Principles of Online Decision-Making (CS-303): Midterm Exam

October 28, 2025

Duration: **1h45**.

Total points: **100**.

Number of pages: **16**.

Allowed documents: **class notes, homeworks with solutions, 2-page cheat sheet**.

There should in general be enough room below every question for intermediate calculations and your answer. However, you are allowed to use additional sheets of paper; please **write your name on every sheet**, number them, and staple them to this document before handing in.

The use of **mobile phones, tablets, smart watches, earphones**, and other communication devices is **prohibited**.

Last name:
First name:
SCIPER number:
Signature:

Please leave blank.

1	2	3	4	5	Total
50	12	12	14	12	100

Question 1: Multiple Choice Questions (50 points)

(50 pts) All questions have a single answer. Check the correct one. Grading:

- Correct answer: +2 points;
- Wrong answer: −1 point;
- No answer or "I don't know": 0 point.

1. Suppose we try to solve the secretary problem, but instead of observing x_i directly, we can only observe whether x_i is the best secretary so far or not (without having access to the actual value x_i). How does this affect the expected performance?
   - ☐ Better than the original secretary problem.
   - ☐ Worse than the original secretary problem.
   - ■ The same as the original secretary problem.
   - ☐ I don't know

2. Let X_1 be a standard normal random variable, and X_2 be an exponential random variable with rate parameter $\lambda = 1$ ($\mathbb{P}(X_2 \geq t) = e^{-t}\ \forall\, t \geq 0$). X_1 and X_2 are independent. Which of the following is true?
   - ☐ $X_1 X_2$ is subgaussian
   - ☐ $X_1 + X_2$ is subgaussian
   - ■ Neither $X_1 X_2$ nor $X_1 + X_2$ is subgaussian
   - ☐ I don't know

3. Let X_1 denote the Bernoulli random variable with parameter $p = 0.5$, and X_2 the Bernoulli random variable with parameter $p = 0.01$. Let σ_1 and σ_2 be the smallest subgaussian parameters of X_1 and X_2 respectively. Which of the following is true?
   - ☐ $\sigma_1 < \sigma_2$
   - ■ $\sigma_1 > \sigma_2$
   - ☐ $\sigma_1 = \sigma_2$
   - ☐ I don't know

4. Let X_1 be σ_1-subgaussian random variable, and X_2 be σ_2-subgaussian random variable. Assume that X_1 and X_2 are independent. Which of the following is true about the random variable $Z = X_1 - 2X_2$?
   - ☐ Z is $|\sigma_1 - 2\sigma_2|$-subgaussian
   - ■ Z is $\sqrt{\sigma_1^2 + 4\sigma_2^2}$-subgaussian
   - ☐ Z is not necessarily subgaussian
   - ☐ I don't know

5. Let $\sigma < \tau$ be two positive real numbers. There are two random variables X and Y. Assume that X is σ-subgaussian and Y is τ-subgaussian, but Y is not σ-subgaussian. Assume that X and Y are independent and symmetric. Which of the following is true?
   - ☐ for any $\epsilon > 0$, $\mathbb{P}(|X| > \epsilon) < \mathbb{P}(|Y| > \epsilon)$
   - ☐ for any $\epsilon > 0$, $\mathbb{P}(|X| > \epsilon) > \mathbb{P}(|Y| > \epsilon)$
   - ■ Neither
   - ☐ I don't know

6. We are given two random variables $X, Y \sim \mathcal{N}(0,1)$ (standard normal). For their sum $Z = X + Y$, which is σ-subgaussian, what is the possible range of σ?

   ■ $\sigma \in [0, 2]$
   ☐ $\sigma \in [\sqrt{2}, 2]$
   ☐ $\sigma \in \{1, 2\}$ (i.e., either exactly 1 or 2)
   ☐ I don't know

7. Which of the following environments of two-armed bandit is a case of unstructured bandit? ($\mathcal{B}(p)$ is a Bernoulli distribution of parameter p).

   ☐ $\mathcal{E} = \{(\mathcal{N}(0,\sigma), \mathcal{N}(0,2\sigma)) | \sigma \in [0,1]\}$
   ☐ $\mathcal{E} = \{(\mathcal{N}(1,1), \mathcal{N}(0,2)), (\mathcal{N}(1,1), \mathcal{N}(0,3)), (\mathcal{N}(0,2), \mathcal{N}(0,3))\}$
   ■ $\mathcal{E} = \{(\mathcal{B}(0), \mathcal{B}(1)), (\mathcal{B}(0), \mathcal{B}(0)), (\mathcal{B}(1), \mathcal{B}(1)), (\mathcal{B}(1), \mathcal{B}(0))\}$
   ☐ I don't know

8. We run the variant of UCB where the variance of each arm is known to the algorithm. We use $\delta = 0.05$ on a Gaussian bandit with two arms. We observe the following plot (Figure 1). Which of the following settings is the most likely to generate this plot?

