Universidade Federal do Rio de Janeiro

The Johnson-Lindenstrauss Lemma

Felipe Pagginelli Patrício

Rio de Janeiro

Maio de 2020

Universidade Federal do Rio de Janeiro

The Johnson-Lindenstrauss Lemma

Felipe Pagginelli Patrício

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática.

Orientador: Prof. Wladimir Augusto das Neves

Coorientador: Prof. Hugo Tremonte de Carvalho

Rio de Janeiro
Maio de 2020

Universidade Federal do Rio de Janeiro

The Johnson-Lindenstrauss Lemma

Felipe Pagginelli Patrício

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática.

Aprovada por:

_____

Presidente, Prof. Wladimir Augusto das Neves (IM-UFRJ)

_____

Prof. Hugo Tremonte de Carvalho (IM-UFRJ)

_____

Prof. Bernardo Freitas Paulo da Costa (IM-UFRJ)

_____

Prof. César Javier Niche Mazzeo (IM-UFRJ)

_____

Prof. Amit Bhaya (COPPE-UFRJ)

Prof. Roberto Imbuzeiro Oliveira (IMPA)

Prof. Carlos Tomei (PUC-Rio)

Rio de Janeiro
Maio de 2020

CIP - Catalogação na Publicação

P71t

Patrício, Felipe Pagginelli
The Johnson-Lindenstrauss Lemma / Felipe Pagginelli Patrício. -- Rio de Janeiro, 2020.
105 f.

Orientador: Wladmir Augusto das Neves.
Coorientador: Hugo Tremonte de Carvalho.
Dissertação (mestrado) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Programa de Pós-Graduação em Matemática, 2020.

1. The Johnson-Lindenstrauss Lemma. 2. Dimensionality Reduction. 3. Concentration of Measure. 4. Random Projections. 5. Data science. I. Neves, Wladmir Augusto das, orient. II. Carvalho, Hugo Tremonte de, coorient. III. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

Resumo

The Johnson-Lindenstrauss Lemma

Felipe Pagginelli Patrício

Resumo da dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática.

Resumo: Os métodos usuais para análise e tomada de decisão são baseados no processamento de informações. Esse campo de estudo, no entanto, tem mudado dramaticamente com o advento dos dados de alta dimensão. O rápido desenvolvimento das tecnologias de armazenamento e aquisição de dados tem possibilitado que dispositivos tomem milhares – ou mesmo milhões – de medições simultaneamente. Os dados em alta dimensão resultam justamente desse tipo de aferição, sendo encontrados comumente em áreas como: processamento de imagens, aprendizado de máquina, reconhecimento de padrões, extração de características, análise de grafos, dentre outros. Entretanto, lidar com esse tipo de informação é muito problemático por vários motivos, em particular, armazenamento e complexidade computacional. Felizmente, dados frequentes em aplicações costumam concentrar-se em estruturas cuja dimensão intrínseca é inferior a que nos é apresentada. Dado isso, podemos nos valer de métodos de pré-processamento para lidar com esse cenário de forma mais palatável. Nesta dissertação, apresentamos um método para redução de dimensão, o Lema de Johnson-Lindenstrauss. Esse resultado surpreendente nos permite projetar um conjunto de dados de M pontos em $\mathbb{R}^N$ quase isometricamente (a menos de um erro pré-fixado), em um subspaço cuja dimensão m possui ordem $\log M$. No mais, m independe da dimensão N do conjunto de dados original.

Palavras–chave. Lema de Johnson-Lindenstrauss, Concentração de Medida, Redução de Dimensão, Alta Dimensionalidade, Ciência de Dados, Projeções Aleatórias, Estatística em Alta Dimensão.

Rio de Janeiro
Maio de 2020

Abstract

The Johnson-Lindenstrauss Lemma

Felipe Pagginelli Patrício

Abstract da dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática.

Abstract: Usual methods for data analysis and decision making are based on information processing. However, this field of study has changed dramatically with the advent of high-dimensional data. The fast development of data storage and acquisition technologies has enabled devices to take thousands – or even millions – of measurements simultaneously. High-dimensional data results precisely from this kind of measurements and are commonly found in fields such as: image processing, machine learning, pattern recognition, feature extraction, graph theory and data streaming, among others. However, dealing with this kind of information is very problematic for several reasons, in particular, storage and computational complexity. Fortunately, datasets that are frequent in applications are usually concentrated in structures whose intrinsic dimension is smaller than the one that we are presented. Therefore, we can use preprocessing methods to deal with this setting in a more reasonable way. In this dissertation, we present a method for dimensionality reduction, the Johnson-Lindenstrauss Lemma. This amazing result allows us to project a dataset of M points in $\mathbb{R}^N$ quasi-isometrically (except for a prefixed error), onto a subspace whose dimension m is of order $\log M$. Moreover, m is independent of the dimension N of the original dataset.

Keywords. Johnson-Lindenstrauss Lemma, Concentration of Measure, Dimensionality Reduction, High-dimensionality, Data science, Random Projections, Statistics in High Dimensions.

Rio de Janeiro
Maio de 2020

Notation & Terminology

Along this text, we shall use the following notations and concepts.

Notation. *Let $N \in \mathbb{N}$. We denote*

$$[N] := \{1, \ldots, N\}.$$

Definition (ℓ_p normed space)**.** *For $p \in [1, \infty)$, we define the normed space ℓ_p as the sequences $x = \{x_i\}_{i=1}^{\infty} \subset \mathbb{R}$ for which the sum $\sum_{i=1}^{\infty} |x_i|^p$ converges. The norm in this space is defined as*

$$\|x\|_p := \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{\frac{1}{p}}.$$

We also define the normed space ℓ_∞ as the sequences $x = \{x_i\}_{i=1}^{\infty} \subset \mathbb{R}$ for which $\sup_{i \geq 1} |x_i| < \infty$. Moreover, the norm in this space is given by

$$\|x\|_\infty := \sup_{i \geq 1} |x_i|.$$

Notation. *We denote the Euclidean space $\mathbb{R}^n$ provided with the norm ℓ_p as ℓ_p^n. In particular, a set $X \subset \mathbb{R}^n$ endowed with the metric ℓ_p will be represented as (X, ℓ_p).*

In the present text, we will represent the set of $(m \times n)$-matrices over a field $\mathbb{K}$ by $\mathcal{M}_{m \times n}(\mathbb{K})$. In particular, when it comes to the *field of Real Numbers*, we will denote it just as $\mathcal{M}_{n \times m}$.

Definition (Frobenius norm of a matrix)**.** *Let $A = (a_{ij})$ be a matrix in $\mathbb{R}^{m \times n}$. We define its Frobenius norm as*

$$\|A\|_{\mathcal{F}} := \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2}.$$

Definition ($\mathcal{L}_{pq}$ matrix norm)**.** *Let A be a matrix representing a linear transformation $\ell_p^n \mapsto \ell_q^m$, with $p, q \geq 1$. We define the norm $\mathcal{L}_{pq}$ of the matrix A as*

$$\|A\|_{p \to q} := \sup_{\|x\|_p = 1} \|A\, x\|_q.$$

Definition (Big-$\mathcal{O}$ notation)**.** *We say that a quantity $g(n)$ is $\mathcal{O}(f(n))$ if, for all $\mathbb{N} \ni n \geq n_0$, we have*

$$|g(n)| \leq k|f(n)|,$$

for some positive constant k.

In the same way we have defined the Big-$\mathcal{O}$ notation to deal with upper asymptotic bounds, we will define Big-Ω notation to deal with lower asymptotic bounds.

Definition (Big-Ω notation)**.** *We say that a quantity $g(n)$ is $\Omega(f(n))$ if, for all $\mathbb{N} \ni n \geq n_0$, we have:*

$$|g(n)| \geq k|f(n)|,$$

for some positive constant k.

Finally, the following notation will be useful throughout the text.

Notation. *Let $\xi, x \in \mathbb{R}$ and $\varepsilon > 0$. The expression*

$$\xi \in [x \pm \varepsilon]$$

means ξ is a quantity in the interval $[x - \varepsilon, x + \varepsilon]$. The notation for other kinds of intervals is made in an analogous fashion.

Contents

Chapter 1

Dimension reduction

1.1 High-dimensional data

Usual procedures for analysis and decision making are based on *data analysis*. However, this field of study has been changing dramatically with the advent of high-dimensional data. Over the last twenty years, the great development of data storage and acquisition technologies has enabled devices to take thousands (or even millions) of measurements simultaneously. High-dimensional data results from such kind of measurements, and images are a basic example of how this kind of data is ubiquitous in our everyday life. For instance, a 300×300 image, that is quite small, is represented by a 90,000-dimensional vector.

Dealing with such wide arrays of data is very problematic since they require high storage, and operating with them leads to unbearable computational burden. Also, methods from *Classical Statistics* are too limited in their capacity to deal effectively with contemporary datasets. Namely, most of Statistics developed during the 20th century focused on data whose number M of experimental units is large compared to the number N of unknown features. Accordingly, most of the classical theory provides results for the asymptotic setting for N fixed and M going to infinity. This approach is very useful for the usual cases, i.e., "large M" and "small N", but it can be seriously misleading for the "large N" case, requiring then a new statistical paradigm. This has led to a new branch of Statistics referred to as *high-dimensional data analysis* [Don2000].

Recently, the efforts of data analysis community have been dealing with these problems by taking advantage of underlying structures of datasets in order to reduce the effective dimension of the original problem. An example is exploring the *sparsity property*, i.e., taking advantage from the fact that usual datasets hold a large number of irrelevant and redundant variables. Another possibility for simplification comes from the fact that high-dimensional data are not usually spread "uniformly" in the Euclidean space, but rather concentrated around some low-dimensional structures.

As a consequence of these simpler settings, a good first step in the analysis of a high-dimensional dataset is to reduce its dimension somehow. A classical statistical approach to dimensionality reduction is the *principal component analysis* (PCA), that projects the dataset in a subspace that maximizes the variance of the projected data. Intuitively, the PCA tries to preserve the global aspect of the original dataset. On the other hand, we

have the method that motivates the present work based on the *The Johnson-Lindenstrauss Lemma*, whose application aims to preserve only the pairwise distances instead of the entire disposition of points in the dataset.

It is important to remark that these two are not the only dimensionality reduction techniques available, and we direct the interested reader to [Bsp2006] for more details. Before we present a little more deeply the PCA and the *multi-dimensional scaling*, let us briefly discuss the phenomenon of the "curse of dimensionality".

1.2 Curse of dimensionality

There is no formal definition of what the curse of dimensionality is. In the literature, this term is used to describe a myriad of unexpected or undesirable events that might occur when dealing with high-dimensional problems and its presentation is usually made through examples [Bsp2006, Gir2014].

According to [Gir2014], the impacts of high-dimensionality can be subdivided in four groups: the first, *high-dimensional spaces are vast and data points are isolated in their immensity*; second, *the accumulation of small noises in many different directions can produce a large global noise*; third, *an event that is an accumulation of rare events may not be rare*; finally, *numerical computations in high-dimensional spaces can be intractable.* The first one is discussed in Section 1.2.1 and the others in Section 1.2.2, both based on [Gir2014].

1.2.1 Geometrical issues of high-dimensionality

- Volume of a N-dimensional ball

  We will point out some counterintuitive geometrical characteristics of the N-dimensional ball when N is high. Firstly, note that *high-dimensional balls with a fixed radius have a vanishing volume.*

  From Multivariate Calculus, we know that the volume of a N-dimensional ball of radius $r > 0$ is equal to

  $$V_N(r) = \frac{\pi^{N/2}}{\Gamma(1 + N/2)} r^N,$$

  with Γ being the Gamma function $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ for $x > 0$. Furthermore, by applying the Stirling approximation

  $$\Gamma(\alpha) = \alpha^{\alpha - 1/2} e^{-\alpha} \sqrt{2\pi} \{1 + \mathcal{O}\left(\alpha^{-1}\right)\} \text{ for } \alpha \to +\infty,$$

  we conclude

  $$V_N(r) \overset{N \to \infty}{\sim} \left(\frac{2\pi e r^2}{N}\right)^{N/2} (N\,\pi)^{-1/2}. \tag{1.2.1}$$

  Note that, for any $r > 0$, this volume goes to zero faster than exponentially with the dimension N. Such behavior in depicted in the plot of $N \mapsto V_N(1)$ in Figure 1.4 from [Gir2014]. Note that for $N = 20$ the volume of the unit ball is almost zero.

As a consequence of this geometrical fact, follows one of the features that characterizes the curse of dimensionality: *a vast space consisting of isolated points*. Namely, suppose we intend to take enough samples $X_1, \ldots, X_M$ from the uniform distribution over the set $[0,1]^N$ such that for any X_i there is another point X_j whose distance to X_i is at most 1. How many points $X_1, \ldots, X_M$ we need to take in this box in order to achieve this goal?

This question can be reformulated as: "how many unit balls centered in $[0,1]^N$ are necessary to cover this set?" If unit balls centered in $X_1, \ldots, X_M \in [0,1]^N$ cover this hypercube, we have

$$[0,1]^N \subset \bigcup_{i=1}^{M} B_N(X_i, 1),$$

with $B_N(x,r) \in \mathbb{R}^N$ being the open ball centered in x with radius $r > 0$. Consequently,

$$\begin{aligned} 1 &= \text{Volume}([0,1]^N) \\ &\leq \text{Volume}\left\{\bigcup_{i=1}^{M} B_N(X_i, 1)\right\} \\ &\leq \sum_{i=1}^{M} \text{Volume}\{B_N(X_i, 1)\} = M\, V_N(1), \end{aligned}$$

that yields

$$M \geq \frac{1}{V_N(1)} = \frac{\Gamma(1 + N/2)}{\pi^{N/2}}.$$

Finally, by making N sufficiently large, we conclude

$$M \geq \left(\frac{N}{2\pi e}\right)^{N/2} \sqrt{N\,\pi},$$

from the Stirling approximation.

We thus arrive at an astonishing result: *the number of unit balls necessary to cover the cube* $[0,1]^N$ *grows more than exponentially fast with the dimension* N. To clarify how unrealistic obtaining these samples can be, we exhibit Table 1.1.

N	20	30	50	100	150	200
M	39	45630	$5.7\,10^{12}$	$4.2\,10^{40}$	$1.28\,10^{72}$	larger than the estimated number of particles in the observable universe

Table 1.1: Lower bound on the number of unitary balls required for covering the hypercube $[0,1]^N$ (from Table 1.1 in [Gir2014]).

Another interesting fact about the volume of N-dimensional balls is that *the volume of a high-dimensional ball is concentrated in its "crust"*. Indeed, we informally consider

as the "crust" of a N-ball $B_N(0,r)$, $r>0$, the set $C_N(r)$ obtained by removing some sufficient inner part of this ball, say

$$C_N(r) := B_N(0,r) - B_N(0,0.99\,r).$$

We than have that

$$\begin{aligned}\frac{\text{Volume}\{C_N(r)\}}{\text{Volume}\{B_N(0,r)\}} &= \frac{\text{Volume}\{B_N(0,r)\} - \text{Volume}\{B_N(0,0.99\,r)\}}{\text{Volume}\{B_N(0,r)\}} \\ &= \frac{V_N(r) - V_N(0.99\,r)}{V_N(r)} = 1 - 0.99^N,\end{aligned}$$

which goes exponentially fast to 1, as seen in the plot of $N \mapsto 1-0.99^N$ in Figure 1.5 from [Gir2014].

- Pairwise distances in a high dimensional spaces

  Let $X=(X_n)_{n\in[N]}$, $Y=(Y_n)_{n\in[N]} \in \mathbb{R}^N$ be i.i.d. random variables uniformly distributed in the hypercube $[0,1]^N$. The mean squared distance between this pair of points is given by

$$\mathbb{E}\{\|Y-X\|_2^2\} = \sum_{n=1}^{N} \mathbb{E}\left\{(Y_n - X_n)^2\right\} = N\,\mathbb{E}\{(Y_1-X_1)^2\} = N/6.$$

  That is, the mean squared pairwise distance between points uniformly sampled in $[0,1]^N$ grows linearly with the dimension N.

  We had already exhibited the immensity of high-dimensional spaces in the previous example in which we tried to cover the N-dimensional unit cube. Now, we have seen that *the higher is the dimension from which the data points are uniformly chosen, larger are their pairwise distances is this space.*

  On the other hand, the standard deviation of this pairwise distance will be

$$\sqrt{\text{Var}\{\|Y-X\|_2^2\}} = \sqrt{\sum_{n=1}^{N} \text{Var}\{(Y_n-X_n)^2\}} \;\approx\; 0.2\sqrt{N}.$$

  That is, even though the standard deviation also grows with the dimension, it does not increase as fast as the mean distance. More precisely, the scaled deviation

$$\frac{\sqrt{\text{Var}\{\|Y-X\|_2^2\}}}{\mathbb{E}\{\|Y-X\|_2^2\}}$$

  shrinks like $1/\sqrt{N}$. Therefore, *for high dimensional settings, uniformly chosen data points have similar pairwise distances.* As we will show next, these unexpected behavior of high dimensional spaces will make some methods from Classical Statistics harder or even impossible, to be applied.

- High dimensional k-Nearest Neighbors estimator

  Consider a dataset of M i.i.d. observations $\{(Y_i, X^{(i)})\}_{i\in[M]}$ from random variables Y and X such that $Y \in \mathbb{R}$ is a *response variable* by N *covariate variables* $X_1, \ldots, X_N \in [0,1]$. Moreover, assuming the X_k's are i.i.d with uniform distribution on $[0,1]$, we can define a random vector $X = (X_1, \ldots, X_N) \in [0,1]^N$ with uniform distribution on the hypercube $[0,1]^N$.

  We describe the relation between X and Y by the *classical regression model*

$$Y_i = f(X^{(i)}) + \varepsilon_i, \quad i \in [M],$$

  with $f : [0,1]^N \to \mathbb{R}$ and $\varepsilon_1, \ldots, \varepsilon_M$ being pairwise independent and *centered*. Now, assuming that f is continuous, it is natural to estimate $f(x)$ by some average of the Y_i's associated to X_i's in some vicinity of $x \in \mathbb{R}^N$. The simplest version of this idea is the *k-Nearest Neighbors estimator*. It is obtained by estimating $f(x)$ through the mean of the Y_i's associated to the k points X_i nearest from x.

  We now recall from our discussion of pairwise distances in high-dimensions that points sampled uniformly in $[0,1]^N$ will have approximately the same pairwise distances. Consequently, *the notion of "nearest points" vanishes*. This phenomenon is illustrated in Figure 1.3 from [Gir2014], where we have plotted the histograms of the distribution of the pairwise distances $\{\|X^{(i)} - X^{(j)}\|_2 : 1 \le i < j \le M\}$ for $M = 100$ and dimensions $N = 2, 10, 100$ and 1000.

1.2.2 Probability and Statistics in high dimensions

- Tails of High-Dimensional Gaussian Distribution

  Gaussian distributions are known to have very thin tails. Indeed, the density

$$g_N(x) = (2\pi)^{-N/2} \exp\{-\|x\|^2/2\}$$

  of a standard Gaussian distribution $\mathcal{N}(0, \boldsymbol{I}_N)$ in $\mathbb{R}^N$ decreases exponentially fast with the square norm of x. Yet, when N is large, most of the mass of the standard Gaussian distribution lies in its tails.

  First, we will show that *the Gaussian distribution in high dimensions is much flatter than in lower ones*, losing the characteristic bell-like shape. In fact, the maximum value of this PDF is $g_N(0) = (2\pi)^{-N/2}$, which decreases exponentially fast toward 0 as N increases.

  Given this result, we expect that the mass around the origin will vanish for high-dimensional settings. More precisely, let $\delta > 0$ be a small positive real number and write the *"bell" set* as

$$B_{N,\delta} = \{x \in \mathbb{R}^N : g_N(x) \ge \delta\, g_N(0)\}.$$

  Our intuition from the one and two-dimensional cases ($N \in \{1, 2\}$) is that a small value of δ would imply that the probability of a standard Gaussian variable being in the bell set $B_{N,\delta}$ is close to one.

Indeed, our intuition of the exponentially decreasing tails on the one and two-dimensional particular cases implies that most of the points will have density much smaller than $g_N(0)$ (and even than $\delta g_N(0)$). On the other hand, we will show that this intuition does not lead to similar conclusions for high-dimensional settings. More precisely, *the probability mass in the bell set decreases exponentially on* N as we can see in Figure 1.7 from [Gir2014].

We shall prove this result through the

Theorem 1.2.1 (Markov inequality)**.** *Let X be a random variable that is non-negative and integrable ($\mathbb{E}X$ exists and is well-defined). Then, the following upper bound on its tail holds:*

$$\mathbb{P}\{X \geq a\} \leq \frac{\mathbb{E}X}{a}, \quad \forall a > 0.$$

Actually, from Markov inequality, we have

$$\begin{aligned}
\mathbb{P}\{X \in B_{N,\delta}\} &= \mathbb{P}\left\{\exp\left(-\frac{\|X\|_2^2}{2}\right) \geq \delta\right\} \\
&\leq \frac{1}{\delta}\,\mathbb{E}\left\{\exp\left(-\frac{\|X\|_2^2}{2}\right)\right\} \\
&= \frac{1}{\delta}\int_{x\in\mathbb{R}^N} (2\pi)^{-N/2}\, e^{-\|x\|_2^2}\, dx = \frac{1}{\delta}\, 2^{-N/2},
\end{aligned}$$

concluding the result. So, if we want to have $\mathbb{P}\{X \in B_{N,\delta}\} \geq 1/2$ for instance, we must choose $\delta \leq 2^{1-N/2}$, which is exponentially small.

- Noise Accumulation

  It is usual to elaborate random models as some parametrized function plus a random variable representing *noise* (measuring errors and model imprecision in general). Such random variable is named *additive noise*. Clearly, it is interesting for the random noise in the model to be "small" in some sense such as zero mean and small variance. The problem we intend to shed light on is that, in high dimensions, the accumulation in many different directions even of "small" noises can produce a large global noise.

  Assume that we intend to evaluate a function F at some point $\theta_1 \in \mathbb{R}$. However, we have only access to a noisy observation of θ_1, denoted by $X_1 = \theta_1 + \varepsilon_1$, with $\mathbb{E}\{\varepsilon_1\} = 0$ and $\mathrm{Var}\{\varepsilon_1\} = \sigma^2$. If F is 1-Lipschitz, then the mean squared error of such approximation is given by

  $$\mathbb{E}\{\|F(X_1) - F(\theta_1)\|_2^2\} \leq \mathbb{E}\{|\varepsilon_1|^2\} = \sigma^2.$$

  Consequently, a small variance yields a small mean squared error.

