Homework 11
High-Dimensional Probability for Data Science, Fall 2023

Hints are in the back of this homework set.

As in the previous homework sets, $C, C_1, C_2, \ldots$ and $c, c_1, c_2, \ldots$ denote positive absolute constants of your choice.

We introduced subgaussian random variables and subgaussian norm before. Here extend these notions to higher dimensions. We say that a random *vector* X taking values in $\mathbb{R}^d$ is subgaussian if all of its one-dimensional projections are subgaussian, i.e. $\langle X, v\rangle$ is a subgaussian random *variable* for any fixed vector $v \in \mathbb{R}^d$. We define the *subgaussian norm* of X as

$$\|X\|_{\psi_2} = \sup_v \|\langle X, v\rangle\|_{\psi_2} \tag{1}$$

where the supremum is over all unit vectors v in $\mathbb{R}^d$.

1. The coordinates of a subgaussian random vector

(a) Let $X_1, \ldots, X_d$ be subgaussian random variables. Check that $X = (X_1, \ldots, X_d)$ is a sub-gaussian random vector in $\mathbb{R}^d$.

(b) Let $X_1, \ldots, X_d$ be *independent* and mean zero subgaussian random variables. Show that that the subgaussian norm of the random vector $X = (X_1, \ldots, X_d)$ is equivalent to the largest subgaussian norm of its coordinates, i.e.

$$\max_{i=1,\ldots,d} \|X_i\|_{\psi_2} \le \|X\|_{\psi_2} \le C_1 \max_{i=1,\ldots,d} \|X_i\|_{\psi_2}. \tag{2}$$

(c) Show by example that the independence assumption is essential in part (b). Specifically, find a random vector X (with dependent coordinates) such that $\|X\|_{\psi_2}$ is much larger than $\max_{i=1,\ldots,d} \|X_i\|_{\psi_2}$.

---

2. Examples of subgaussian random vectors

(a) Show that the uniform distribution on the unit cube $[-1, 1]^d$ is subgaussian, with subgaussian norm bounded by an absolute constant. Repeat this problem for the Boolean cube $\{-1, 1\}^d$.

(b) Consider a normal random vector X with mean 0 and covariance matrix Σ, i.e. $X \sim N(0, \Sigma)$. Show that X is a subgaussian random vector, and

$$\|X\|_{\psi_2} \le C_2 \sqrt{\|\Sigma\|} \tag{3}$$

where we have the operator norm of Σ in the right hand side.

In Lecture 20, we gave a solution to the covariance estimation problem for normally distributed random vectors $X \sim N(0, \Sigma)$. In the following problem, we generalize the solution for all high-dimensional *subgaussian* distributions. Thereby we provide guarantees for PCA for more realistic types of data. Specifically, we will assume that X is a subgaussian that satisfies a bound like (3). In this general situation, we will deduce the same covariance estimation guarantees for X as we did in Lecture 20 for normal distributions.

The result you will establish was first proved only in 2000's. Congratulations on having advanced to the forefront of mathematical data science! Keep doing a great job.

3. Covariance estimation for subgaussian distributions

(a) Let X be a subgaussian random vector taking values in $\mathbb{R}^d$ with mean zero and covariance matrix $\Sigma = \mathbb{E}\, XX^\mathsf{T}$. Consider the sample covariance matrix $\Sigma_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\mathsf{T}$, where X_i are independent random vectors with the same distribution as X. Prove that

$$\|\Sigma_n - \Sigma\| \le C_3 \sqrt{\frac{d}{n}}\, \|X\|_{\psi_2}^2 \tag{4}$$

with probability at least $1 - 2e^{-d}$, as long as $n \ge d$.

(b) In particular, for $X \sim N(0, \Sigma)$, explain how the bound (4) becomes

$$\|\Sigma_n - \Sigma\| \le C_4 \sqrt{\frac{d}{n}}\, \|\Sigma\|\,.$$

In other words, for normal distribution we can achieve covariance estimation with multiplicative error (which is small if the sample size $n \gg d$).

TURN OVER FOR HINTS

Hints

Hint for Problem 1.

(a) Express $\langle X, v\rangle$ as a sum of random variables $X_i v_i$, and use triangle inequality. (Recall that the subgaussian norm is indeed a norm, so triangle inequality holds.)

(b) For the lower bound, use definition (1) for v chosen from the standard basis of $\mathbb{R}^d$. For the upper bound, express $\langle X, v\rangle$ as a sum of *independent* random variables $X_i v_i$, and use Proposition on p.2 from Lecture 11 for these random variables.

(c) Choose the random vector X whose coordinates are identical, e.g. $X = (g, g, \dots, g)$ where $g \sim N(0, 1)$.

Hint for Problem 2.

(a) This should follow immediately from Problem 1(b).

(b) Check that $\langle X, v\rangle$ is a normal random variable with mean 0 and variance $\sigma^2 = v^\mathsf{T}\Sigma v$; use the definition of the operator norm to bound σ^2; recall the gaussian tail. A similar computation was made on p.2 of Lecture 20.

Hint for Problem 3.

(a) Argue like in the proof of the covariance estimation theorem for normal distributions (Lecture 20). Argue that, without loss of generality, we can assume that $\|X\|_{\psi_2} = 1$. (Replace X with X/M where M is the subgaussian norm of X.) Follow the proof and use Bernstein's inequality for general δ. This should yield a probability bound of the form $2\exp[-c \cdot \min(\delta^2, \delta)n]$. Now choose $\delta = C\sqrt{d/n}$ with a sufficiently large absolute constant C, and the probability bound above becomes $2\exp(-10d)$ (use the assumption that $d \le n$). This is sufficient to finish the proof as in Lecture 20.