Name: _____ SUID: _____

Stanford University

AA228/CS238: Decision Making under Uncertainty

Autumn 2020

Prof. Mykel J. Kochenderfer • Remote • email: *mykel@stanford.edu*

MIDTERM 3 **Due date: November 13, 2020 (5pm)**

You have **90 minutes** to complete this exam. This exam is electronically timed; you do not have to keep track of your own time. To accommodate those in other time-zones and complex working situations, you may choose any 90 minute window between 5pm PST Nov 12th, 2020 and 5pm PST Nov 13th, 2020 to take the exam. Answer all questions. You may consult any material (e.g., books, calculators, computer programs, and online resources), but you may not consult other people inside or outside of the class. If you need clarification on a question, please make a private post on Piazza. **Only what is submitted prior to the deadline will be graded.**

Question 1. (1 pt) I understand the above instructions and will adhere to them.

Question 2. (5 pts) You enter a casino and there are two slot machines A and B, each with potentially different winning probabilities. You have a uniform prior over the machines' winning probabilities (i.e. you do not have any prior belief that one is more likely than the other). You have been playing for a few rounds and keeping track of the outcomes of each attempt, and want to use UCB1 exploration to select your actions. Use the exploration parameter $c = 10$. Use a Bayesian approach to estimate your winning probabilities—do not use a maximum likelihood approach.

a) (2 pts) If you have played on machine A a total of 10 times, winning 8 of those times, and you have played on machine B a total of 4 times, winning 3 of those times, which machine would you play on next? What is its estimated upper confidence bound?

b) (2 pts) If you have played on machine A a total of 10 times, winning 9 of those times, and you have played on machine B a total of 10 times, winning 8 of those times, which machine would you play on next? What is its estimated upper confidence bound?

c) (1 pts) In which of the above situations (part (a) and/or part (b)) did the UCB1 exploration strategy lead to a non-greedy action? Why? Provide a quantitative argument.

Solution:

Recall that we compute the upper confidence bound for an action a using

$$\text{UCB1}(a) = \rho_A + c\sqrt{\frac{\log N(s)}{N(s,a)}}$$

Here we have a single state (so the s in $N(s)$ and $N(s,a)$ will never change and we will drop it from our experssions going forward).

a) Because we have a uniform prior we start by modeling ρ_A and ρ_B as distributed according to a beta distribution $\text{Beta}(1,1)$. We then update this based on our experience to get posterior distributions of the winning probabilities for each machine:

$$\rho_A \sim \text{Beta}(9,3)$$
$$\rho_B \sim \text{Beta}(4,2)$$

With a Bayesian approach N comes from our pseudocounts with $N = 12+6 = 18$, $N(a_a) = 9+3 = 12$, and $N(a_b) = 4+2 = 6$. With this, we can now compute the upper confidence bound for both actions:

$$\text{UCB1}(a_A) = \frac{9}{9+3} + c\sqrt{\frac{\log(18)}{12}} \approx 5.66$$
$$\text{UCB1}(a_B) = \frac{4}{4+2} + c\sqrt{\frac{\log(18)}{6}} \approx 7.61$$

This would lead us to select action a_b since $\text{UCB1}(a_B) \geq \text{UCB1}(a_A)$. The estimated upper confidence bound is 7.61.

Alternate Solution: We will also count for full credit solutions which did not use the pseudocounts in determining $N(a)$ and N. In this approach, we have $N = 10+4 = 14$, $N(a_a) = 10$, and $N(a_b) = 4$, which are the actual counts of the pulls that we performed. With these values we get:

$$\text{UCB1}(a_A) = \frac{9}{9+3} + c\sqrt{\frac{\log(14)}{10}} \approx 5.89$$
$$\text{UCB1}(a_B) = \frac{4}{4+2} + c\sqrt{\frac{\log(14)}{4}} \approx 8.79$$

In this case we still select action a_b since $\text{UCB1}(a_B) \geq \text{UCB1}(a_A)$. The estimated upper confidence bound is 8.79.

b) Because we have a uniform prior we start by modeling ρ_A and ρ_B as distributed according to a beta distribution $\text{Beta}(1,1)$. We then update this based on our experience to get posterior distributions of the winning probabilities for each machine:

$$\rho_A \sim \text{Beta}(10,2)$$
$$\rho_B \sim \text{Beta}(9,3)$$

With a Bayesian approach we use the pseudocounts, giving us $N = 12+12 = 24$, $N(a_a) = 10+2 = 12$, and $N(a_b) = 9+3 = 12$. With this, we can now compute the upper confidence bound for both actions:

$$\text{UCB1}(a_A) = \frac{10}{10+2} + c\sqrt{\frac{\log(24)}{12}} \approx 5.98$$
$$\text{UCB1}(a_B) = \frac{9}{9+3} + c\sqrt{\frac{\log(24)}{12}} \approx 5.90$$

This would lead us to select action a_a since $\text{UCB1}(a_B) \geq \text{UCB1}(a_A)$. The estimated upper confidence bound is 5.98.