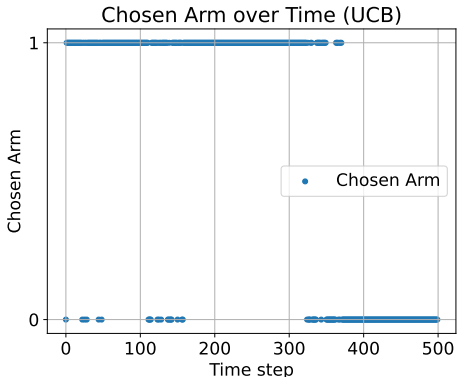


Figure 1: Arm chose for each step

   ■ Arm 0 is $\mathcal{N}(1,1)$ and arm 1 is $\mathcal{N}(0,50)$
   ☐ Arm 0 is $\mathcal{N}(0,1)$ and arm 1 is $\mathcal{N}(0,50)$
   ☐ Arm 0 is $\mathcal{N}(0,50)$ and arm 1 is $\mathcal{N}(1,50)$
   ☐ I don't know

9. What happens as the horizon n goes to infinity if we run the UCB algorithm with a fixed δ (which does not depend on n or t)?

   ■ The cumulative regret is asymptotically linear.
   ☐ The cumulative regret is asymptotically sublinear.
   ☐ The cumulative regret is asymptotically superlinear.
   ☐ I don't know

10. You wish to apply UCB with confidence parameter δ (δ is a constant) on a bandit environment with σ-subgaussian arms, but only have an implementation of UCB for 1-subgaussian arms (for convenience, call this UCB1). You are too lazy to make another implementation for σ-subgaussian arms, and wish to obtain the same results with using UCB1. What can you do?
    - ☐ Run UCB1 with a specific confidence parameter δ_1 depending on σ and δ
    - ☐ Run UCB1 but observed reward is multiplied by some specific factor α depending on σ
    - ■ Both work
    - ☐ I don't know

11. For the (standard) UCB algorithm, suppose we have two optimal arms $\mu_1 = \mu_2 = \mu_*$. What is most likely to happen?
    - ■ In the long run, both arms get pulled approximately the same number of times.
    - ☐ The algorithm eventually favors one of the two and plays it until time n.
    - ☐ The one that gets pulled first will be favored and played until time n.
    - ☐ I don't know

12. Suppose you are running the asymptotic UCB algorithm for the multi-armed bandit problem with k arms, assuming that the reward distributions are subgaussian with parameter 1. After running the algorithm for a few rounds, you find that the empirical variance of the rewards from arm i is much higher than 1. What should you do to improve the cumulative regret of the algorithm?
    - ■ Increase the width of the confidence interval for arm i
    - ☐ Decrease the width of the confidence interval for arm i
    - ☐ Increase the width of the confidence intervals for all arms
    - ☐ I don't know

13. For the UCB algorithm, we first ensured that the probability of a bad event occurring in one time step was bounded from above by δ_n. Then we needed to ensure that the probability that the bad event ever occurred (over n time steps) goes to zero. Which is the largest δ_n that ensures this?
    - ☐ $\delta_n = 1/n$.
    - ■ $\delta_n = (n \ln n)^{-1}$.
    - ☐ $\delta_n = n^{-2}$.
    - ☐ I don't know

14. Consider the following bandit environment with k arms. Let $\theta_1, \cdots, \theta_k$ be some real values. For each round, a random variable X_t is drawn from a standard normal distribution $\mathcal{N}(0, 1)$, and the reward of arm i is 1 if $X_t > \theta_i$ and 0 otherwise. Which of the following algorithm is suitable for this bandit environment?
    - ☐ Thompson Sampling with Gaussian prior
    - ■ Thompson Sampling with Beta prior
    - ☐ Neither of the above
    - ☐ I don't know

15. What is true about the variance of the posterior distribution for each arm in the Thompson Sampling algorithm?

- ☐ It converges to the variance of the reward distribution of the arm as the number of pulls of the arm goes to infinity
- ■ It goes to zero as the number of pulls of the arm goes to infinity
- ☐ It converges to the variance of the prior distribution as the number of pulls of the arm goes to infinity
- ☐ I don't know

16. We want to perform perform Thompson Sampling for 1000 steps on a Bernoulli bandit with k arms. After 100 steps, our random generator (used for Thompson Sampling) breaks down.

    What is our best option from here out of the following for the remaining 900 steps, in order to minimize the cumulative regret?

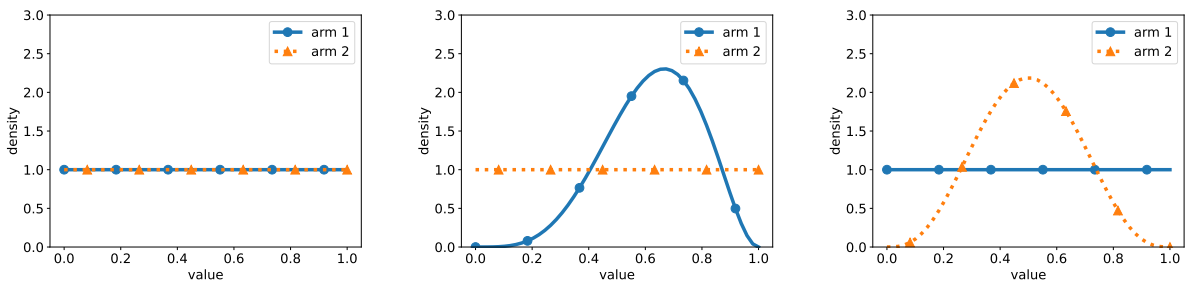
- ■ Use UCB for 900 steps, and include the first 100 samples to compute the confidence intervals.
- ☐ Ignore the first 100 samples and do UCB from scratch for the 900 steps.
- ☐ Cycle through the k arms for the 900 steps.
- ☐ I don't know

17. We perform Bernoulli Thompson sampling with 3 arms for 300 steps and observe the posterior distributions for each arm. If we continue the experiment, how will these three distributions evolve asymptotically?

