Homework 7
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.

In Lecture 11, we discovered that the standard Gaussian random vector in $\mathbb{R}^n$ is very likely to be near the sphere of radius $\sqrt{n}$. This "thin shell phenomenon" is not particular to the Gaussian distribution. You should be able to extend our argument to all subgaussian distributions.

1. Thin shell phenomenon for subgaussian distributions

Consider a random vector $X = (X_1, \dots, X_n)$ in $\mathbb{R}^n$ whose all coordinates X_i are independent random variables that satisfy

$$\mathbb{E}\, X_i = 0, \quad \operatorname{Var}(X_i) = 1, \quad \|X_i\|_{\psi_2} \le K.$$

Show that

$$\mathbb{P}\left\{0.99\sqrt{n} \le \|X\|_2 \le 1.01\sqrt{n}\right\} \ge 1 - 2\exp(-c_1 n).$$

Our proof of Johnson-Lindenstrauss lemma utilized a *Gaussian random matrix* – a matrix whose entries are $N(0,1)$ – to project the data onto a space of lower dimension. Here you will check that a *Bernoulli random matrix* – a matrix with ± 1 entries – works as well. Bernoulli matrices take less memory to store – one bit per entry – so they are preferred in practice. The result you are about to prove in part (b) was first established by D. Achlioptas[1] in 2003.

2. Johnson-Lindenstrauss lemma with binary coins

Let B be an $n \times d$ Bernoulli random matrix, i.e. a matrix whose entries are i.i.d. symmetric Bernoulli random variables (that is, each entry takes values ± 1 with probability $1/2$).

(a) Fix any unit vector $z \in \mathbb{R}^d$. Prove that the random vector Bz satisfies the thin-shell phenomenon:

$$\mathbb{P}\left\{0.99\sqrt{n} \le \|Bz\|_2 \le 1.01\sqrt{n}\right\} \ge 1 - 2\exp(-c_1 n).$$

---

[1] D. Achlioptas, *Database-friendly random projections: Johnson-Lindenstrauss with binary coins*, Journal of Computer and System Sciences, Volume 66, Issue 4, June 2003, Pages 671-687.

(b) Let $x_1, \ldots, x_N$ be any fixed vectors in $\mathbb{R}^d$. Let B be an $n \times d$ Bernoulli random matrix, and set $T = \frac{1}{\sqrt{n}} B$. Prove that if $n = C \log N$, then the map $T : \mathbb{R}^d \to \mathbb{R}^n$ approximately preserves the pairwise geometry of the data set, namely that following event holds with positive probability:

$$0.99 \left\|x_i - x_j\right\|_2 \leq \left\|T(x_i) - T(x_j)\right\|_2 \leq 1.01 \left\|x_i - x_j\right\|_2 \quad \text{for all } i, j = 1, \ldots, N. \tag{1}$$

The most surprising feature of Johnson-Lindenstrauss lemma is its ability to compress the data into such a small dimension, namely $n = O(\log N)$. One may wonder whether this can be further improved: can one always compress the data into dimension $n = o(\log N)$? We will show that this is not the case: the logarithmic dimension is optimal. Moreover, it is optimal even if we allow the compression map to be arbitrary and possibly nonlinear.

3. No Johnson-Lindenstrauss into $o(\log N)$ dimension

(a) Let $z_1, \ldots, z_N$ be vectors in $\mathbb{R}^n$ that satisfy

$$1 < \left\|z_i - z_j\right\|_2 \leq 2 \quad \text{for all distinct } i, j \in \{1, \ldots, N\}. \tag{2}$$

Show that $N \leq 5^n$.

(b) Let $n < \frac{1}{2} \log N$. Find vectors $x_1, \ldots, x_N$ in $\mathbb{R}^N$ for which there does not exist any map $T : \mathbb{R}^N \to \mathbb{R}^n$ that satisfies (1).

Life in high dimensions is full of surprises. We know from linear algebra that the space $\mathbb{R}^n$ can not accommodate more than n orthogonal vectors. However, $\mathbb{R}^n$ can accommodate exponentially many *almost* orthogonal vectors, for large n. This is another manifestation of how much more room there is in high-dimensional worlds than in our three-dimensional world.

4. There are exponentially many almost orthogonal vectors

(a) Prove that there exist unit vectors $z_1, \ldots, z_N$ in $\mathbb{R}^n$ such that $N \geq \exp(cn)$ and

$$\left|\langle z_i, z_j \rangle\right| \leq 0.01 \quad \text{for all distinct } i, j \in \{1, \ldots, N\}. \tag{3}$$

(b) Show that if unit vectors $z_1, \ldots, z_N$ in $\mathbb{R}^n$ that satisfy (3), then $N \leq 5^n$.

TURN OVER FOR HINTS

Hints for Problem 2

(a) Show that the random vector $X = Bz$ satisfies the assumption of Problem 1. You will need to use subgaussian Hoeffding inequality proved in Lecture 11.

(b) Argue like in the proof of Johnson-Lindenstrauss lemma in Lecture 12.

Hints for Problem 3

(a) Homework 2 Problem 5(a) implies that that all points z_i lie in some ball of radius 2. Consider Euclidean balls centered at points z_i and with radii $1/2$. Show that these balls are disjoint and lie in some Euclidean ball of radius $2+1/2$. Thus the total volume of those balls is bounded by the ball in which they lie. Now if you remember how the volume scales in dimension n, you should be able to complete the proof.

(b) Choose $x_1, \ldots, x_N$ to be the standard basis of $\mathbb{R}^N$ and let $z_i = T(x_i)$.

Hints for Problem 4

(a) Let all coordinates of all vectors z_{ij} be independent symmetric Bernoulli random variables multiplied by $1/\sqrt{n}$. Fix a pair (i, j) and use Hoeffding's inequality to show that the bad event $|\langle z_i, z_j \rangle| > 0.01$ occurs with exponentially small probability. Conclude by taking the union bound over all pairs (i, j). The logic of this argument is similar to the proof of Johnson-Lindenstrauss lemma.

(b) Check that (3) for unit vectors implies (2).