Worksheet 2

Q1

Suppose a game where you choose to flip one of two (possibly unfair) coins. You win $1 if your chosen coin shows heads and lose $1 if it shows tails.

1. Model this as a K–armed bandit problem: define the action set.
2. Is the reward a deterministic or stochastic function of your action?
3. You do not know the coin flip probabilities. Instead, you are able to view 6 sample flips for each coin respectively: (T,H,H,T,T,T) and (H,T,H,H,H,T). Use the sample average formula (equation 2.1 in the book) to compute the estimates of the value of each action.
4. Decide on which coin to flip next! Assume it's an exploit step.

Answer:

1. This problem can be modeled as a 2–armed bandit problem. The two actions here are: a_1: flipping the first coin and a_2: flipping the second coin. The optimal action value of each action is $q_*(a_i) = \mathbb{E}[R_t|A_t = a] = (+1) * p_i + (-1) * (1 - p_i) = 2p_i - 1$ for $i \in \{1, 2\}$ with p_i denoting the probability of getting a heads for the ith coin.
2. The reward is a stochastic function of the action. Choosing a coin fixes the probability of getting a head/tail from that coin and the randomness comes from the flipping action itself.
3. The reward sequence of the first coin is $\{-1, +1, +1, -1, -1, -1\}$ and that of the second coin $\{+1, -1, +1, +1, +1, -1\}$. Taking the average, we get $Q_7(a_1) = -0.33$ and $Q_7(a_2) = +0.33$.
4. Given the above action value estimates, we'll choose the second coin.

Q2

Consider a problem where an agent is trying to get to school and must choose how long to wait at the bus stop. The agent can walk to school, but wants to catch the bus if possible. At the same time, the agent doesn't want to wait too long because of delays. Unfortunately, the time it takes for a bus to arrive is effectively random.

1. This is not a K-armed bandit problem because your action set, how long to wait, is not a positive integer. How could you reformulate the bus-waiting problem as a K-armed bandit?
2. In problems with continuous random variables, we rarely know the distribution of a variable. Instead, we often make assumptions on its distribution. One commonly assumed distribution for continuous random variables is the Gaussian (or Normal) distribution. Is the Gaussian assumption in this bus-waiting problem reasonable? Justify your answer using properties of the Gaussian distribution and other assumptions about the distribution of time spent waiting at the bus stop.

Answer:

1. One way could be to fix a maximal waiting time and then discretize the waiting time into a finite number of intervals. For example, we could choose the maximal waiting time to be 60 minutes and discretize the waiting time to 0 minutes, 5 minutes, 10 minutes, ... , 60 minutes to get a total of 13 actions.

   An alternate could be to model the problem as continuous bandit.

2. Yes, assuming a Gaussian distribution for the arrival time (assume that the agent waits until the bus arrives, so that the waiting time of the agent is equal to the arrival time of the bus) is reasonable:

   (a) The arrival time is continuous and Gaussian distribution is a continuous distribution as well.

   (b) The Gaussian distribution (because of Central Limit Theorem) is often a suitable modeling choice for processes that depend on a number of other probabilistic factors. For example, the arrival time here may depend on multiple factors such as the traffic situation in the city.

   (c) More intuitively, the arrival time of the bus would be close to the scheduled time, and the probability of the bus arriving too early or too late than this scheduled time should be low – this is modeled by the tails in the Gaussian distribution.

   Although, in order to use a Gaussian distribution here, we'll have to restrict the range of the random variable to non–negative real numbers in Gaussian distribution. Another popular choice for modeling waiting times is the exponential distribution.

Q3

Challenge Problem: Imagine your agent is solving a 3-armed bandit problem. Unlike usual, you get extra information: you know that the reward for each action is randomly distributed according to a Gaussian distribution with unknown mean, and a variance of 1. Each of the three actions might have a different mean, μ_a for Gaussian $\mathcal{N}(\mu_a, 1)$. How might the action rule for the UCB algorithm change, given this information? Hint: Recall that a 95% confidence interval assuming a Gaussian distribution with variance $\sigma = 1$, for sample average $Q_t(a)$ using $N_t(a)$ samples, is $(Q_t(a) - 1.96\frac{\sigma}{\sqrt{N_t(a)}}, Q_t(a) + 1.96\frac{\sigma}{\sqrt{N_t(a)}})$.

Answer:

Once we know that the reward for each action is randomly distributed according to a Gaussian distribution, we have a direct way of computing the uncertainty in our action value estimates: use 1.96 times the sample variance for a 95% confidence. The UCB algorithm can then be written as:

$$A_t := \arg\max_{a \in \{1,2,3\}} \left[Q_t(a) + 1.96\frac{\sigma}{\sqrt{N_t(a)}} \right],$$

where σ is the variance of action values (preferrably computed incrementally), $N_t(a)$ is the count of how many times each action has been taken, and $Q_t(a)$ is the action value estimate of action a at time t. Using the value of 1.96 in the above expression ensures that there is just a $\frac{1-0.95}{2} = 0.025 \equiv 2.5\%$ chance that actual action value $q^*(a) > \left[Q_t(a) + 1.96\frac{\sigma}{\sqrt{N_t(a)}} \right]$. We could use constants other than 1.96 for a different confidence level, computed with the help of properties for the Gaussian distribution. This removes the need to find a the hyperparameter c.