- ■ The posteriors of all three arms will always concentrate at their true expected reward
- ☐ The posterior of the suboptimal arms will change, but they may not concentrate at a specific point
- ☐ It is possible that the posterior of a suboptimal arm remains the same indefinitely
- ☐ I don't know

18. We perform Thompson Sampling with a two-armed Bernoulli bandit of average returns 0.2 and 0.8. After 6 steps, which of the following posteriors are possible (Figure 2)?

- ☐ Only plot c
- ■ Plots b and c
- ☐ Plots a, b and c
- ☐ I don't know



(a) (b) (c)

Figure 2: Posterior distribution of the arms expected rewards

19. If we have an infinite number of Bernoulli arms and the bandit is unstructured, which of the following is true?

- ☐ We can obtain a sublinear regret using a variation of UCB.
- ☐ We can obtain a sublinear regret using a variation of Thompson Sampling.
- ■ No algorithm can guarantee a sublinear regret.
- ☐ I don't know

20. Consider the best arm identification problem, with two arms. You are given that both arms have Gaussian rewards with variance 1, but unknown means. You are given a fixed time horizon n, so you run the Sequential Halving algorithm for n rounds. Which of the following statements best describes the behaviour of the algorithm in this scenario?

- ☐ Play both arms equally often till round $n/2$, then play the best arm for the remaining rounds (explore-then-commit strategy)
- ☐ Play both arms equally often until some adaptively chosen round t, then play the best arm for the remaining rounds (adaptive-explore-then-commit strategy)
- ■ Play both arms equally often till round n (pure exploration)
- ☐ I don't know

21. Consider the best arm identification problem, with two arms. You are given that both arms have Gaussian rewards with variance 1, but unknown means. You want to find the best arm with probability at least $1-\delta$, for a fixed $\delta \in (0,1)$. As a function of the suboptimality gap Δ, what is the least number of times you will have to pull each arm (in expectation)?

- ☐ proportional to $(1/\Delta)$
- ■ proportional to $(1/\Delta^2)$
- ☐ proportional to $(\exp(1/\Delta))$
- ☐ I don't know

22. We have been working with the $\|\cdot\|_V$ norm in the context of linear bandits. Is there a V for which this norm actually becomes L_2 (Euclidean) norm?

- ☐ Yes, when V is the matrix where every entry is equal to 1.
- ■ Yes, when V is the identity matrix.
- ☐ No, this is not possible.
- ☐ I don't know

23. We have a linear bandit with fixed set of arms $\mathcal{A}_t = \mathcal{A}$. Suppose that two arms $a_1, a_2 \in \mathbb{R}^d$ are collinear, and both have positive inner product $\langle a_1, \theta_* \rangle > 0, \langle a_2, \theta_* \rangle > 0$. Which arm's UCB is larger?

- ☐ The arm that has been pulled more often.
- ■ The one with larger norm $\|\cdot\|$.
- ☐ The one with smaller Euclidean distance $\|\cdot - \theta_*\|$
- ☐ I don't know

24. Suppose you are running a bandit algorithm for displaying advertisements on a website. You have k different ads (arms), and you want to maximize the total number of clicks (rewards) over a time horizon of n website visits (rounds). Consider the following two scenarios:

    (a) You don't know the features of the advertisements, so you run an unstructured bandit algorithm.

    (b) You know the features of the advertisements, and you run a linear bandit algorithm.

    Suppose a new advertisement option becomes available. Which of the following statements best describes the behaviour of the bandit algorithm when the new ad is introduced?

    - [x] In scenario (a), the algorithm will recommend the new ad a few times in order to learn if it is good or not, while in scenario (b), it will judge whether to run the ad based on its features.
    - [ ] In scenario (b), the algorithm will recommend the new ad a few times in order to learn if it is good or not, while in scenario (a), it will judge whether to run the ad based on the prior performance of the existing ads.
    - [ ] In both scenarios, the algorithm will recommend the new ad a few times in order to learn if it is good or not.
    - [ ] I don't know

25. Consider a linear bandit problem with d-dimensional arms and unknown parameter θ_*. You observe that the last component of all arm feature vectors is zero. Which of the following statements is FALSE?

