Solution of Minor 2

Q1: Multi-Armed Bandits

Time slot	1	2	3	4	5	6	7	8	9	10
Action	2	3	2	2	1	1	1	2	3	3
Reward	4.5	2.0	1.0	4.0	2.5	2.6	7.0	5.0	1.6	3.2

To solve this problem, we first need to calculate $Q_t(a)$ and $N_t(a)$ for $t = 11$ and all $a \in \{1,2,3\}$. Using the above table we have,

$$N_{11}(1) = 3, \qquad Q_{11}(1) = \frac{2.5 + 2.6 + 7.0}{3} = 4.033$$
$$N_{11}(2) = 4, \qquad Q_{11}(2) = \frac{4.5 + 1.0 + 4.0 + 5.0}{4} = 3.625$$
$$N_{11}(3) = 3, \qquad Q_{11}(3) = \frac{2.0 + 1.6 + 3.2}{3} = 2.267$$

(a) UCB algorithm chooses the action with the largest,

$$\eta_t(a) = Q_t(a) + \sqrt{\frac{2ln(t)}{N_t(a)}}$$

Using the value of $N_t(a)$ and $Q_t(a)$ calculated above,

$$\eta_{11}(1) = 4.033 + \sqrt{\frac{2ln(11)}{3}} = 5.297$$

$$\eta_{11}(2) = 3.625 + \sqrt{\frac{2ln(11)}{4}} = 4.720$$

$$\eta_{11}(3) = 2.267 + \sqrt{\frac{2ln(11)}{3}} = 3.531$$

The agent will choose **action 1** in time $t = 11$ because it has the highest $\eta_{11}(a)$ value.

(b) Using the value of $Q_t(a)$ calculated above, we can see that in time $t = 11$, action 2 DOES NOT have the highest $Q_t(a)$ value. Hence, action 2 will get selected at time $t = 11$ only when**:**

- the agent decides to explore which has the probability of $\boldsymbol{\varepsilon = 0.1}$, **AND**
- action 2 gets selected during uniform random sampling among the three actions. This has a probability of $\mathbf{1/3}$.

The probability of **both** these conditions being satisfied is $\mathbf{0.1 \cdot \frac{1}{3} = 0.033}$.

(c) We want to calculate $P[r_{11} = x]$, the probability distribution of reward for $t = 11$. Using *law of total expectation*, we get,

$$P[r_{11} = x] = \sum_{a_{11}=1}^{3} P[r_{11} = x \mid a_{11}]P[a_{11}] \qquad (1)$$

Now, using part (b) we know that $P[a_{11} = 2] = 0.033$. Similarly, $P[a_{11} = 3] = 0.033$, and $P[a_{11} = 1] = 0.9 + \frac{0.1}{3} = 0.933$. Also,

$$P[r_{11} = x \mid a_{11} = 1] = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{(x-3.5)^2}{2}\right)$$
$$P[r_{11} = x \mid a_{11} = 2] = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{(x-4)^2}{2}\right)$$
$$P[r_{11} = x \mid a_{11} = 3] = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{(x-2)^2}{2}\right)$$

Substituting these values in (1) we get,

$$P[r_{11} = x] = \frac{0.933}{\sqrt{2\pi}} exp\left(-\frac{(x-3.5)^2}{2}\right) + \frac{0.033}{\sqrt{2\pi}} exp\left(-\frac{(x-4)^2}{2}\right) + \frac{0.033}{\sqrt{2\pi}} exp\left(-\frac{(x-2)^2}{2}\right)$$

Q2: Markov Decision Process

(a) The state is the question number n because it decides the current reward and the probability of transitioning to the next question. The state space is $S = \{1,2,\cdots,N,end\}$. end is the terminal state that is required because we are dealing with an episodic task.

(b) The action is to quit ($a = 0$), or to answer question ($a = 1$).

(c) The average reward for the state-action pairs, $r(n,a)$, can be expressed as,

$$r(n,a) = \begin{cases} 0 & ; n \neq end, a = 0 \\ \theta_n f(n) - (1-\theta_n)\sum_{i=1}^{n-1} f(i) & ; n \neq end, a = 1 \\ 0 & ; n = end \end{cases}$$

Explanation of $n \neq end, a = 0$: If the contestant quits, he/she does not win the money for question number n. Hence, the average **immediate** reward is 0.

Explanation of $n \neq end, a = 1$: If the contestant answers question, he/she wins $f(n)$ with probability θ_n and loses all the money he/she has won till now, $\sum_{i=1}^{n-1} f(i)$, with probability $(1-\theta_n)$. Note that for $n = 1$, $\sum_{i=1}^{n-1} f(i) = 0$.

(PTO)

(d) The Bellman optimality equation for this problem is as follows.

$$V^*(n) = \max_{a\in\{0,1\}} Q^*(n,a) \quad where,$$

$$Q^*(n,0) = 0 + V^*(end) \qquad ; n \neq end$$

$$Q^*(n,1) = \left(\theta_n f(n) - (1-\theta_n)\sum_{i=1}^{n-1} f(i)\right) + \theta_n V^*(n+1) + (1-\theta_n)V^*(end) \; ; n \notin \{N, end\}$$

$$Q^*(n,1) = \left(\theta_n f(n) - (1-\theta_n)\sum_{i=1}^{n-1} f(i)\right) + V^*(end) \qquad ; \; n = N$$

$$Q^*(end,a) = 0 \qquad ; a \in \{0,1\}$$

Explanation of $Q^*(n,0)$ for $n \neq end$: If the contestant quits, the immediate average reward is 0 and the game ends (hence the transition to the **end** state).

Explanation of $Q^*(n,1)$ for $n \notin \{N, end\}$: If the contestant answers the question, and he/she is **not** in the **last question**, the immediate average reward is $\theta_n f(n) - (1-\theta_n)\sum_{i=1}^{n-1} f(i)$. If the answer is correct it transitions to the next question (this is captured using the term $\theta_n V^*(n+1)$) and if the answer is wrong the game ends (this is captured using the term $(1-\theta_n)V^*(end)$).

Explanation of $Q^*(n,1)$ for $n = N$: If the contestant answers the question, and he/she is in the **last question**, the immediate average reward is $\theta_n f(n) - (1-\theta_n)\sum_{i=1}^{n-1} f(i)$ and the game ends.

Explanation of $Q^*(end,a)$ for $a \in \{0,1\}$: The net reward starting in the end state is always zero.