BIG DATA COMPUTING

Homework 4
Random Projections

Leonardo Di Nino : 1919479

March 15, 2024

Assignment 2

Question 1

We want to prove that given a random projection matrix $S = \frac{1}{\sqrt{k}}U$, being U a matrix whose entries are all independent and similar standard gaussians ($U_{ij} \sim N(0,1)$), and defining a mapping $f : \mathbb{R}^d \to \mathbb{R}^k$ such that $f(v) = Sv$, the following holds for every $x, y \in \mathbb{R}^d$:

$$\mathbb{E}[f(x)^T f(y)] = x^T y \tag{1}$$

Let's start expanding the expression within brackets:

$$\mathbb{E}[f(x)^T f(y)] = \mathbb{E}[(Sx)^T Sy] = \mathbb{E}[x^T S^T Sy] = \mathbb{E}[x^T \frac{1}{\sqrt{k}} U^T \frac{1}{\sqrt{k}} Uy] = \frac{1}{k}\mathbb{E}[x^T U^T U y] \tag{2}$$

Since $U^T U \in R^{d \times d}$ and we are going to consider the expected value of a matrix to be the matrix of the expected values of its entries, we have:

$$\mathbb{E}[f(x)^T f(y)] = \frac{1}{k} x^T \mathbb{E}[U^T U] y \tag{3}$$

In particular we have that $(U^T U)_{rs} = \sum_{i=1}^{k} u_{ri} u_{si}$ since each entry of $U^T U$ is the dot product between the r-th row of U^T, i.e. the r-th column of U, and the s-th column of U.
So now we can figure out the expected value of $U^T U$:

$$\mathbb{E}[(U^T U)_{rs}] = \mathbb{E}[\sum_{i=1}^{k} u_{ri} u_{si}] = \sum_{i=1}^{k} \mathbb{E}[u_{ri} u_{si}] \tag{4}$$

Now we should distinguish two cases taking into account the distribution of the entries of the matrix and that under IID assumption expected value can be factorized:

$$\begin{cases} \mathbb{E}[u_{ri} u_{si}] = \mathbb{E}[u_{ri}^2] = 1 & \text{if } r = s \\ \mathbb{E}[u_{ri} u_{si}] = \mathbb{E}[u_{ri}]\mathbb{E}[u_{si}] = 0 & \text{if } r \neq s \end{cases}$$

This implies that:

$$\mathbb{E}[(U^T U)_{rs}] = \mathbb{E}[\sum_{i=1}^{k} u_{ri} u_{si}] = \sum_{i=1}^{k} \mathbb{E}[u_{ri} u_{si}] = \begin{cases} k & \text{if } r = s \\ 0 & \text{if } r \neq s \end{cases} \tag{5}$$

Rearranging a bit the expression in eq.(3) we have:

$$\mathbb{E}[f(x)^T f(y)] = x^T \frac{1}{k} \mathbb{E}[U^T U] y \tag{6}$$

And finally we have the following:

$$\frac{1}{k}\mathbb{E}[(U^T U)_{rs}] = \frac{1}{k}\mathbb{E}[\sum_{i=1}^{k} u_{ri} u_{si}] = \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}[u_{ri} u_{si}] = \begin{cases} 1 & \text{if } r = s \\ 0 & \text{if } r \neq s \end{cases} \tag{7}$$

This clearly shows that $\mathbb{E}[U^T U] = \mathbb{I}$ and this ends the proof, since:

$$\mathbb{E}[f(x)^T f(y)] = x^T \frac{1}{k} \mathbb{E}[U^T U] y = x^T \mathbb{I} y = x^T y \tag{8}$$

Question 2

Since the cosine of the angle θ between two vectors is the following:

$$cos(\theta) = \frac{x^T y}{||x||_2 ||y||_2} \tag{9}$$

In our case we can simplify the expression since the vectors are L_2 normalized:

$$d = cos(\theta) = x^T y \tag{10}$$

This means that a reasonable estimator for this quantity is the following:

$$\hat{d} = \frac{1}{N} \sum_{i=1}^{N} f_i(x)^T f_i(y) \tag{11}$$

We could guess this expression from what we have done in the previous point, leveraging a standard repeated sampling and averaging scheme based on the IID assumption which the matrix S is built on. In particular it is easy to show that:

$$\mathbb{E}[\hat{d}] = \mathbb{E}[\frac{1}{N} \sum_{i=1}^{N} f_i(x)^T f_i(y)] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[f_i(x)^T f_i(y)] = \frac{1}{N} N x^T y = x^T y \tag{12}$$