  Regarding the high-dimensional setting, assume now that we intend to evaluate $F(\theta_1, \ldots, \theta_N)$ from noisy observations $X_j = \theta_j + \varepsilon_j$ of $\theta_j, j \in [N]$. Also, assume that the

noise variables are all centered and have variance σ^2. Again, endowed with the 1-Lipschitz regularity condition, we have

$$\begin{aligned}\mathbb{E}\{\|F(X_1,\dots,X_N)-F(\theta_1,\dots,\theta_N)\|_2^2\} &\leq \mathbb{E}\{\|(\varepsilon_1,\dots,\varepsilon_N)\|_2^2\}\\ &= \sum_{j=1}^{N}\mathbb{E}\{\varepsilon_j^2\} = N\sigma^2.\end{aligned}$$

Notice that, for a large $N \in \mathbb{N}$, the upper bound above gives no guarantee of small error even for a small variance. Furthermore, if F satisfies $\|F(x+h)-F(x)\|_2 \geq C\|h\|_2$ for some $C>0$, then the mean squared error will have $C\,N\sigma^2$ as a lower bound assuring that it will not necessarily be small for larger dimensions. A central example of such situation arises in the *linear regression model with high dimensional covariates*, that we exhibit next.

- High dimensional linear regression

  Assume that we have M pairs of observations $\{(Y_i, x^{(i)})\}_{i\in[M]}$, with $Y_i = \langle x^{(i)} : \beta^*\rangle + \varepsilon_i$, being the *response* with *additive noise* to the *covariates vector* $x^{(i)} \in \mathbb{R}^N$. Moreover, we consider $\varepsilon_1,\dots,\varepsilon_M$ to be i.i.d centered, with variance σ^2. Our goal is to estimate $\beta^* \in \mathbb{R}^N$.

  Writing

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_M \end{bmatrix} \in \mathbb{R}^M, \quad \boldsymbol{X} = \begin{bmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(M)})^T \end{bmatrix} \in \mathbb{R}^{M\times N} \quad \text{and} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_M \end{bmatrix} \in \mathbb{R}^M,$$

  we have $Y = \boldsymbol{X}\beta^* + \varepsilon$. Such random model is a particular case of the *classical regression* already mentioned, called *linear regression*.

  A classical estimator of β^* is the least-square estimator

$$\hat{\beta} \in \underset{\beta\in\mathbb{R}^N}{\arg\min}\|Y - \boldsymbol{X}\beta\|_2^2,$$

  which is uniquely defined when the rank of $\boldsymbol{X}$ is N. We shall focus on this case, that yields $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$ as the unique solution. Now, we will prove that

$$\mathbb{E}\{\|\hat{\beta}-\beta^*\|_2^2\} = C\sigma^2,$$

  with $C \in \mathbb{R}$ being a constant that equals N for the particular case when $\boldsymbol{X}$ is orthogonal.

  Indeed,

$$\begin{aligned}\hat{\beta} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y\\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\beta^*+\varepsilon)\\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})\beta^* + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\varepsilon\\ &= \beta^* + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\varepsilon.\end{aligned}$$

Consequently,

$$\mathbb{E}\{\|\hat{\beta} - \beta^*\|_2^2\} = \mathbb{E}\{\|(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\varepsilon\|_2^2\}.$$

Finally, using that,

$$\mathbb{E}\{\|\boldsymbol{A}\varepsilon\|_2^2\} = \sigma^2 \operatorname{tr}\{\boldsymbol{A}^T\boldsymbol{A}\}, \quad \forall \boldsymbol{A} \in \mathbb{R}^{N\times M},$$

we conclude

$$\begin{aligned}\mathbb{E}\{\|\hat{\beta} - \beta^*\|_2^2\} &= \sigma^2 \operatorname{tr}\left\{[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\right\} \\ &= \sigma^2 \operatorname{tr}\left\{\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right\} \\ &= \sigma^2 \operatorname{tr}\left\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right\}.\end{aligned}$$

Therefore, the higher is the dimension of the covariates, the larger is the mean squared error.

- An accumulation of rare events may not be rare

  Consider we have access to an observation Z_1 of a quantity θ_1 through the following model

  $$Z_1 = \theta_1 + \varepsilon_1,$$

  with $\varepsilon_1 \sim \mathcal{N}(0,1)$ being a Gaussian noise. Also, consider the following result:

  Lemma 1.2.1 (Tails of the Gaussian distribution)**.** *Let Z be a standard Gaussian random variable. For any $x \geq 0$, we have*

  $$\mathbb{P}\{|Z| \geq x\} \leq e^{-x^2/2}.$$

  Proof. Define a function $\phi : [0,\infty) \to \mathbb{R}$ such that

  $$\phi(x) = e^{-x^2/2} - \mathbb{P}\{|Z| \geq x\}.$$

  At first, note that $\phi(0) = 0$. Furthermore,

  $$\phi(x) = e^{-x^2/2} - \sqrt{\frac{2}{\pi}} \int_x^\infty e^{-t^2/2}\, dt$$

  from what

  $$\phi'(x) = \left(\sqrt{2/\pi} - x\right) e^{-x^2/2},$$

  which is a non-negative value for $x \leq \sqrt{2/\pi}$. Consequently, $\phi(x)$ is non-negative on $[0, \sqrt{2/\pi}]$.

  Finally, for $x \geq \sqrt{2/\pi}$,

  $$\sqrt{\frac{2}{\pi}} \int_x^\infty e^{-t^2/2}\, dt \leq \int_x^\infty t e^{-t^2/2}\, dt = e^{-x^2/2},$$

  concluding the proof that ϕ is non-negative on $\mathbb{R}_+$. □

Now, from Lemma 1.2.1, we have

$$\mathbb{P}\{|\varepsilon_1| \leq x\} = 1 - \mathbb{P}\{|\varepsilon_1| \geq x\} \geq 1 - e^{-x^2/2}, \quad x > 0.$$

It follows that, with probability at least $1-\alpha$, the noise ε_1 has an absolute value smaller than $\sqrt{2\log(1/\alpha)}$.

Regarding the high-dimensional setting, consider we are observing N quantities $\theta_1, \ldots, \theta_N$ blurred by $\varepsilon_1, \ldots, \varepsilon_N \sim \mathcal{N}(0,1)$ i.i.d. We have that

$$\mathbb{P}\left\{\max_{j\in[N]} |\varepsilon_j| \geq x\right\} = 1 - \prod_{j\in[N]} (1 - \mathbb{P}\{|\varepsilon_j| \geq x\}) = 1 - (1 - \mathbb{P}\{|\varepsilon_1| \geq x\})^N.$$

We now recall that

$$\lim_{x\to 0} \frac{(1+x)^r}{1+rx} = 1$$

for any $r \in \mathbb{R}$ fixed. It follows that $(1+x)^r$ approaches $1+rx$ for arbitrarily small values of x and, since Lemma 1.2.1 yields $\mathbb{P}\{|\varepsilon_1| \geq x\} \to 0$ as $x \to 0$, we have

$$(1 - \mathbb{P}\{|\varepsilon_1| \geq x\})^N \sim 1 - N\,\mathbb{P}\{|\varepsilon_1| \geq x\},$$

concluding then

$$\mathbb{P}\left\{\max_{j\in[N]} |\varepsilon_j| \geq x\right\} \sim N\,\mathbb{P}\{|\varepsilon_1| \geq x\}. \tag{1.2.2}$$

Therefore,

$$\mathbb{P}\left\{\max_{j\in[N]} |\varepsilon_j| \leq x\right\} \sim 1 - N\,\mathbb{P}\{|\varepsilon_1| \geq x\} \geq 1 - Ne^{-x^2/2}.$$

Thus, if we want to bound simultaneously the absolute values $|\varepsilon_1|, \ldots, |\varepsilon_N|$ with probability $1-\alpha$, we can only guarantee that $\max_{j\in[N]} |\varepsilon_j|$ is smaller than $\sqrt{2\log(N/\alpha)}$. This extra $\log(N)$ factor can be a serious issue for large N as illustrated in the example *"False Discoveries"* in Chapter 1 of [Gir2014].

1.3 Dimensionality reduction

Before presenting dimensionality reduction as a paradigm to overcome the curse of dimensionality, we briefly discuss some points about dimension itself. However, since this is not the main subject of this text, we direct the interested reader to [LV2007] for more details.

1.3.1 What is dimension?

The main goal of statistical inference is to collect data and develop models that may be used to make claims about a population of interest. Mathematically, a dataset consisting of M observations, each characterized by N covariates, will be represented by an $M \times N$ matrix

$\boldsymbol{X}$. That is, $\boldsymbol{X}$ is viewed as a collection of M points in $\mathbb{R}^N$ and we call N the *dimension* of this data, saying that it is N-dimensional.

As said earlier, we can often reduce the dimension of the data from its original recording, called *extrinsic dimensionality* to a smaller one called *intrinsic dimensionality*.

Such dimensionality reduction can be made since the data is not completely random but reflects underlying structures that generates them. For instance, pixel intensity signals contain information about the image they compose; marketing data consider social structures in a population, and so on. Indeed, in many cases, the data have an intrinsic low complexity and, when the low complexity structures are known, our problem might become solvable by classical statistical methods. The major issue in Statistics in high dimensions is that these structures are usually unknown.

Therefore, our main task is to identify these structures, at least approximately. There are a number of statistical approaches to dimensionality reduction like the *Principal Component Analysis* (PCA) and the *Multidimensional Scaling* (MDS), and despite being not the main subject of the text, both methods will be briefly discussed.

1.3.2 Principal Component Analysis - PCA

Principal component analysis (PCA) (or Karhunen-Loève transform in Signal Processing) is a widely used technique for dimension reduction of a dataset consisting of a large number of interrelated variables, while retaining as much as possible of the "information" present in the dataset. This technique was invented in 1901 by Karl Pearson [KP1901] and independently derived in 1933 by Harold Hotelling [Htl1933]. There are two commonly used definitions of PCA that generate the same algorithm.

The first stands for an orthogonal projection of a centered dataset in a lower dimensional linear space such that the variances of the projected data is maximized among all choices of subspaces of this fixed dimension [Htl1933]. This subspace is known as *principal subspace*, and this is made as follows: we determine the direction whose variance of the projected data is the largest, and such direction is called the *first principal component*; each further component is chosen in order to have the largest projected variance constrained that it is orthogonal to the preceding ones.

Alternatively, it can be defined as the linear projection of such dataset that minimizes the average projection cost [KP1901]. More precisely, let $\mathcal{P}_{N,m}$ represent the set of the rank m projections $P : \mathbb{R}^N \to \mathbb{R}^N$. This second definition yields the following optimization problem:

$$P^* = \underset{P \in \mathcal{P}_{N,m}}{\arg\min} \|P\boldsymbol{X} - \boldsymbol{X}\|_{2\to 2}^2.$$

This is illustrated by the figures in [KP1901]).

- Motivation

  As said before, several problems of decision-making are solved by means of the analysis of a dataset represented by a matrix $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ whose each of its M rows, $x_1, \ldots, x_M$, represents an N-dimensional data vector, i.e., an N-covariate observation. A possible scheme towards such problems consists in describing the data cloud by the variance of its components and the correlations between them.

Consequently, this setup requires the analysis of M variances and $M(M-1)/2$ correlations of N dimensional vectors, i.e, MN scalar variances and $N^2\,M(M-1)/2$ scalar correlations. Not only these operations will be unbearable for high-dimensional problems but also these descriptive statistics might give little information about the dataset itself. As an example of such behavior, we have the *Ascombe Datasets* [Asc1973], that consists of four quite different datasets with essentially identical descriptive statistics. Such problems motivate the usage of techniques to reduce dimensionality from N to $m \ll N$.

- Intuition

  In order to project the data while maximizing the variance on the projected space, we make a change of coordinates. The first element of the new basis will have the direction of the greatest data variance and will be known as *first principal component*. The further principal components will be defined analogously with a decreasing order of variance for the projected data.

- Mathematical model

  Let $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ be the data matrix with N-dimensional rows $x_1, \ldots, x_M$. We intend to project the data onto a space with dimensionality $m \ll N$ while maximizing the variance of the projected data. For the moment, we shall assume that dimension m is given.

- Deriving the principal components from maximum variance formulation

  We begin obtaining the first principal component, i.e., the direction of a unit vector u_1 that maximizes the variance of the projected data. The mean of the projected data in this direction is given by $u_1^T\bar{x}$, with $\bar{x}$ being the sample mean

$$\bar{x} := \frac{1}{M}\sum_{k=1}^{M} x_k.$$

  Also, the variance of the projected data in this direction is

$$\frac{1}{M}\sum_{k=1}^{M}(u_1^T x_k - u_1^T\bar{x})^2 = u_1^T\boldsymbol{S}u_1, \tag{1.3.1}$$

  with $\boldsymbol{S}$ being the data covariance matrix defined by

$$\boldsymbol{S} := \frac{1}{M}\sum_{k=1}^{M}(x_k-\bar{x})(x_k-\bar{x})^T.$$

  Now, the first principal component will result from the following constrained optimization problem.

$$\max_{u_1\in\mathbb{R}^N}\left\{u_1^T\boldsymbol{S}u_1\right\} \qquad \text{s.t.} \qquad \|u_1\|_2 = 1.$$

Aiming to solve it, we shall use the Lagrange multipliers method. Namely, we will maximize

$$u_1^T \boldsymbol{S} u_1 + \lambda_1 (1 - u_1^T u_1),$$

with $\lambda_1 \in \mathbb{R}$.

At first, we make

$$\frac{\partial}{\partial u_1} \left\{ u_1^T \boldsymbol{S} u_1 + \lambda_1 (1 - u_1^T u_1) \right\} = 0,$$

from which we obtain

$$\boldsymbol{S} u_1 = \lambda_1 u_1,$$

which says that u_1 must be an eigenvector of $\boldsymbol{S}$. Finally, we multiply both sides of the equation above by u_1^T, which yields

$$u_1^T \boldsymbol{S} u_1 = \lambda_1,$$

concluding λ_1 will be the largest eigenvalue of $\boldsymbol{S}$ associated to u_1 and that it will also be the largest value for the variance of projected data.

The further principal components will be derived by applying the reasoning above recursively. That is, we maximize the variance regarding the directions that are orthogonal to those already considered. Indeed, we roughly sketched the proof of the following result:

Theorem 1.3.1 (Maximum variance formulation of PCA)**.** *The m principal components of a dataset $x_1, \ldots, x_M$ with $M \geq m$ with correlation matrix $\boldsymbol{S}$ will be given by the orthonormal subset of eigenvectors of $\boldsymbol{S}$, $u_1, \ldots, u_m$, that are respectively associated to eigenvalues such that*

$$\lambda_1 \geq \ldots \geq \lambda_m.$$

For completeness, we also state without proof the minimum-error formulation of the PCA:

Theorem 1.3.2 (Minimum-error formulation of PCA)**.** *Let*

$$J = \sum_{i=m+1}^{N} u_i^T \boldsymbol{S} u_i$$

be the mean squared error between the dataset and its projection onto the subspace generated by m elements, $u_1, \ldots, u_m$, from an orthonormal basis $\{u_1, \ldots, u_N\} \subset \mathbb{R}^N$. Such error is minimized when $u_1, \ldots, u_m$ are eigenvectors of the dataset correlation matrix, $\boldsymbol{S}$, respectively associated to its m highest eigenvalues.

1.3.3 Multidimensional Scaling - MDS

Multidimensional Scaling (MDS) is a technique of *data visualization* and *non-linear dimensionality reduction*. Namely, given the pairwise distances among a set of $M \in \mathbb{N}$ objects on which a distance function is defined, the MDS intends to translate this information into a configuration of M points in a m-dimensional Euclidean space, for a prefixed m, that preserves as much as possible these pairwise distances. In particular, for $m \in \{1, 2, 3\}$, the resulting points can be visualized on a scatter plot.

More precisely, let $d_{i,j}$, for $i, j \in [M]$, be the pairwise distances among the set of $M \in \mathbb{N}$ objects. These distances are the entries of a matrix $\boldsymbol{D} = [d_{i,j}]$ called *dissimilarity matrix*. Given the matrix $\boldsymbol{D}$ and a prefixed m, the MDS intends to determine a set of points $\{x_1, \ldots, x_M\} \in \mathbb{R}^m$ such that

$$\|x_i - x_j\| \approx d_{i,j}, \quad \text{with} \quad i, j \in [M].$$

In classical MDS, the norm $\|\cdot\|$ is the Euclidean distance; in different formulations, it may be an arbitrary metric. Note that the solution for MDS with the Euclidean distance is not unique since rigid transformations preserve distances.

Usually, MDS is formulated as an optimization problem and $\{x_1, \ldots, x_M\}$ is found as a minimizer of some cost function. For example, the *Classical multidimensional scaling* (also known as *Principal Coordinates Analysis* – PCoA) uses a cost function called *Strain*. Namely,

$$\text{Strain}_D(x_1, \ldots, x_N) := \sqrt{\frac{\sum_{i,j}(b_{i,j} - \langle x_i : x_j \rangle)^2}{\sum_{i,j} b_{i,j}^2}},$$

with $\boldsymbol{B}$ being the result of applying the *double centering* to $\boldsymbol{D}^{(2)} = [d_{i,j}^2]$.

1.3.4 Motivation of the JL-Lemma

In modern algorithm design, data is often high dimensional. Thus, one seeks to first preprocess the data via some dimensionality reduction scheme that preserves geometry in such a way that is acceptable for particular applications.

So consider a set $X = \{x_1, \ldots, x_M\}$ in a high dimensional vector space representing the data (say ℓ_2^N, usually with $N \gg M$). If we intend to derive a result or implement an algorithm involving the mutual distances among these vectors, we may have problems by the *computational storage and transmission burdens and by the curse of dimensionality.*

We can, of course, project these vectors isometrically in a M-dimensional vector space. Namely, it suffices to take $span\{X\}$. It, therefore, arises the following question:

"What if we relax the isometry constraint in our claim?
Could we obtain a dimension reduction to ℓ_2^m with m smaller than M?"

The idea in such investigation is to transform a high dimensional problem into a lower dimensional one such that the optimal solution to the lower dimensional problem can be lifted to a nearly optimal solution to the original problem. For further applications see [Ind2001, Mat2008, Vem2004].

The following definition will be essential along this text.

Definition 1.3.1 (Quasi-isometry)**.** *Let $f : X \subset \ell_2^N \to \ell_2^m$ and $\varepsilon > 0$. We say f is a quasi-isometry if:*

$$\forall u, v \in X, \quad (1-\varepsilon)\|u-v\|_2^2 \leq \|f(u)-f(v)\|_2^2 \leq (1+\varepsilon)\|u-v\|_2^2.$$

Remark. *In the literature, the equation above is also read by saying that "X is embedded in ℓ_2^m with a distortion of at most ε".*

Provided with this definition, we ask a further question:

"How small can we make m by accepting a bigger ε (i.e., by worsening the quasi-isometry)?"

The goal of the present work is to exhibit the answers given so far to these questionings and how they have been improved. More precisely, we shall discuss the Johnson-Lindenstrauss Lemma, that claims that if we relax the isometry constraint to a quasi-isometry, the m as above will not only be smaller than M, but it will have its logarithmic order.

Chapter 2

The outset of JL-Lemma

2.1 Introduction

The goals of this Chapter are: present the original statement of the JL-Lemma, introduce some of key concepts and reproduce its first proof due to W. B. Johnson and J. Lindenstrauss in their work [JL1984]. Finally, we discuss the limitations of this pioneering version to applications, that motivates further developments of the Lemma that will be discussed in Chapter 3. A non-comprehensive discussion about technical points related to the random projection argument used within the proof are presented in Appendix A.

2.2 Origin

Since the JL-Lemma has been discussed in a myriad of fields along its history, this text is also an attempt to make a linkage between these approaches. Indeed, we start this section by presenting the JL-Lemma as it was stated in [JL1984]. In that paper, the result was presented merely as a tool to prove a result about extension of Lipschitz maps into Hilbert spaces. Consequently, the authors did not present it in a manner that is suitable for applications. Finally, we briefly motivate the usage of the JL-Lemma in [JL1984] and relate it to its modern intuition.

2.2.1 Main result

The JL-Lemma was first proved in a 1984 paper [JL1984] by William B. Johnson and Joram Lindenstrauss. At that point, the Lemma was viewed as a tool to prove a result about Lipschitz' extensions of functions into a Hilbert space.

Namely, let (X, d_1) and (Y, d_2) be metric spaces. We say that a function $f : X \to Y$ is Lipschitz when there is a $K > 0$ such that

$$d_2\{f(y), f(x)\} \leq K d_1(y, x), \quad \forall x \neq y \in X.$$

Moreover, the least such K satisfying the equation above is called Lipschitz constant of f

and it is denoted by

$$\begin{aligned}\|f\|_{lip} = &\inf\left\{K>0 : d_2\{f(y),f(x)\} \leq K d_1(y,x),\ \forall x \neq y \in X\right\} \\ = &\sup_{x\neq y\in X} \frac{d_2\{f(y),f(x)\}}{d_1(y,x)}.\end{aligned}$$

Remark (Is the Lipschitz constant a norm?)**.** *The answer is clearly no, since the metrics d_1 and d_2 not necessarily satisfy $\|\lambda f\|_{lip} = |\,\lambda\,|\,\|f\|_{lip}$ for any real λ. On the other hand, even for the simplest case, with d_1 and d_2 being the Euclidean norm, the answer is still no, since the Lipschitz norm of any constant function is zero even when it is not the null function.*

For instance, any function over a finite domain is Lipschitz. More precisely, let (X,d) be a metric space and $A = \{x_1,\ldots,x_N\}$ a finite subset of X. Any function

$$f : A \to \ell_2^N$$

is Lipschitz with constant

$$\|f\|_{lip} = \max_{x\neq y\in A} \frac{\|f(y)-f(x)\|_2}{d(y,x)}.$$

The Lipschitz extension problem consists in determining the smallest $L = L(X,N)$ such that, for any such f, there exists a Lipschitz extension $\tilde{f} : X \to \ell_2^N$ of f satisfying:

$$\|\tilde{f}\|_{lip} \leq L\|f\|_{lip}.$$

Intuitively, we intend to determine a Lipschitz map whose "distance distortions" in all X are as similar as possible to the ones made by f in the set $\{x_1,\ldots,x_N\}$. The main result of Johnson and Lindenstrauss' work is that, for any metric space X,

$$L \leq C\sqrt{\log N},$$

for some constant $C \geq 0$ and the JL-Lemma is a tool to achieving this result.

