ECE 493, Spring 2020, Assignment 1
Due: Friday June 19, 11:59pm

Submit to the UWaterloo Crowdmark site using the link you received via email. Your answer can be handwritten converted to an electronic file by and scanner or camera; or the answers can be typed up in a word processor or LaTeX and submitted as a pdf.

1 Basics of Probability

1. Let X and Y be two random variables. If X and Y are uncorrelated, does it imply X and Y are independent? If yes, show how, else provide a counterexample.

   Definitions:

   - $Cov(X,Y) = E[XY] - E[X]E[Y]$
   - $Corr(X,Y) = \frac{Cov(X,Y)]}{\sqrt{Var(X)*Var(Y)}}$

   To be uncorrelated means to have a *Covariance* of 0:

   $$E[XY] = E[X]E[Y]. \tag{1}$$

   $\therefore$ X and Y are independent.

2. A satellite is sending a binary message/code, i.e. a sequence of 1s and 0s. Suppose 70% of the message to be sent is 0s. There is a 80% chance of a given 0 or 1 being received correctly. Find the probability that 0 was sent if 1 was received.

   Definitions:

   - $P(A|B) = \frac{P(B|A)*P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)}$

   Let the event $transmission = 0_2$ be represented as $t0_2$ and the event $received = 1_2$ be represented as $r1_2$.

   $$P(transmission = 0_2|received = 1_2) = P(t0_2|r1_2) \tag{2}$$

   $$\begin{aligned} P(t0_2|r1_2) &= \frac{P(r1_2|t0_2)P(t0_2)}{P(r1_2|t0_2)P(t0_2) + P(r1_2|t1_2)P(t1_2)} \\ &= \frac{(1-0.8)*0.7}{(1-0.8)*0.7 + 0.8*0.3} \\ &= 0.37 \end{aligned} \tag{3}$$

   $\therefore$ The probability that a 0_2 was sent given that a 1_2 was received is 37%.

3. **Basic Inference**: The probability of getting a headache is 1/10, getting the flu is 1/40, and if you have a flu, there's 1/2 chance of getting the headache. One day you wake up with a headache and say 'Dang, I have a headache. Since 50% of flues are associated with headaches,

I must have a 50-50 chance of coming down with the flu'. Is this line of reasoning correct? Why/Why not?

Definitions:

- $P(A|B) = \frac{P(B|A)*P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)}$

This line of reasoning is not correct because the problem states that the of getting the a headache given you have the flu is 50%. The problem **does not** state that the probability of getting the flu given you have a headache is 50%. These two statements are not equivalent.

Let h be the event that you have a headache, and f be the event that you have the flu. The problem describes the following conditional probability:

$$\begin{aligned} P(f|h) &= \frac{P(h|f) * P(f)}{P(h|f)P(f) + P(h|f^c)P(f^c)} \\ &= \frac{0.5 * 0.025}{0.5 * 0.025 + 0.1 * (1 - 0.025)} \\ &= 0.11 \end{aligned} \tag{4}$$

$\therefore$ There is only an 11% chance of getting the flu given you have a headache, not 50%.

4. **Bayesian Learning**: We are interested in forming a belief (hypothesis) about our environment based on what we have observed (evidence). Bayes rule is important when we want to assign a probability to our hypothesis after observing some evidence. Recall Bayes Rule:

$$\mathbb{P}(H|e) = \frac{\mathbb{P}(e|H)\mathbb{P}(H)}{\mathbb{P}(e)},$$

where H is the hypothesis, e is the evidence, $\mathbb{P}(H)$ is our prior distribution over the different hypotheses, $\mathbb{P}(e|H)$ is the likelihood of the evidence given our hypothesis, $\mathbb{P}(H|e)$ is the posterior distribution of the different hypotheses given the evidence, and $\mathbb{P}(e)$ is the probability of the evidence.

Suppose there are five kinds of bags of marbles:

(a) 10% are bags of type h_1 with 100% red marbles.

(b) 20% are bags of type h_2 with 75% red marbles and 25% green marbles.

