB instead of ε-greedy?

E3. Suppose $\gamma = 0.5$ and the following sequence of rewards is received R_1=-2, R_2=2, R_3=6, R_4=3, and R_5=2, with T=5. What are the discounted return values G_0, G_1, … , G_5? (Hint: work backwords)
$G_5 = 0$
$G_4 = R_5 + \gamma G_5 = 2$
$G_3 = R_4 + \gamma G_4 = 3+0.5*2 = 4$
$G_2 = R_3 + \gamma G_3 = 6+0.5*4 = 8$
$G_1 = R_2 + \gamma G_2 = 2+0.5*8 = 6$
$G_0 = R_1 + \gamma G_1 = -2+0.5*6 = 1$

E4. Explain the meaning of the following equation (use a graph if helpful):

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s, a)$$

E5. Consider the MDP with discount factor γ=0.5, with uniform random policy $\pi_1(s, a)$ that takes all actions from state s with equal probability. Starting with an initial value function of $V_1(A) = V_1(B) = V_1(C) = 1$ apply one iteration of iterative policy evaluation to compute a new value function $V_2(A)$



$V_2(A) = 2 + 0.5*(1/3 * V_1(B) + 2/3 * V_1(A)) = 2.5$

E6.

Consider a MDP with three states [A, B, C] and two actions [Wait, Transmit]. We are given the following samples generated from taking actions in the MDP. For the following problems, assume γ = 1 and α = 0.5. All Q-values are initialized to 0.

We run Q-learning on the following samples:

s	a	s'	r
A	Transmit	B	2
C	Wait	A	0
B	Wait	A	-2
B	Transmit	C	-6
C	Transmit	A	2
A	Transmit	A	-2

What are the estimates for the following Q-values as obtained by Q-learning?

a) Q(C,Wait) = 0.5

b) Q(C,Transmit) = 1.5

1. $Q(A, Transmit) \leftarrow (1 - \alpha)Q(A, Transmit) + \alpha(r + \gamma \max Q(B, a)) = 0.5(0) + 0.5(2) = 1$
2. $Q(C, Wait) \leftarrow (1 - \alpha)Q(C, Wait) + \alpha(r + \gamma \max Q(A, a)) = 0.5(0) + 0.5(1) = 0.5$
3. $Q(C, Transmit) \leftarrow (1 - \alpha)Q(C, Transmit) + \alpha(r + \gamma \max Q(A, a)) = 0.5(0) + 0.5(3) = 1.5$