# Homework 6
## 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.

Sparse random graphs tend to have lots of isolated vertices – vertices with no edges incident to them. Let us show this by working out a simple but powerful *first moment method*.

### 1. The first moment method

(a) Consider any events $E_1, \dots, E_N$. Let $X$ denote the "counting random variable", which equals the number of events $E_i$ that occur. Prove that

$$\mathbb{E}\, X = \sum_{i=1}^{N} \mathbb{P}(E_i).$$

(b) Consider a random graph $G(N,p)$ whose expected degree $d \coloneqq (N-1)p$ satisfies $d < c \log N$. Show that the expected number of isolated vertices is at least $N^{0.99}$.

In data science, one often needs to handle a large number of random variables at the same time. Suppose sample $N$ random points from the standard normal distribution. Then, on average, the entire sample lies within $O(\sqrt{\log N})$ from the origin. This is quite a good bound, since the logarithm grows slowly. You will now prove this bound in a general framework, for all subgaussian distributions:

### 2. Maximum of subgaussians

Let $X_1, \dots, X_N$ be sub-gaussian random variables, which are not necessarily independent. Assume that $\|X_i\|_{\psi_2} \le K$ for all $i$. Show that

$$\mathbb{E} \max_{i=1,\dots,N} |X_i| \le CK\sqrt{\log N}.$$

## 3. Set Balancing

A number of players sign up to play at an amateur soccer club. Each player submits the list of dates he is available to play during the next $D$ days. The coach collects all the lists and sees that at least $C \log D$ players will be available to play on each of the $D$ days. The coach now wants to give each player either a red or a green jersey. He is hoping that on each day, roughly the same number of people will come in jerseys of each color, so they can form two teams and play against each other. Prove that the coach can indeed give out jerseys in such a way that on each of the $D$ days there will be between 49% and 51% players in red jerseys (and thus also between 49% and 51% players in green jerseys).

Let us prove this result by a probabilistic method. Let the coach give each player a red or green jersey independently with probability 1/2. Let us show that the desired conclusion holds with positive probability. To do so, follow these steps:

(a) Denote by $R \subset \{1, \dots, N\}$ the (random) set of players who received red jerseys. Denote by $A_d \subset \{1, \dots, N\}$ the (deterministic) set of players who are available on day $d$, where $d = 1, \dots, D$. We are interested in the size of $R_d := R \cap A_d$, the set of players in red jerseys who come on day $d$. Note that $|R_d|$ has binomial distribution.

(b) Use Chernoff inequality for small deviations (Lecture 7 p.3) to bound the probability of the bad event where $|R_d|$ deviates more than 1% from its mean.

(c) Use the union bound over days $d = 1, \dots, D$.

---

The classical laws of probability theory, such as the law of large numbers, demonstrate the benefit of *averaging independent observations* $X_i$: the more observations we have, the more confident we become about the mean of the distribution. The next problem gives one more example of the benefit of averaging. Here you will find yourself in an unfamiliar territory where almost nothing is assumed about continuous random variables $X_i$. They may have infinite variance, and even their means do not need to exist!

## 4. Small ball probabilities

Let $X_1, \dots, X_N$ be *non-negative* independent random variables. Assume that the PDF (probability density function) of each $X_i$ is uniformly bounded by 1.

(a) Check that each $X_i$ satisfies

$$\mathbb{P}\{X_i \le \varepsilon\} \le \varepsilon \quad \text{for all } \varepsilon > 0.$$

(b) Show that the MGF (moment generating function) of each $X_i$ satisfies

$$\mathbb{E} \exp(-tX_i) \le \frac{1}{t} \quad \text{for all } t > 0.$$

(c) Deduce that averaging increases the strength of (a) dramatically. Namely, show that

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N} X_i \le \varepsilon\right\} \le (C\varepsilon)^N \quad \text{for all } \varepsilon > 0.$$

TURN OVER FOR HINTS

## Hints for Problem 1

(a) Express $X$ as the sum of indicator random variables $\sum_{i=1}^{N} \mathbf{1}_{E_i}$, and use linearity of expectation.

(b) Consider the events $E_i = \{\text{vertex } i \text{ is isolated}\}$. Show that $\mathbb{P}(E_i) \geq N^{-0.01}$ if the absolute constant $c > 0$ in the assumption is chosen sufficiently small. Then use the first moment method from part (a).

## Hints for Problem 2

Bound $\mathbb{E} \max_i |X_i|$ by $\mathbb{E} \left( \sum_{i=1}^{n} |X_i|^p \right)^{1/p}$; move the expected value inside the sum (how?); use the subgaussian bound on the moments (page 1 of Lecture 11); and finally optimize in $p$, or just guess any value of $p$ that works.

## Hints for Problem 3

(a) You should get $R_d \sim \text{Binom}(|A_d|, \frac{1}{2})$.

(b) The probability bound you should get is $2\exp(-0.01^2 \cdot \frac{1}{2}|A_d|/6)$. Use the lower bound on $|A_d|$ from the assumption to further bound this probability by $< 1/D$, if the absolute constant $C > 0$ is chosen sufficiently large.

(c) Since the probability of each bad event from (b) is less than $1/D$, the probability of the union of $D$ bad events is less than 1.

## Hints for Problem 4

(b) Recall the formula for the expectation of a function of a random variable, in terms of the probability density function.

(c) Rewrite the inequality $\sum X_i \leq \varepsilon N$ as $\sum(-X_i/\varepsilon) \geq -N$ and use the MGF method as in the proof of Hoeffding's inequality. Use part (b) to bound the MGF.