Homework 5
High-Dimensional Probability for Data Science, Fall 2023

Hints are in the back of the homework set.

As before, $C, C_1, C_2, \dots$ and $c, c_1, c_2, \dots$ denote *positive absolute constants of your choice*. See more explanation in Homework 3.

The version of Chernoff's inequality we proved in class provides a bound for the *upper* tail on a sum S_N of independent Bernoulli random variables, i.e. we showed that $\mathbb{P}\{S_N \geq t\}$ is small for large t. Let us complement this result by bounding the *lower tail* of S_N, i.e. show that $\mathbb{P}\{S_N \leq t\}$ is small for small t.

1. The lower tail in Chernoff's inequality

Let X_i be independent Bernoulli random variables with parameters p_i. Consider their sum $S_N = \sum_{i=1}^N X_i$ and denote its mean by $\mu = \mathbb{E}\, S_N$. Prove that

$$\mathbb{P}\{S_N \leq t\} \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t \quad \text{for any } 0 < t \leq \mu.$$

Chernoff's inequality is remarkably sharp. It can be reversed up to a factor e in the base of the exponent:

2. Reverse Chernoff inequality

Let S_N be a binomial random variable with mean μ, that is $S_N \sim \mathrm{Binom}(N, \mu/N)$. Show that

$$\mathbb{P}\{S_N \geq t\} \geq e^{-\mu}\left(\frac{\mu}{t}\right)^t$$

for any integer $t \in \{1, \dots, N\}$ such that $t \geq \mu$.

The definition of Poisson distribution involves factorials, which are not very convenient in computations. An alternative approach to such computations is provided by the MGF method.

3. Tail bounds for Poisson distribution

Let X be a random variable that has the Poisson distribution with parameter μ.

(a) Compute the moment generating function (MGF) of X.

(b) Use the MGF method to prove the following tail bounds:

$$\mathbb{P}\left\{X \geq t\right\} \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t, \quad t \geq \mu$$

$$\mathbb{P}\left\{X \leq t\right\} \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t, \quad 0 < t \leq \mu$$

$$\mathbb{P}\left\{|X-\mu| \geq t\sqrt{\mu}\right\} \leq 2\exp\left(-\frac{t^2}{3}\right), \quad 0 \leq t \leq \sqrt{\mu}.$$

---

4. Boosting randomized algorithms

Imagine we have an algorithm for solving some decision problem. (For example, the algorithm may answer the question: “is there a motorcycle in a given image?”). Suppose each time the algorithm runs, it gives the correct answer independently with probability $\frac{1}{2}+\delta$ with some small $\delta \in (0,1/2)$. In other words, the algorithm performs just marginally better than a random guess.

To improve the performance, the following “boosting” procedure is often employed. Run the algorithm N times and take the majority vote. Show that the new algorithm gives the correct answer with probability at least $1-2\exp(-c\delta^2 N)$. This is good because the confidence rapidly (exponentially!) approaches 1 as N grows.

In class, I gave a flawed proof of the following observation, but we figured out how to fix the flaw (see the hint). Let us make the proof right.

5. Irregularity of sparse random graphs

Consider a random graph $G(N,p)$ whose expected degree $d := (N-1)p$ satisfies $d < c\log N$. Then, with probability at least 0.9, at least one vertex has degree at least $10d$.

TURN OVER FOR HINTS

Hint for Problem 1

Note that $\mathbb{P}\{S_N \le t\} = \mathbb{P}\{-S_N \ge -t\}$. Proceed as in the proof of Chernoff's inequality.

Hint for Problem 2

Since S_N has a binomial distribution, the smaller probability $\mathbb{P}\{S_N = t\}$ can be expressed using binomial coefficients. To lower bound the binomial coefficient, use the result of Problem 1 from Homework 3. To handle one of the remaining terms $(1-\mu/N)^{N-t}$, check that the smaller quantity $(1-\mu/N)^{N-\mu}$ is bounded below by $e^{-\mu}$.

Hints for Problem 3.

(a) The Taylor series for the exponential function will help you to simplify the computation.

(b) The same bounds were already proved for sums of Bernoulli random variables S_N. Proceed as in those proofs, but use part (a) whenever a bound on MGF is needed.

Hints for Problem 4.

Apply a convenient form of Chernoff's inequality for S_N being the number of the wrong answers.

Hint for Problem 5.

As we noted in Lecture 8, it was wrong to say that the degrees $\deg(i)$ and $\deg(j)$ are independent because of the possible edge connecting the vertices i and j. Try to remove that obstacle as follows.

Split the set of N vertices into two subsets A and B, each having $N/2$ vertices. (Assume N is even for simplicity.) For each vertex $i \in A$, define the *quasi-degree* $\deg'(i)$ to be the number of edges connecting i to the vertices in B.

Next, check that (a) the quasi-degrees are independent; (b) the quasi-degrees are bounded above by the true degrees; (c) there exists at least one vertex $i \in A$ with a disproportionally large quasi-degree. Part (c) can be obtained by modifying the proof in Lecture 8.