    - [ ] The last component of θ_* cannot be learned by any algorithm.
    - [x] The cumulative regret of LinUCB will grow sublinearly only if the last component of θ_* is zero.
    - [ ] The cumulative regret of LinUCB will grow sublinearly regardless of the value of the last component of θ_*.
    - [ ] I don't know

Question 2: The Secretary Problem (12 points)

In homework 1 question 1(a), we considered a variant of the secretary problem where the scores are drawn independently from the uniform distribution: $x_i \sim \text{Unif}(0, 1)$, i.i.d. In this question, we revisit this setting but use a different selection rule. Specifically, we use the simple thresholding scheme from homework 1 question 1(b). Let n be the number of secretaries, and let θ $(0 < \theta < 1)$ be a fixed threshold. At each step $i \leq n$, the algorithm accepts the secretary if $x_i > \theta$. If no secretary is accepted before the last step, the algorithm accepts the final secretary.

1. (2 pts) What is the probability that exactly k secretaries have scores greater than θ $(1 \leq k \leq n)$?

   Let n be the number of secretaries. The probability that exactly k secretaries have $x_i > \theta$ and all others $x_j \leq \theta$ $(j \neq i)$ is:

   $$P = \binom{n}{k} \cdot (1-\theta)^k \cdot \theta^{n-k}.$$

   Here, careful not to forget to count the possible choice of k secretaries.

2. (2 pts) Conditioned on the event that there are exactly k $(1 \leq k \leq n)$ secretaries with scores greater than θ, what is the conditional probability of winning (i.e., picking the best secretary) with this algorithm?

   When $k \geq 1$, the first secretary with $x_i > \theta$ is selected. The probability that this secretary is the best k secretaries (who have higher score than θ) (since the best among the k is equally likely to appear in any position). Therefore, conditioned on the eventthat there are exactly k $(1 \leq k \leq n)$ secretaries with scores greater than θ, the probability of winning is $1/k$.

3. (4 pts) What is the probability of winning (i.e., picking the best secretary) with this algorithm? (You can leave your answer in the summation form.)

   Let k be the number of secretaries with $x_i > \theta$. The probability that exactly k secretaries have scores above θ is $\binom{n}{k}(1-\theta)^k\theta^{n-k}$ as we've seen in (1). When $k \geq 1$, the probability of winning given k secretaries have scores above θ is $1/k$. If $k = 0$, we select the last secretary, who is the best with probability $\frac{1}{n}$. Therefore, the total probability of winning is:

   $$P_{\text{win}} = \sum_{k=1}^{n} \binom{n}{k} \cdot (1-\theta)^k \cdot \theta^{n-k} \cdot \frac{1}{k} + \theta^n \cdot \frac{1}{n}.$$

4. (4 pts) We now consider the standard secretary problem where the scores of secretaries are arbitrary and unknown. We focus on the case when there are only two secretaries ($n = 2$). Surprisingly, even when there are only two candidates, there is a stopping rule to pick the best secretary with probability higher than 1/2. The strategy works as follows: first, pick a random thereshold θ and if the score of the first secretary is higher than θ, you pick the first secretary, and if not pick the second one. Prove that this stopping rule has actually higher winning chance than 1/2.

   We consider three cases: (1) when both secretaries' scores are greater than θ, (2) when θ is between two scores, and (3) when both secretaries' scores are smaller than θ. When (1), the first secatary is chosen by the algorithm, and he/she is the best secretary with probability 1/2. When (3), the second secretary is with probability 1/2, the best one (conditioned on (1) and (3)), and the algorithm choose him/her. In case of (2) you pick the best one with probability 1 (conditioned on (2)). So, the probability of winning with this policy is $0.5 * (p_1 + p_3) + p_2 > 0.5$, where p_i is the probability of case (1) happens.

Question 3: Concentration of Measure (12 points)

In this question, we shall derive a concentration bound for Bernoulli random variables. Recall that a Bernoulli random variable X with parameter p is a random variable that takes value 1 with probability p and value 0 with probability $1-p$. Also recall that a random variable X is said to be sub-Gaussian with parameter σ if for all $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{\frac{\sigma^2\lambda^2}{2}}.$$

Finally, recall that a sub-Gaussian random variable X with parameter σ satisfies the following concentration bound: for all $t > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

1. (4 pts) Prove that a Bernoulli random variable X with parameter p is sub-Gaussian with parameter $\frac{1}{2}$. You should derive this from first principles; do NOT use the fact that bounded random variables are subgaussian. (Hint: You may use the fact that for all $t \in \mathbb{R}$, we have $\frac{e^t+e^{-t}}{2} \leq \exp\left(\frac{t^2}{2}\right)$.)

   We prove that a Bernoulli random variable is sub-Gaussian with parameter $1/2$. Let $X \sim$ Bernoulli(p) and set $Y := X - p$. Then Y takes values $1-p$ (with probability p) and $-p$ (with probability $1-p$), and $\mathbb{E}[Y] = 0$. Define

   $$M(\lambda) := \mathbb{E}\left[e^{\lambda Y}\right], \qquad \psi(\lambda) := \log M(\lambda).$$

   We will show that $\psi(\lambda) \leq \lambda^2/8$ for all $\lambda \in \mathbb{R}$, which implies $\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leq \exp(\lambda^2/8)$, i.e. X is sub-Gaussian with parameter $1/2$.

