CS 747 (Autumn 2025) Week 1 Test (Batch 1)

5.35 p.m. – 6.00 p.m., August 8, 2025, LA 001

Name: _____ Roll number: _____

Note. There is one question in this test. You can use the space on both pages for your answer. Draw a line (either vertical or horizontal) and do all your rough work on one side of it.

Question 1. A bandit has two arms: a and a'. Both yield 0 or 1 (Bernoulli) rewards stochastically, with arm a having a mean reward of $p = \frac{1}{3}$, and arm a' having a mean reward of $p' = \frac{1}{2}$.

An algorithm L selects an arm uniformly at random (that is, with probability 50%) at each step, and pulls the selected arm. L is independently run *two* times, both times starting from a null history. In the first run, it generates a history (that is, an action-reward sequence)

$$a_1^0, r_1^0, a_1^1, r_1^1, a_1^2, r_1^2, \ldots$$

wherein we have used superscripts to denote the time step, and subscripts to indicate the run. Adopting the same notation, the second run generates a history

$$a_2^0, r_2^0, a_2^1, r_2^1, a_2^2, r_2^2, \ldots.$$

What is the probability that the histories generated by the two runs are *identical* for the first time step (that is, time step 0)? In other words, what is

$$\mathbb{P}\{a_1^0 = a_2^0 \text{ and } r_1^0 = r_2^0\}?$$

Show your steps and calculations; provide your answer as a fraction. [3 marks]

Answer 1. The sequence (a^0, r^0) can take four possible values on any run, listed below with their corresponding probabilities.

Outcome	Probability
$(a, 0)$	$\frac{1}{2} \cdot (1 - p) = \frac{1}{3}$
$(a, 1)$	$\frac{1}{2} \cdot p = \frac{1}{6}$
$(a', 0)$	$\frac{1}{2} \cdot (1 - p') = \frac{1}{4}$
$(a', 1)$	$\frac{1}{2} \cdot p' = \frac{1}{4}$

For run 1 and run 2 to produce the same sequence (a^0, r^0), they must both produce $(a, 0)$, or they must both produce $(a, 1)$, or they must both produce $(a', 0)$, or they must both produce $(a', 1)$. Since the runs are independent, the probability that runs 1 and 2 produce the same sequence a^0, r^0 is

$$\left(\frac{1}{3}\right)^2 + \left(\frac{1}{6}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 = \frac{19}{72}.$$

CS 747 (Autumn 2025) Week 1 Test (Batch 2)

6.15 p.m. – 6.40 p.m., August 8, 2025, LA 001

Name: _____ Roll number: _____

Note. There is one question in this test. You can use the space on both pages for your answer. Draw a line (either vertical or horizontal) and do all your rough work on one side of it.

Question 1. A multi-armed bandit has 3 arms: a_1, a_2, and a_3. In class we had considered the case that the rewards from bandit arms would be either 0 or 1. However, each arm in this question can generate 3 possible rewards, each with an associated probability. The table below describes the three reward distributions.

Arm	Reward	Probability
a_1	0	0.3
	5	0.6
	10	0.1
a_2	3	0.4
	7	0.4
	12	0.2
a_3	0	0.5
	10	0.1
	12	0.4

Notice, for example, that when a_2 is pulled, the reward is (i) 3 with probability 0.4; (ii) 7 with probability 0.4; and (iii) 12 with probability 0.2.

Consider an algorithm that pulls each arm once, and thereafter, at each step, probabilistically pulls an arm based on the arms' empirical mean rewards. Precisely, an arm with the highest empirical mean reward (picked arbitrarily in case of a tie) is selected with probability 0.9, and each of the other two arms is selected with probability 0.05. The selected arm is pulled.

For horizon $T \geq 1$, let $\text{rew}(T)$ denote the total reward after T pulls. What is

$$\lim_{T \to \infty} \frac{\mathbb{E}[\text{rew}(T)]}{T}?$$

Provide an informal justification, along with calculations, to accompany your answer. [3 marks]

Answer 1.

The expected reward from the arm a_1 is

$$x_1 = 0 \times 0.3 + 5 \times 0.6 + 10 \times 0.1 = 4.$$

The expected reward from the arm a_2 is

$$x_2 = 3 \times 0.4 + 7 \times 0.4 + 12 \times 0.2 = 6.4.$$

The expected reward from the arm a_3 is

$$x_3 = 0 \times 0.5 + 10 \times 0.1 + 12 \times 0.4 = 5.8.$$

Clearly a_2 is the sole optimal arm. By guaranteeing each arm at least some constant probability of being pulled on each step, our algorithm will give each arm an infinite number of pulls with high probability, so their empirical means will eventually be in the same sequence as their true means. As $T \to \infty$, arm a_2 will get 90% of the pulls, while arms a_1 and a_3 will each get 5% of the pulls. Thus

$$\begin{aligned}\lim_{T\to\infty} \frac{\mathbb{E}[\text{rew}(T)]}{T} &= (90\% \text{ of } x_2) + (5\% \text{ of } x_1) + (5\% \text{ of } x_3) \\ &= 0.9 \times 6.4 + 0.05 \times 4 + 0.05 \times 5.8 \\ &= 6.25.\end{aligned}$$