(c) 40% are bags of type h_3 with 50% red marbles and 50% green marbles.

(d) 20% are bags of type h_4 with 25% red marbles and 75% green marbles.

(e) 10% are bags of type h_5 with 100% green marbles.

We buy a bag at random, pick 5 marbles from the bag and observe them to be all green. We are interested in knowing what type of bag we bought and what colour the next marble will be.

Let e be the event that the 5 marbles drawn are green and let $\mathbb{P}(e) = \alpha$.

ASSUMPTION: Each marble pick is independent (picking marbles then putting thme back in the bag).

(a) Given that the 5 marbles drawn were green, compute the probability of each type of bag i.e. what is the probability of bag type h_i for $i = \{1, 2, 3, 4, 5\}$, given that 5 marbles picked were green?

Definitions:

- $P(A|B) = \frac{P(B|A)*P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)}$

$$P(h_i|5g) = \frac{P(5g|h_i) * P(h_i)}{P(5g|h_i) * P(h_i) + P(5g|h_i^c)P(h_i^c)} \tag{5}$$

$$\begin{aligned} P(h_1|5g) &= \frac{0 * 0.1}{0 * 0.1 + 0.25^5 * 0.2 + 0.5^5 * 0.4 + 0.75^5 * 0.2 + 1^5 * 0.1} \\ &= 0 \end{aligned} \tag{6}$$

$$\begin{aligned} P(h_2|5g) &= \frac{0.25^5 * 0.2}{0 * 0.1 + 0.25^5 * 0.2 + 0.5^5 * 0.4 + 0.75^5 * 0.2 + 1^5 * 0.1} \\ &= 0.00122 \end{aligned} \tag{7}$$

$$\begin{aligned} P(h_3|5g) &= \frac{0.5^5 * 0.4}{0 * 0.1 + 0.25^5 * 0.2 + 0.5^5 * 0.4 + 0.75^5 * 0.2 + 1^5 * 0.1} \\ &= 0.078 \end{aligned} \tag{8}$$

$$\begin{aligned} P(h_4|5g) &= \frac{0.75^5 * 0.2}{0 * 0.1 + 0.25^5 * 0.2 + 0.5^5 * 0.4 + 0.75^5 * 0.2 + 1^5 * 0.1} \\ &= 0.30 \end{aligned} \tag{9}$$

$$\begin{aligned} P(h_5|5g) &= \frac{1^5 * 0.1}{0 * 0.1 + 0.25^5 * 0.2 + 0.5^5 * 0.4 + 0.75^5 * 0.2 + 1^5 * 0.1} \\ &= 0.62 \end{aligned} \tag{10}$$

(b) Use the results from part (a) to find the value of α using the axioms of probability.

Definitions:

- $P(A) = P(A|B) * P(B) + P(A|B^c) * P(B^c)$

$$\begin{aligned} P(e) &= \sum_{i=1}^{i=5} P(e|h_i) * P(h_i) \\ &= 0 * 0.1 + 0.25^5 * 0.2 + 0.5^5 * 0.4 + 0.75^5 * 0.2 + 1^5 * 0.1 \quad = 0.16 \end{aligned} \tag{11}$$

$\therefore \alpha = 0.16$

(c) Let X be a random variable that represents the outcome of the next pick from the bag i.e. X can be either green or red. How would you make a prediction about the outcome of X, given that the first 5 marbles picked were green? (Give an equation for $\mathbb{P}(X|e)$) Hint: Think about how you can use the concept of joint and marginal distributions with the events of picking different types of bags.

Let Y represent a random variable for the number of green marbles drawn from the bag. Then to solve for

$$p_{X|Y}(x|e) = \frac{p_{YX}(e, x)}{p_Y(e)} \tag{12}$$

(d) Use the result from part (c) to compute $\mathbb{P}(X = \text{"green"}|e)$.