   Note that $\psi(0) = 0$ and $\psi'(0) = \mathbb{E}[Y] = 0$. We compute

   $$\psi'(\lambda) = \frac{\mathbb{E}[Ye^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]}, \qquad \psi''(\lambda) = \frac{\mathbb{E}[Y^2e^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]} - \left(\frac{\mathbb{E}[Ye^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]}\right)^2.$$

   Introduce a tilted random variable Z, supported on $\{-p, 1-p\}$, by

   $$\mathbb{P}(Z = 1-p) = \frac{p\,e^{\lambda(1-p)}}{M(\lambda)}, \qquad \mathbb{P}(Z = -p) = \frac{(1-p)\,e^{-\lambda p}}{M(\lambda)}.$$

   Then $\psi''(\lambda) = \mathrm{Var}(Z)$. Since Z takes only the two values $1-p$ and $-p$, whose difference is 1, we have $\mathrm{Var}(Z) \leq 1/4$ for all λ. Thus $\psi''(\lambda) \leq 1/4$.

   Because $\psi'(0) = 0$,

   $$\psi'(\lambda) = \int_0^\lambda \psi''(t)\,dt \leq \int_0^\lambda \frac{1}{4}\,dt = \frac{\lambda}{4},$$

   and since $\psi(0) = 0$,

   $$\psi(\lambda) = \int_0^\lambda \psi'(s)\,ds \leq \int_0^\lambda \frac{s}{4}\,ds = \frac{\lambda^2}{8}.$$

   Therefore,

   $$\mathbb{E}\left[e^{\lambda(X-p)}\right] = e^{\psi(\lambda)} \leq \exp\left(\frac{\lambda^2}{8}\right),$$

   so X is sub-Gaussian with parameter $1/2$.

Remark on the "maximizing over p" approach. It is initially tempting to try to bound

$$M(\lambda, p) = \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] = p\, e^{\lambda(1-p)} + (1-p)\, e^{-\lambda p}$$

by first maximizing over $p \in [0,1]$. A Taylor expansion near $\lambda = 0$ shows

$$M(\lambda, p) = 1 + \frac{\lambda^2}{2} p(1-p) + O(\lambda^3),$$

so for small $|\lambda|$ the largest value indeed occurs at $p = 1/2$. This aligns with the fact that $p(1-p)$ is maximized at $p = 1/2$.

However, for general $\lambda \neq 0$, $p = 1/2$ does *not* maximize $M(\lambda, p)$. Differentiating in p and setting the derivative to zero gives

$$p^*(\lambda) = \frac{e^\lambda - 1 - \lambda}{\lambda\,(e^\lambda - 1)},$$

which depends on λ. Thus there is no fixed "worst" choice of p, and one would have to analyze $M(\lambda, p^*(\lambda))$ itself, which is unnecessarily complicated.

The argument above avoids this issue entirely. Since the tilted variable Z always takes two values at distance 1, its variance is always at most $1/4$, uniformly in p and λ. This uniform control on the curvature of $\psi(\lambda)$ yields the desired bound $\psi(\lambda) \leq \lambda^2/8$ directly, and hence the sub-Gaussian parameter $1/2$.

2. (2 pts) Prove that the sum of two independent sub-Gaussian random variables with parameters σ_1 and σ_2 is a sub-Gaussian random variable with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.

   Let X and Y be independent sub-Gaussian random variables with parameters σ_1 and σ_2, and write $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. Then

$$\mathbb{E}\left[e^{\lambda\left((X+Y)-(\mu_X+\mu_Y)\right)}\right] = \mathbb{E}\left[e^{\lambda(X-\mu_X)} e^{\lambda(Y-\mu_Y)}\right].$$

   Independence of X and Y implies

$$\mathbb{E}\left[e^{\lambda(X-\mu_X)} e^{\lambda(Y-\mu_Y)}\right] = \mathbb{E}\left[e^{\lambda(X-\mu_X)}\right] \cdot \mathbb{E}\left[e^{\lambda(Y-\mu_Y)}\right].$$

   Since X is sub-Gaussian with parameter σ_1 and Y is sub-Gaussian with parameter σ_2,

$$\mathbb{E}\left[e^{\lambda(X-\mu_X)}\right] \leq \exp\left(\frac{\sigma_1^2\lambda^2}{2}\right), \qquad \mathbb{E}\left[e^{\lambda(Y-\mu_Y)}\right] \leq \exp\left(\frac{\sigma_2^2\lambda^2}{2}\right).$$

   Multiplying these bounds gives

$$\mathbb{E}\left[e^{\lambda\left((X+Y)-(\mu_X+\mu_Y)\right)}\right] \leq \exp\left(\frac{(\sigma_1^2+\sigma_2^2)\lambda^2}{2}\right).$$

   Thus $X + Y$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.

3. (2 pts) Prove that if X is a sub-Gaussian random variable with parameter σ, then for all $\alpha \in \mathbb{R}$, αX is a sub-Gaussian random variable with parameter $|\alpha|\sigma$.

Let X be sub-Gaussian with parameter σ, so

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}X)}\right] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right) \qquad \text{for all } \lambda \in \mathbb{R}.$$

For any $\alpha \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda(\alpha X-\mathbb{E}[\alpha X])}\right] = \mathbb{E}\left[e^{\lambda\alpha(X-\mathbb{E}X)}\right] = \mathbb{E}\left[e^{(\alpha\lambda)(X-\mathbb{E}X)}\right] \leq \exp\left(\frac{\sigma^2(\alpha\lambda)^2}{2}\right) = \exp\left(\frac{(|\alpha|\sigma)^2\lambda^2}{2}\right).$$

Thus αX is sub-Gaussian with parameter $|\alpha|\sigma$.