2.2.2 Role of the JL-Lemma

The JL-Lemma deals with a particular case of the main result of [JL1984] for which $X = \ell_2^N$ and $f(\{x_1,\ldots,x_N\})$ lies in an m-dimensional subspace of ℓ_2^N, which we will henceforth denote as ℓ_2^m. In this case, not only the extension $\tilde{f} : \ell_2^N \to \ell_2^m$ yields a distortion that is comparable to $\|f\|_{lip}$ but also $\tilde{f}$ will be a *quasi-isometry,* i.e., there will be K, $\tilde{K} > 0$ such that

$$\frac{1}{\tilde{K}}\|y-x\|_2 \leq \|\tilde{f}(y)-\tilde{f}(x)\|_2 \leq K\|y-x\|_2, \quad \forall x \neq y \in \ell_2^N. \tag{2.2.1}$$

As we have also mentioned, the least K that satisfies the right hand side of Equation 2.2.1 is the Lipschitz constant $\|\tilde{f}\|_{lip}$. In a analogous fashion, we shall make an abuse of notation (since $\tilde{f}$ is not necessarily bijective) and denote the least $\tilde{K}$ in the left hand side of Equation 2.2.1 as $\|\widetilde{f^{-1}}\|_{lip}$, obtaining the equation below:

$$\frac{1}{\|\widetilde{f^{-1}}\|_{lip}}\|y-x\|_2 \leq \|\tilde{f}(y)-\tilde{f}(x)\|_2 \leq \|\tilde{f}\|_{lip}\|y-x\|_2, \quad \forall x \neq y \in \ell_2^N. \tag{2.2.2}$$

Moreover, the distortion of this quasi-isometry can be quantified as follows. Given $A = \{x_1, \ldots, x_N\} \subset (X, d)$ and $\varepsilon \in (0, 1)$, the JL-Lemma guarantees that any map $f : A \to \ell_2^m$ has a Lipschitz extension $\tilde{f} : \ell_2^N \to \ell_2^m$ such that

$$\|\tilde{f}\|_{lip}\|\widetilde{f^{-1}}\|_{lip} \leq \frac{1+\varepsilon}{1-\varepsilon}$$

since $m \geq \text{floor}(m_0 \log N)$, with $m_0 > 0$ depending only on ε.

The inequality above means that there exists a positive constant $M > 0$ such that

$$M(1-\varepsilon)\|y - x\|_2 \leq \|\tilde{f}(y) - \tilde{f}(x)\|_2 \leq (1+\varepsilon)M\|y - x\|_2, \quad \forall x \neq y \in \ell_2^N.$$

More precisely, it suffices to take

$$M = \frac{1}{(1-\varepsilon)\|\widetilde{f^{-1}}\|_{lip}}.$$

The map $\tilde{f} : \ell_2^N \to \ell_2^m$ satisfies a special kind of quasi-isometry with distortion $\varepsilon > 0$. These maps will be henceforth called *ε-isometries*.

We end the present subsection with the JL-Lemma as it was originally stated by Johnson and Lindenstrauss in [JL1984].

Theorem 2.2.1 (Lemma 1 of [JL1984])**.** *For each $0 < \varepsilon < 1$, there is an expression $m_0(\varepsilon) > 0$ depending only on ε so that, if $A \subset \ell_2^N$ and $\#A = N$ for some $N > 2$, then there is a mapping $f : A \to \ell_2^m$, with $m = \text{floor}\{m_0(\varepsilon) \log N\}$ which satisfies*

$$\|\tilde{f}\|_{lip}\|\widetilde{f^{-1}}\|_{lip} \leq \frac{1+\varepsilon}{1-\varepsilon}. \tag{2.2.3}$$

2.2.3 The turnaround

As discussed in the previous subsection, from the JL-Lemma, we conclude that any set $A = \{x_1, \ldots, x_N\} \subset \ell_2^N$, can be quasi-isometrically projected into a subspace whose dimension is of order $\log N$. Also, for a fixed N, the dimension of the subspace depends only of the quasi-isometry's distortion $\varepsilon \in (0, 1)$ and not of the set A being projected.

What Johnson and Lindenstrauss might not have noticed in 1984 was that this Lemma yields a tool to reduce dimensionality that is very suitable for applications involving high-dimensional data. Indeed, it allows us to project a dataset into a space whose dimension is of logarithmic order of its cardinality. Consequently, the JL-Lemma is nowadays being used in a wide range of applications such as: *genetic algorithms ([BV2005]), data streaming ([JW2013]), nearest neighbor search ([AlCh2006, DeBm2006, DeBm2007, IndMot1998]) and compressed sensing ([BDDW2006, Ward2008, Don2006, CT2005]).*

Such powerful result might suggest a fearful proof and an even worse construction. Both of these problems are brilliantly avoided due to a probabilistic argument used by Johnson and Lindenstrauss called *random projection.*

Briefly, the random projection argument goes as follows. Instead of defining a general map f and extending it to a ε-isometric projection, we make a guess of $\tilde{f}$ as a rank m linear

projection. As one might argue, there is no reason for such map to satisfy the JL-Lemma. Here is where the randomness comes in. We select at random an m-dimensional subspace of ℓ_2^N and take $\tilde{f}$ as the projection into such space. Through a concentration of measure result, Johnson and Lindenstrauss proved that for a $m \in \mathbb{N}$, with $m \geq \text{floor}\{m_0(\varepsilon) \log N\}$, this random $\tilde{f}$ will satisfy the Lemma with positive probability. Furthermore, as m gets larger, this probability approaches 1 exponentially fast.

Finally, we still need to formalize what does it mean to select a random subspace and to explain what is concentration of measure. The next sections of the present Chapter will be dedicated to formalize these concepts that will be necessary to prove the JL-Lemma.

2.3 Concentration of measure

Some finite dimensional objects such as convex sets may present an unexpected behavior when the dimension of its ambient space is large or tends to infinity, i.e., in a *asymptotic setting*. For instance, a key result towards the proof of the JL-Lemma is that, as the dimension N grows, the relative surface area of the N-sphere becomes exponentially concentrated around any of its equatorial strips. The study of the geometrical properties of sets in a high dimensionality setting motivates a field called *Asymptotic Geometric Analysis*. As illustrated by our main result, a key concept in this field is one of these unexpected properties called *concentration of measure*.

The phenomenon of concentration of measure started receiving attention around the end of the 60's and the beginning of the 70's and is a key concept in Asymptotic Geometric Analysis. It may be informally described by the tendency of functions depending on a sufficiently large number of variables, under very weak assumptions, to concentrate around its mean or median. The concentration of measure can be thought as a generalization of the *Law of Large Numbers* and will be the keystone to our random projection argument towards the JL-Lemma.

2.3.1 A brief on concentration of measure

We start the present subsection by defining a main concept to the concentration of measure theory.

Definition 2.3.1 (External set)**.** *Let (X, d) be a metric space endowed with a Borel measure μ, with $\mu(X) = 1$. For any Borel subset $A \subset X$, we define its ε-extension, ε-neighborhood, ε-external set or ε-blowup as*

$$A_\varepsilon := \{x \in X : d(x, A) < \varepsilon\}.$$

Intuitively, we say that (X, d) has a *concentration of measure* when there is a value (often $1/2$) such that, for any $A \subset X$, with $\mu(A) \geq 1/2$, its ε-blowup A_ε is measurable and concentrates most of the measure of X for any $\varepsilon > 0$.

In particular, to prove the JL-Lemma, we will need a concentration of measure result for the *Haar measure* μ_{N-1} on the sphere $(\mathbb{S}^{N-1}, \rho)$ endowed with its *geodesic distance* ρ (see Appendix B for more details). In this case, we shall prove that, any Borel set $A \subset \mathbb{S}^{N-1}$ with

$\mu_{N-1}(A) \geq 1/2$, satisfies

$$\mu_{N-1}(A_\varepsilon) \geq 1 - C\exp\{-cN\varepsilon^2\},$$

with C and c being positive universal constants, i.e., they are independent of the Borelian $A \subset \mathbb{S}^{N-1}$, of the perturbation $\varepsilon > 0$ and of the dimension $N > 2$. As we shall see next, this inequality follows from an important result called *isoperimetric inequality on the sphere*.

2.3.2 The isoperimetric inequality on the sphere

The origin of the *isoperimetric problem* and also of its name goes back to Antiquity, one of its first versions being the *Dido's Problem*, which intends to determine the shape of a figure with maximum area given its perimeter. Roughly speaking, the three-dimensional isoperimetric problem consists in determining the figure that minimizes the surface area given a fixed volume. For higher dimensions, stating the problem is a bit more complicated. Indeed, there are many definitions of surface area, each of them being more suited for a particular purpose. In this text, we will choose the one of *Minkowski content*. We shall not discuss on the good properties that motivate this choice and we direct the interested reader to [Fdr1996] or [MiApAv2015].

Provided with this choice, we may generalize the isoperimetric problem for a metric space (X, d) endowed with a measure μ. In particular, [MiApAv2015] argues that this generalization yields the following:

Definition 2.3.2 (Isoperimetric problem on the sphere)**.** *Let $(\mathbb{S}^{N-1}, \rho)$ be the sphere endowed with its geodesic metric and its unique rotation invariant measure μ_{N-1}. Now, fix $\alpha \in (0,1)$. Among all Borel sets that satisfy $\mu_{N-1}(A) \geq \alpha$, we intend to determine the ones for which $\mu_{N-1}(A_\varepsilon)$ is minimal for all $\varepsilon > 0$.*

This problem is solved by the *Lévy-Schmidt isoperimetric inequality on the sphere* below. Such inequality claims that, for a fixed volume, the spherical caps are the sets that solve the isoperimetric problem for the sphere.

Notation. *Throughout this Chapter, we shall represent a spherical cap centered in $z \in \mathbb{S}^{N-1}$ with angle $\phi \in [0, \pi]$ as $K_{N-1}(z, \phi)$. Moreover, when it is not necessary, its center may be omitted. For more details, see Appendix B.*

Theorem 2.3.1 (Lévy-Schmidt)**.** *Let $\alpha \in (0,1)$ and $K_{N-1}(\phi) \subset \mathbb{S}^{N-1}$ be the spherical cap with polar angle $\phi \in [0, \pi]$ whose measure is $\mu_{N-1}\{K_{N-1}(\phi)\} = \alpha$. Then, for every $A \subset \mathbb{S}^{N-1}$ with this same measure $\mu_{N-1}(A) = \alpha$ and every $\varepsilon > 0$, we have*

$$\mu_{N-1}(A_\varepsilon) \geq \mu_{N-1}\{[K_{N-1}(\phi)]_\varepsilon\} = \mu_{N-1}\{K_{N-1}(\phi + \varepsilon)\}.$$

We shall not prove this Theorem, but we direct the interested reader to [FLM1977].

This result is specially useful since we may bound the measure of any Borel set by a

spherical cap, whose measure, as shown in Appendix B, can be easily calculated:

$$\mu_{N-1}\{K_{N-1}(\phi)\} := \gamma_N \int_0^\phi \sin^{N-2}\theta\, d\theta$$

$$= \frac{1}{\sqrt{\pi}} \frac{\Gamma(N/2)}{\Gamma\{(N-1)/2\}} \int_0^\phi \sin^{N-2}\theta\, d\theta \quad (\phi \in [0, \pi/2]).$$

2.3.3 Concentration of measure on $\mathbb{S}^{N-1}$

The concentration of measure inequality on the sphere that will be used to prove the JL-Lemma follows directly from Equation 2.3.4. Indeed, what we will prove is that, for a high dimensional setting, the measure of any spherical cap $K_{N-1}(\phi)$ with $\phi > \pi/2$ will approach the total measure of $\mathbb{S}^{N-1}$ exponentially on the dimension N as ϕ increases.

In order to exhibit the concentration phenomenon for the sphere, we need a lower bound on the measure of a spherical caps that approaches 1 exponentially. The following results shall be useful.

Lemma 2.3.1. *For any $N > 2$, we have*

$$\gamma_N^{-1} := \int_0^\pi \sin^{N-2}\theta\, d\theta = \sqrt{\pi}\, \frac{\Gamma\{(N-1)/2\}}{\Gamma(N/2)} \geq \sqrt{\frac{2\pi}{N-1}}.$$

Proof. This result is a direct consequence of the following recurrence relation of the Gamma function:

$$\Gamma(z+1) = z\Gamma(z), \quad \mathcal{R}(z) > 0; \tag{2.3.1}$$

and of the fact that Γ, when restricted to positive real numbers, is a strictly logarithmically convex function, i.e., for any pair of positive numbers, x_1, x_2, we have

$$\Gamma\{(1-t)x_1 + tx_2\} \leq \Gamma^{1-t}(x_1)\Gamma^t(x_2), \quad \forall t \in [0,1],$$

with the inequality being strict for $t \in (0,1)$. In particular, for $t = 1/2$,

$$\Gamma\{\mathrm{AM}(x_1, x_2)\} \leq \sqrt{\Gamma(x_1)\Gamma(x_2)}, \tag{2.3.2}$$

with AM standing for *arithmetic mean*.

Now, since

$$\frac{N}{2} = \mathrm{AM}\left\{\frac{N-1}{2}, \frac{N+1}{2}\right\},$$

we conclude that

$$\Gamma\left(\frac{N}{2}\right) \leq \sqrt{\Gamma\left(\frac{N-1}{2}\right)\Gamma\left(\frac{N+1}{2}\right)}$$

$$= \sqrt{\Gamma\left(\frac{N-1}{2}\right)\frac{N-1}{2}\,\Gamma\left(\frac{N-1}{2}\right)} = \Gamma\left(\frac{N-1}{2}\right)\sqrt{\frac{N-1}{2}},$$

from Equations 2.3.2 and 2.3.1.

Finally, we apply the inequality above to conclude that

$$\frac{\Gamma\{(N-1)/2\}}{\Gamma(N/2)} \geq \sqrt{\frac{2}{N-1}},$$

from what

$$\gamma_N^{-1} \geq \sqrt{\frac{2\pi}{N-1}},$$

as desired. □

Lemma 2.3.2. *Let ε be a positive number. We have that*

$$\int_{\varepsilon}^{\pi/2} \cos^{N-2}\theta \, d\theta \leq \sqrt{\frac{\pi}{2(N-2)}} \exp\left\{-\frac{(N-2)\varepsilon^2}{2}\right\}.$$

Proof. Consider a function $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(\theta) = e^{\theta^2/2}\cos\theta.$$

This function decreases for $\theta \in [0, \pi/2]$. Furthermore, $f(0) = 1$ and $f(\theta) \geq 0$, for $\theta \in [0, \pi/2]$. It follows that

$$f(\theta) = \cos\theta \, e^{\theta^2/2} \leq 1, \quad \forall\theta \in [0, \pi/2];$$

and, equivalently, that

$$\cos\theta \leq e^{-\theta^2/2}, \quad \forall\theta \in [0, \pi/2].$$

We may apply this result to prove the present Lemma as follows. For any $N > 2$, we have that

$$\cos^{N-2}\theta \leq \exp\left\{-\frac{(N-2)\theta^2}{2}\right\}, \quad \forall\theta \in [0, \pi/2].$$

Thus, for $\varepsilon \in [0, \pi/2]$

$$\begin{aligned}
\int_{\varepsilon}^{\pi/2} \cos^{N-2} d\theta \leq & \int_{\varepsilon}^{\pi/2} \exp\left\{-\frac{(N-2)\theta^2}{2}\right\} d\theta \\
= & \int_{0}^{\pi/2-\varepsilon} \exp\left\{-\frac{(N-2)(\theta+\varepsilon)^2}{2}\right\} d\theta \\
\leq & \exp\left\{-\frac{(N-2)\varepsilon^2}{2}\right\} \int_{0}^{\pi/2-\varepsilon} \exp\left\{-\frac{(N-2)\theta^2}{2}\right\} d\theta, \qquad (2.3.3)
\end{aligned}$$

with the last inequality being justified by the fact that

$$\exp\big\{-(N-2)\theta\varepsilon\big\} \leq 1, \quad \forall\theta \geq 0.$$

Now, since the exponential map is positive over its entire domain, we have

$$\int_{0}^{\pi/2-\varepsilon} \exp\left\{-\frac{(N-2)\theta^2}{2}\right\} d\theta \leq \int_{0}^{\infty} \exp\left\{-\frac{(N-2)\theta^2}{2}\right\} d\theta = \sqrt{\frac{\pi}{2(N-2)}}.$$

Consequently, we conclude from Equation 2.3.3 that

$$\int_{\varepsilon}^{\pi/2} \cos^{N-2} \, d\theta \leq \sqrt{\frac{\pi}{2(N-2)}} \exp\left\{-\frac{(N-2)\varepsilon^2}{2}\right\}.$$

□

Theorem 2.3.2. *Let $(\mathbb{S}^{N-1}, \rho)$ be the sphere endowed with the geodesic metric ρ and the Haar measure μ_{N-1}. Given a natural $N > 2$ and $\varepsilon \in (0, \pi/2]$, we have*

$$\mu_{N-1}\left\{K_{N-1}\left(\frac{\pi}{2}+\varepsilon\right)\right\} \geq 1 - \frac{1}{2}\sqrt{\frac{N-1}{N-2}} \exp\{\varepsilon^2\} \exp\left\{-\frac{N\varepsilon^2}{2}\right\}.$$

In particular, when $\varepsilon \in (0, 1]$,

$$\mu_{N-1}\left\{K_{N-1}\left(\frac{\pi}{2}+\varepsilon\right)\right\} \geq 1 - 2\exp\left\{-\frac{N\varepsilon^2}{2}\right\}.$$

That is, the ε-neighborhood of any hemisphere concentrates all but a exponentially small measure of the sphere.

Proof. Indeed,

$$\begin{aligned}
\mu_{N-1}\left\{K_{N-1}\left(\frac{\pi}{2}+\varepsilon\right)\right\} &= \gamma_N \int_0^{\pi/2+\varepsilon} \sin^{N-2}\theta \, d\theta \\
&= \gamma_N \int_{-\pi/2}^{\varepsilon} \cos^{N-2}(\theta) \, d\theta \\
&= \gamma_N \left\{\int_{-\pi/2}^{\pi/2} \cos^{N-2}(\theta) \, d\theta - \int_{\varepsilon}^{\pi/2} \cos^{N-2}(\theta) \, d\theta\right\} \\
&= \gamma_N \left\{\int_0^{\pi} \sin^{N-2}(\theta) \, d\theta - \int_{\varepsilon}^{\pi/2} \cos^{N-2}(\theta) \, d\theta\right\} \\
&= 1 - \gamma_N \int_{\varepsilon}^{\pi/2} \cos^{N-2}(\theta) \, d\theta.
\end{aligned}$$

Finally, using the Lemmas 2.3.1 and 2.3.2, we conclude that

$$\begin{aligned}
\mu_{N-1}\left\{K_{N-1}\left(\frac{\pi}{2}+\varepsilon\right)\right\} &\geq 1 - \sqrt{\frac{N-1}{2\pi}}\sqrt{\frac{\pi}{2(N-2)}} \exp\left\{-\frac{(N-2)\varepsilon^2}{2}\right\} \\
&= 1 - \frac{1}{2}\sqrt{\frac{N-1}{N-2}} \exp\left\{-\frac{(N-2)\varepsilon^2}{2}\right\} \\
&= 1 - \frac{1}{2}\sqrt{\frac{N-1}{N-2}} \exp\{\varepsilon^2\} \exp\left\{-\frac{N\varepsilon^2}{2}\right\}.
\end{aligned}$$

Also, when $\varepsilon \leq 1$,

$$\sqrt{\frac{N-1}{N-2}} \exp\{\varepsilon^2\} \leq \sqrt{2}e \leq 4,$$

concluding the result. □

Finally, we can now directly exhibit the concentration of measure on the sphere.

Corollary 2.3.1 (Concentration of measure on the sphere)**.** *Any equatorial strip of* $\mathbb{S}^{N-1}$ *with convex opening angle of* $2\varepsilon \in (0, 2]$*, here denoted as* $Eq(2\varepsilon)$*, will concentrate all but an exponentially small measure of the sphere. More precisely,*

$$\mu_{N-1}\{Eq(2\varepsilon)\} \geq 1 - 4\exp\left\{-\frac{N\varepsilon^2}{2}\right\}.$$

Proof. Indeed, let $z \in \mathbb{S}^{N-1}$. The referred equatorial strip can be written as

$$Eq(2\varepsilon) = K_{N-1}\left(z, \frac{\pi}{2} + \varepsilon\right) \cap K_{N-1}\left(-z, \frac{\pi}{2} + \varepsilon\right).$$

From Theorem 2.3.2, both caps in the intersection above have a measure of at least

$$1 - 2\exp\left\{-\frac{N\varepsilon^2}{2}\right\}.$$

Consequently,

$$\begin{aligned}
\mu_{N-1}\left\{[Eq(2\varepsilon)]^c\right\} &= \mu_{N-1}\left\{\left[K_{N-1}\left(z, \frac{\pi}{2} + \varepsilon\right) \cap K_{N-1}\left(-z, \frac{\pi}{2} + \varepsilon\right)\right]^c\right\} \\
&= \mu_{N-1}\left\{\left[K_{N-1}\left(z, \frac{\pi}{2} + \varepsilon\right)\right]^c \cup \left[K_{N-1}\left(-z, \frac{\pi}{2} + \varepsilon\right)\right]^c\right\} \\
&\leq \mu_{N-1}\left\{\left[K_{N-1}\left(z, \frac{\pi}{2} + \varepsilon\right)\right]^c\right\} + \mu_{N-1}\left\{\left[K_{N-1}\left(-z, \frac{\pi}{2} + \varepsilon\right)\right]^c\right\} \\
&\leq 4\exp\left\{-\frac{N\varepsilon^2}{2}\right\}.
\end{aligned}$$

□

2.3.4 Concentration of Lipschitz maps on $\mathbb{S}^{N-1}$

An important consequence of the concentration of measure and the isoperimetric inequality on the sphere is that any Lipschitz map $f : \mathbb{S}^{N-1} \to \mathbb{R}$ (relative to the geodesic distance) concentrates most of its measure about a number $M_f \in \mathbb{R}$ called *median* or *Lévy's mean of* f. That is, for any $\varepsilon > 0$, we have the following concentration of measure inequality:

$$\mu_{N-1}\left\{x \in \mathbb{S}^{N-1} : |f(x) - M_f| \leq \varepsilon\|f\|_{lip}\right\} \geq 1 - C\exp\left\{-cN\varepsilon^2\right\},$$

with C and c being positive universal constants, i.e., they are independent of the perturbation $\varepsilon > 0$ and of the dimension $N > 2$. This result is essential to the proof of JL-Lemma and will be discussed in more details.