Q4

In class we saw how the bandit problem can be formulated as a MDP. Suppose we have a bandit problem with two arms, with mean rewards $\mu_1 = 10, \mu_2 = 5$ for arm 1 and arm 2 respectively. Suppose we have an episodic task where an agent plays the above bandit problem twice. However, if they pull arm 1 (take action 1) then the mean rewards for each arm switch, that is $\mu_1 = 5, \mu_2 = 10$. If arm 2 is pulled the bandit problem is replayed without change. If the agent plays the policy $\pi(\text{arm } 1|s) = 0.3$ at both time steps then what is the value function? In other words, find $v_\pi(S)$ for both states S_1 and S_2.

Answer:

THE QUESTION DESCRIPTION IS UNCLEAR TO ME.

Let us model this problem as an MDP. Let there be four different states:

1. s_1 corresponding to the starting state with $\mu_1^{(s_1)} = 10, \mu_2^{(s_1)} = 5$,
2. s_2 with $\mu_1^{(s_2)} = 5, \mu_2^{(s_2)} = 10$ (this is the state we reach after taking action a_1 in state s_1),
3. s_3 with $\mu_1^{(s_3)} = 10, \mu_2^{(s_3)} = 5$ (this is the state we reach after taking action a_2 in state s_1), and
4. s_T, the terminal state, which we reach after taking either action a_1 or a_2 from either of the states s_1 and s_2.

All the transitions are deterministic, and the expected reward from a state is given by $\mu_1^{(s_i)}$ and $\mu_2^{(s_i)}$ for state s_i. Now let us find $v_\pi(S)$ for all the four states, for the general policy: $\pi(a_1 \mid S = s_i) = p_i$ and $\pi(a_2 \mid S = s_i) = 1 - p_i$ for state s_i with $i \in \{1, 2, 3\}$.

- $v_\pi(s_T) = 0$ by definition.
- For state s_2, we'll transition into s_T irrespective of the action we take. Thus, we have

$$\begin{aligned}
v_\pi(s_2) &= \sum_a \pi(a|s_2) \sum_{s',r} p(s', r|s_2, a)(r + \gamma v_\pi(s')) \\
&= \pi(a_1|s_2) \sum_r p(s_T, r|s_2, a_1)(r + \gamma v_\pi(s_T)) + \pi(a_2|s_2) \sum_r p(s_T, r|s_2, a_2)(r + \gamma v_\pi(s_T)) \\
&= p_2 \sum_r p(s_T, r|s_2, a_1)(r + \gamma \times 0) + (1 - p_2) \sum_r p(s_T, r|s_2, a_2)(r + \gamma \times 0) \\
&= p_2 \times \mu_1^{(s_2)} + (1 - p_2) \times \mu_2^{(s_2)} \\
&= p_2 \times 5 + (1 - p_2) \times 10 \\
&= 10 - 5p_2.
\end{aligned}$$

- For state s_3 we have, we'll again transition into s_T irrespective of the action we take. So, we have

$$\begin{aligned}
v_\pi(s_3) &= \sum_a \pi(a|s_3) \sum_{s',r} p(s', r|s_3, a)(r + \gamma v_\pi(s')) \\
&= \pi(a_1|s_3) \sum_r p(s_T, r|s_3, a_1)(r + v_\pi(s_T)) + \pi(a_2|s_3) \sum_r p(s_T, r|s_3, a_2)(r + v_\pi(s_T)) \\
&= p_3 \sum_r p(s_T, r|s_3, a_1)(r + \gamma \times 0) + (1 - p_3) \sum_r p(s_T, r|s_3, a_2)(r + \gamma \times 0) \\
&= p_3 \times \mu_1^{(s_3)} + (1 - p_3) \times \mu_2^{(s_3)} \\
&= p_3 \times 10 + (1 - p_3) \times 3 \\
&= 3 + 7p_3.
\end{aligned}$$