4. (2 pts) Let $X_1, X_2, \ldots, X_n$ are independent Bernoulli random variables with parameter p, and let $\bar{X}_n = (\sum_{i=1}^n X_i)/n$. Using the results from the previous parts, prove that for all $t > 0$, we have

$$\mathbb{P}\left(|\bar{X}_n - p| \geq t\right) \leq 2\exp\left(-2nt^2\right).$$

1. gives us that each Bernoulli random variable X_i is sub-Gaussian with parameter $1/2$, i.e.

$$\mathbb{E}\left[e^{\lambda(X_i-p)}\right] \leq \exp\left(\frac{\lambda^2}{8}\right).$$

Since the X_i are independent, the sum $\sum_{i=1}^n (X_i - p)$ is sub-Gaussian with parameter $\sqrt{n}\cdot(1/2)$. Dividing by n to form the sample mean,

$$\bar{X}_n - p = \frac{1}{n}\sum_{i=1}^n (X_i - p),$$

the sub-Gaussian parameter gets scaled by $1/n$, so

$$\bar{X}_n - p \quad \text{is sub-Gaussian with parameter} \quad \frac{1}{2\sqrt{n}}.$$

Using the standard sub-Gaussian tail bound

$$\mathbb{P}(|Z| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right),$$

with $Z = \bar{X}_n - p$ and $\sigma = \frac{1}{2\sqrt{n}}$, we get

$$\mathbb{P}\left(|\bar{X}_n - p| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\cdot(1/(4n))}\right) = 2\exp(-2nt^2).$$

So the average of n Bernoulli(p) samples concentrates around p with probability at least $1 - 2e^{-2nt^2}$.

5. (2 pts) The upper bound derived in the previous question is independent of p. How do you expect the actual concentration behavior of $\bar{X}_n$ to depend on p? In other words, keeping n and t fixed, how does $\mathbb{P}\left(|\bar{X}_n - p| \geq t\right)$ to vary as a function of p? At what values of p does it achieve its minima and maxima? Explain your answer.

Even though the above bound does not show it, the actual concentration of $\bar{X}_n$ depends on p. Since

$$\mathrm{Var}(\bar{X}_n) = \frac{p(1-p)}{n},$$

the spread of $\bar{X}_n$ is largest when $p(1-p)$ is largest, which happens at $p = \frac{1}{2}$, and becomes smaller as p moves toward 0 or 1.

By the CLT, for large n, $\bar{X}_n$ is approximately normal with mean p and standard deviation $\sqrt{p(1-p)/n}$. For fixed t, the event $|\bar{X}_n - p| \geq t$ is more likely when this standard deviation is larger. Therefore, the tail probability is *largest* when $p = 1/2$ and *smallest* when p is near 0 or 1.

Question 4: Multi-Armed Bandits (14 points)

In class, we briefly talked about sequential elimination (SE) as an extension of ETC, where each arm may be eliminated as soon as we are reasonably sure that it cannot be the best arm. One way to achieve this is by proceeding in phases, using the algorithm below. At all times, this algorithm maintains an *active set* of contenders to become the winning arm. In the ℓ-th phase, the algorithm tries to eliminate from the current active set all arms i for which $\Delta_i \geq 2^{-\ell}$.

We assume that arm $i = 1$ is the unique optimal arm.The noise for every arm is 1-subgaussian.

1. **INPUT**: k, sequence $m_1, m_2, m_3, ...$
2. $A_1 = \{1, 2, 3, \ldots, k\}$ (initialize active set)
3. FOR $\ell = 1, 2, 3, \ldots$ DO:
4. Pull each arm $i \in A_\ell$ exactly m_ℓ times, collect reward samples $X_{i,\ell,1}, X_{i,\ell,2}, \ldots, X_{i,\ell,m_\ell}$
5. Let $\hat{\mu}_{i,\ell} = m_\ell^{-1} \sum_{m=1}^{m_\ell} X_{i,\ell,m}$ be the average reward for arm i from this phase only
6. Update the active set: $A_{\ell+1} = \left\{i : \hat{\mu}_{i,\ell} + 2^{-\ell} \geq \max_{j \in A_\ell} \hat{\mu}_{j,\ell}\right\}$
7. END FOR

Note that because the elimination process is random, the number of rounds until there is a single "survivor" is random with infinite support.

1. (2 pts) Given the definition of the algorithm, are $\hat{\mu}_{i,\ell}$ and $\hat{\mu}_{i,\ell'}$ (for $\ell \neq \ell'$) independent? Why or why not?

   Yes, they are independent, because $\hat{\mu}_{i,\ell}$ and $\hat{\mu}_{i,\ell'}$ are functions of two disjoint sets of i.i.d. random variables.

2. (2 pts) Prove that $(\hat{\mu}_{i,\ell} - \mu_i) - (\hat{\mu}_{1,\ell} - \mu_1)$ is a $\sqrt{2/m_\ell}$-subgaussian random variable. (You may use the properties of subgaussian random variables covered in class).