Let the event of the 6th marble picked being green be represented as g.

$$\begin{aligned} p_{X|Y}(g|e) &= \frac{p_{XY}(g,e)}{p_Y(e)} \\ p_{X|Y}(g|e) &= \frac{p_{Y|X}(e,g) * p_X(g)}{p_Y(e)} \end{aligned} \tag{13}$$

From 4.b, it can be seen that:

$$p_Y(e) = 0.16. \tag{14}$$

Following a similar process as 4.a & 4.b, it can be seen that:

$$\begin{aligned} p_X(g) &= 0^1 * 0.1 + 0.25^1 * 0.2 + 0.5^1 * 0.4 + 0.75^1 * 0.2 + 1^1 * 0.1 \\ &= 0.5 \end{aligned} \tag{15}$$

Due to the assumption stated above, of marble pickings being independent:

$$\begin{aligned} p_{XY}(g,e) &= p_X(g) * p_Y(e) \\ &= 0.5 * 0.16 \\ &= 0.08 \end{aligned} \tag{16}$$

Putting it all together:

$$\begin{aligned} p_{X|Y}(g|e) &= \frac{p_{Y|X}(e,g) * p_X(g)}{p_Y(e)} \\ &= \frac{0.5 * 0.16 * 0.5}{0.16} \\ &= 0.25 \end{aligned} \tag{17}$$

$\therefore$ The probability of picking a green marble after picking 5 green marbles is 25%.

2 Multi-Armed Bandits

1. Consider a k-armed bandit problem with k = 4 actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of Q1(a) = 0, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1$, $R_1 = 1$, $A_2 = 2$, $R_2 = 1$, $A_3 = 2$, $R_3 = 2$, $A_4 = 2$, $R_4 = 2$, $A_5 = 3$, $R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

Table 1: k-Armed Bandit Actions

Step	Action	Reward	Action with Largest Reward	ϵ action	Greedy action
1	1	1		Maybe	Maybe
2	2	1	1	Yes	No
3	2	2	1,2	Maybe	Maybe
4	2	2	2	Maybe	Maybe
5	3	0	2	Yes	No

Note: All actions can potentially be an ϵ action, as an ϵ action choice is independent of action values (meaning it can pick the best action), " ...instead select randomly from among all the actions with equal probability, independently of the action-value estimates ".

2. Suppose you face a 2-armed bandit task whose true action values change randomly from time step to time step. Specifically, suppose that, for any time step, the true values of actions 1 and 2 are respectively 10 and 20 with probability 0.5 (case A), and 90 and 80 with probability 0.5 (case B). If you are not able to tell which case you face at any step, what is the best expected reward you can achieve and how should you behave to achieve it? Now suppose that on each step you are told whether you are facing case A or case B (although you still don't know the true action values). This is an associative search task. What is the best expected reward you can achieve in this task, and how should you behave to achieve it?

   For the first part of the question, it does not matter which initial bandit is chosen. After choosing an initial bandit, the greedy approach should be employed (exploiting and never exploring). If the greedy approach is taken, 50% of the choices will lead to un-optimal cases, and 50% of the choices will lead to optimal cases, but the reward difference between optimal and sub-optimal is the same for action 1 and action 2.

   For example, if we iterate on $2x$ number of actions, we get the following rewards depending on the initial bandit chosen.

   - A1 initial bandit: x*10 + x*90 = x*100
   - A2 initial bandit: x*20 + x*80 = x*100

   For the second part of the question, if we know whether the environment is in state A or state B, the first set of actions should be spent exploring the environment to completely map out a single case, either A or B. This will take at most 3 actions. Once a single case has been mapped, the agent can choose the greedy action for that state, and the opposite action for the second case. A drawback to this approach is that it is not a scalable solution to problems with more than 2 bandits.

   For example, if we find that action 1 is optimal in case A, for all case A environments the agent will choose action 1, and action 2 for case B.

3 MDPs

1. Let $G_t = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{T-1} R_T$. Suppose $\gamma = 0.5$ and the following sequence of rewards is received: $R_1 = 1$, $R_2 = 2$, $R_3 = 5$, $R_4 = -2$, $R_5 = 2$, and $T = 5$. Solve for G_0, G_1, G_2, G_3, G_4, and G_5.

