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

5.35 p.m. – 6.00 p.m., August 19, 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. In this question, we consider TS2, which is a variant of Thompson Sampling (itself henceforth abbreviated TS). As in class, we assume that our underlying bandit instance yields Benoulli rewards, with the mean reward of arm a being $p_a \in (0, 1)$.

Recall that while running TS, the number of successes and failures of arm a after t total pulls are denoted s^t_a and f^t_a, respectively. To implement TS2, we define $\bar{s}^t_a$ and $\bar{f}^t_a$ to be the number of successes and failures, respectively, from only the preceding two pulls of arm a. Indeed suppose arm a has been pulled u^t_a times in the first t pulls, and the sequence of rewards it has obtained is $r_1, \dots, r_{u^t_a}$. The subscripts in this sequence do not indicate the points of time when arm a was pulled, but rather the "pull number". Thus, for example, r_5 is the reward from the fifth pull of arm a.

- If $u^t_a = 0$: then $\bar{s}^t_a \overset{\text{def}}{=} \bar{f}^t_a \overset{\text{def}}{=} 0$.
- If $u^t_a = 1$: then $\bar{s}^t_a = r_1$ and $\bar{f}^t_a \overset{\text{def}}{=} 1 - \bar{s}^t_a$.
- If $u^t_a \geq 2$: then $\bar{s}^t_a \overset{\text{def}}{=} r_{u^t_a - 1} + r_{u^t_a}$ and $\bar{f}^t_a \overset{\text{def}}{=} 2 - \bar{s}^t_a$.

By contrast, recall that $s^t_a = r_1 + r_2 + \cdots + r_{u^t_a}$ and $f^t_a = u^t_a - s^t_a$.

TS2 is identical to TS, except that it uses $\bar{s}^t_a$ and $\bar{f}^t_a$ in place of s^t_a and f^t_a, respectively, to construct the belief distribution of each arm a at time t. However, like TS, at each step TS2 obtains samples for these belief distributions and pulls the arm corresponding to the largest sample.

For a given 2-armed bandit instance I, let $R^T(\text{TS2}, I)$ and $R^T(\text{TS}, I)$ denote the expected cumulative regrets of these algorithms, respectively, on I, for horizon T. Describe the dependence of $R^T(\text{TS2}, I)$ and $R^T(\text{TS}, I)$ on T, along with supporting mathematical justification. Based on this reasoning, work out the value of

$$\lim_{T\to\infty} \frac{R^T(\text{TS2}, I)}{R^T(\text{TS}, I)}.$$

If relevant, you can use the following facts.

- The pdf of the beta distribution with integer parameters $\alpha \geq 1$ and $\beta \geq 1$ evaluated at $x \in [0, 1]$ is given by

$$\text{beta}(x; \alpha, \beta) \overset{\text{def}}{=} \frac{(\alpha + \beta - 1)!}{(\alpha - 1)!(\beta - 1)!} \cdot x^{\alpha-1} \cdot (1 - x)^{\beta-1}.$$

- If the arms of our instance are a_1 and a_2, then the probability that TS2 pulls arm a_1 at time step t is given by

$$\int_{x=0}^{1} \text{beta}(x; \bar{s}^t_{a_1} + 1, \bar{f}^t_{a_1} + 1) \int_{y=0}^{x} \text{beta}(y; \bar{s}^t_{a_2} + 1, \bar{f}^t_{a_2} + 1)\, dy\, dx.$$

The same expression holds for TS with $\bar{s}^t_a$ replaced by s^t_a and $\bar{f}^t_a$ replaced by f^t_a.

[3 marks]

Answer 1. Since both $\bar{s}_a^t$ and $\bar{f}_a^t$ are constrained to come from the finite set $\{0, 1, 2\}$ for each arm a, we notice that the probability that any particular arm is selected at any time step comes from a finite set of positive probabilities $Q = \{q_1, q_2, \dots, q_m\}$. Indeed the size of Q is at most $3 \times 3 \times 3 \times 3$, which would account for every possible combination of $\bar{s}_{a_1}^t$, $\bar{f}_{a_1}^t$, $\bar{s}_{a_2}^t$, and $\bar{f}_{a_2}^t$.