   First, $\hat{\mu}_{i,\ell} - \mu_i$ is zero-mean (including for $i = 1$). Second, $m_\ell(\hat{\mu}_{i,\ell} - \mu_i)$ is $\sqrt{m_\ell}$-subgaussian, because it is the sum of m_ℓ independent 1-subgaussian RVs. Third, $(\hat{\mu}_{i,\ell} - \mu_i)$ is therefore $\sqrt{1/m_\ell}$-subgaussian. Finally, the difference between the two terms is $\sqrt{2/m_\ell}$-subgaussian because the two terms are independent.

3. (3 pts) Using the above observations, show that for any phase $\ell \geq 1$, the probability that the best arm gets freshly eliminated from the active set satisfies

$$\mathbb{P}[1 \notin A_{\ell+1}, 1 \in A_\ell] \leq k \exp\left(-\frac{m_\ell 2^{-2\ell}}{4}\right).$$

   Hint: the event $\{1 \notin A_{\ell+1}\}$ means that in phase ℓ there was at least one arm $i \neq 1$ that "beat" arm 1 by a large enough margin.

$$\begin{aligned}
\mathbb{P}[1 \notin A_{\ell+1}, 1 \in A_\ell] &\leq \mathbb{P}[1 \in A_\ell, \exists i \in A_\ell \setminus \{1\} : \hat{\mu}_{i,\ell} \geq \hat{\mu}_{1,\ell} + 2^{-\ell}] && (1)\\
&= \mathbb{P}[1 \in A_\ell, \exists i \in A_\ell \setminus \{1\} : \hat{\mu}_{i,\ell} - \hat{\mu}_{1,\ell} \geq 2^{-\ell}] && (2)\\
&\leq k\mathbb{P}[1 \in A_\ell, \hat{\mu}_{2,\ell} - \hat{\mu}_{1,\ell} \geq 2^{-\ell}] && (3)\\
&\leq k\mathbb{P}[1 \in A_\ell, (\hat{\mu}_{2,\ell} - \mu_2) - (\hat{\mu}_{1,\ell} - \mu_1) \geq 2^{-\ell}] && (4)\\
&\leq k \exp\left(-\frac{m_\ell 2^{-2\ell}}{4}\right). && (5)
\end{aligned}$$

4. (3 pts) Now bound the probability that arm 1 and some other arm i with large enough gap ($\Delta_i \geq 2^{-\ell}$) are active in ℓ, and i does not get eliminated (i.e., is still in the active set in round $\ell+1$).

$$\mathbb{P}[i \in A_{\ell+1}, 1 \in A_\ell, i \in A_\ell] \leq \exp\left(-\frac{m_\ell \left(\Delta_i - 2^{-\ell}\right)^2}{4}\right).$$

Hint: for arm i to survive to phase $\ell+1$, its estimator in phase ℓ cannot have been too far below that of arm 1.

$$\begin{aligned}
\mathbb{P}[i \in A_{\ell+1}, 1 \in A_\ell, i \in A_\ell] &\leq \mathbb{P}[1 \in A_\ell, i \in A_\ell, \hat{\mu}_{i,\ell} + 2^{-\ell} \geq \hat{\mu}_{1,\ell}] && (6)\\
&= \mathbb{P}[1 \in A_\ell, i \in A_\ell, (\hat{\mu}_{i,\ell} - \mu_i) - (\hat{\mu}_{1,\ell} - \mu_1) \geq \Delta_i - 2^{-\ell}] && (7)\\
&\leq \exp\left(-\frac{m_\ell \left(\Delta_i - 2^{-\ell}\right)^2}{4}\right). && (8)
\end{aligned}$$

5. (4 pts) Now we set the number of samples m_ℓ per arm in phase ℓ to $m_\ell = 2^{4+2\ell}\ln(\ell/\delta)$, where $0 < \delta < 1$ is a parameter.

With this choice, find an upper bound on the probability that the best arm 1 will ever be eliminated, i.e., $\mathbb{P}[\exists \ell : 1 \notin A_\ell]$. Hint: use the bounds on the events defined above. Note that $\sum_{\ell=1}^{\infty} \ell^{-4} = c$, where $c < \infty$ is a constant that you do not need to compute. You should obtain an expression involving only c, k and δ.

$$\begin{aligned}
\mathbb{P}[\exists \ell : 1 \notin A_\ell] &\leq \sum_{\ell}^{\infty} \mathbb{P}[1 \notin A_{\ell+1}, 1 \in A_\ell] && (9)\\
&\leq k \sum_{\ell=1}^{\infty} \exp\left(-\frac{m_\ell 2^{-2\ell}}{4}\right) && (10)\\
&= k \sum_{\ell=1}^{\infty} \exp\left(-4\ln(\ell/\delta)\right) && (11)\\
&\leq k\delta^4 \sum_{\ell=1}^{\infty} \frac{1}{\ell^4} = k\delta^4 c. && (12)
\end{aligned}$$

Question 5: Linear Bandits (12 points)

We consider a linear bandit setting. Recall that for linear bandits, the payoff at step t is

$$X_t = A_t^T \theta^* + \eta_t$$

where A_t is the arm played at step t, and η_t is some independent, zero-mean noise.
We assume that η_t is 1-subgaussian. Here $\theta^* \in \mathbb{R}^d$ can be any vector of dimension d.