Definition 2.3.3 (Lévy's mean)**.** *Let* (X, d) *be a metric space endowed with a measure* μ*. A median, also called Lévy's mean of a measurable function* $f : X \to \mathbb{R}$ *is a real number* M_f *such that*

$$\mu\{x \in X : f(x) \geq M_f\} \geq 1/2 \quad \text{and} \quad \mu\{x \in X : f(x) \leq M_f\} \geq 1/2.$$

Now, to simplify the notation, let us name the following sets:

$$\begin{aligned} A_+^f &:= \{x \in X : f(x) \geq M_f\}, \\ A_-^f &:= \{x \in X : f(x) \leq M_f\}, \\ A^f &:= \{x \in X : f(x) = M_f\} = A_+^f \cap A_-^f. \end{aligned}$$

That being settled, when (X, d) is $\mathbb{S}^{N-1}$ with the geodesic metric, we have:

$$\mu_{N-1}\{A_+^f\},\ \mu_{N-1}\{A_-^f\} \geq 1/2 = \mu_{N-1}\{K_{N-1}(\pi/2)\},$$

with $K_{N-1}(\pi/2)$ being an hemisphere of $\mathbb{S}^{N-1}$. Moreover, for any $\varepsilon > 0$, the isoperimetric inequality for the sphere (Theorem 2.3.1) yields

$$\mu_{N-1}\{(A_+^f)_\varepsilon\},\ \mu_{N-1}\{(A_-^f)_\varepsilon\} \geq \mu_{N-1}\{K_{N-1}(\pi/2 + \varepsilon)\}. \tag{2.3.4}$$

Equation 2.3.4 above provides us with a lower bound on the measures of $(A_+^f)_\varepsilon$ and $(A_-^f)_\varepsilon$ that is dependent of the measure of a spherical cap. It is specially useful since we may use the concentration of measure from Theorem 2.3.2 to conclude that

$$\mu_{N-1}\{(A_+^f)_\varepsilon\},\ \mu_{N-1}\{(A_-^f)_\varepsilon\} \geq 1 - 2\exp\left\{-\frac{N\varepsilon^2}{2}\right\}, \quad \forall \varepsilon \in (0, 1). \tag{2.3.5}$$

The concentration of a Lipschitz map about its median follows from Equation 2.3.5 above. However, in order to prove it we will need the following result:

Lemma 2.3.3. *Let $f : (\mathbb{S}^{N-1}, \rho) \to \mathbb{R}$ be a Lipschitz function and M_f be the Lévy mean of f. We have that, for any $\varepsilon > 0$,*

$$(A_+^f)_\varepsilon \cap (A_-^f)_\varepsilon = (A^f)_\varepsilon.$$

We shall not prove this result and we direct the interested reader to [FLM1977].

Theorem 2.3.3. *Let $f : (\mathbb{S}^{N-1}, \rho) \to \mathbb{R}$ be a Lipschitz function and M_f be the Lévy mean of f. We have, for any $\varepsilon > 0$, that*

$$\mu_{N-1}\left\{(A^f)_\varepsilon\right\} = \mu_{N-1}\left\{|f(x) - M_f| \leq \varepsilon \|f\|_{lip}\right\} \geq 1 - 4\exp\left\{-\frac{N\varepsilon^2}{2}\right\}.$$

Proof. We have, from Lemma 2.3.3, that

$$(A^f)_\varepsilon = (A_+^f)_\varepsilon \cap (A_-^f)_\varepsilon, \quad \forall \varepsilon > 0.$$

Equivalently,

$$\{\mathbb{S}^{N-1} \setminus (A^f)_\varepsilon\} = \{\mathbb{S}^{N-1} \setminus (A_+^f)_\varepsilon\} \cup \{\mathbb{S}^{N-1} \setminus (A_-^f)_\varepsilon\}, \quad \forall \varepsilon > 0.$$

By applying the subadditive property of the measure μ_{N-1} to these sets, we have

$$\mu_{N-1}\{\mathbb{S}^{N-1} \setminus (A^f)_\varepsilon\} \leq \mu_{N-1}\{\mathbb{S}^{N-1} \setminus (A_+^f)_\varepsilon\} + \mu_{N-1}\{\mathbb{S}^{N-1} \setminus (A_-^f)_\varepsilon\}, \quad \forall \varepsilon > 0.$$

Thus,

$$1 - \mu_{N-1}\{(A^f)_\varepsilon\} \leq (1 - \mu_{N-1}\{(A^f_+)_\varepsilon\}) + (1 - \mu_{N-1}\{(A^f_-)_\varepsilon\}), \quad \forall \varepsilon > 0,$$

from what

$$\mu_{N-1}\{(A^f)_\varepsilon\} \geq \mu_{N-1}\{(A^f_+)_\varepsilon\} + \mu_{N-1}\{(A^f_-)_\varepsilon\} - 1, \quad \forall \varepsilon > 0. \tag{2.3.6}$$

At this point of the proof, we could handle the problem of bounding the measure of $(A^f)_\varepsilon$ from $\mu_{N-1}\{(A^f_+)_\varepsilon\}$, $\mu_{N-1}\{(A^f_-)_\varepsilon\} \geq 1/2$. However, this results in nothing more than $\mu_{N-1}\{(A^f)_\varepsilon\} \geq 0$. It becomes clear that we need a better approximation. Such goal will be achieved through the application of the isoperimetric inequality for the sphere (Theorem 2.3.1).

Namely, since $\mu_{N-1}(A^f_+)$, $\mu_{N-1}(A^f_-) \geq 1/2$, the isoperimetric inequality implies that the spherical cap, or more precisely, the hemisphere $K_{N-1}(\pi/2)$ is such that

$$\begin{aligned}
\mu_{N-1}\{(A^f_+)_\varepsilon\},\, \mu_{N-1}\{(A^f_-)_\varepsilon\} &\geq \mu_{N-1}\{[K_{N-1}(\pi/2)]_\varepsilon\} \\
&= \mu_{N-1}\{K_{N-1}(\pi/2 + \varepsilon)\} \\
&\geq \quad 1 - 2\exp\left\{-\frac{N\varepsilon^2}{2}\right\}, \quad \forall \varepsilon \in (0,1),
\end{aligned}$$

with the last inequality being due to the concentration of measure from Theorem 2.3.2, as done in Equation 2.3.5. Now, by substitution in Equation 2.3.6, we conclude the desired result. □

2.4 Random projections

2.4.1 Introduction

Random projection refers to the technique of projecting a set of points from a high-dimensional space, say ℓ_2^N, into a *randomly chosen* m-dimensional subspace, with $m < N$ as small as possible. However, we still need to define precisely what we mean by *"choosing a random subspace"* or *"choosing a random projection"*.

As we shall recall in this section, any rank m linear projection in the canonical basis has the matrix form $\boldsymbol{U}^t\boldsymbol{Q}_m\boldsymbol{U}$, with $\boldsymbol{U} \in O(N)$; and $\boldsymbol{Q} \in \mathcal{M}_{N\times N}$ being the matrix that vanishes all but the first m coordinates of a vector. Consequently, a first approach to define a random projections would be define a probability measure σ on $O(N)$ and selecting $\boldsymbol{U}^t\boldsymbol{Q}_m\boldsymbol{U}$ accordingly to it. Such measure need to satisfy some conditions to yield an intuitive and useful model of randomness.

Namely, we would like this probability measure σ to be "uniform" in the Borel subsets of $(O(N), \|\cdot\|_{\mathcal{F}})$, with the Frobenius norm, in the following sense. For any Borel subset $S \subset (O(N), \|\cdot\|_{\mathcal{F}})$, we have

$$\sigma(S) = \sigma(\boldsymbol{T}S), \quad \forall \boldsymbol{T} \in O(N).$$

That is, such measure is *invariant by left translations* of the group $O(N)$, being intuitively, uniform in $O(N)$. From now on, we endow $O(n)$ with the Haar measure.

In Appendix A, we present an important result called *Haar Theorem*. Such Theorem claims that there is a unique probability measure that satisfy the *lattus sensu uniformity* that we motivated previously, called *(normalized) Haar measure.* Moreover, that measure is Radon, countably summable and finite on compact subsets of $O(N)$.

2.4.2 Remarks from Linear Algebra

In the present subsection, we intend to formalize the characterization of any rank m linear projection as matrix product in the canonical basis. By definition, a rank m linear projection is an idempotent map $P \in \mathcal{L}(\ell_2^N, \ell_2^N)$ such that $P(\ell_2^N)$ is a m-dimensional vector subspace of ℓ_2^N. Consequently, ℓ_2^N may be represented by the following direct sum:

$$\ell_2^N = P(\ell_2^N) \oplus \mathrm{Ker}(P),$$

with $\mathrm{Ker}(P)$ being the kernel of P.

Furthermore, if P is an orthonormal projection and α and β are orthonormal ordered bases of $P(\ell_2^N)$ and $\mathrm{Ker}(P)$ respectively, we may define an orthonormal ordered basis $\alpha \cup \beta$ of ℓ_2^N such that the elements of $P(\ell_2^N)$ and $\mathrm{Ker}(P)$ have, respectively, its last $N-m$ and its first m coordinates equal to zero. Also, in this basis, P will be represented by the rank m matrix $\boldsymbol{Q}_m \in \mathcal{M}_{N\times N}$ that vanishes all but the first m coordinates of any $x \in \mathbb{R}^N$. We shall henceforth refer to such ordered basis as $\boldsymbol{\xi} = \{\xi_i\}_{i\in[N]}$ and to the canonical basis of $\mathbb{R}^N$ as $\boldsymbol{e} = \{e_i\}_{i\in[N]}$.

Now, let $\boldsymbol{U} \in \mathcal{M}_{N\times N}$ be the matrix that changes the basis $\boldsymbol{e}$ into the basis $\boldsymbol{\xi}$, that is,

$$\xi_i = \boldsymbol{U} e_i, \quad \forall i \in [N].$$

Note that $\boldsymbol{U}$ is an orthogonal matrix since it transforms an orthonormal basis to another one. Therefore, we conclude that the linear projection P is represented by the rank m matrix $\boldsymbol{U}^T\boldsymbol{Q}_m\boldsymbol{U} \in \mathcal{M}_{N\times N}$ in the canonical basis. Conversely, this matrix product is clearly a rank m projection.

2.4.3 Uniformly distributed random projection

Finally, from the discussion made so far in this section, $P : \mathbb{R}^N \to \mathbb{R}^N$ is a rank m projection if, and only if, it has a matrix representation of the form $\boldsymbol{U}^t\boldsymbol{Q}_m\boldsymbol{U}$ in the canonical basis. Consequently, we have the following

Definition 2.4.1 (Uniformly distributed rank m random projection)**.** *We say that a random rank m projection $P : \mathbb{R}^N \to \mathbb{R}^N$ is uniformly distributed if its matrix form representation in the canonical basis is $\boldsymbol{U}^t\boldsymbol{Q}_m\boldsymbol{U}$ with $\boldsymbol{U}$ being uniformly distributed in $O(N)$.*

Theorem 2.4.1. *The random variable $\|Px\|_2$ has the same distribution in both following cases:*

1. *$P : \mathbb{R}^N \to \mathbb{R}^N$ is a fixed rank m projection, and x is uniformly distributed on $\mathbb{S}^{N-1}$;*

2. *$x \in \mathbb{S}^{N-1}$ is fixed, and the random rank m projection $P : \mathbb{R}^N \to \mathbb{R}^N$ is uniformly distributed.*

That is, for any Borel set $S \subset \mathbb{R}$,

$$\mu_{N-1}\{x \in \mathbb{S}^{N-1} : \|\boldsymbol{U}^t\boldsymbol{Q}_m\boldsymbol{U}x\|_2 \in S\} = \sigma\{\boldsymbol{U} \in O(N) : \|\boldsymbol{U}^t\boldsymbol{Q}_m\boldsymbol{U}x\|_2 \in S\}.$$

Proof. Indeed, let $S \subset \mathbb{R}$ be a Borel set. Also, the continuous map

$$f : \mathbb{S}^{N-1} \to \mathbb{R}, \text{ with } f(x) = \|\boldsymbol{Q}_m x\|_2$$

will be useful. Also, we have that

$$\|\boldsymbol{U}_0^t\boldsymbol{Q}_m\boldsymbol{U}_0x\|_2 = \|\boldsymbol{Q}_m\boldsymbol{U}_0x\|_2$$

since orthogonal maps do not change the norm of a vector. Consequently,

$$\mu_{N-1}\{x \in \mathbb{S}^{N-1} : \|\boldsymbol{U}_0^t\boldsymbol{Q}_m\boldsymbol{U}_0x\|_2 \in S\} = \mu_{N-1}\{x \in \mathbb{S}^{N-1} : \|\boldsymbol{Q}_m\boldsymbol{U}_0x\|_2 \in S\}.$$

Moreover, the rotation invariance of μ_{N-1} yields

$$\begin{aligned}\mu_{N-1}\{x \in \mathbb{S}^{N-1} : \|\boldsymbol{Q}_m\boldsymbol{U}_0x\|_2 \in S\} &= \mu_{N-1}\{\boldsymbol{U}_0x \in \mathbb{S}^{N-1} : \|\boldsymbol{Q}_m(\boldsymbol{U}_0x)\|_2 \in S\} \\ &= \mu_{N-1}\{y \in \mathbb{S}^{N-1} : \|\boldsymbol{Q}_m y\|_2 \in S\} \\ &= \mu_{N-1}\{y \in \mathbb{S}^{N-1} : f(y) \in S\} \\ &= \mu_{N-1}\{f^{-1}(S)\}. \end{aligned} \tag{2.4.1}$$

In Appendix B, we proved that, for any fixed $x_0 \in \mathbb{S}^{N-1}$, μ_{N-1} is the push-forward of the (normalized) Haar measure σ in $O(N)$ through the map

$$g : O(N) \to \mathbb{S}^{N-1}, \text{ with } g(\boldsymbol{U}) = \boldsymbol{U}x_0.$$

That is, for any Borel set $A \subset \mathbb{S}^{N-1}$,

$$\mu_{N-1}(A) = \sigma\{g^{-1}(A)\}.$$

Applying this result to Equation 2.4.1, we get

$$\mu_{N-1}\{f^{-1}(S)\} = \sigma\{g^{-1}[f^{-1}(S)]\}.$$

Finally,

$$\begin{aligned}\sigma\{g^{-1}[f^{-1}(S)]\} &= \sigma\{\boldsymbol{U} \in O(N) : g(\boldsymbol{U}) \in f^{-1}(S)\} \\ &= \sigma\{\boldsymbol{U} \in O(N) : f(\boldsymbol{U}x_0) \in S\} \\ &= \sigma\{\boldsymbol{U} \in O(N) : \|\boldsymbol{Q}_m\boldsymbol{U}x_0\|_2 \in S\} \\ &= \sigma\{\boldsymbol{U} \in O(N) : \|\boldsymbol{U}^t\boldsymbol{Q}_m\boldsymbol{U}x_0\|_2 \in S\},\end{aligned}$$

concluding the proof. □

2.5 The original proof of the JL-Lemma

2.5.1 Introduction

We start by recalling the main result that we wish to prove. Consider a N-point set $A = \{x_1, \dots, x_N\} \subset \ell_2^N$ and a pre-fixed quasi-isometry distortion $\varepsilon \in (0, 1)$. There is a positive number $m_0(\varepsilon)$ such that any $m \geq \text{floor}\{m_0(\varepsilon) \log N\}$ yields a rank m random linear projection $f : \ell_2^N \to \ell_2^N$ satisfying

$$M(1-\varepsilon)\|v-u\|_2 \leq \|f(v)-f(u)\|_2 \leq M(1+\varepsilon)\|v-u\|_2, \quad \forall u \neq v \in A,$$

with positive probability, and with $M \in \mathbb{R}$ being a positive constant. This result will be a consequence of the concentration of a Lipschitz function on the sphere about its Lévy mean, as discussed next.

More precisely, we introduce the following map

$$\begin{aligned} F : (O(N), \sigma) &\to \mathcal{L}(\ell_2^N, \ell_2^N) \\ \boldsymbol{U} &\mapsto \boldsymbol{U}^T \boldsymbol{Q}_m \boldsymbol{U}. \end{aligned}$$

As previously discussed, the random matrix $F(\boldsymbol{U})$ determines the notion of *rank m random projection*. Also, we claim that the map $f = F(\boldsymbol{U})$ will satisfy the JL-Lemma with positive probability.

2.5.2 The induced concentration of measure on $O(N)$

In the present subsection, we intend to show that the concentration of Lipschitz maps on $\mathbb{S}^{N-1}$ induces a concentration of the Haar measure σ on $O(N)$. Indeed, since our choice of f – and consequently its Lipschitz extension, $\tilde{f}$ – is linear, we may represent its quasi-isometric property as a norm distortion of a new set $B \subset \ell_2^N$ of unit vectors such that

$$B := \left\{ \frac{u-v}{\|u-v\|_2} : u, v \in A \text{ and } u \neq v \right\} \subset \mathbb{S}^{N-1}.$$

In this setting, our goal is to select $\boldsymbol{U} \in O(N)$ such that

$$M(1-\varepsilon) \leq \|\boldsymbol{U}^t \boldsymbol{Q}_m \boldsymbol{U} z\|_2 \leq M(1+\varepsilon), \quad \forall z \in B, \tag{2.5.1}$$

for some constant $M > 0$.

Next, define a map $r_m : \mathbb{S}^{N-1} \to \mathbb{R}$ such that

$$r_m(x) := \sqrt{N} \sqrt{\sum_{i=1}^{m} x_i^2} = \sqrt{N}\, \|\boldsymbol{Q}_m x\|_2.$$

Since any orthogonal matrix is norm preserving, we may rewrite Equation 2.5.1 as

$$M(1-\varepsilon) \leq \|\boldsymbol{Q}_m \boldsymbol{U} z\|_2 \leq M(1+\varepsilon), \quad \forall z \in B,$$

and, consequently, to prove the JL-Lemma, it suffices to choose $\boldsymbol{U} \in O(N)$ such that

$$M\sqrt{N}\,(1-\varepsilon) \leq r_m(\boldsymbol{U}z) \leq M\sqrt{N}\,(1+\varepsilon), \quad \forall z \in B. \tag{2.5.2}$$

Our goal is to prove Equation 2.5.2 through a probabilistic argument. Namely, we show that, for a sufficiently large rank m,

$$\sigma\left\{\boldsymbol{U} \in O(N) : M\sqrt{N}\,(1-\varepsilon) \leq r_m(\boldsymbol{U}z) \leq M\sqrt{N}\,(1+\varepsilon),\ \forall z \in B\right\} > 0.$$

That is, that the probability of a $\boldsymbol{U}$ selected from the uniform (normalized Haar) distribution on $O(N)$ satisfy Equation 2.5.2 is strictly positive.

More precisely, we have shown previously in this Chapter (Theorem 2.4.1) that the distribution of $\|\boldsymbol{U}^t Q \boldsymbol{U} x\|_2$ is the same for: a constant $\boldsymbol{U} \in O(N)$ and a uniformly distributed $x \in \mathbb{S}^{N-1}$; a constant $x \in \mathbb{S}^{N-1}$ and a uniformly distributed $\boldsymbol{U} \in O(N)$. Thus, the same holds for $r_m(\boldsymbol{U}z)$. Consequently, for any fixed $y \in \mathbb{S}^{N-1}$, the measure

$$\sigma\left\{\boldsymbol{U} \in O(N) : M\sqrt{N}\,(1-\varepsilon) \leq r_m(\boldsymbol{U}y) \leq M\sqrt{N}\,(1+\varepsilon)\right\}$$

is equal to

$$\mu_{N-1}\left\{x \in \mathbb{S}^{N-1} : M\sqrt{N}\,(1-\varepsilon) \leq r_m(\boldsymbol{U}x) \leq M\sqrt{N}\,(1+\varepsilon)\right\},$$

for a fixed $\boldsymbol{U} \in O(N)$.

Now, since $x \mapsto r_m(\boldsymbol{U}x)$ is a $\sqrt{N}$-Lipschitz map, the concentration of measure on the sphere (Theorem 2.3.3) guarantees that r_m concentrates about its median, $M_r > 0$. That is, for a fixed $\boldsymbol{U} \in O(N)$ and $\xi \in (0,1)$,

$$\mu_{N-1}\left\{x \in \mathbb{S}^{N-1} : |r_m(\boldsymbol{U}\,x) - M_r| \leq \sqrt{N}\,\xi\right\} \geq 1 - 4\exp\left\{-\frac{N\xi^2}{2}\right\}.$$

Therefore, the interchangeability between the measures μ_{N-1} and σ, yields, for each fixed $x \in \mathbb{S}^{N-1}$ and $\xi \in (0,1)$, that

$$\sigma\left\{\boldsymbol{U} \in O(N) : |r_m(\boldsymbol{U}x) - M_r| \leq \sqrt{N}\,\xi\right\} \geq 1 - 4\exp\left\{-\frac{N\xi^2}{2}\right\},$$

that is a concentration of measure inequality on $O(N)$. Now, applying the union bound relative to $z \in B$ in the inequality above, we conclude that

$$\sigma\left\{\boldsymbol{U} \in O(N) : |r_m(\boldsymbol{U}z) - M_r| \leq \sqrt{N}\,\xi,\ \forall z \in B\right\} \geq 1 - 2N(N-1)\exp\left\{-\frac{N\xi^2}{2}\right\}, \tag{2.5.3}$$

since $\#B = \binom{N}{2}$. For a fixed $N > 2$, this probability is positive when

$$\xi > \sqrt{\frac{2}{N}\log\{2N(N-1)\}}. \tag{2.5.4}$$

In order to conclude the JL-Lemma's proof, we need to determine a sufficiently small ε so that Equation 2.5.2 is satisfied with positive probability. That is, a small enough ξ such that

$$[M_r - \sqrt{N}\xi, M_r + \sqrt{N}\xi] \subset [M\sqrt{N}\,(1-\varepsilon), M\sqrt{N}\,(1+\varepsilon)],$$

for a constant $M > 0$.