Alternate Solution: We will also count for full credit solutions which did not use the pseudocounts in determining $N(a)$ and N. In this approach, we have $N = 10+10 = 20$, $N(a_a) = 10$, and $N(a_b) = 10$, the actual counts of the pulls that we performed. With these values we get:

$$\text{UCB1}(a_A) = \frac{10}{10+2} + c\sqrt{\frac{\log(20)}{10}} \approx 6.31$$
$$\text{UCB1}(a_B) = \frac{9}{9+3} + c\sqrt{\frac{\log(20)}{10}} \approx 6.22$$

In this case we still select action a_a since $\text{UCB1}(a_A) \geq \text{UCB1}(a_B)$. The estimated upper confidence bound is 6.31.

c) In part a, it led to a non-greedy action. At each step, the greedy action will be that which maximizes our expected utility. Since we represent the true probability of success as distributed according to

a beta distribution, the expected value of this distribution will be the expected utility of taking that action — e.g. if $\rho_A \sim \text{Beta}(9, 3)$, and we receive a reward of 1 for a successful pull, our expected reward for taking action a will be equal to the mean of this beta distribution $\rho_A = \frac{9}{3}$. So to see whether we took a greedy action we look at ρ_A and ρ_B, and see whether we took the larger of the two.

In part a, we had $\rho_A = \frac{3}{4}$ and $\rho_B = \frac{2}{3}$. With UCB1 exploration we decided to take action b. Since $\rho_A > \rho_B$, we took the non-greedy action.

In part b, we had $\rho_A = \frac{5}{6}$ and $\rho_B = \frac{3}{4}$. With UCB1 exploration we decided to take action a. Since $\rho_A > \rho_B$, we took the greedy action.

For full credit, you need both (i) to say that the scenario in part (a) led to a non-greedy action and (ii) to justify this by comparing the ρ values in the scenarios.

Question 3. (6 pts) You are standing in front of two closed doors. Behind one of the doors is a tiger and behind the other is a large reward of 100. If you open the door with the tiger, then you receive a penalty of -10. Instead of opening one of the two doors, you can listen, in order to gain some information about the location of the tiger. Unfortunately, listening is not free (it has a cost of -1). In addition, listening is also not entirely accurate. There are only two states s_L and s_R, which correspond to the tiger being behind the left door and the tiger being behind the right door respectively.

The transition and observation models can be described in detail as follows. We can open one of the doors with action a_L (open left door) or a_R (open right door). After we open a door and collect the associated reward, the problem "resets", meaning that we transition into state s_L or state s_R with equal probability. The listen action a_N does not change the state of the world. When the world is in state s_L, the listen action results in observation o_L (i.e. hearing the tiger behind the left door) with probability 0.85 and the observation o_R with probability 0.15; similarly in world state s_R, the listen action results in observation o_R with probability 0.85 and the observation o_L with probability 0.15.

a) (2 pts) Draw the alpha vectors (as drawn in the book and in class) for each action, a_L, a_R, and a_N for a one step horizon. Be sure to label your axes.

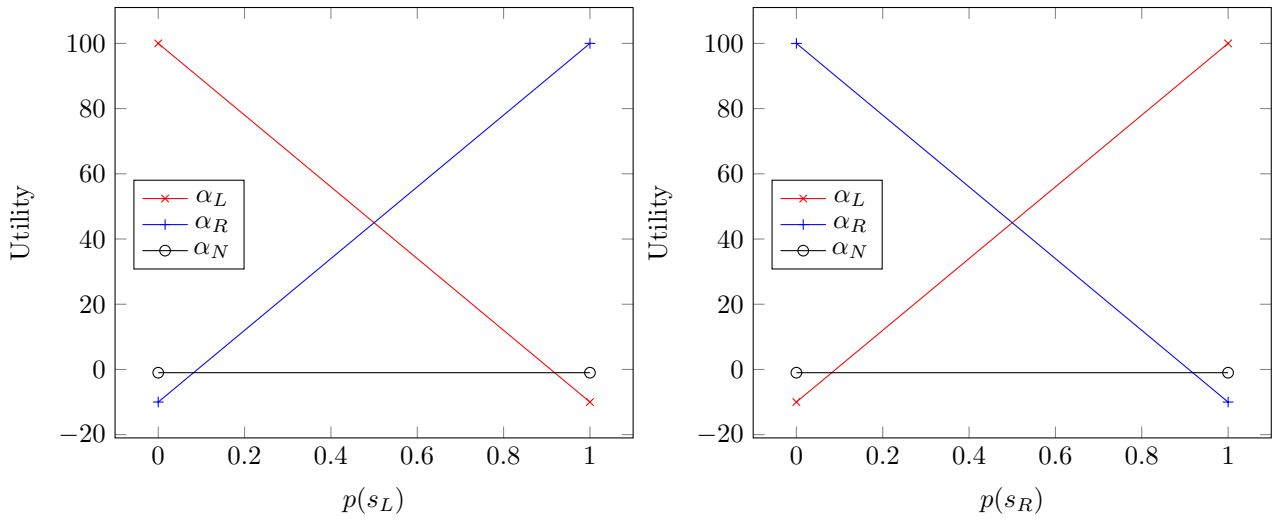
b) (4 pts) Update the s_L element of the a_N alpha vector (corresponding to the listen action) using Fast Informed Bound (i.e. calculate the value of $\alpha_N^{(2)}(s_L)$) using your alpha vectors from (a) as α_L^1, α_R^1, α_N^1. Use discount factor $\gamma = 0.9$.