We consider two sets of arms (*i.e* two different bandits):

$$\mathcal{A} = \{a \in \mathbb{R}^d : ||a||_2 \leq 1\} \text{ and } \mathcal{A}' = \{e_1, ..., e_d\}$$

where $e_i = (0, ..., 0, 1, 0, ..., 0)$ is the vector with a one at the i-th coordinate and zeros elsewhere.

1. (2 pts) Are these two bandits structured or unstructured? Justify your answer.

   For $\mathcal{A}$, the bandit is structured, because for instance we can learn information on arm $(1/2, 0, ..., 0)$ by playing arm $(1/4, 0, ..., 0)$ (because we learn information on the first coordinate of θ^*).
   For $\mathcal{A}'$ the bandit is unstructured, because playing e_i only gives information of the i-th coordinate of θ^*, which is independent of the reward given by the other arms.

2. (2 pts) Express the optimal arm and the optimal reward as a function of θ^*, in the case of $\mathcal{A}$ and $\mathcal{A}'$.

   The best arm in both cases is the arm a one which maximises $a^T\theta^*$.
   For $\mathcal{A}$, it is the vector of $\mathcal{A}$ which is in the same direction as θ^* and of maximal norm, so $\frac{\theta^*}{||\theta^*||_2}$.
   The corresponding reward is $\theta^T \frac{\theta^*}{||\theta^*||_2} = ||\theta^*||_2$.
   For $\mathcal{A}'$, the best arm is the e_{i^*}, where i^* is the index of the biggest coordinate of θ^*. The corresponding reward is $\theta^*_{i^*}$.

3. (2 pts) For $\mathcal{A}$, in the case where $\eta_t = 0$ for all t, what is the smallest number of steps needed to learn θ^*?
   Using this, design an algorithm whose regret does not depend on the horizon n. Compute the corresponding upper bound on the regret and prove that your algorithm respects it.

   If there is no noise, we can learn θ^* exactly in d steps, by playing arm e_t (which also belongs to $\mathcal{A}$) at step t for $1 \leq t \leq n$. Indeed, this way we have $A_t = \theta^*_t$, such that we know θ^* fully after step d. After step d, we can play the optimal arm found in 2. because we know θ^*.
   This way, the cumulative regret of the algorithm for any horizon n is the cumulative suffered from the first d steps, which is at most $2d||\theta^*||_2$.

4. (3 pts) We now fix $d = 2$ and we use the set of arms $\mathcal{A}'' = \{a, b, c\}$, where $a = (1, 0)$, $b = (0, 1)$, $c = (1/\sqrt{2}, 1/\sqrt{2})$. For the three first steps, we have played the following arms: $A_1 = a$, $A_2 = a$, $A_3 = b$ and we have observed the following rewards: $X_1 = 2$, $X_2 = 4$, $X_3 = 1$.

   Compute $\hat{\theta}_3$ using the least squares estimator (without regularisation).

$$V_3 = A_1A_1^T + A_2A_2^T + A_3A_3^T = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\hat{\theta}_3 = V_3^{-1}(X_1A_1 + X_2A_2 + X_3A_3) = \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 6 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$

5. (3 pts) Under the same assumptions as the previous question, compute A_4, the arm which will be played by the `LinUCB` algorithm at step 4 (it can be a, b or c). We assume $\beta_4 = 2$ (the radius of the confidence set).

We compute the best reward of each arm inside the confidence set.

$$a^T\hat{\theta}_3 + \sqrt{\beta_4}||a||_{V_3^{-1}} = a^T\hat{\theta}_3 + \sqrt{\beta_4 a^T V_3^{-1} a} = 3 + \sqrt{2 \times 1/2} = 4$$

$$b^T\hat{\theta}_3 + \sqrt{\beta_4}||b||_{V_3^{-1}} = b^T\hat{\theta}_3 + \sqrt{\beta_4 b^T V_3^{-1} b} = 1 + \sqrt{2 \times 1} = 1 + \sqrt{2}$$

$$c^T\hat{\theta}_3 + \sqrt{\beta_4}||c||_{V_3^{-1}} = c^T\hat{\theta}_3 + \sqrt{\beta_4 c^T V_3^{-1} c} = 4/\sqrt{2} + \sqrt{2 \times \frac{3}{4}} = 2\sqrt{2} + \sqrt{\frac{3}{2}}$$

The arm selected is clearly not b. We need to compare the values for a and c.
We show by equivalences that $A_4 = c$:

$$4 < 2\sqrt{2} + \sqrt{\frac{3}{2}} \Leftrightarrow 16 < (2\sqrt{2} + \sqrt{\frac{3}{2}})^2 \Leftrightarrow 16 < 8 + \frac{3}{2} + 4\sqrt{3}$$
$$\Leftrightarrow \frac{13}{2} < 4\sqrt{3} \Leftrightarrow \frac{169}{4} < 48 \Leftrightarrow 169 < 192$$