As in the paper [JL1984], this will be done following [FLM1977]. That is, in the next subsection, we shall present a result from [FLM1977] that bounds the median M_r of r_m in terms of its mean

$$\mathbb{E}_{\mathbb{S}^{N-1}}\{r_m(x)\} = \int_{\mathbb{S}^{N-1}} r_m(x)\, d\mu_{N-1}(x).$$

This is an uncommon approach since the calculation of the median is usually simpler than the mean. However, we are not interested in calculating $\mathbb{E}_{\mathbb{S}^{N-1}}\{r_m(x)\}$ directly. We instead present a technique used in [JL1984] to bound the mean of r_m on $\mathbb{S}^{N-1}$ through the *Khintchine inequality*.

2.5.3 Approximating the mean by the median on the sphere

The paper [FLM1977] presents an important result that allows us to approximate the mean of $\|\cdot\|_2$ on the sphere by its Lévy mean, $\text{med}\{\|\cdot\|_2\}$. This result is very convenient since, in general, the median of a function on $\mathbb{S}^{N-1}$ is more easily calculated than its mean. Namely, we have the following:

Lemma 2.5.1. *Let $\|\cdot\|$ be a norm in a N-dimensional Banach space X and $|||\cdot|||$ be an inner product norm on X such that*

$$a|||x||| \leq \|x\| \leq b|||x|||, \quad \forall x \in X,$$

for suitable $0 < a \leq b < \infty$. There is an absolute constant $c > 0$ so that whenever the equation above holds with $b \leq \sqrt{N}$, with $N = \dim X$, then

$$\left| \int_{|||x|||=1} \|x\|\, d\mu_{N-1}(x) - med\{\|\cdot\|\} \right| < c.$$

Moreover, in the proof of such result, it was stated that

$$c = \sum_{m=1}^{\infty} 4(m+1)e^{-m^2/2} \lesssim \frac{20}{3}.$$

In particular, we may apply this result to r_m since

$$\|\boldsymbol{Q}_m x\|_2 \leq \|x\|_2, \quad \forall x \in \mathbb{S}^{N-1}.$$

from what

$$r_m(x) = \sqrt{N}\,\|\boldsymbol{Q}_m x\|_2 \leq \sqrt{N}\,\|x\|_2 = \sqrt{N}, \quad \forall x \in \mathbb{S}^{N-1}.$$

Consequently, we obtain

$$\left| \int_{\mathbb{S}^{N-1}} r_m(z)\, d\mu_{N-1}(z) - M_r \right| \leq \frac{20}{3}.$$

Or equivalently,

$$-\frac{20}{3}+\int_{\mathbb{S}^{N-1}} r_m(z)\,d\mu_{N-1}(z)\leq M_r\leq \frac{20}{3}+\int_{\mathbb{S}^{N-1}} r_m(z)\,d\mu_{N-1}(z). \tag{2.5.5}$$

In the next subsection, we present a classical method to bound the mean of $\|\cdot\|_2$, and consequently of $r_m(\cdot)$, on $\mathbb{S}^{N-1}$. From this result, we may find bounds on the median M_r depending only on the rank $m>0$ of the random projection. Finally, we show that for a sufficiently large m, the JL-Lemma will be satisfied.

2.5.4 Bounding the mean through Khintchine's inequality

In this subsection, we will present a technique to bound the mean

$$\mathbb{E}_{\mathbb{S}^{N-1}}\{r_m(z)\}=\int_{\mathbb{S}^{N-1}} r_m(z)\,d\mu_{N-1}(x),$$

for $m\in[N]$. Namely,

$$\sqrt{\frac{m}{2}}\leq\int_{\mathbb{S}^{N-1}} r_m(z)\,d\mu_{N-1}(z)\leq\sqrt{2m}.$$

In order to do so, we need to introduce the following concepts.

Definition 2.5.1 (Rademacher random variable)**.** *We say that a discrete random variable X has the Rademacher distribution iff.*

$$\mathbb{P}\{X=k\}=\begin{cases}1/2, & k\in\{-1,1\}\\ 0, & \textit{otherwise}\end{cases}.$$

Theorem 2.5.1 (Khintchine inequality)**.** *Let $\eta=(\eta_1,\ldots,\eta_N)\in[-1,1]^N$ be a random vector whose entries are i.i.d. Rademacher random variables. Then, for $0<p<\infty$ and $x\in\mathbb{C}^N$, we have that*

$$A_p\|x\|_2\leq\left(\mathbb{E}_\eta|\langle x:\eta\rangle|^p\right)^{1/p}\leq B_p\|x\|_2,$$

for some constants $A_p,B_p>0$ depending only on p.

Remark (Signal average)**.** *It is common in the literature, including [JL1984], to call a expectation as in the Theorem above of signal average since the randomness comes from the signal that multiplies each coordinate of $x\in\mathbb{R}^N$ With this terminology, the expectation would be denoted as*

$$\mathbb{E}_\eta|\langle x:\eta\rangle|^p=Av_\pm\left\{\sum_{i=1}^N\pm x_i\right\}$$

The sharp values for the constants A_p and B_p were determined by Haagerup (see [Hrup1981]) and are exhibited below

$$A_p=\begin{cases}2^{1/2-1/p}, & 0<p\leq p_0\\ 2^{1/2}\left[\frac{1}{\sqrt{\pi}}\Gamma\left(\frac{1+p}{2}\right)\right]^{1/p}, & p_0<p<2\\ 1, & 2\leq p<\infty\end{cases}$$

and

$$B_p = \begin{cases} 1, & 0 < p \leq 2 \\ 2^{1/2} \left[\frac{1}{\sqrt{\pi}} \Gamma\left(\frac{1+p}{2} \right) \right]^{1/p}, & 2 < p < \infty, \end{cases}$$

with $p_0 \approx 1.847$.

The particular case that is interesting for us is the one for $p = 1$. In that case, $A_1 = 1/\sqrt{2}$ and $B_1 = 1$. Consequently, the Khintchine inequality may be rewritten as

$$\frac{1}{\sqrt{2}} \|x\|_2 \leq \mathbb{E}_\eta \big\{ |\langle x : \eta \rangle| \big\} \leq \|x\|_2. \tag{2.5.6}$$

From the Equation above, it follows a method to bound the mean of r_m on the sphere, i.e., the integral

$$\int_{\mathbb{S}^{N-1}} r_m(z) \, d\mu_{N-1}(z).$$

Indeed, take a natural $m \in [N]$ and a set $\{\eta_1, \ldots, \eta_m\}$ of independent Rademacher random variables. Next, fix the canonical basis $\{e_1, \ldots, e_N\}$ and define the random vector

$$\eta = \sum_{i=1}^{m} \eta_i e_i = (\eta_1, \ldots, \eta_m, 0, \ldots, 0) \in \mathbb{R}^N.$$

Also, we define the following number:

$$\alpha_N = \int_{\mathbb{S}^{N-1}} |z_1| \; d\mu_{N-1}(z),$$

with z_1 being the first coordinate of $z \in \mathbb{S}^{N-1}$ in the canonical basis. In order to estimate the integral of r_m on the sphere, we need to prove the following Equation:

$$\sqrt{m} \, \alpha_N = \mathbb{E}_\eta \left\{ \int_{\mathbb{S}^{N-1}} |\langle z : \eta \rangle| \; d\mu_{N-1}(z) \right\}. \tag{2.5.7}$$

Intuitively, Equation 2.5.7 follows from the symmetry of $\mathbb{S}^N$ and the rotation invariance of the measure μ_{N-1}. From these properties, the integral on $\mathbb{S}^{N-1}$ of $x \mapsto |\langle x : \eta \rangle|$ will depend only on the norm of the random vector η, whose value is $\sqrt{m}$ for any choice of η. More precisely, the random vector $\eta_1 e_1$ has norm 1. Thus, we may take a rotation $\boldsymbol{T} \in O(N)$ such that

$$\eta = \boldsymbol{T} \left(\|\eta\|_2 \eta_1 e_1 \right) = \sqrt{m} \, \boldsymbol{T} \left(\eta_1 e_1 \right).$$

We conclude, from the rotation invariance of μ_{N-1}, that:

$$\begin{aligned}
\mathbb{E}_\eta\left\{\int_{\mathbb{S}^{N-1}} |\langle z:\eta\rangle|\, d\mu_{N-1}(z)\right\} &= \sqrt{m}\,\mathbb{E}_\eta\left\{\int_{\mathbb{S}^{N-1}} |\langle z:\boldsymbol{T}(\eta_1 e_1)\rangle|\, d\mu_{N-1}(z)\right\}\\
&= \sqrt{m}\,\mathbb{E}_\eta\left\{\int_{\mathbb{S}^{N-1}} \left|\langle \boldsymbol{T}^t z:\eta_1 e_1\rangle\right|\, d\mu_{N-1}(z)\right\}\\
&= \sqrt{m}\,\mathbb{E}_\eta\left\{\int_{\mathbb{S}^{N-1}} \left|\langle \boldsymbol{T}^t z:\eta_1 e_1\rangle\right|\, d\mu_{N-1}(\boldsymbol{T}^t z)\right\}\\
&= \sqrt{m}\,\mathbb{E}_\eta\left\{\int_{\mathbb{S}^{N-1}} |\langle z:\eta_1 e_1\rangle|\, d\mu_{N-1}(z)\right\}\\
&= \sqrt{m}\,\mathbb{E}_\eta\left\{|\,\eta_1\,|\int_{\mathbb{S}^{N-1}} |z_1|\, d\mu_{N-1}(z)\right\} = \sqrt{m}\,\alpha_N,
\end{aligned}$$

concluding then the proof of Equation 2.5.7.

Now, we bound the mean of r_m on $\mathbb{S}^{N-1}$ as follows. Apply the Khintchine inequality (Equation 2.5.6) to $\mathbb{E}_\eta\{|\langle \boldsymbol{Q}_m z:\eta\rangle|\}$. Doing so, we obtain

$$\frac{1}{\sqrt{2}}\|\boldsymbol{Q}_m z\|_2 \le \mathbb{E}_\eta\{|\langle \boldsymbol{Q}_m z:\eta\rangle|\} \le \|\boldsymbol{Q}_m z\|_2, \quad \forall z\in\mathbb{S}^{N-1}.$$

Equivalently, we have

$$\frac{1}{\sqrt{2N}}\, r_m(z) \le \mathbb{E}_\eta\{|\langle z:\eta\rangle|\} \le \frac{1}{\sqrt{N}}\, r_m(z), \quad \forall z\in\mathbb{S}^{N-1}.$$

By the monotonicity property for the integral, we have

$$\begin{aligned}
\frac{1}{\sqrt{2N}}\int_{\mathbb{S}^{N-1}} r_m(z)\, d\mu_{N-1}(z) &\le \int_{\mathbb{S}^{N-1}} \mathbb{E}_\eta\{|\langle z:\eta\rangle|\}\, d\mu_{N-1}(z)\\
&= \mathbb{E}_\eta\left\{\int_{\mathbb{S}^{N-1}} |\langle z:\eta\rangle|\, d\mu_{N-1}(z)\right\}\\
&= \sqrt{m}\,\alpha_N\\
&\le \frac{1}{\sqrt{N}}\int_{\mathbb{S}^{N-1}} r_m(z)\, d\mu_{N-1}(z). \qquad (2.5.8)
\end{aligned}$$

Note that we can interchange the integral with the expectation above since η is a discrete random vector.

Finally, by applying Equations 2.5.6 and 2.5.7 to Equation 2.5.8, for any $m\in[N]$, we have

$$\sqrt{Nm}\,\alpha_N \le \int_{\mathbb{S}^{N-1}} r_m(z)\, d\mu_{N-1}(z) \le \sqrt{2Nm}\,\alpha_N. \qquad (2.5.9)$$

In particular, we may choose $m=N$, from what

$$r_N(z) = \sqrt{N}\,\|z\|_2 = \sqrt{N}$$

and

$$\int_{\mathbb{S}^{N-1}} r_N(z)\, d\mu_{N-1}(z) = \sqrt{N}.$$

Applying this result to Equation 2.5.9, we get

$$\frac{1}{\sqrt{2N}} \leq \alpha_N \leq \frac{1}{\sqrt{N}}.$$

Thus, for any natural $m \in [1, N]$,

$$\sqrt{\frac{m}{2}} \leq \int_{\mathbb{S}^{N-1}} r_m(z)\, d\mu_{N-1}(z) \leq \sqrt{2m}.$$

2.5.5 The lower bound on the rank m

In the previous subsection, we proved that, for any natural $1 \leq m \leq N$,

$$\sqrt{\frac{m}{2}} \leq \int_{\mathbb{S}^{N-1}} r_m(z)\, d\mu_{N-1}(z) \leq \sqrt{2m}.$$

Applying the inequality above to Equation 2.5.5, we conclude the following:

$$M_r \geq \sqrt{\frac{m}{2}} - \frac{20}{3}.$$

Furthermore, in Equation 2.5.3, we choose

$$\xi = \frac{\varepsilon}{\sqrt{N}} \left(\sqrt{\frac{m}{2}} - \frac{20}{3} \right) \leq \frac{\varepsilon}{\sqrt{N}} M_r,$$

that yields

$$\sigma \left\{ \boldsymbol{U} \in O(N) : (1-\varepsilon) M_r \leq r_m(\boldsymbol{U}z) \leq (1+\varepsilon) M_r,\ \forall z \in B \right\}$$

$$\geq 1 - 2N(N-1) \exp\left\{ - \left(\sqrt{\frac{m}{2}} - \frac{20}{3} \right)^2 \frac{\varepsilon^2}{2} \right\}.$$

By making $M = M_r > 0$ in the sufficient condition for the JL-Lemma (Equation 2.5.2), we conclude the proof since the probability above is positive. From Equation 2.5.4, we know that it happens when

$$\xi = \frac{\varepsilon}{\sqrt{N}} \left(\sqrt{\frac{m}{2}} - \frac{20}{3} \right) > \sqrt{\frac{2}{N} \log\{2N(N-1)\}}.$$

Consequently, the JL-Lemma is satisfied since

$$m > 2 \left(\frac{20}{3} + \frac{1}{\varepsilon} \sqrt{2 \log\{2N(N-1)\}} \right)^2.$$

Even though the JL-Lemma is satisfied for the lower bound in the previous equation, this expression is exaggeratedly large. Moreover, it does not exhibit in a clear manner a key characteristic of this Lemma: a lower bound that has the logarithmic order of N.

In order to obtain a more pleasant expression, we shall take a rougher value for M_r, namely,

$$M_r \geq \sqrt{\frac{m}{2}} - \frac{20}{3} > \sqrt{\frac{m}{3}},$$

for a sufficiently large m. We shall also make a new choice of ξ to be $\varepsilon\sqrt{m/3N}$. In this setting, we conclude from Equation 2.5.4 that the JL-Lemma will be satisfied with positive probability when

$$m > \frac{6}{\varepsilon^2}\log\{2N(N-1)\}.$$

Finally, it suffices to prove that there exists a positive number $m_0(\varepsilon)$ depending only on ε such that

$$m_0(\varepsilon)\log N \geq \frac{6}{\varepsilon^2}\log\{2N(N-1)\},$$

or, equivalently,

$$m_0(\varepsilon) \geq \frac{6}{\varepsilon^2}\frac{\log\{2N(N-1)\}}{\log N}.$$

Indeed, we may take $m_0(\varepsilon) \geq 18\varepsilon^{-2}$ since

$$\frac{\log\{2N(N-1)\}}{\log N} \leq 3, \quad \forall N > 2,$$

concluding the proof of the JL-Lemma.

2.6 The parameter space in the JL-Lemma

In the statement of the JL-Lemma, we claimed that, for any dimension $N > 2$ and any quasi-isometry distortion $\varepsilon \in (0,1)$, there is a sufficiently large natural number $0 < m < N$ with the logarithmic order of N such that we may define, with positive probability, a ε-isometric projection from ℓ_2^N to ℓ_2^m by a rank m linear random projection. Along the proof, however, we made some bounds that may not be valid for the entire parameter domain described in the JL-Lemma's statement.

More precisely, most of the bounds used during the proof are valid just for sufficiently large values of N and ε. If these parameters are not large enough, we may get for example a lower bound on m that is bigger than N itself. Consequently, we shall dedicate this section to specify the parameter domain on which the presented proof of the JL-Lemma holds. On the other hand, we recall the reader that since the JL-Lemma is usually applied in a high-dimensional setting, it is not a problem to have a lower bound on N bigger than 2.

During the proof, we defined a constant $\xi \in (0,1)$. It was used to obtain a lower bound on the probability of the $\sqrt{N}$-Lipschitz map $r_m(\boldsymbol{U}x)$ being sufficiently near from its median M_r on $\mathbb{S}^{N-1}$ for a given $x \in \mathbb{S}^{N-1}$ and a uniformly chosen $\boldsymbol{U} \in O(N)$. As seen in Equation 2.5.3, this lower bound is

$$\underline{\mathbb{P}} = 1 - 2N(N-1)\exp\left\{-\frac{N\xi^2}{2}\right\}.$$

From Equation 2.5.4, such probability will be positive iff. ξ is bigger than a $\underline{\xi}$ such that

$$\underline{\xi} := \sqrt{\frac{2}{N}\log\{2N(N-1)\}}.$$

Moreover, since $\xi \in (\underline{\xi}, 1)$, we must have

$$\sqrt{\frac{2}{N}\log\{2N(N-1)\}} < 1,$$

which holds for $N \geq 11$.

2.6.1 Sharper bounds

In a first moment, we made the following choice for ξ

$$\xi_1 = \frac{\varepsilon}{\sqrt{N}}\left(\sqrt{\frac{m}{2}} - \frac{20}{3}\right).$$

This value comes from a lower bound on the median of r_m on $\mathbb{S}^{N-1}$,

$$M_r \geq \sqrt{\frac{m}{2}} - \frac{20}{3},$$

obtained from the *Khintchine inequality*. During the proof, we concluded that, for $\xi = \xi_1$, the concentration of r_m around its median implies that the JL-Lemma is satisfied with probability bigger than $\underline{\mathbb{P}}$. Since this lower bound on the probability is positive just for $\xi \in (\underline{\xi}, 1)$, the JL-Lemma will be satisfied with positive probability when

$$\underline{\xi} < \xi_1 < 1,$$

i.e., for values of the parameters N, m and ε such that

$$\sqrt{\frac{2}{N}\log\{2N(N-1)\}} < \frac{\varepsilon}{\sqrt{N}}\left(\sqrt{\frac{m}{2}} - \frac{20}{3}\right) < 1.$$

From those inequalities, we obtain the following bounds on m :

$$2\left(\frac{20}{3} + \frac{1}{\varepsilon}\sqrt{2\log\{2N(N-1)\}}\right)^2 < m < 2\left(\frac{\sqrt{N}}{\varepsilon} + \frac{20}{3}\right)^2.$$

In particular, we obtain a lower bound $\underline{m} < m$ such that

$$\underline{m} = 2\left(\frac{20}{3} + \frac{1}{\varepsilon}\sqrt{2\log\{2N(N-1)\}}\right)^2.$$

This bound is completely useless if it is bigger than N. Consequently, we intend to determine the parameter values such that $\underline{m} < N$. It may be done by solving the following inequality:

$$2\left(\frac{20}{3} + \frac{1}{\varepsilon}\sqrt{2\log\{2N(N-1)\}}\right)^2 < N.$$

Finally, the inequality above yields the following lower bound on ε :

$$\underline{\varepsilon} := \frac{\sqrt{2\log\{2N(N-1)\}}}{\sqrt{\frac{N}{2}} - \frac{20}{3}}.$$

Since $\varepsilon \in (0,1)$, we must have

$$0 < \frac{\sqrt{2\log\{2N(N-1)\}}}{\sqrt{\frac{N}{2}} - \frac{20}{3}} < 1,$$

that holds for any natural $N \geq 267$.

In sum, note that the presented proof for the JL-Lemma is not valid for the entire parameter space $N > 2$ and $\varepsilon \in (0,1)$. Instead, we must have

$$N \geq 267 \quad \text{and} \quad \frac{\sqrt{2\log\{2N(N-1)\}}}{\sqrt{\frac{N}{2}} - \frac{20}{3}} < \varepsilon < 1.$$

For these parameter values, the JL-Lemma will be satisfied with a probability bigger than

$$\underline{\mathbb{P}}_1 = 1 - 2N(N-1)\exp\left\{-\frac{\varepsilon^2}{2}\left(\sqrt{\frac{m}{2}} - \frac{20}{3}\right)^2\right\},$$

that is positive for any natural $m > \underline{m}$ such that

$$\underline{m} = 2\left(\frac{20}{3} + \frac{1}{\varepsilon}\sqrt{2\log\{2N(N-1)\}}\right)^2.$$

However, the expression above does not exhibit a clear logarithmic dependence of $\underline{m}$ on $\log N$, as stated in the JL-Lemma. This behavior is more evident when we opt for rougher bounds and, consequently, for a smaller parameter space.

2.6.2 Rougher bounds

In the proof for the JL-Lemma presented in the previous section, we choose a rougher value

$$\xi_2 := \varepsilon\sqrt{\frac{m}{3N}}$$

for ξ, i.e., when m is sufficiently large, we have $\xi_2 \leq \xi_1$. This choice is useful since it provides us with a easy proof that the lower bound $\underline{m}$ for m may be presented in the form $m_0(\varepsilon)\log N$. Furthermore, along the proof of the JL-Lemma, it was shown that, for $\xi = \xi_2$, the concentration of r_m about its median M_r implies that the JL-Lemma is satisfied with a probability bigger than

$$\underline{\mathbb{P}} = 1 - 2N(N-1)\exp\left\{-\frac{N\xi^2}{2}\right\}.$$

As explained in the previous subsection, the sharper value ξ_1 for ξ comes from a lower bound on M_r obtained from the *Kintchine inequality*:

$$M_r \geq \sqrt{\frac{m}{2}} - \frac{20}{3}.$$

On the other hand, the rougher one ξ_2 comes from the fact that

$$\sqrt{\frac{m}{2}} - \frac{20}{3} \geq \sqrt{\frac{m}{3}} \quad \text{(and consequently } \xi_1 \geq \xi_2\text{)},$$

for a sufficiently large natural m, or more precisely, for any natural $m \geq 2640$.