Solution:

a) We found the following alpha vectors by considering the reward from taking a single step for each action for each of the two states we could be in. For instance, $\alpha_L = [R(s_L, a_L), R(s_R, a_L)]$.

$$\alpha_L = \begin{bmatrix} -10 \\ 100 \end{bmatrix} \qquad \alpha_R = \begin{bmatrix} 100 \\ -10 \end{bmatrix} \qquad \alpha_N = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

Depending on whether s_L or s_R is shown on the x axis we get one of the two following equally valid plots. We draw a line corresponding to $\alpha_L^T \begin{bmatrix} p(s_L) \\ p(s_R) \end{bmatrix}$, a line corresponding to $\alpha_R^T \begin{bmatrix} p(s_L) \\ p(s_R) \end{bmatrix}$ and a line corresponding to $\alpha_N^T \begin{bmatrix} p(s_L) \\ p(s_R) \end{bmatrix}$ for different values of $p(s_L)$ or $p(s_R)$ on the x axis.



b) Recall that the Fast Informed Bound (FIB) update equation is given by:

$$\alpha_a^{(k+1)}(s) = R(s,a) + \gamma \sum_o \max_{a'} \sum_{s'} O(o \mid a, s') T(s' \mid s, a) \alpha_{a'}^{(k)}(s')$$

So we'll apply this for $a = a_N$ and s $= s_L$. We first note that the sum over s' will disappear, as with the listening action we will remain in the same state — so $T(s' \mid s, a)$ will be 1 when $s' = s_L$ and 0 otherwise. Removing the sum and plugging in $a = a_N$ and s_L gives us the simplified expression

$$\alpha_N^{(2)}(s_L) = R(s_L, a_N) + \gamma \sum_o \max_{a'} O(o \mid a_N, s_L) \alpha_{a'}^{(1)}(s_L)$$

Now, we can expand the sum with our two observations o_L and o_R

$$\begin{aligned}\alpha_N^{(2)}(s_L) &= R(s_L, a_N) + \gamma \left(\max_{a'} O(o_L \mid a_N, s_L) \alpha_{a'}^{(1)}(s_L) + \max_{a'} O(o_R \mid a_N, s_L) \alpha_{a'}^{(1)}(s_L) \right) \\ &= R(s_L, a_N) + \gamma \left(\max_{a'} 0.85 \alpha_{a'}^{(1)}(s_L) + \max_{a'} 0.15 \alpha_{a'}^{(1)}(s_L) \right)\end{aligned}$$

Now the only thing left to evaluate is $0.85\alpha_{a'}^{(1)}(s_L)$ and $0.15\alpha_{a'}^{(1)}(s_L)$ for all possible values of a': a_L, a_R, and a_N. This is in order to evaluate the $\max_{a'}$ terms. We observe that in both cases, since we are evaluating the alpha vector at s_L, we will choose a_R whose alpha vector is $(100, -10)$. Choosing $a' = a_R$ our expression becomes

$$\begin{aligned}\alpha_N^{(2)}(s_L) &= R(s_L, a_N) + \gamma \left(\max_{a'} 0.85 \alpha_{a'}^{(1)}(s_L) + \max_{a'} 0.15 \alpha_{a'}^{(1)}(s_L) \right) \\ &= -1 + \gamma \left(0.85 \cdot 100 + 0.15 \cdot 100 \right) \\ &= -1 + 0.9 \cdot 100 = 89\end{aligned}$$

So this will be our updated value. Interestingly, our observation probabilities did not affect the final result.

Question 4. (3 pts) You built a little robot and its first task is to localize itself. It has a map of the world, but only a noisy lidar signal to detect distances to surrounding objects. Name a method it can use to perform belief updates, and explain why it would be better than some of the other methods discussed in class.
Solution: You could use a particle filter to perform belief updates which works with continous state spaces and does not impose restrictions on the dynamics of the system. This has an advantage over a Kalman filter (which would impose a more restrictive model on the transition and observation functions). Particle filters also can represent multimodal beliefs, while Kalman filters are restricted to a unimodal belief.

