Mathematics of Data Science, Fall 2025 - Homework 2
Due at 11:49PM on Friday Oct/3 (Gradescope)

Your submitted solutions to assignments should be your own work. You are allowed to discuss homework problems with other students, but should carry out the execution of any thoughts/directions discussed independently, on your own. Acknowledge any source you consult. Do not use any type of Large Language Model, e.g., ChatGPT, to blindly answer this assignment. If you do, your submission will be voided and you will get zero as a grade.

Problem 1 - Gaussian wonderland

Let $X \sim N(0, I_n)$ be a standard normal random vector in $\mathbf{R}^n$.

(a) Pick a matrix $Q \in O(n) = \{A \in \mathbf{R}^{n\times n} \mid AA^\top = I_n\}$. Show that $QX \sim N(0, I)$.

(b) Show that $X \sim \sqrt{\xi}U$ where $\xi \sim \chi_n^2$ and $U \sim \text{Unif}(\mathbb{S}^{n-1})$.[1]

(c) Let G be a random $m \times n$ Gaussian matrix with i.i.d. entries where $G_{11} \sim N(0,1)$. Let $u, v \in \mathbb{S}^{n-1}$ be unit orthogonal vectors. Prove that Gu and Gv are i.i.d. random vectors with $Gu \sim N(0, I_m)$.

Problem 2 - Arguments that we missed

Show the following two things we did not prove in class.

(a) Let $X \in \mathbf{R}^n$ be a vector. Its order statistics are given by the reordering

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}.$$

Prove that for any $X, Y \in \mathbf{R}^n$ and $k \in [n]$, we have that

$$|X_{(k)} - Y_{(k)}| \le \|X - Y\|_2.$$

(b) Let $X_1, \ldots, X_n$ be i.i.d. random variables with $X_1 \sim N(0, 1/n)$. Pick a $\delta > 0$ and consider $R_\delta = \{x : |\|x\| - 1| \le \delta \text{ and } |X_1| \le \delta\}$. Prove an upper bound on $\mathbb{P}(X \notin R_\delta)$ that goes to zero as $n \to \infty$. Thus, with high probability, a Gaussian vector concentrates on a ring.

(c) (**Bonus**) Let $\Psi \colon \mathbf{R}_+ \to \mathbf{R}_+$ be an increasing, convex function such that $\psi(0) = 0$, and $\psi(x) \to \infty$ as $x \to \infty$. Define the Orlicz norm of a random variable X as

$$\|X\|_\psi = \inf\{t > 0 \mid \mathbb{E}\psi(|X|/t) \le 1\}.$$

The *Orlicz space* $L_\psi = L_\psi(\Omega, \Sigma, \mathbb{P})$ consists of all random variables X in the probability space $(\Omega, \Sigma, \mathbb{P})$ with finite Orlicz norm, i.e., $L_\psi = \{X \mid \|X\|_\psi < \infty\}$. Show that $\|\cdot\|_\psi$ is indeed a norm on the space L_ψ.

[1] The symbol $\sim$ denotes "has the same distribution as," ${\chi_n}^2$ is a chi-squared distribution with n degrees of freedom, and $\mathbb{S}^{n-1} = \{u \in \mathbf{R}^n \mid \|u\|_2 = 1\}$.

Problem 3 - Second-order Gaussian chaos

Fix a symmetric matrix $A \in \mathbf{R}^{n\times n}$ with zero diagonal. Let $X \sim N(0, I_n)$ be a standard normal random vector. The quadratic form $Z = X^\top AX$ is called a second-order Gaussian chaos.

(a) Compute $\mathbb{E}Z$ and $\operatorname{Var} Z$.

(b) Explain why $Z \sim \sum_{i=1}^n \lambda_i ({X_i}^2 - 1)$ where $(\lambda_1, \dots, \lambda_n)$ are the eigenvalues of A.

(c) Prove the upper tail bound

$$\mathbb{P}(Z \geq t) \leq \exp\left(-\min\left\{\frac{c_1 t}{\|A\|_{\mathrm{op}}}, \frac{c_2 t^2}{\|A\|_F^2}\right\}\right)$$

where $c_1, c_2 > 0$ are universal constants (find them explicitly), and $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_F$ denote the operator and Frobenius norm, respectively.

Problem 4 - Radamacher processes

Let $\varepsilon_1, \dots, \varepsilon_n$ be independent symmetric Bernoulli random variables (also known as Radamacher), i.e., $\mathbb{P}(\epsilon_1 = \pm 1) = 1/2$, and let $T \subseteq \mathbf{R}^n$ be a set. Define

$$Z = \sup_{t\in T} \sum_{k=1}^n \varepsilon_k t_k.$$

Prove a bound of the form

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq 2\exp(-t^2/2\sigma^2) \qquad \text{with} \qquad \sigma^2 = \sum_{k=1}^n \sup_{t\in T} t_k^2.$$

Show with an example that the variance proxy σ^2 can exhibit a vastly incorrect scaling as a function of the dimension n.