EE5603: Concentration Inequalities **2024**

Exam 2: September 2024

Instructor: Shashank Vatedka

Instructions: *This is a closed-book exam. You are not permitted to refer to any material or discuss the problem with anyone. Malpractice will be severely punished. Please mention your ROLL Number and name clearly in the answer sheet.*

Justify all your statements clearly. You may use any result proved in class (but clearly state which results you are using), but everything else needs to be proved.

Question 2.1. Find whether the following random variables are subgaussian and/or subexponential:

- Exponential distribution $f_X(x) = \alpha e^{-\alpha x} 1_{\{x \geqslant 0\}}$ with mean $\alpha > 0$ (5pts)
- Bounded random variable with support $[-a, a]$, i.e., $\Pr[X \in [-a, a]] = 1$ (5pts)

Question 2.2. For any real number a, let $g_a : \mathbb{R}^n \to \mathbb{R}$ be the function such that $g_a(\underline{x})$ is the number of elements in $x_1, \ldots, x_n$ which are less than or equal to a.

- What is $\mathbb{E}[g_a(\underline{X})]$ if $X_1, \ldots, X_n$ are iid(f_X) for any arbitrary density f_X? (3pts)
- Is $g_a(\underline{X})$ subexponential? subgaussian? (3pts)
- Derive a tail bound on the probability $\Pr[|g_a(\underline{X}) - \mathbb{E}[g_a(\underline{X})]| \geqslant \delta]$ for any $\delta > 0$ (2pts)

Question 2.3. Let us analyze the performance of a linear hypothesis tester/classifier. Suppose that there are two classes/hypotheses H_1, H_2:

- If H_1 is true, then you observe a random vector $\underline{X} \in \mathbb{R}^n$ which is Gaussian distributed with mean $\underline{a} \in \mathbb{R}^n$ and covariance matrix $\sigma^2 I$, where I is the $n \times n$ identity matrix
- If H_2 is true, then you observe a random vector $\underline{X} \in \mathbb{R}^n$ which is Gaussian distributed with mean $-\underline{a} \in \mathbb{R}^n$ and covariance matrix $\sigma^2 I$, where I is the $n \times n$ identity matrix

Consider the following classifier: Given $\underline{X}$, the classifier outputs H_1 if $\underline{X}^T \underline{a} \geqslant 0$, and H_2 if $\underline{X}^T \underline{a} < 0$. Find an upper bound on the probability of error (i.e., probability that the algorithm outputs H_1 when actually H_2 is true and vice versa) of this classifier. (10pts)

Question 2.4. For any sequence $x_1, \ldots, x_n$, let $f(x_1, \ldots, x_n)$ denote the number of distinct elements in $x_1, \ldots, x_n$. For example, $f(01101010) = 2$ and $f(10026001) = 4$.

Let p_X be an arbitrary pmf over any finite alphabet, and $X_1, \ldots, X_n$ be iid random variables with pmf p_X.

- Does f satisfy the bounded differences property? If so, find $c_1, \ldots, c_n$. Use this to get an upper bound for the variance of $Z = f(X_1, \ldots, X_n)$. (5pts)
- Compute an expression for $\mathbb{E}f(X_1, \ldots, X_n)$ in terms of p_X. How does this scale with n? Linearly? Sublinearly? (4pts)
- Show that $\mathrm{Var}(Z) \leqslant \mathbb{E}[Z]$, and hence derive an upper bound on $\Pr[|Z - \mathbb{E}[Z]| \geqslant \delta \mathbb{E}[Z]]$. For what values of δ does this give a meaningful bound? (5pts)

2.1 Results that you may use without proof in the exam

You may use any of the concentration inequalities or results proved in class.

A random variable is subgaussian if it satisfies the following properties:

- $\Pr[|X| \geq \delta] \leq 2e^{-\delta^2/K_1^2}$ for all $\delta > 0$
- $(\mathbb{E}|X|^p)^{1/p} \leq K_2\sqrt{p}$ for all $p \geq 1$
- $\mathbb{E}e^{t^2X^2} \leq e^{K_3^2t^2}$ for all $|t| \leq 1/K_3$
- $\mathbb{E}e^{X^2/K_4^2} \leq 2$
- If $\mathbb{E}[X] = 0$, then $\mathbb{E}e^{tX} \leq e^{K_5^2t^2}$ for all $t \in \mathbb{R}$

A random variable is subexponential if it satisfies the following properties:

- $\Pr[|X| \geq \delta] \leq 2e^{-\delta/K_1}$ for all $\delta > 0$
- $(\mathbb{E}|X|^p)^{1/p} \leq K_2 p$ for all $p \geq 1$
- $\mathbb{E}e^{t|X|} \leq e^{K_3 t}$ for all $0 \leq t \leq 1/K_3$
- $\mathbb{E}e^{|X|/K_4} \leq 2$
- If $\mathbb{E}[X] = 0$, then $\mathbb{E}e^{tX} \leq e^{K_5^2t^2}$ for all $|t| \leq 1/K_5$

2.1.0.1 Configuration functions

Let $\mathcal{X}$ be an arbitrary set, and Π be a property defined on finite products of $\mathcal{X}$, i.e., Π can be defined as a collection of sets $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \ldots$, where $\mathcal{S}_k \subset \mathcal{X}^k$ for $k = 1, 2, 3, \ldots$. We say that a sequence/vector $\underline{x} \in \mathcal{X}^m$ satisfies property Π if $\underline{x} \in \mathcal{S}_m$. Property Π is said to be hereditary if it satisfies the following property: if $\underline{x} = (x_1, \ldots, x_m)$ satisfies property Π, then every subsequence of $\underline{x}$ also satisfies Π, or equivalently,

$$(x_1, \ldots, x_m) \in \mathcal{S}_m \implies (x_{i_1}, \ldots, x_{i_k}) \in \mathcal{S}_k \quad \text{for all } 1 \leq k \leq m \text{ and } i_1 < i_2 < \ldots < i_k$$

The configuration function associated with such a Π, denoted f_Π, is the function that maps every $\underline{x}$ to the size of the largest subsequence of $\underline{x}$ that satisfies Π.

If f is a configuration function and $X_1, \ldots, X_n$ are independent random variables, then

$$\mathrm{Var}(f(X_1, \ldots, X_n)) \leq \mathbb{E}[f(X_1, \ldots, X_n)]$$

If $X_1, \ldots, X_n$ are Bernoulli(p) random variables, then

$$\Pr[|\frac{1}{n}\sum_{i=1}^{n} X_i - p| \geq p\delta] \leq 2e^{-n\delta^2 p/3}$$