Homework 3
High-Dimensional Probability for Data Science, Fall 2023

Hints are in the back of the homework set.

In all problems of this homework as well as in the future homework sets, $C, C_1, C_2, \ldots$ and $c, c_1, c_2, \ldots$ denote *positive absolute constants of your choice*. The rationale is that we usually do not care about values of constants; we care more about how the result depends on critical quantities like dimension, sample size, etc. Thus, whenever you see C_1, you can replace it any positive constant you like, for example 10 or 100.

Usually, you will find it easier to choose big values for $C, C_1, C_2, \ldots$ and small values for $c, c_1, c_2, \ldots$. For instance, you are free to choose $C = 100$ and $c = 0.01$, or even leave the values of C and c unspecified as long as it is clear that they are absolute constants, which do not depend on anything. For example $C_2 = \sqrt{n}$ is not a valid choice but $C_2 = 1000000$ is.

Binomial coefficients are often awkward to work with because they are expressed in terms of factorials. One can approximately simplify factorials using Stirling's formula, but the result can still be a little complicated for practical purposes. In this problem, we note a simple and popular two-sided bound on the binomial coefficients. Basically, it says that $\binom{n}{m}$ is approximately $\left(\frac{n}{m}\right)^m$. The same approximation holds even for a more complicated object – the partial sums of the binomial coefficients, which we denote

$$\binom{n}{\leq m} := \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{m}.$$

Problem 1 (Binomial coefficients)

Prove the following inequalities:

$$\left(\frac{n}{m}\right)^m \leq \binom{n}{m} \leq \binom{n}{\leq m} \leq \left(\frac{en}{m}\right)^m$$

for all integers $m \in [1, n]$.

In Lecture 3, we established an upper bound on the volume of any polytope with m vertices contained in the unit Euclidean ball B in $\mathbb{R}^n$. We showed that the ratio of the volumes of P and B is always bounded by $(3\sqrt{\log(m)/n})^n$. In the next two problems, we strengthen this bound by replacing m with m/n. This stronger bound was first proved by Carl and Pajor [1] in 1988. Dafnis, Giannopoulos and Tsolomitis [2] showed in 2009 that Carl-Pajor's bound is optimal for the entire range of m and n

by considering random polytopes. Congratulations: you are proving some serious and relatively modern results!

Problem 2 (Covering numbers of polytopes)

Let P be a polytope with m vertices contained in the unit Euclidean ball B in $\mathbb{R}^n$. Prove that the covering numbers of P satisfy

$$N(P, \varepsilon) \le (C_1 m \varepsilon^2)^{1/\varepsilon^2}$$

for any ε such that $m\varepsilon^2 \ge 1$. (You may assume for simplicity that $1/\varepsilon^2$ is an integer.)

Problem 3 (Volume of polytopes)

Let P be a polytope with m vertices contained in the unit Euclidean ball B in $\mathbb{R}^n$. Deduce that the volume of P satisfies

$$\frac{\mathrm{Vol}(P)}{\mathrm{Vol}(B)} \le \left(C \sqrt{\frac{\log(em/n)}{n}} \right)^n .$$

Berry-Eseen central limit theorem states that for any i.i.d (independent and identically distributed) random variables $X_1, \dots, X_n$ with zero mean, unit variance, the normalized sum $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^N X_i$ satisfies

$$\sup_{x \in \mathbb{R}} |\mathbb{P}\{S_n \le x\} - \mathbb{P}\{g \le x\}| \le \frac{C\rho}{\sqrt{N}} \quad \text{for all } N \in \mathbb{N},$$

Here $g \sim N(0,1)$ is a standard normal random variable and $\rho = \mathbb{E}|X_1|^3$. In Lecture 4 we gave a heuristic explanation why the error $O(1/\sqrt{N})$ is optimal. Now let us prove this formally.

Problem 4 (The error in CLT is at least $1/\sqrt{N}$)

Find random variables X_i that satisfy the assumptions of the Berry-Eseen central limit theorem, and for which

$$\sup_{x \in \mathbb{R}} |\mathbb{P}\{S_N \le x\} - \mathbb{P}\{g \le x\}| \ge \frac{c\rho}{\sqrt{N}} \quad \text{for all } N \in \mathbb{N}.$$

Problem 5 (Heavy-tailed distributions)

Give an example of a random variable X that has finite expectation (i.e. $\mathbb{E}\, X < \infty$) but infinite variance (i.e. $\mathrm{Var}(X) = \infty$).

References

[1] B. Carl, A. Pajor, *Gelfand numbers of operators with values in a Hilbert space,* Inventiones Mathematicae 94 (1988), 479–504.

[2] N. Dafnis, A. Giannopoulos, A. Tsolomitis, *Asymptotic shape of a random polytope in a convex body,* Journal of Functional Analysis 257 (2009), 2820–2839.

TURN OVER FOR HINTS

Hints

Hints for Problem 1.

To prove the upper bound, multiply both sides by the quantity $(m/n)^m$, replace this quantity by $(m/n)^k$ in the left side, and use the binomial theorem (a.k.a. Newton's binomial). To prove the lower bound, use the definition of the binomial coefficient to express it as a product of m fractions; check that each fraction is lower bounded by n/m.

Hints for Problem 2.

In Lecture 3 we proved a slightly weaker bound. Proceed similarly but with a sharper bound on the cardinality of the set $\mathcal{N}$. You can use without proof that the number of ways to choose an *unordered* subset of k elements from a set of m elements equals $\binom{m+k-1}{k}$. Simplify this binomial coefficient using Problem 1.

Hints for Problem 3.

in Lecture 3 we proved a slightly weaker bound. Proceed similarly. At the end, you will need to optimize ε. If this becomes a challenging task, you may guess a good value for ε instead. Recall that it was $\varepsilon = \sqrt{2\log(m)/n}$ in Lecture 3; now you will easily guess how to modify it.

Hints for Problem 4. Make each X_i take value 1 and -1 with probabilities $1/2$, and estimate $\mathbb{P}\{g = 0\}$ (this should be trivial) and $\mathbb{P}\{S_n = 0\}$ (express the latter probability in terms of binomial coefficients, write them out via factorials, and use a convenient form of Stirling's formula). Now slightly perturb x around the zero value; how does the perturbation affect the values of $\mathbb{P}\{S_n \le x\}$ and $\mathbb{P}\{g \le x\}$?