Without loss of generality, suppose q_1 is the smallest element of Q. It follows that on each time step, each arm has a probability of at least q_1 of being pulled. On instances with different mean rewards, this implies TS2 will not be greedy in the limit (although it will perform an infinite amount of exploration). In turn, this implies TS2 will incur $\Omega(T)$ (that is, linear-in-T) regret. On the other hand, we know that the regret of TS is $O(\log T) = o(T)$. Thus,

$$\lim_{T \to \infty} \frac{R^T(\mathrm{TS2}, I)}{R^T(\mathrm{TS}, I)} = \infty.$$

In TS, note that the probability of the empirically-inferior arm being pulled goes down (to zero) as this arm gathers more and more pulls. The linear regret of TS2 is on account of this probability being *lower-bounded* by a positive constant. The student is encouraged to implement both TS and TS2 and validate the claims made above.

On bandit instances in which all arms are optimal, regret is 0 for all algorithms, and hence the limit in the question is not well-defined.

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

6.15 p.m. – 6.40 p.m., August 19, 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. Thompson Sampling is implemented on a 2-armed bandit instance, in which both arms yield Bernoulli rewards. The mean reward of arm a_1 is p_1, while the mean reward of arm a_2 is p_2.

What is the probability that the first two pulls are both of the same arm *and* both corresponding rewards are 1? In other words, what is the probability that the "arm-reward" sequence after 2 pulls is either $a_1, 1, a_1, 1$ or $a_2, 1, a_2, 1$?

- The pdf of the beta distribution with integer parameters $\alpha \geq 1$ and $\beta \geq 1$ evaluated at $x \in [0, 1]$ is given by
$$\text{beta}(x; \alpha, \beta) \overset{\text{def}}{=} \frac{(\alpha + \beta - 1)!}{(\alpha - 1)!(\beta - 1)!} \cdot x^{\alpha - 1} \cdot (1 - x)^{\beta - 1}.$$
- At time t, let s^t_a and f^t_a denote the number of successes and failures of arm $a \in \{a_1, a_2\}$. The probability that arm a_1 is pulled at time step t is given by
$$\int_{x=0}^{1} \text{beta}(x; s^t_{a_1} + 1, f^t_{a_1} + 1) \int_{y=0}^{x} \text{beta}(y; s^t_{a_2} + 1, f^t_{a_2} + 1)\ dy\ dx.$$

You can apply these formulae suitably in your working; simplify and provide your final answer in terms of p_1 and p_2. [3 marks]

Answer 1. Let the random action-reward sequence in the the first two steps be a^0, r^0, a^1, r^1, with the superscipts indicating time step.

$$\begin{aligned}\mathbb{P}\{a^0 = a_1 \text{ and } r^0 = 1 \text{ and } a^1 = a_1 \text{ and } r^1 = 1\} = \alpha_1 \cdot \alpha_2 \cdot \alpha_3 \cdot \alpha_4, \text{ where}\\ \alpha_1 = \mathbb{P}\{a^0 = a_1\},\\ \alpha_2 = \mathbb{P}\{r^0 = 1 | a^0 = a_1\},\\ \alpha_3 = \mathbb{P}\{a^1 = a_1 | a^0 = a_1 \text{ and } r^0 = 1\},\\ \alpha_4 = \mathbb{P}\{r^1 = 1 | a^1 = a_1\}.\end{aligned}$$

We work out each of these factors.

$$\begin{aligned}\alpha_1 &= \int_{x=0}^{1} \text{beta}(x; 1, 1) \int_{y=0}^{x} \text{beta}(y; 1, 1)\ dy\ dx = \frac{1}{2}.\\ \alpha_2 &= p_1.\\ \alpha_3 &= \int_{x=0}^{1} \text{beta}(x; 2, 1) \int_{y=0}^{x} \text{beta}(y; 1, 1)\ dy\ dx = \frac{2}{3}.\\ \alpha_4 &= p_1.\end{aligned}$$

Consequently we infer that

$$\mathbb{P}\{a^0 = a_1 \text{ and } r^0 = 1 \text{ and } a^1 = a_1 \text{ and } r^1 = 1\} = \frac{(p_1)^2}{3}.$$

In the calculations above, we have nowhere assumed any relationship between p_1 and p_2—so the same steps apply to arm a_2 to yield

$$\mathbb{P}\{a^0 = a_2 \text{ and } r^0 = 1 \text{ and } a^1 = a_2 \text{ and } r^1 = 1\} = \frac{(p_2)^2}{3}.$$

The required probability is $\frac{(p_1)^2 + (p_2)^2}{3}$.