Again, the probability's lower bound $\underline{\mathbb{P}}$ is positive only if $\xi \in (\underline{\xi}, 1)$. Moreover, when $\xi = \xi_2$, $\underline{\mathbb{P}}$ is a lower bound on the probability of the JL-Lemma being satisfied. Consequently, the JL-Lemma will be satisfied with positive probability when

$$\underline{\xi} < \xi_2 < 1,$$

i.e., for values of the parameters N, m and ε such that

$$\sqrt{\frac{2}{N}\log\{2N(N-1)\}} < \varepsilon \sqrt{\frac{m}{3N}} < 1.$$

From these inequalities, we obtain the following bounds on m :

$$\frac{6}{\varepsilon^2}\log\{2N(N-1)\} < m < \frac{3N}{\varepsilon^2}.$$

Next, we verify for which parameter values the lower bound on m will be useful, i.e.,

$$2640 < \frac{6}{\varepsilon^2}\log\{2N(N-1)\} < N.$$

These inequalities define the following bounds on ε :

$$\underline{\varepsilon} = \sqrt{\frac{6}{N}\log\{2N(N-1)\}} < \varepsilon < \sqrt{\frac{\log\{2N(N-1)\}}{440}} = \overline{\varepsilon}.$$

Finally, we must determine which N implies $(\underline{\varepsilon}, \overline{\varepsilon}) \subset (0, 1)$.

At first, we have $\underline{\varepsilon} < \overline{\varepsilon}$ for any natural $N \geq 2640$. Now, differently from the sharper bounds in the previous section, N is also bounded from above. Indeed, in order to have $\overline{\varepsilon} < 1$, it suffices to take any natural $N < 2.4 \times 10^{95}$. Fortunately, such a bound is *(way) bigger than the estimated amount of atoms in the universe* and will not be a problem in any application. On the other hand, this value depicts a remarkable characteristic of the interval $(\underline{\varepsilon}(N), \overline{\varepsilon}(N))$: it is way too small for reasonable values of N. For example, it approaches the size of $(0, 1/2)$ just for $N > 10^{25}$. Consequently, these rougher bounds, despite of their theoretical usefulness, are virtually useless for applications.

2.7 Conclusion

In this Chapter we introduced the necessary notation and presented the original proof of the JL-Lemma. Moreover, we exhibited how limited is this first version of the Lemma for applications. A discussion about its theoretical improvements is the goal of the next Chapter.

Chapter 3

Theoretical improvements to the JL-Lemma

3.1 Introduction

In Chapter 2, we presented the original proof to the Johnson-Lindenstrauss Lemma [JL1984]. This Lemma gives us a tool for *dimensionality reduction* through *random projections*. Namely, datasets may be almost isometrically projected on a lower dimension with, for example, the goal of *simplifying data analysis, solve computation and storage problems of algorithms and avoid the curse of dimensionality*, previously described in Chapter 1.

However, methods such as PCA intend to project the data such that the resulting dataset is "similar" to the original one, in the sense that the variance of the projected dataset is maximized, avoiding then distinct points of the original dataset to collapse. This search for a similarity in the global conformation of the original dataset is displayed by the fact that the PCA is *adaptive*, that is, dependent of the dataset that we intend to project. The JL-Lemma has a different goal: it projects the data aiming the preservation of the pairwise distances instead the global conformation of the dataset, and in a non-adaptative fashion.

More precisely, the JL-Lemma is an amazing result that expresses an important property of random projections. If we substitute the isometry requirement of the projection by an ε-isometry one, the Lemma yields a projection onto a dimension $m(\varepsilon) > 0$ that is not only independent from the projected set (*non-adaptive*), but also has the logarithmic order of the dimension N. Indeed, m has a lower bound of the form

$$m \geq m_0(\varepsilon) \log N.$$

Even though the JL-Lemma yields a strong property of random projections, there is still room for *theoretical and practical improvements*.

3.1.1 What do we mean by improvement?

The first proof of the JL-Lemma was presented in [JL1984] by W. B. Johnson and J. Lindenstrauss, two analysts that used this result as a tool for achieving a result about extensions of Lipschitz functions into a Hilbert space. As a consequence, they may not have

noticed it was the inception of a new paradigm towards dimensionality reduction and had no clear motivation to present the result in a "sharper" form that is more suitable for applications. More precisely, we may split the historical treatment in two kinds of improvements: *the theoretical and the practical ones*, with this first one being the subject of this Chapter.

Firstly, these theoretical results refers to theorems that guarantee a *tighter (larger) lower bound* on the projected dimension m. Moreover, since the first proof for the JL-Lemma was embedded in a more abstract context, it is quite harder than the modern ones, that use techniques from Statistics and Probability Theory. Indeed, any new proof is usually easier to follow up than the previous ones. In particular, we cite the one from Dasgupta and Gupta made in 1999 and published in [DasGup2003], that yields a bound that stayed the sharper for almost 10 years and that can be understood by any undergraduate student with a first Probability or Statistics course. Another important point is that the pioneer proof to the JL-Lemma demonstrated just that we can select the JL-embedding from a certain distribution with positive probability. In modern statements of the result, the authors gave an *explicit construction of the sampling method to determine the JL-embedding*. Beyond computational and storage improvements, that will not be discussed in the present text, this new form of the JL-Lemma, named *distributional JL-Lemma*, guarantees that not only the probability of building the right JL-transform is positive, but also that *it can also be made arbitrarily close to* 1.

Finally, the original statement of the JL-Lemma yields that, for a prefixed $A \subset \mathbb{R}^N$, with $N > 2$ and $\varepsilon \in (0, 1)$, we may choose a random rank m linear projection $P : \ell_2^N \to \ell_2^m$ that projects A in ℓ_2^m ε-isometrically with positive probability as long as

$$m \geq m_0(\varepsilon) \log N, \text{ with } m_0(\varepsilon) = 18\, \varepsilon^{-2}.$$

However, in the transcription of the original proof exhibited in Chapter 2, we discussed how loose are the bounds used to conclude this result. Consequently, a natural question is if we can achieve a sharper expression for $m_0(\varepsilon)$ in order to obtain a tighter (i.e., a larger), and consequently more precise, lower bound on m.

3.2 The geometric era of the JL-Lemma

We start the present section explaining its very title. Namely, we may split historically the treatment towards the JL-Lemma in three epochs: *the geometric era, the Gaussian era and the sub-Gaussian era*.

In this first period, the JL-projection is not explicitly specified, although it is always a random linear orthogonal projection. Also, the results were proved through: geometric arguments, as the concentration of the Haar measure in the unit sphere (or the Grassmannians set); direct computations of the Haar measure of subsets of the unit sphere (or the Grassmannians set); estimations for the Haar integral of certain functions on these sets, and so on.

As the reader may have noticed, this was the treatment made in the original proof [JL1984] in 1984; furthermore, as we may explain in deeper details in this section, a very similar reasoning was made in [FklMae1988] in 1988 in a proof that obtains a sharper bound and is way simpler than the previous one by a technical detail. The last work based on

such paradigm of proof was [IndMot1998] in 1998. Although this proof still uses geometrical arguments and does not improve the bound from [FklMae1988], it gives birth to a new look on the Lemma as we shall see in the next section.

3.2.1 Original lower bound on m – 1984

In Chapter 2, we presented the first proof of the JL-Lemma from [JL1984]. This proof is made through the following steps: exhibiting the *concentration of measure* of the relative surface area of the sphere $\mathbb{S}^{N-1}$ about its equatorial strips through a geometrical approach; the use of another geometrical result, *the isoperimetric inequality on the sphere* to conclude that any Lipschitz function $f : \mathbb{S}^{N-1} \to \mathbb{R}$ concentrates about its *Lévy's mean (or median)* M_f; guessing as a candidate for JL-embedding a *random rank* m *linear projection* $P : \ell_2^N \to \ell_2^m$; obtaining an approximation to the median M_r of the following Lipschitz map

$$\begin{aligned} r_m : \mathbb{S}^{N-1} &\longrightarrow \quad \mathbb{R} \\ z \quad &\mapsto \sqrt{N}\,\|Pz\|_2 \end{aligned}$$

by applying the Lemma 2.7 from [FLM1977], that bounds the difference between the median and the expectation of r_m in $\mathbb{S}^{N-1}$, and the *signal average*, a widely known technique to estimate $\mathbb{E}_{\mathbb{S}^{N-1}}\{r_m(z)\}$ using only the *Khintchine inequality*. Through those steps, we conclude the JL-Lemma with a lower bound

$$m = \text{floor}(18\,\varepsilon^{-2}\log N).$$

Finally, despite of the several loose bounds and approximations made during that proof, in the pioneering paper [JL1984], Johnson and Lindenstrauss argued that we cannot determine a JL-projection into a space whose dimension m has a dependence in N that is less than logarithmic. In fact, they used a result about ε-nets to conclude that in a ball with radius 2 in ℓ_2^m, there are at most 4^m vectors whose pairwise distance is at least 1. Consequently, for a sufficiently small distortion ε, there is no ε-isometry that projects an orthonormal set with more than 4^m vectors in a m-dimensional subspace of ℓ_2^N (for more details see [JL1984]). That being said, the improvements on this lower bound will be done through determining better expressions on $\varepsilon > 0$ in the definition of m.

3.2.2 Frankl and Maehara's proof – 1988

In [FklMae1988], P. Frankl and H. Maehara improved the lower bound on m and exhibited a simplified proof of the JL-Lemma than the one in [JL1984]. Their goal in that work was to apply this enhanced form of the Lemma to the *sphericity problem on graphs* [FklMae1986, Mae1984, HMae1984, Mae1986, HMae1986, Pach1980].

In short, let $G = (V, E)$ be a graph with an N-vertex set V. The adjacency matrix $\mathcal{A}(G) \in \mathcal{M}_{N\times N}$ of this graph determines which vertices are connected by edges in E. More precisely, for any $i, j \in [N]$,

$$[\mathcal{A}(G)]_{ij} = \begin{cases} 1, \text{ if } i \neq j \text{ and } (v_i, v_j) \in E \\ 0, \text{ otherwise.} \end{cases} .$$

Consequently, we can describe $G = (V, E)$ by the column set $A \subset [0,1]^N$ of this symmetric matrix. Thus, to simplify the terminology, we shall not make a difference between the points of V and A.

In this setting, we define the sphericity, $\mathrm{spg}(G)$, of this N-vertex graph as the smallest $m \in \mathbb{N}$ such that there is an embedding $f : A \to \mathbb{R}^m$ satisfying $0 < \|f(x) - f(y)\|_2 < 1$ if, and only if, x and y are linked by an edge in E. Recall that this yields a tool to represent the original graph $G = (V, E)$, or more precisely its vector form $A \subset \mathbb{R}^N$, through the set $f(A) \subset \mathbb{R}^m$, with m being hopefully way smaller than N.

Provided with this brief introduction, the reader might already have guessed what was the contribution of the JL-Lemma to the sphericity problem. In fact, the sphericity is smaller than any dimension m with an embedding of the vertex set into $\mathbb{R}^m$, as previously exhibited; on the other hand, from [JL1984] the JL-Lemma yields such an embedding for an $m(N, \varepsilon)$ with the form $m_0(\varepsilon) \log N$, with $\varepsilon \in (0, 1)$. Therefore, the main result from Frankl and Maehara in [FklMae1988] was that the sphericity of a graph has an upper bound with the logarithmic order of the number N of vertices.

Furthermore, they also stated an enhanced version of JL-Lemma that not only has a tighter lower bound $m(N, \varepsilon)$, but also has a simpler proof than the original one. The Frankl and Maehara's version of the JL-Lemma is:

Theorem 3.2.1 (Frankl and Maehara – 1988)**.** *Given $N \in \mathbb{N}$ and $\varepsilon \in (0, 1/2)$, define*

$$m(N, \varepsilon) := ceil\left\{ \frac{27}{\varepsilon^2(3 - 2\varepsilon)} \log(N) \right\} + 1.$$

If $N > m(N, \varepsilon)^2$, then for any N-point set $A \subset \mathbb{R}^N$, there exists a map $f : A \to \mathbb{R}^{m(N,\varepsilon)}$ such that

$$(1 - \varepsilon)\|u - v\|^2 < \|f(u) - f(v)\|^2 < (1 - \varepsilon)\|u - v\|^2, \quad \forall u, v \in A.$$

The proof starts just as in [JL1984]. We assume f to be a rank m linear projection in the canonical basis, i.e.,

$$\boldsymbol{U}^t \boldsymbol{Q}_m \boldsymbol{U} \in \mathcal{M}_{N \times N},$$

with $\boldsymbol{U} \in O(N)$ and $\boldsymbol{Q}_m \in \mathcal{M}_{N \times N}$, the matrix such that

$$\boldsymbol{Q}_m(x_1, \ldots, x_N) = (x_1, \ldots, x_m, 0, \ldots, 0) \in \mathbb{R}^N, \quad \forall x = (x_1, \ldots, x_N) \in \mathbb{R}^N.$$

Thus, for a fixed $\varepsilon \in (0, 1/2)$, we intend to choose $\boldsymbol{U} \in O(N)$ such that

$$M(1 - \varepsilon)\|v - u\|_2 \leq \|\boldsymbol{U}^t \boldsymbol{Q}_m \boldsymbol{U} v - \boldsymbol{U}^t \boldsymbol{Q}_m \boldsymbol{U} u\|_2 \leq M(1 + \varepsilon)\|v - u\|_2, \quad \forall u, v \in A,$$

for some real constant $M > 0$.

To guarantee that f is indeed a JL-empedding, in both works [JL1984] and [FklMae1988] this problem was dealt through a technique called *random projection*. Indeed, we take the random map f obtained by sampling $\boldsymbol{U}$ from the uniform distribution on $O(N)$. Consequently, it suffices to prove that f is a JL-embedding with positive probability for a dimension $m \geq m_0(\varepsilon) \log N$. This is done through a concentration of measure argument.

Here is where the modern proof breaks apart. In both works, a key argument was the fact that the norm $\|f(x)\|_2$ for any $x \in \mathbb{S}^{N-1}$ concentrates with high probability about a fixed number. However, the proof in [FklMae1988] is way simpler than the original one, because Johnson and Lindenstrauss had to compute several bounds to conclude the concentration of $\|f(x)\|_2$ in a useful manner to obtain the required lower bound $m(N, \varepsilon)$. On the other hand, Frankl and Maehara take a shortcut by concluding that the squared norm, $\|f(x)\|_2^2$, has a simpler concentration about m/N. More precisely, they proved the following result:

Proposition 3.2.1. *Let $z \in \mathbb{R}^N$ be a unit vector and $H \subset \mathbb{R}^N$ be a random m-dimensional subspace through the origin. Also, define the random variable X as the squared length of the projection of z onto H. By choosing $\varepsilon \in (0, 1/2)$, $N > m^2$, and $m > 24 \log N + 1$, we have*

$$\mathbb{P}\left\{\left| X - \frac{m}{N} \right| > \varepsilon \frac{m}{N}\right\} < 2\sqrt{m} \exp\left\{-\frac{\varepsilon^2(3-2\varepsilon)}{12}(m-1)\right\}.$$

Now, for N and m as in the Proposition above and for any previously fixed $\varepsilon \in (0, 1/2)$ and $z \in B$, the event

$$\frac{m}{N}(1-\varepsilon) \leq \|\boldsymbol{U}^t \boldsymbol{Q}_m \boldsymbol{U} z\|_2^2 \leq \frac{m}{N}(1+\varepsilon),$$

occurs with a probability larger than

$$1 - 2\sqrt{m} \exp\left\{-\frac{\varepsilon^2(3-2\varepsilon)}{12}(m-1)\right\}.$$

Thus, by applying the union bound relatively to $z \in B$, we conclude that the projection

$$\sqrt{\frac{N}{m}}\, \boldsymbol{U}^t \boldsymbol{Q}_m \boldsymbol{U},$$

satisfies the JL-Lemma with probability larger than

$$1 - 2\binom{N}{2}\sqrt{m} \exp\left\{-\frac{\varepsilon^2(3-2\varepsilon)}{12}(m-1)\right\}.$$

Finally, note that this probability is positive if

$$m > 1 + \frac{12}{\varepsilon^2(3-2\varepsilon)} \log\{N(N-1)\sqrt{m}\}.$$

Moreover, since $N > m^2$, we have that

$$N(N-1)\sqrt{m} < N^{9/4}$$

and, consequently,

$$m > 1 + \frac{27}{\varepsilon^2(3-2\varepsilon)} \log N.$$

Remark. *We must recall that the proof's overview here is not historically accurate. As indicated in the previous proposition, Frankl and Maehara did not directly select a random rank m linear projection. Instead, they selected a random m-dimensional vector subspace of $\mathbb{R}^N$. However, since these models for random projections are equivalent, yielding the same proof, we made an analogy with the original proof in Chapter 2, for simplicity.*

For completeness, we must say that Frankl and Maeraha published an improved version of the Lemma in 1990 [**?**]. We shall not make a deeper discussion about it, but they used a result about the fast decay of the tails of the Beta distribution. This new version of the Lemma dropped the constraint of $N > m(N, \varepsilon)^2$ and obtained a lower bound that was only obtained again in Dasgupta and Gupta's work in 1998.

3.2.3 Indyk and Motwani's proof – 1998

In [IndMot1998][1], Piotr Indyk and Rajeev Motwani exhibited a simpler proof to the Frankl and Maehara's version of the JL-Lemma, with the goal of addressing the ε-NNS problem, that we will discuss briefly. This new statement of the Lemma is valid for any norm $\|\cdot\|_p$, with $p \in [1, 2]$. Moreover, they considered the JL-transform as a Gaussian matrix, i.e., a matrix $T \in \mathcal{M}_{m\times N}$ whose entries are are i.i.d random variables following a $\mathcal{N}(0, 1)$ distribution.

More precisely, consider a prefixed M-point dataset $X = \{X_1, \ldots, X_M\}$ in some metric space, say the Euclidean space ℓ_2^N, for simplicity. Now, suppose that we intend to answer a certain distance dependent query whose options are only the points in X. The *nearest neighbor search (NNS)* consists of determining a *query point*, that is, a point $q \in \mathcal{M}$ satisfying the query and then returning as a solution the point $X^* \in X$ with minimal distance to q. An intuitive approach to this problem is through the *brute force algorithm* that stands to a myriad of *naïve* problem solution methods. In this context, it means the exhaustive distance calculation between the points in X and the query point in order to determine which yields the smaller distance. Obviously, this strategy becomes impractical as the dimension N or number M of points increase.

As an alternative approach, we may preprocess the set X in order to efficiently solve the NNS problem. Indeed, the low dimensional case was already solved [Herb1987], however, the preprocessing itself has a high computational cost. Consequently, despite of decades of effort, the solutions to the NNS problem were far from satisfactory in theory or in practice until the publication of [IndMot1998]. Indeed, for high values of N or M, those solutions provided little improvement over the brute force algorithm. As a consequence to the failure in addressing the NNS, a relaxed version of this problem took place: *the approximate or ε-approximate nearest neighbors problem (ε-NNS)*. In this new setting, given a prefixed $\varepsilon > 0$, we intend to preprocess X in order to efficiently find a point in $X^* \in X$ whose distance to a query point $q \in \mathcal{M}$ is the least one, up to a $(1 + \varepsilon)$ multiplicative constant. That is, for a fixed $\varepsilon > 0$, we want to find $X^* \in X$ such that

$$d(X^*, q) \leq (1 + \varepsilon)\, d(X_j, q) \quad \forall j \in [M].$$

The introduction of the ε-NNS yielded a significant improvement with respect to the algorithms known until 1998 for the classical NNS problem. Indeed, they were of two kinds: low preprocessing cost, but linear query time in M and N; sublinear query time in M and polynomial in N, but exponential preprocessing cost M^N. On the other hand, in [IndMot1998], two algorithms were presented to address the ε-NNS problem. The first one preprocess X with polynomial cost in N and M and a query search time that is truly sublinear in N and M,

[1] This paper was republished later, in 2012 [IndMot2012].

more precisely, for $\varepsilon > 1$, it has time $\mathcal{O}\{(M^{1+1/\varepsilon} + NM)\ \text{poly}(\log M)\}$. The second one has a mildly exponential preprocessing cost $\mathcal{O}(M \log M) \times \mathcal{O}(\varepsilon^{-N})$ for $\varepsilon \in (0,1)$ and a query time polynomial in $\log M$ and N, i.e., $\mathcal{O}(N \log M)$. In this setting, the use of the Frankl and Maehara's JL-Lemma [FklMae1988] in the preprocessing phase of the second algorithm for $(\mathcal{M}, d) = \ell_p^N$, with $p \in [1,2]$, resulted in the best of both worlds: the first known algorithm with preprocessing and query time polynomial in N and $\log M$. Namely, they proved the following:

Theorem 3.2.2 (Proposition 3 from [IndMot1998]). *For any $\varepsilon > 0$, there is an algorithm for the ε-NNS in ℓ_p^N, with $p \in [1,2]$, whose preprocessing time is $(NM)^{\mathcal{O}(1)}$ and requires a $\mathcal{O}\{M\,poly(\log N)\}$ query time.*

Remark. *A more direct treatment of how the JL-Lemma is applied to the preprocessing phase of the ε-NNS is way too technical and will not be exhibited in this text. Hence, we direct the interested reader to the paper [IndMot1998].*

Finally, the statement of the JL-Lemma due to Indyk and Motwani was the following:

Theorem 3.2.3 (Indyk and Motwani – 1998). *For any $p \in [1,2]$, any M-point set $S \subset \ell_p^N$, and any $\varepsilon > 0$, there exists a map $f : S \to \ell_2^m$ with $m = \mathcal{O}(\log M)$ such that for all $u, v \in S$,*

$$(1-\varepsilon)\|u-v\|_p < \frac{N}{m}\|f(u)-f(v)\|_2^2 < (1+\varepsilon)\|u-v\|_p.$$

Note that this is the first result towards generalizing the JL-Lemma for other norms than just $\|\cdot\|_2$. In this regard, they have also proved that the JL-Lemma does not apply for the ℓ_∞ norm[2].