An extended Kalman filter can also be used to account for nonlinear dynamics, but it would still impose assumptions around the noise being gaussian. Similarly, an unscented Kalman filter can be used, which has the advantage that it is a derivative free approach. This, like the EKF, assumes gaussian noise.

Question 5. (6 pts) Consider an environment with 2 states and 2 actions. An agent performs actions and observes the rewards and transitions listed below. Each step consists of the current state s which can be s_1 or s_2, reward r, action a which can be a_1 or a_2, and next state s' which can be s_1 or s_2.

Perform Q-learning after each step, and provide the values of the Q-function after each update. Use learning rate $\alpha = 0.5$ and a discount factor $\gamma = 0.5$ for each step. Initialize your Q-function with zeros for all states and actions.

step	s	a	r	s'
1	s_1	a_1	-10	s_1
2	s_1	a_2	-10	s_2
3	s_2	a_1	20	s_1
4	s_1	a_2	-10	s_2

a) (4 pts) Report the values of the Q-function after each of the steps described in the table above.

b) (2 pts) What is the optimal policy after the last step?

Solution:

a) Recall the update rule for Q-learning

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left(r + \gamma \max_{a'} Q(a', s') - Q(s,a) \right)$$

Step 1:

$$Q(s_1, a_1) \leftarrow 0 + 0.5\,(-10 + 0.5 \max(0,0) - 0) = -5$$

Q function:

$$\begin{bmatrix} -5 & 0 \\ 0 & 0 \end{bmatrix}$$

Step 2:

$$Q(s_1, a_2) \leftarrow 0 + 0.5\,(-10 + 0.5 \max(0,0) - 0) = -5$$

Q function:

$$\begin{bmatrix} -5 & -5 \\ 0 & 0 \end{bmatrix}$$

Step 3:

$$Q(s_2, a_1) \leftarrow 0 + 0.5\,(20 + 0.5 \max(-5,-5) - 0) = 8.75$$

Q function:

$$\begin{bmatrix} -5 & -5 \\ 8.75 & 0 \end{bmatrix}$$

Step 4:

$$Q(s_1, a_2) \leftarrow -5 + 0.5\,(-10 + 0.5 \max(8.75,0) - (-5)) = -5.3125$$

$$\begin{bmatrix} -5 & -5.3125 \\ 8.75 & 0 \end{bmatrix}$$

b) Given that after Step 4

$$Q(s_1, a_1) = -5 \qquad Q(s_1, a_2) = -5.3125$$
$$Q(s_2, a_1) = 8.75 \qquad Q(s_2, a_2) = 0$$

We conclude that the optimal policy is

$$\pi_Q(s_1) = a_1$$
$$\pi_Q(s_2) = a_1$$

since $Q(s_1, a_1) > Q(s_1, a_2)$ and $Q(s_2, a_1) > Q(s_2, a_2)$.

Question 6. (6 pts) Assume you are interested in an MDP with 5 discrete states, 3 discrete actions, and unknown transition and reward functions. You decide to take a model-based approach and learn the transition and reward functions.

a) (2 pts) What is an advantage of Bayesian model-based methods over maximum likelihood model-based methods for reinforcement learning?

b) (2 pts) If you model your belief over the transition probabilities using a collection of independent Dirichlet distributions, how many parameters will you need in total to describe these Dirichlet distributions?

c) (2 pts) If you have a strong belief that $P(s' \mid s_1, a_1)$ is a uniform distribution, give one possible initialization of the pseudocounts for the Dirichlet distribution over $P(s' \mid s_1, a_1)$ that captures this belief.

Solution:

a Possible answers:

- A Bayesian approach will not assign 0 probability to rare transitions that are not observed in the dataset.
- Some algorithms (like posterior sampling) that depend on a Bayesian approach build in exploration into the process so you don't have to pick an exploration strategy and/or tune exploration hyper parameters.

Other advantages of a Bayesian over maximum-likelihood approach will be accepted for full credit.

b There are $|S|^2|A| = 75$ parameters. Each state-action pair will require $|S| = 5$ parameters to describe its Dirichlet distribution, we have $|S||A| = 5 \cdot 3$ state action pairs, so we will need

$$5 \cdot (5 \cdot 3) = 75$$

parameters total.

c We would want to initialize all pseudocounts to be the same large number. For example, $\text{Dir}(100, 100, 100, 100, 100)$. We will accept any solution where (i) all counts are equal and (ii) all counts are greater than or equal to 2, although to fit the description "strong" belief you would typically want higher counts than 2.