   Definitions:

   - $G_t = R_{t+1} + \gamma G_{t+1}$
   - $G_T = 0$

$$G_5 = 0 \tag{18}$$

$$\begin{aligned} G_4 &= R_5 + \gamma G_5 \\ &= 2 + 0.5 * 0 \\ &= 2 \end{aligned} \tag{19}$$

$$
\begin{aligned}
G_3 &= R_4 + \gamma G_4 \\
&= -2 + 0.5 * 2 \\
&= -1
\end{aligned} \tag{20}
$$

$$
\begin{aligned}
G_2 &= R_3 + \gamma G_3 \\
&= 5 + 0.5 * -1 \\
&= 4.5
\end{aligned} \tag{21}
$$

$$
\begin{aligned}
G_1 &= R_2 + \gamma G_2 \\
&= 2 + 0.5 * 4.5 \\
&= 4.25
\end{aligned} \tag{22}
$$

$$
\begin{aligned}
G_0 &= R_1 + \gamma G_1 \\
&= 1 + 0.5 * 4.25 \\
&= 3.125
\end{aligned} \tag{23}
$$

2. Let $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$. Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$, followed by an infinite sequence of 6s. What are G_0 and G_1?

   Definitions:

   - $\sum_{k=0}^{\infty} ar^k = \frac{a}{1-r}$ for $|r| < 1$

$$
\begin{aligned}
G_1 &= \sum_{k=0}^{\infty} 0.9^k 6 \\
&= \frac{6}{1 - 0.9} \\
&= 60
\end{aligned} \tag{24}
$$

$$
\begin{aligned}
G_0 &= R_1 + \gamma G_1 \\
&= 2 + 0.9 * 60 \\
&= 56
\end{aligned} \tag{25}
$$

3. Suppose you treated pole-balancing as an episodic task but also used discounting, with all rewards zero except for -1 upon failure. What then would the return be at each time? How does this return differ from that in the discounted, continuing formulation of this task? (For explanation of pole-balancing see Example 3.4 on Sutton's RL Book)

   The return at each instance will be either -1 for failure and 0 for success. This differs from the continuous case as the return of each episode is independent of the length of the episode. As the return of each episode is independent of the length of the episode, this episodic task is not focused on maximizing time spent balancing the pole, but instead searches for a way to balance the pole forever (equilibrium).

4. Figure 1 (left) shows a rectangular grid world representation of a simple finite MDP. The cells of the grid correspond to the states of the environment. At each cell, four actions are

possible: north, south, east, and west, which deterministically cause the agent to move one cell in the respective direction on the grid. Actions that would take the agent off the grid leave its location unchanged, but also result in a reward of -1. Other actions result in a reward of 0, except those that move the agent out of the special states A and B. From state A, all four actions yield a reward of +10 and take the agent to A'. From state B, all actions yield a reward of +5 and take the agent to B'.

Assuming discount factor $\gamma = 0.9$ and equal probabilities to all four actions, value functions seen in Figure 1 (right) obtained. Show numerically that the Bellman expectation equation holds for the cell at the middle of top row, valued at +4.4, with respect to its four neighboring states, valued at +8.8, +2.3, +5.3 and off-grid (These numbers are accurate only to one decimal place.) (**Hint:** Try to understand what happens when agent takes action to north, off-grid)

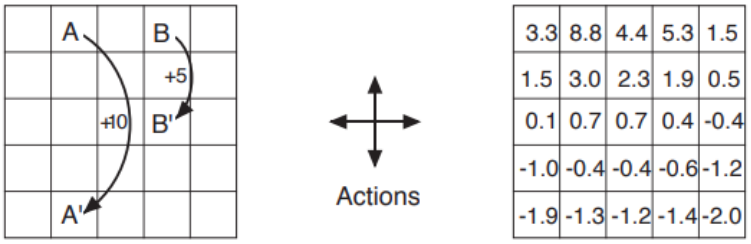

Figure 1: Gridworld example: exceptional reward dynamics (left) and state-value function for the equiprobable random policy (right)