3.3 The Gaussian era of the JL-Lemma

3.3.1 Revisiting Indyk and Motwani's work

Beyond presenting a new proof for the JL-Lemma, it was also proved in [IndMot1998] a result that is the keystone to the Dasgupta and Gupta's proof [DasGup2003] and the inception of a new paradigm of proof that is still used nowadays. Namely,

Lemma 3.3.1 (Lemma 7 from [IndMot1998]). *Let u be a unit vector in $\mathbb{R}^N$. For any even positive integer m, let $U_1, \ldots, U_m$ be random vectors chosen independently from the standard N-dimensional Gaussian distribution (i.e., each of its components is $\mathcal{N}(0,1)$). Now, define*

$$W = W(u) = (X_1, \ldots, X_m), \text{ with } X_i = \langle u : U_i \rangle;$$

and $L(u) = \|W\|_2^2$. Then, for any $\beta > 1$,

1. $\mathbb{E}(L) = m$,

[2]See the Theorem 6 from [IndMot1998].

2. $\mathbb{P}\{L \geq \beta m\} < \mathcal{O}(m) \exp\left\{-\frac{k}{2}(\beta - 1 - \log\beta)\right\}$

3. $\mathbb{P}\{L \leq m/\beta\} < \mathcal{O}(m) \exp\left\{-\frac{k}{2}(\frac{1}{\beta} - 1 + \log\beta)\right\}$.

Sketch of the proof. Recall that each X_i is distributed as $\mathcal{N}(0,1)$. Now, for $i \in [m/2]$, define

$$Y_i = X_{2i-1}^2 + X_{2i}^2.$$

Each Y_i follows the exponential distribution with parameter $\lambda = 1/2$ [Feller1991]. Thus,

$$\mathbb{E}(L) = \sum_{i=1}^{m/2} \mathbb{E}(Y_i) = \frac{m}{2} \times 2 = m;$$

also L follows the Gamma distribution with parameters $\alpha = 1/2$ and $\upsilon = m/2$ [Feller1991]. Since this distribution is dual to the Poisson, we have that

$$\mathbb{P}\{L \geq \beta m\} = \mathbb{P}\{P_{\beta m}^{1/2} \leq \upsilon - 1\},$$

with P_t^α being a random variable following the Poisson distribution with parameter αt. Bounding the later quantity is a matter of simple calculation. □

It is important to note that the assumption of the projection in the JL-Lemma as a random linear orthogonal projection that was made in previous results ([JL1984, FklMae1988]) is dropped. It is proved instead that random $(N \times m)$-matrices whose columns are random Gaussian vectors $\{U_1, \ldots, U_m\} \subset \mathbb{R}^N$ satisfy the ε-isometry with high probability, which justifies the name of this section.

3.3.2 Dasgupta and Gupta's proof of JL-Lemma – 2003

In 2003, it was the turn of Dasgupta and Gupta to make further improvements in the JL-Lemma [DasGup2003]. The lower bound on m derived by them remained as the best one known until Matoušek's 2008 work [Mat2008]. Surprisingly, this proof is simpler than the previous ones, maybe the simplest proof for the JL-Lemma.

Theorem 3.3.1 (Dasgupta and Gupta – 2003)**.** *Given a set V of M points in $\mathbb{R}^N$, for some M, $N \in \mathbb{N}$, choosing $\varepsilon \in (0,1)$ and $m \geq$ floor $\{24\,(3\varepsilon^2 - 2\varepsilon^3)^{-1} \log M\}$ yields a linear map $T : \mathbb{R}^N \to \mathbb{R}^m$ satisfying*

$$(1-\varepsilon)\|u-v\|_2^2 \leq \|T(u) - T(v)\|_2^2 \leq (1+\varepsilon)\|u-v\|_2^2, \quad \forall u, v \in X.$$

In order to prove this version of the JL-Lemma, we shall present the Lemma 3.3.1 from Indyk and Motwani as it was stated by Dasgupta and Gupta in their work. To be accurate, this analogous result is stronger than its previous version since that one yields a lower bound on m that is larger by an additive factor of roughly $\mathcal{O}(\log\log M)$.

At first, let $X_1, \ldots, X_N$ be independent standard Gaussian random variables and define the Gaussian random vector $X = (X_1, \ldots, X_N) \in \mathbb{R}^N$. Next, let $\pi_m(X) \in \mathbb{R}^m$ be the projection of X on its first m coordinates and define

$$Z = \frac{\pi_m(X)}{\|X\|_2} \in \mathbb{R}^m$$

and denote $L = \|Z\|_2^2$. Now, we have the following:

Proposition 3.3.1. *With the notation presented so far, we have that*

$$\mathbb{E}(L) = \frac{m}{N}.$$

Proof. In fact,

$$1 = \mathbb{E}(1) = \mathbb{E}\left\{\frac{\|X\|_2^2}{\|X\|_2^2}\right\} = N\,\mathbb{E}\left\{\frac{X_1^2}{\|X\|_2^2}\right\},$$

with this last equality being justified by the linearity of the expectation and by the $\{X_i\}_{i\in[N]}$ being an i.i.d. set. Consequently,

$$\mathbb{E}\left\{\frac{X_1^2}{\|X\|_2^2}\right\} = \frac{1}{N}.$$

Finally,

$$\mathbb{E}(L) = \mathbb{E}\left\{\frac{\sum_{i=1}^m X_i^2}{\|X\|_2^2}\right\} = m\,\mathbb{E}\left\{\frac{X_1^2}{\|X\|_2^2}\right\} = \frac{m}{N},$$

concluding the proof. □

Now, we are able to state the Dasgupta and Gupta's version of Lemma 3.3.1. Such result states that not only $\mathbb{E}(L) = m/N$, but also that L is tight concentrated around its mean.

Lemma 3.3.2. *Let $m < N$. Then:*

1. *if $\beta < 1$,*

$$\mathbb{P}\left\{L \leq \frac{\beta m}{N}\right\} \leq \beta^{m/2}\left(1 + m\frac{1-\beta}{N-m}\right)^{(N-m)/2} \leq \exp\left\{\frac{m}{2}(1-\beta+\log\beta)\right\}; \quad (3.3.1)$$

2. *if $\beta > 1$,*

$$\mathbb{P}\left\{L \geq \frac{\beta m}{N}\right\} \leq \beta^{m/2}\left(1 + m\frac{1-\beta}{N-m}\right)^{(N-m)/2} \leq \exp\left\{\frac{m}{2}(1-\beta+\log\beta)\right\}. \quad (3.3.2)$$

Proof of the Lemma 3.3.2. Let us start proving Equation 3.3.1; more precisely, that

$$\mathbb{P}\left\{L \leq \frac{\beta m}{N}\right\} \leq \beta^{m/2}\left(1 + m\frac{1-\beta}{N-m}\right)^{(N-m)/2}. \quad (3.3.3)$$

In fact,

$$\begin{aligned}\mathbb{P}\left\{L \leq \frac{\beta m}{N}\right\} &\leq \mathbb{P}\left\{NL \leq \beta m\right\} \\ &= \mathbb{P}\left\{N\|Z\|_2^2 \leq \frac{\|X\|_2^2}{\|X\|_2^2}\,\beta m\right\} \\ &= \mathbb{P}\{N(X_1^2 + \ldots + X_m^2) \leq (X_1^2 + \ldots + X_N^2)m\beta\}.\end{aligned}$$

Moreover,

$$\begin{aligned}&\mathbb{P}\{(X_1^2+\ldots+X_m^2)N \leq (X_1^2+\ldots+X_N^2)m\beta\}\\ &\quad = \mathbb{P}\{(X_1^2+\ldots+X_N^2)m\beta - (X_1^2+\ldots+X_m^2)N \geq 0\},\end{aligned}$$

which is equal to

$$\mathbb{P}\{\exp\{t[(X_1^2+\ldots+X_N^2)m\beta - (X_1^2+\ldots+X_m^2)N]\} \geq 1\}, \tag{3.3.4}$$

for $t > 0$.

Next, we apply Markov's inequality to Equation 3.3.4, concluding that

$$\mathbb{P}\left\{L \leq \frac{\beta m}{N}\right\} \leq \mathbb{E}\big\{\exp\{t[(X_1^2+\ldots+X_N^2)m\beta - (X_1^2+\ldots+X_m^2)N]\}\big\}.$$

Now, note that for $X \sim \mathcal{N}(0,1)$, we have

$$\begin{aligned}&\mathbb{E}\big\{\exp\{t[(X_1^2+\ldots+X_N^2)m\beta - (X_1^2+\ldots+X_m^2)N]\}\big\}\\ &\qquad = \mathbb{E}\left\{e^{tm\beta X^2}\right\}^{(N-m)} \mathbb{E}\left\{e^{t(m\beta-N)X^2}\right\}^m.\end{aligned}$$

We now use the easily proved fact that if $X \sim \mathcal{N}(0,1)$, then

$$\mathbb{E}\exp\{sX^2\} = \frac{1}{\sqrt{1-2s}} \qquad (-\infty < s < 1/2).$$

Consequently,

$$\mathbb{E}\left\{e^{tm\beta X^2}\right\}^{(N-m)} \mathbb{E}\left\{e^{t(m\beta-N)X^2}\right\}^m = (1-2tm\beta)^{-(N-m)/2}[1-2t(m\beta-N)]^{-m/2}. \tag{3.3.5}$$

We now set the right hand side of Equation 3.3.5 as

$$g(t) := (1-2tm\beta)^{-(N-m)/2}[1-2t(m\beta-N)]^{-m/2}.$$

This is a positive expression since the left hand side of Equation 3.3.5 is the product of positive numbers. From this fact, this expression gives us two additional constraints

$$tm\beta < 1/2 \quad \text{and} \quad t(m\beta - N) < 1/2,$$

with this later being included in the former since $t \geq 0$, yielding $t \in (0, \frac{1}{2m\beta})$.

Now, to minimize $g(t)$, we maximize

$$f(t) := (1-2tm\beta)^{(N-m)}[1-2t(m\beta-N)]^m$$

for $t \in (0, \frac{1}{2m\beta})$. Differentiating f, we get that the maximum is achieved at

$$t_0 = \frac{(1-\beta)}{2\beta(N-m\beta)},$$

which lies in the allowed range. Hence, we have

$$f(t_0) = \left(\frac{N-m}{N-m\beta}\right)^{N-m} \left(\frac{1}{\beta}\right)^{m};$$

and the fact that $g(t_0) = 1/\sqrt{f(t_0)}$ proves Equation 3.3.3. Finally,

$$\begin{aligned} g(t_0) &= \beta^{m/2} \left(1 + m\frac{1-\beta}{N-m}\right)^{(N-m)/2} \\ &= \exp\left\{\left[\beta^{m/2}\left(1 + m\frac{1-\beta}{N-m}\right)^{(N-m)/2}\right]\right\} \\ &= \exp\left\{\frac{m}{2}\log\beta + \left(\frac{N-m}{2}\right)\log\left(1 + m\frac{1-\beta}{N-m}\right)\right\} \\ &\leq \exp\left\{\frac{m}{2}\log\beta + \left(\frac{N-m}{2}\right)\left[m\frac{1-\beta}{N-m} - \frac{m^2}{2}\left(\frac{1-\beta}{N-m}\right)^2\right]\right\}, \end{aligned}$$

since, for all $x \geq 0$,

$$\log(1+x) \leq x - \frac{x^2}{2}. \tag{3.3.6}$$

Consequently,

$$\begin{aligned} \beta^{m/2}\left(1 + m\frac{1-\beta}{N-m}\right)^{(N-m)/2} &\leq \frac{\exp\left\{\frac{m}{2}(1-\beta+\log\beta)\right\}}{\exp\left\{\frac{m^2}{4}\frac{(1-\beta^2)}{(N-m)}\right\}} \\ &\leq \exp\left\{\frac{m}{2}(1-\beta+\log\beta)\right\}, \end{aligned}$$

proving the first part of Lemma 3.3.2.

Regarding Equation 3.3.2, the proof is almost exactly the same as that of Equation 3.3.1. In fact, the same calculations will show that

$$\begin{aligned} \mathbb{P}\big\{N(X_1^2 + \ldots + X_m^2) \leq m\beta(X_1^2 + \ldots + X_N^2)\big\} \\ \leq (1 + 2tm\beta)^{-(N-m)/2}[1 + 2t(m\beta - N)]^{-m/2} \end{aligned}$$

for $t \in \left(0, \frac{1}{2(N-m\beta)}\right)$. Also, the right hand side of the inequality above is minimized at $-t_0$, with t_0 defined as previously. This value does lie in the desired range $\left(0, \frac{1}{2(N-m\beta)}\right)$ for $\beta > 1$, which yields

$$\mathbb{P}\{N(X_1^2 + \ldots + X_m^2) \leq m\beta(X_1^2 + \ldots + X_N^2)\} \leq \beta^{m/2}\left(1 + m\frac{1-\beta}{N-m}\right)^{(N-m)/2}$$

and the proof of Equation 3.3.2 follows in an analogous fashion. □

We now recall a central result used to prove the JL-Lemma in Chapter 2: if we randomly choose a point $x \in \mathbb{S}^{N-1}$ and a fixed rank m linear orthogonal projection $P : \ell_2^N \to \ell_2^N$, the distribution of the norm of $Px \in \ell_2^N$ would be the same as if we have chosen a fixed $x \in \mathbb{S}^{N-1}$ and a uniformly chosen projection P.

Our next step is Theorem 3.3.2, which claims that the normalization of a N-standard Gaussian vector has an uniform distribution on the sphere. As a consequence, $L \in \mathbb{R}$ will have the same distribution of the norm of the projection on a fixed m-subspace of ℓ_2^N of a uniformly chosen point in $\mathbb{S}^{N-1}$. Finally, by the result in the paragraph above, L will also have the same distribution of the norm of a fixed point in $\mathbb{S}^{N-1}$ projected on a uniformly chosen m-subspace of ℓ_2^N. Namely,

Theorem 3.3.2. *Let we represent a Gaussian vector $g \in \mathbb{R}^N$ in polar form as*

$$g = r\theta,$$

where $r = \|g\|_2$ is the length and $\theta = g/\|g\|_2$ is the direction of g. We have that r and θ are independent random variables. Moreover, θ is uniformly distributed on the unit sphere $\mathbb{S}^{N-1}$.

We are now able to present the following proof:

Proof of the Theorem 3.3.1. If $N \leq m$, the result follows directly. Else, take a random m-dimensional subspace $S \in \mathbb{R}^N$ and let $v_i' \in S$ be the projection of $v_i \in V$ into S. Then, by the discussion motivated by Theorem 3.3.2, L has the same distribution of $\|v_i' - v_j'\|_2^2$. Moreover, by setting $\beta = 1 - \varepsilon$ and $\mu = (m/N)\|v_i - v_j\|_2^2$ and applying Equation 3.3.1 from Lemma 3.3.2, we get

$$\begin{aligned}
\mathbb{P}\{\|v_i' - v_j'\|_2^2 \leq (1-\varepsilon)\mu\} &= \mathbb{P}\{L \leq (1-\varepsilon)\mu\} \\
&\leq \exp\left\{\frac{m}{2}[1 - (1-\varepsilon) + \log(1-\varepsilon)]\right\} \\
&\leq \exp\left\{\frac{m}{2}\left[\varepsilon - \left(\varepsilon + \frac{\varepsilon^2}{2}\right)\right]\right\} = \exp\left\{-\frac{m\varepsilon^2}{4}\right\} \\
&\leq \exp\{-2\log M\} = 1/M^2,
\end{aligned}$$

by applying the fact that

$$\log(1-x) \leq -x - \frac{x^2}{2}, \tag{3.3.7}$$

for $0 \leq x < 1$, in the second line.

Similarly, we can apply Equations 3.3.2 and 3.3.6 to get

$$\begin{aligned}
\mathbb{P}\{L \geq (1+\varepsilon)\mu\} &\leq \exp\left\{\frac{m}{2}[1 - (1+\varepsilon) + \log(1+\varepsilon)]\right\} \\
&\leq \exp\left\{\frac{m}{2}\left[-\varepsilon + \left(\varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3}\right)\right]\right\} \\
&= \exp\left\{-\frac{m\varepsilon^2(3-2\varepsilon)}{12}\right\} \\
&= \exp\{-2\log M\} = \frac{1}{M^2}.
\end{aligned}$$

Now, set the map $f(v_i) = \sqrt{N/m}\, v_i'$. By the above calculations, for some fixed pair (i, j), the chance that the distortion

$$\frac{\|f(v_i) - f(v_j)\|_2^2}{\|v_i - v_j\|_2^2}$$

does not lie in the range $[1-\varepsilon, 1+\varepsilon]$ is at most $2/M^2$. Using the trivial union bound, the chance that some pair of points suffers a large distortion is at most

$$\binom{M}{2} \times 2/M^2 = 1 - 1/M.$$

Hence, f has the desired properties with probability at least $1/M$. □

Recall that during their proof Dasgupta and Gupta used the Markov inequality. This yields a loose bound and will give room to an improved version of this proof that will be presented in 2010 by Javier Rojo and Tuan S. Nguyen [RojNg2010].

3.3.3 Rojo and Nguyen – 2010

In this next work, Rojo and Nguyen [RojNg2010] revisit the proof presented by Dasgupta and Gupta [DasGup2003]. Namely, they noticed that even though this previous work was based on the the assumption of a Gaussian random matrix, the proof uses a loose bound on the probability for the JL-Lemma to hold since it uses the *Markov inequality.* Rojo and Nguyen avoid this by working directly over the distribution of the squared norm of the projected vectors.

Moreover, we must say that Rojo and Nguyen further investigated, also in [RojNg2010], a version of the JL-Lemma that projects ℓ_2^N to ℓ_1^m. However, we shall not make a deeper discussion about it in this text.

Regarding the Rojo and Nguyen's version of the JL-Lemma, it is based on the following results:

Lemma 3.3.3 (Lemma 3.1 from [RojNg2010])**.** *Let m be an even integer, and $0 < \varepsilon < 1$. Let $\lambda_1 = m(1+\varepsilon)/2$ and $d = m/2$. Then,*

$$g(m, \varepsilon) := e^{-\lambda_1} \frac{\lambda_1^{d-1}}{(d-1)!}$$

is a decreasing function in m for a fixed ε.

Theorem 3.3.3 (Theorem 3.3 from [RojNg2010])**.** *Let d be a positive integer.*

1. *Let $1 \leq d < \lambda_1$. Then,*

$$\sum_{k=0}^{d-1} \frac{\lambda_1^k}{k!} = \left(\frac{\lambda_1}{\lambda_1 - d}\right) \left\{ \frac{\lambda_1^{d-1}}{(d-1)!} \right\}.$$

2. *Let $0 < \lambda_2 < d$. Then,*

$$\sum_{k=d}^{\infty} \frac{\lambda_2^k}{k!} = \left(\frac{\lambda_2}{d - \lambda_2}\right) \left\{ \frac{\lambda_2^{d-1}}{(d-1)!} \right\}.$$

We're now in the position of stating and proving the Rojo and Nguyen version of the JL-Lemma.

Theorem 3.3.4. *For any $0 < \varepsilon < 1$ and integer M, let m be the smallest even integer satisfying*

$$\frac{1+\varepsilon}{\varepsilon}\, g(m,\varepsilon) \leq \frac{1}{M^2}.$$

Then, for any set V of M points in $\mathbb{R}^N$, we may choose a linear map $f : \mathbb{R}^N \to \mathbb{R}^m$ such that

$$(1-\varepsilon)\|u-v\|_2^2 \leq \|f(u)-f(v)\|_2^2 \leq (1+\varepsilon)\|u-v\|_2^2, \;\; \forall\, u, v \in V \tag{3.3.8}$$

with a probability of at least

$$1 - \frac{2}{M^2}.$$

Proof. Firstly, since f is assumed linear, Equation 3.3.8 is equivalent to

$$\mathbb{P}\{\|f(z)\|_2^2 \geq (1+\varepsilon)\|z\|_2^2\} + \mathbb{P}\{\|f(z)\|_2^2 \leq (1-\varepsilon)\|z\|_2^2\} \leq \frac{2}{M^2}, \tag{3.3.9}$$

for $z \in V - V$. Furthermore, the bound in Equation 3.3.9 can be obtained by separately bounding the left and right tail probabilities. That is, by finding f such that simultaneously

$$\mathbb{P}\{\|f(z)\|_2^2 \geq (1+\varepsilon)\|z\|_2^2\} \leq \frac{1}{M^2} \tag{3.3.10}$$

and

$$\mathbb{P}\{\|f(z)\|_2^2 \leq (1-\varepsilon)\|z\|_2^2\} \leq \frac{1}{M^2} \tag{3.3.11}$$

holds.
Also, let $\boldsymbol{R}$ be a $N \times m$ random matrix with independent standard Gaussian entries whose column set will be denoted by $\{r_1, \ldots, r_m\} \subset \mathbb{R}^N$. For x in the M-point set V, define

$$f(x) = \frac{1}{\sqrt{m}}\, x^T \boldsymbol{R} \quad \text{and} \quad y = \sqrt{m}\, \frac{f(x)}{\|x\|_2}.$$

Then, $y_j = \langle x/\|x\|_2 : r_j \rangle \sim \mathcal{N}(0,1)$ and consequently $y_j^2 \sim \chi_1^2$ with $\mathbb{E}\{\|y\|_2^2\} = m$. Next, recall that if $X \sim \Gamma(d,1)$ and $Y \sim \text{Poisson}(x)$, we have $\mathbb{P}\{X \geq x\} = \mathbb{P}\{Y \leq d-1\}$. That is,

$$\int_x^\infty \frac{1}{\Gamma(d)} z^{d-1} e^{-z}\, dz = \sum_{j=0}^{d-1} \frac{x^j e^{-x}}{j!}, \tag{3.3.12}$$

for $d = 1, 2, 3, \ldots$

Now, let us refer to $m(1+\varepsilon)$ in Equation 3.3.10 as α_1 to simplify the notation. Since

$$\|y\|_2^2 = \sum_{j=1}^{m} y_j^2 \sim \chi_m^2 \equiv \Gamma(m/2, 2),$$

we may apply Equation 3.3.12 to Equation 3.3.10 with $d = m/2$ and write the right-tail probability as

$$\mathbb{P}\{\|y\|_2^2 \geq \alpha_1\} = e^{-\alpha_1/2}\sum_{j=0}^{d-1}\frac{(\alpha_1/2)^j}{j!}. \tag{3.3.13}$$

In the same fashion, let $\alpha_2 := m(1-\varepsilon)$ in Equation 3.3.11. Then, the left-tail probability can be written as

$$\mathbb{P}\{\|y\|_2^2 \leq \alpha_2\} = e^{-\alpha_2}\sum_{j=d}^{\infty}\frac{(\alpha_2/2)^j}{j!}. \tag{3.3.14}$$

The proof is concluded by applying Theorem 3.3.3 to Equations 3.3.13 and 3.3.14. For illustration, let us apply it to Equation 3.3.13 with

$$\lambda_1 = \alpha_1/2 = m(1+\varepsilon)/2 \quad \text{and} \quad d = m/2.$$

Indeed, the right-tail probability is bounded as follows:

$$\mathbb{P}\{\|y\|_2^2 \geq \alpha_1\} = e^{-\lambda_1}\sum_{j=0}^{d-1} \leq \left(\frac{1+\varepsilon}{\varepsilon}\right)\left[\frac{\lambda_1^{d-1}}{(d-1)!}\right]e^{-\lambda_1}.$$

On the other hand, by setting

$$\lambda_2 = \alpha_2 = m(1-\varepsilon)/2 \quad \text{and} \quad d = m/2$$

in Equation 3.3.14, it follows from Theorem 3.3.3 that

$$\begin{aligned}\mathbb{P}\{\|y\|_2^2 \leq \alpha_2\} &= e^{-\lambda_2}\sum_{j=d}^{\infty}\frac{\lambda_2^j}{j!} \\ &= \left(\frac{1-\varepsilon}{\varepsilon}\right)\left[\frac{\lambda_2^{d-1}}{(d-1)!}\right]e^{-\lambda_2} \\ &\leq \left(\frac{1+\varepsilon}{\varepsilon}\right)\left[\frac{\lambda_1^{d-1}}{(d-1)!}\right]e^{-\lambda_1},\end{aligned}$$

with the last inequality being due to the fact that

$$e^{\lambda_1-\lambda_2} \leq \left(\frac{\lambda_1}{\lambda_2}\right)^d.$$

Finally, note that the bound on the left-tail probability is the same as the one for the right-tail probability. Therefore,

$$\mathbb{P}\{\|y\|_2^2 \geq \alpha_1\} + \mathbb{P}\{\|y\|_2^2 \leq \alpha_2\} \leq 2\left(\frac{1+\varepsilon}{\varepsilon}\right)g(m,\varepsilon).$$

Moreover, we can obtain, for a given ε, the lower bound on m by numerically obtaining the smallest even integer m such that

$$\left(\frac{1+\varepsilon}{\varepsilon}\right)g(m,\varepsilon) \leq \frac{1}{M^2},$$

concluding then the proof. □

3.4 The sub-Gaussian era of the JL-Lemma

3.4.1 Introduction

Before exploring further developments in the JL-Lemma, we have to introduce some new concepts that justify the name of the present section.