- Finally, for state s_1 we transition into s_2 upon taking action a_1 and into state s_3 upon taking action a_2. Therefore,

$$\begin{aligned}
v_\pi(s_1) &= \sum_a \pi(a|s_1) \sum_{s',r} p(s', r|s_1, a)(r + \gamma v_\pi(s')) \\
&= \pi(a_1|s_1) \sum_r p(s_2, r|s_1, a_1)(r + \gamma v_\pi(s_2)) + \pi(a_2|s_1) \sum_r p(s_3, r|s_1, a_2)(r + \gamma v_\pi(s_3)) \\
&= p_1 \sum_r p(s_2, r|s_1, a_1)(r + \gamma(10 - 5p_2)) + (1 - p_1) \sum_r p(s_3, r|s_1, a_2)(r + \gamma(3 + 7p_3)) \\
&= p_1 \sum_r r \cdot p(s_2, r|s_1, a_1) + \gamma p_1(10 - 5p_2) + (1 - p_1) \sum_r r \cdot p(s_3, r|s_1, a_2) + \gamma(1 - p_1)(3 + 7p_3) \\
&= p_1 \mu_1^{(s_1)} + \gamma p_1(10 - 5p_2) + (1 - p_1)\mu_2^{(s_1)} + \gamma(1 - p_1)(3 + 7p_3) \\
&= 10p_1 + \gamma p_1(10 - 5p_2) + 5(1 - p_1) + \gamma(1 - p_1)(3 + 7p_3).
\end{aligned}$$

For this question, assume $\gamma = 1$ and we are given that $p_1 = p_2 = p_3 = 0.3$. Putting these values in the above expressions we obtain:

$$v_\pi(s_1) = 12.62, v_\pi(s_2) = 8.5, v_\pi(s_3) = 5.1, \text{and } v_\pi(s_T) = 0.$$

Q5

Consider the above MDP where the agent plays the same bandit problem twice and the action values are switched if the agent selects action 1 at time 1. Suppose we now have a discount factor $\gamma = 0.5$, and the agent select a policy π that is the same for both time steps. What is the optimal policy? Now suppose that the agent can play a different policy at each time step, what would be the optimal policy?

Answer:

- **Part 1:** Continuing from the previous answer, we have $p_1 = p_2 = p_3 = p$ and $\gamma = 0.5$. Now we need to find p such that $v_\pi(s_1)$ is maximized. We have

$$\begin{aligned}
v_\pi(s_1) &= 10p_1 + \gamma p_1(10 - 5p_2) + 5(1 - p_1) + \gamma(1 - p_1)(3 + 7p_3) \\
&= 10p + 0.5p(10 - 5p) + 5(1 - p) + 0.5(1 - p)(3 + 7p) \\
&= -6p^2 + 12p + 6.5.
\end{aligned}$$

  This expression attains the maximum value of 12.5 at $p = 1$. Therefore, the optimal policy in this case would be to take a_1 at each state with probability 1.

- **Part 2:** Continuing from answer to the previous question, we get the maximum value of $v_\pi(s_2) = 10$ and $v_\pi(s_3) = 10$ for $p_2 = 0$ and $p_3 = 1$. Putting these into the expression for $v_\pi(s_1)$ we get:

$$\begin{aligned}
v_\pi(s_1) &= 10p_1 + \gamma p_1(10 - 5p_2) + 5(1 - p_1) + \gamma(1 - p_1)(3 + 7p_3) \\
&= 10p_1 + 0.5p_1 \times 10 + 5(1 - p_1) + 0.5(1 - p_1) \times 10 \\
&= 10 + 5p_1.
\end{aligned}$$

  This expression attains a maximal value of $v_\pi(s_1) = 15$ for $p_1 = 1$. Therefore, the optimal policy is to take action a_1 in state s_1, action a_2 in state s_2, and action a_1 in state s_3.

Q6

1. Show that if X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

2. Recall that the correlation between two variables X and Y is

$$\frac{E[(X - E[X]) \cdot (Y - E[Y])]}{\sigma_X \sigma_Y}.$$

   Show that if X and Y are independent, then they are uncorrelated (correlation is zero). Provide an example of X and Y that are uncorrelated but not independent.

Answer:

1. We have

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_x \sum_y xy \Pr(X = x, Y = y) \\
&= \sum_x \sum_y xy \Pr(X = x) \Pr(Y = y) \\
&= \sum_x x \Pr(X = x) \sum_y y \Pr(Y = y) \\
&= \mathbb{E}[X]\,\mathbb{E}[Y],
\end{aligned}$$

   where in the second line, we have used the fact that $\Pr(X = x, Y = y) = \Pr(X = x) \Pr(Y = y)$ when X and Y are independent.

2. We show that the numerator of the correlation is zero, when X and Y are independent:

$$\begin{aligned}
\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] &= \mathbb{E}[XY - X\,\mathbb{E}[Y] - Y\,\mathbb{E}[X] + \mathbb{E}[X]\,\mathbb{E}[Y]] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[Y]\,\mathbb{E}[X] + \mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] \\
&= 0.
\end{aligned}$$

   I'm unable to come up with a good example for X and Y which are uncorrelated and dependent. On the internet, the following example is quite widely used (`http://mathforum.org/library/drmath/view/69928.html`): Let $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$, then their correlation is zero because the numerator of the correlation term is $\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[X^3] - \mathbb{E}[X]^2 = 0 - 0 = 0$, whereas the variables are clearly dependent on each other.