DUE DATE: SEPTEMBER 28 2020.

TYPE YOUR SOLUTION USING WORD OR $\LaTeX$,
SUBMIT A PDF DOCUMENT ON CANVAS
HANDWRITTEN SOLUTIONS WILL NOT BE GRADED.

The purpose of this homework is to prove a first version of the Johnson-Lindenstrauss lemma. We will prove a second version in class. The lemma has immense application in data-science to reduce the dimensionality of large datasets.

Let $x_1, \dots, x_N$ be N points in $\mathbb{R}^n$. We consider a $d \times n$ random $\mathbf{G}$ matrix where the entries are independent zero-mean, unit variance, sub-Gaussian random variables, $g_{i,j}, 1 \le i \le d; 1 \le j \le n$. By choosing ν such that

$$\nu = \max_{i,j} \|g_{i,j}\|_{\psi_2}, \tag{1}$$

we can show that there exists $\nu > 0$, such that

$$\forall 1 \le i \le d, \quad \forall 1 \le j \le n, \qquad \mathbb{P}\left(\left|g_{i,j}\right| > t\right) \le 2\exp\left\{-t^2/\nu^2\right\}. \tag{2}$$

Finally we construct the following $d \times n$ matrix,

$$\mathbf{Q} = \frac{1}{\sqrt{d}}\mathbf{G}. \tag{3}$$

In this problem set, we will show that $\mathbf{Q}$ behaves as an isometry, with high probability.

Assignment $[20 = 10{+}10]$ Let $x \in \mathbb{R}^n$. We apply the matrix $\mathbf{G}$ to x, and define

$$y = \mathbf{G}x = \begin{bmatrix} y_1 \\ \vdots \\ y_d \end{bmatrix}. \tag{4}$$

1. Prove that $\mathbf{Q}$ is an isometry "in expectation",

$$\mathbb{E}\left[\|\mathbf{Q}x\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{\sqrt{d}}y\right\|^2\right] = \|x\|^2. \tag{5}$$

2. Prove that the $y_i, 1 \le i \le d$ are independent sub-Gaussian random variables, and that there exists $c_1 > 0$ such that,

$$\|y_i\|_{\psi_2} \le \|x\| \frac{\nu}{\sqrt{c_1}} \tag{6}$$

HINT: you may use the general Hoeffding inequality for sub-Gaussian random variables (see theorem 2.6.3 in the textbook).

Assignment [60=10+10+20+20]

3. Compute $\mathbb{E}[y_i]$, and $Var[y_i]$ as function of $\|\mathbf{x}\|$.

4. Prove that y_i^2 is sub-exponential, and

$$\|y_i^2\|_{\psi_1} = \|y_i\|_{\psi_2}^2. \tag{7}$$

5. We define the centered random variable

$$z_i = y_i^2 - \mathbb{E}\left[y_i^2\right]. \tag{8}$$

We can show that z_i is also sub-exponential (see Excercise 2.7.10 of the textbook), and there exists $c_2 > 0$, such that

$$\|z_i\|_{\psi_1} = \left\|y_i^2 - \mathbb{E}\left[y_i^2\right]\right\|_{\psi_1} \le c_2 \|y_i^2\|_{\psi_1}. \tag{9}$$

Use the corollary of the general Bernstein inequality (corollary 2.8.3 in the textbook), which was presented in class, to show that there exist $\alpha > 0$ and $\beta > 0$ such that

$$\mathbb{P}\left(\left|\frac{1}{d}\sum_{i=1}^{d} z_i\right| \ge \varepsilon\right) \le 2\exp\left\{-\alpha \min\left(\frac{\varepsilon^2}{\beta^2}, \frac{\varepsilon}{\beta}\right) d\right\}, \tag{10}$$

where

$$\beta = \max_1^d \|z_i\|_{\psi_1} \le \frac{c_2}{c_1}\|\mathbf{x}\|^2 \nu^2. \tag{11}$$

We are interested in the small deviation regime (a similar analysis can be performed for large deviation), and we assume therefore

$$\frac{\varepsilon}{\beta} < 1. \tag{12}$$

6. Let $0 < \delta < 1$. Prove that if we choose

$$d = \frac{2\beta^2}{\alpha \varepsilon^2} \log\left(N/\sqrt{\delta}\right) \tag{13}$$

then

$$\mathbb{P}\left(\left|\frac{1}{d}\sum_{i=1}^{d} z_i\right| \ge \varepsilon\right) \le \frac{2\delta}{N^2}, \tag{14}$$

Assignment [20]

7. We now consider the dataset $\mathbf{x}_1, \ldots, \mathbf{x}_N$, and select d according to (13).

   Prove that if we choose $\mathbf{G} = (g_{i,j})$ at random in the set of random sub-Gaussian matrices, then with probability $1 - \delta$, the matrix $\mathbf{Q} = \frac{1}{\sqrt{d}}\mathbf{G}$ is a ε-isometry for this dataset,

$$\mathbb{P}\left(\mathbf{G} = (g_{i,j})\,; \quad \forall\, 1 \le k < l \le N, \quad \left|\frac{\|\mathbf{Q}(\mathbf{x}_k - \mathbf{x}_l)\|^2}{\|\mathbf{x}_k - \mathbf{x}_l\|^2} - 1\right| \ge \varepsilon\right) \le \delta. \tag{15}$$

   HINT: use a union bound.

In other words, with probability $1 - \delta$, the pairwise distance between any two points $\mathbf{x}_k$ and $\mathbf{x}_l$ is preserved (within a factor of $1 \pm \varepsilon$), after applying the random matrix $\mathbf{Q}$. There are two remarkable features about this result:

- The dimension d onto which we project using $\mathbf{Q}$ is independent of the ambient dimension n;
- $\mathbf{Q}$ is chosen completely at random.