Definition 3.4.1 (Sub-Gaussian tail)**.** *Let X be a real random variable with $\mathbb{E}X = 0$. We say that X has a sub-Gaussian upper tail if there exists a constant $a > 0$ such that for all $\lambda > 0$,*

$$\mathbb{P}\{X > \lambda\} \leq e^{-a\lambda^2}.$$

Moreover, we say that X has a sub-Gaussian upper tail up to λ_0, if the previous bound holds for all $\lambda \leq \lambda_0$. Finally, we say that X has a sub-Gaussian tail , if both X and $-X$ have sub-Gaussian upper tails.

Definition 3.4.2 (Uniform sub-Gaussian tail)**.** *Given the previous Definition, we say that a sequence $X_1, \ldots, X_N$ of random variables has a uniform sub-Gaussian tail when all of them have sub-Gaussian tails with the same constant.*

The next proof we present is based on taking a matrix whose entries are independent random variables with sub-Gaussian entries. More precisely,

Definition 3.4.3 (Rademacher random variable)**.** *Let X be a random variable attaining values 1 and -1 each with probability $1/2$. We say that X is a Rademacher random variable.*

Proposition 3.4.1. *Rademacher random variables have a sub-Gaussian tail.*

Proof. In fact, let X be a Rademacher random variable. Then, $\mathbb{E}X = 0$ and

$$\mathbb{P}\{X > \lambda\} = \begin{cases} 0, \text{ if } \lambda \geq 1 \\ 1/2, \text{ if } -1 \leq \lambda < 1 \\ 1, \text{ if } \lambda < -1 \end{cases}.$$

Consequently, if $0 < \lambda < 1$,

$$\mathbb{P}\{X > \lambda\} = 1/2 = e^{-\log 2} \leq e^{-a\lambda^2},$$

for $a = \log 2$. On the other hand, if $\lambda \geq 1$,

$$\mathbb{P}\{X > \lambda\} = 0 \leq e^{-a\lambda^2},$$

for any $a > 0$. □

3.4.2 Achlioptas – 2001

In his 2001 work, [Ach2003][3], Achlioptas presented a new sampling method of the random projection matrix in the JL-Lemma. This new method has a way smaller computational complexity and it is very easy to implement. Indeed, it drops the assumption of Gaussian entries by substituting them for independent Rademacher random variables.

Another variant of his result has the entries of the projection matrix attaining value 0 with probability 2/3 and values $\sqrt{3}$ and $-\sqrt{3}$ with probability 1/6 each. This later setting allows for computing the projection about 3 times faster than the former. This can be justified since this matrix is *sparse*, since only about one third of its entries are nonzero.

He also proves that surprisingly this simplification comes without *any* sacrifice in the dimensionality reduction of the embedding. In fact, Achlioptas pointed out that the loss of information is minimal since M vectors chosen uniformly in $\mathbb{S}^{N-1}$ are nearly orthonormal in a high-dimensional space. Moreover, the random projection matrix is close to orthogonal in such spaces [Mrk1994]. Finally, for every fixed value N of the original dimension of the dataset, there is a slightly better bound than all current methods.

The Achlioptas' version of the JL-Lemma is now presented:

Theorem 3.4.1 (Achlioptas – 2001 [Ach2003])**.** *Let P be an arbitrary set of M points in $\mathbb{R}^N$, represented as an $(M \times N)$-matrix $\boldsymbol{A}$. Given $\varepsilon, \beta > 0$, let*

$$m_0 = \frac{12(2+\beta)}{\varepsilon^2(3-2\varepsilon)} \log M.$$

For a integer $m \geq m_0$, let $\boldsymbol{R}$ be a $N \times m$ random matrix with $[\boldsymbol{R}_{i,j}]$ being independent random variables from either of the following probability distributions:

$$[\boldsymbol{R}_{i,j}] = \begin{cases} 1, & \textit{with probability } 1/2, \\ -1, & \textit{with probability } 1/2 \end{cases}$$

or

$$[\boldsymbol{R}_{i,j}] = \sqrt{3} \times \begin{cases} 1, & \textit{with probability } 1/6, \\ 0, & \textit{with probability } 2/3, \\ -1, & \textit{with probability } 1/6 \end{cases}.$$

Let

$$\boldsymbol{E} = \frac{1}{\sqrt{m}} \boldsymbol{A}\boldsymbol{R}$$

and let $f : \mathbb{R}^N \to \mathbb{R}^m$ map the i-th row of $\boldsymbol{A}$ to the i-th row of $\boldsymbol{E}$. Thus, with probability at least $1 - M^{-\beta}$, for all $u, v \in P$, we have

$$(1-\varepsilon)\|u-v\|_2^2 \leq \|f(u)-f(v)\|_2^2 \leq (1+\varepsilon)\|u-v\|_2^2.$$

The work of Achlioptas motivated the further speed up methods for the JL-Lemma implementation when the projection matrix is *sparse*. For a matter of brevity we will not discuss this new branch in the area, that leads to the FJLT *(fast Johnson-Lindenstrauss transform)*. The interested reader may direct himself to [AilChz2006][4].

[3] A conference paper on 2001, published in a Journal in 2003.

[4] This conference paper was later published in a Journal in 2009 [AilChz2009].

3.4.3 Matoušek – 2008

In his work [Mat2008], Matoušek obtained a very simple proof of the JL-Lemma using the concept of random variables with sub-Gaussian tails. Moreover, he improved its computational burden by using a *random sparse matrix* inspired by the work [AilChz2006], but we will not discuss these practical results and we direct the interested reader to [Mat2008].

Theorem 3.4.2 (Theorem 3.1 from [Mat2008]). *Let $\varepsilon \in (0, 1/2]$ and $\delta \in (0, 1)$, and define $m = C\varepsilon^{-2} \log \frac{\delta}{2}$, with C being a constant. Define a random linear map $\boldsymbol{T} : \mathbb{R}^N \to \mathbb{R}^m$ by*

$$[\boldsymbol{T}x]_i = \frac{1}{\sqrt{m}} \sum_{j=1}^{N} [\boldsymbol{R}]_{i,j} x_j, \quad i \in [m],$$

with the entries of $\boldsymbol{R}$ being centered random variables with unitary variance and uniform sub-Gaussian tail. Then, for every $x \in \mathbb{R}^N$, we have

$$\mathbb{P}\{(1-\varepsilon)\|x\|_2 \leq \|\boldsymbol{T}x\|_2 \leq (1-\varepsilon)\|x\|_2\} \geq 1 - \delta.$$

This version of the Lemma can be easily proved through the following results:

Lemma 3.4.1 (Lemma 2.2 from [Mat2008]). *Let $X_1, \ldots, X_N$ be independent random variables with zero mean, unitary variance and uniform sub-Gaussian tail. Let also $\alpha_1, \ldots, \alpha_N$ be real coefficients satisfying*

$$\alpha_1^2 + \ldots + \alpha_N^2 = 1.$$

Then the sum

$$Y = \alpha_1 X_1 + \ldots + \alpha_N X_N$$

also has zero mean, unitary variance and a sub-Gaussian tail.

Proposition 3.4.2 (Proposition 3.2 from [Mat2008]). *Let $m \geq 1$ be an integer. Let $Y_1, \ldots, Y_m$ be independent random variables with zero mean, unitary variance and with uniform sub-Gaussian tail. Then*

$$Z = \frac{1}{\sqrt{m}} \left(\sum_{i=1}^{m} Y_i^2 - m \right)$$

has a sub-Gaussian tail up to $\sqrt{m}$.

More precisely, we have that:

Matoušek's proof for the JL-Lemma. Let $x \in \mathbb{R}^N$ be a fixed unit vector and let us set

$$Y_i = \sum_{j=1}^{N} [\boldsymbol{R}]_{i,j} x_j.$$

Then by Lemma 3.4.1, the variables Y_i are centered, have unitary variance and uniform sub-Gaussian tail with constant that we will call a. So Proposition 3.4.2 applies and shows that

$$Z = \frac{1}{\sqrt{m}} \left(\sum_{i=1}^{m} Y_i^2 - m \right)$$

has a sub-Gaussian tail up to $\sqrt{m}$. We note that for a fixed unitary x, the quantity $\|\boldsymbol{T}x\|_2^2 - 1$ has the same distribution as $Z/\sqrt{m}$. Thus, still with a fixed unitary x,

$$\mathbb{P}\{\|\boldsymbol{T}x\|_2 \geq 1+\varepsilon\} \leq \mathbb{P}\{\|\boldsymbol{T}x\|_2^2 \geq 1+2\varepsilon\} = \mathbb{P}\{Z \geq 2\varepsilon\sqrt{m}\}.$$

Since we assume $\varepsilon \leq 1/2$, we are in the allowed range and the last probability is at most

$$\exp\{-a(2\varepsilon\sqrt{m})^2\} = \exp\{-4a\varepsilon^2 C\varepsilon^{-2}\log(2/\delta)\} \leq 1/2\delta,$$

for $C \geq 1/(2a)$. The calculation showing that $\mathbb{P}\{\|\boldsymbol{T}x\|_2 \leq 1-\varepsilon\} \leq 1/2\delta$ is almost the same and will be omitted. □

For brevity, we will not discuss the suitability for applications of the JL-Lemma's versions presented so far. However, we must direct the reader to M.Sc. dissertation of John Fedoruk [Fdk2016][5], that inspired many parts of this text, and discuss in deeper details the practical aspects of the Matoušek's JL-Lemma and also present improved versions of the results in [Mat2008].

3.5 Closing the circle

Let us review what we have presented so far in this Chapter. The pioneering proofs of the JL-Lemma were based on tough geometrical arguments, namely the concentration of measure on certain subsets of the unit sphere. Next, many assumption on the JL-embedding were dropped by selecting its entries from simple probability distributions and the standard Gaussian one, yielding way easier proofs and better lower bounds on the dimension of the projected space. Finally, Matoušek present the linkage among such "simple distributions": their tails have a sub-Gaussian decay.

We thus finish this Chapter by highlighting the interesting fact that the *sub-Gaussian decay* was exactly the core of the geometrical proofs of the JL-Lemma. That is, the first approaches had already found this underlying idea behind the Lemma, but they were very difficult and poorly suitable for applications. Then, their successors took a safety distance from the JL-Lemma's theoretical aspects, substituting them for simple results from Probability and Statistics. The circle is closed when a more formal treatment of these probabilistic ideas direct us back to the same central idea but with a simpler and more practical treatment: the sub-Gaussian concentration of measure inequalities.

Consequently, we may revisit the past proofs reaching the hard parts through this backdoor of the sub-Gaussian paradigm. Indeed, this is a rough summary of the JL-Lemma due to one of the main texts that has inspired this one: Vershinyn's *High-Dimensional Probability: An Introduction with Applications in Data Science* [Vsh2018]. The approach through sub-Gaussian variables done in that text yields a simple treatment of the results that were exhibited here and also of some that we will not discuss.

[5]This text was also published as a paper in 2018 [FSJH2018]

Chapter 4

JL-Lemma optimality

4.1 Introduction

Let us recall some important facts about the JL-Lemma. Firstly presented just as a secondary result by W. B. Johnson and J. Lindenstrauss in [JL1984], the JL-Lemma is nowadays a paradigm for *dimensionality reduction*. In fact, the dimensionality reduction aims to project an M-point dataset in $\mathbb{R}^N$ into a subspace of lower dimension $\mathbb{R}^m$ so that the projected data is similar to the original in a certain aspect. A first idea is trying to project the data isometrically, but it is potentially problematic since not all sets can be isometrically projected in a lower dimension subspace, for instance, the N-simplex in $\mathbb{R}^N$. That being said, the main idea of the JL-Lemma as a dimensionality reduction paradigm is that, if we substitute the isometric projection for a *quasi-isometric (or ε-isometric, with $\varepsilon > 0$)* one, we may achieve a projection into a subspace whose dimension m has the logarithmic order of the cardinality of the dataset, M, i.e., $m = \mathcal{O}\left(\varepsilon^{-2} \log M\right)$.

Along the third chapter of the present text we exhibited an almost comprehensive sequence of versions of the JL-Lemma. Throughout time, these versions resulted in projections into subspaces whose dimension was smaller than the one of their predecessors, but always with $m = \mathcal{O}\left(\varepsilon^{-2} \log M\right)$. This raises the question if the JL-Lemma yields an *optimal projection* or, equivalently, if the dimension m is a *sharp, tight or optimal value*, i.e, if there is a dataset $X \subset \mathbb{R}^N$ that attains such m, concluding that $m = \Omega(\varepsilon^{-2} \log M)$.

In the next section, we shall stress some assumptions made on the JL-embedding along time that simplified a lot the proof of the JL-Lemma. The outcome of such discussion is the distinction between the JL-Lemma and the Distributional JL-Lemma, important but frequently set aside in the literature. In the third section, we explain in more details the optimality problem for the JL-Lemma and exhibit a noncomprehensive list of the attempts to solve it along the time from the inception of the JL-Lemma [JL1984] to the proof of its optimality [LN2017]. We start the fourth section by presenting a brief overview of Larsen and Nelson's optimality result for a linear JL-embedding. Later, we present the proof of this Theorem as in [LN2016]. Finally, in the last section, we make an overview of the Larsen and Nelson's optimality result for a non-linear JL-embedding [LN2017]. Since a formal proof of this result require a lot of concepts that will not be presented in this text like *Coding Theory and the Geometry of Convex Bodies*, we shall not present such proof and we direct

the interested reader to the paper of Larsen and Nelson [LN2017].

4.2 Important remarks on the JL-Lemma

Let us recall the JL-Lemma statement.

Theorem 4.2.1 (JL-Lemma)**.** *Let $X \in \mathbb{R}^N$ be a M-points dataset. For a fixed $\varepsilon > 0$, and a natural $m = \mathcal{O}\left(\varepsilon^{-2} \log M\right)$, there exists an embedding $f : \mathbb{R}^N \to \mathbb{R}^m$ such that*

$$(1-\varepsilon)\|u-v\|_2 \leq \|f(u)-f(v)\|_2 \leq (1+\varepsilon)\|u-v\|_2, \quad \forall u, v \in X.$$

This surprising result yields a powerful paradigm towards dimensionality reduction. Although one should expect this result to have a frightening and highly theoretical proof, the approach to the JL-Lemma, since its first appearance in [JL1984], was always the one of proving a simpler result that implies it. In this section, we shall discuss these workarounds more deeply.

4.2.1 Random projection argument

Instead of determining, for any given $X \subset \mathbb{R}^N, \varepsilon > 0$ and $m = \mathcal{O}\left(\varepsilon^{-2} \log M\right)$, an ε embedding $f : \mathbb{R}^N \to \mathbb{R}^m$ of the vectors in X, all the proofs exhibited in the third chapter of this text were based in an argument called *random projection.* In fact, the idea is not to define an f satisfying the JL-Lemma for given a fixed X, but instead to select f from a probability distribution that yields a JL-embedding of X with positive probability, which proves its existence.

Namely, in [JL1984, FklMae1988, IndMot1998], the embedding f was assumed to be an orthogonal linear projection into a randomly chosen m-dimensional subspace of $\mathbb{R}^N$. Also, in [DasGup2003, Alo2003], the embedding $f \in \mathcal{M}_{N\times m}$ was assumed to be a random matrix with standard Gaussian entries, dropping the orthogonality assumption since its columns are orthogonal with high probability [Mrk1994]. Finally, in [Ach2003, Mat2008], the entries of f are selected from a probability distribution with sub-Gaussian tails.

4.2.2 On the linearity of JL-embedding

We must recall that all known proofs of JL-Lemma presented in Chapter 3 have made the assumption that the JL-embedding $f : \ell_2^N \to \ell_2^m$ is linear. However, it does not diminish its relevance for two main reasons. Firstly, the linearity assumption on f is important in several applications of the JL-Lemma. We may consider, for example, its application to *Compressed Sensing*. In this area, one wishes to (approximately) recover (approximately) sparse signals using few linear measurements [Don2006, CT2005]. We have that the map f, representing a fixed set of measurements of the signal, allows good signal recovery if f satisfies the JL guarantee for the set of all k-sparse vectors, being $k \in \mathbb{N}$ a fixed constant [CT2005]. Secondly, in [LN2017], Larsen and Nelson proved that surprisingly the linearity assumption is not necessary for sharpness.

4.2.3 Distributional JL-Lemma (DJL-Lemma)

It is common in the literature not to make distinction between JL-Lemma and DJL-Lemma. However, they are somewhat distinct: in the JL-Lemma, our goal is to determine a map $f : \ell_2^N \to \ell_2^m$ with m as small as possible such that, given a dataset $X \subset \mathbb{R}^N$ and a distortion $\varepsilon > 0$, f projects X into ℓ_2^m ε-isometrically; on the other hand, in the DJL-Lemma, our goal is to provide a distribution $\mathcal{D}_{\varepsilon,\delta}$, with $\delta < 1/2$ over the set $\mathcal{L}(\ell_2^N, \ell_2^m)$ of linear maps $f : \ell_2^N \to \ell_2^m$ with m as small as possible such that any vector in $\mathbb{R}^N$ has, with a probability greater than $1 - \delta$, its norm distorted by a factor of at most $1 + \varepsilon$.

More precisely, we have the following:

Theorem 4.2.2 (DJL-Lemma)**.** *For any $N > 1$, $\varepsilon > 0$ and $\delta < 1/2$, there is a probability distribution $\mathcal{D}_{\varepsilon,\delta}$ over $\mathcal{L}(\ell_2^N, \ell_2^m)$ for some $m = \mathcal{O}\{\varepsilon^{-2} \log(1/\delta)\}$ such that*

$$\mathop{\mathbb{P}}_{f\sim\mathcal{D}_{\varepsilon,\delta}}\big\{(1-\varepsilon)\|x\|_2 \le \|f(x)\|_2 \le (1+\varepsilon)\|x\|_2\big\} > 1-\delta, \quad \forall x \in \mathbb{R}^N.$$

Also, note that:

Proposition 4.2.1. *The JL-Lemma is a particular case of the DJL-Lemma.*

Proof. In fact, with the assumption of a linear $f : \ell_2^N \to \ell_2^m$, the condition of a ε-isometry in the JL-Lemma, i.e.,

$$(1-\varepsilon)\|u-v\|_2 \le \|f(u)-f(v)\|_2 \le (1+\varepsilon)\|u-v\|_2, \quad \forall u, v \in X$$

is equivalent to the norm distortion, i.e.,

$$(1-\varepsilon)\|x\|_2 \le \|f(x)\|_2 \le (1+\varepsilon)\|x\|_2, \quad \forall x \in X - X.$$

To conclude the proof, it suffices to choose a suitable $\delta < 1/2$ so that the original JL-Lemma holds with positive probability. In order to do so, take $\delta < 1/2 \times 1/\binom{M}{2}$ and then perform a union bound over all

$$x \in X - X = \{u - v \,:\, u \neq v \in X\}.$$

More precisely, let $X = \{x_1, \ldots, x_M\} \subset \mathbb{R}^N$. We have that

$$\begin{aligned}
&\mathop{\mathbb{P}}_{f\sim\mathcal{D}_{\varepsilon,\delta}}\big\{(1-\varepsilon)\|x\|_2 \le \|f(x)\|_2 \le (1+\varepsilon)\|x\|_2, \forall x \in X - X\big\} \\
&= 1 - \mathop{\mathbb{P}}_{f\sim\mathcal{D}_{\varepsilon,\delta}}\big\{\|f(x)\|_2 \notin ((1-\varepsilon)\|x\|_2, (1+\varepsilon)\|x\|_2), \text{ for some } x \in X - X\big\} \\
&\ge 1 - \sum_{1\le i<j\le M} \mathop{\mathbb{P}}_{f\sim\mathcal{D}_{\varepsilon,\delta}}\big\{\|f(x_j)-f(x_i)\|_2 \notin ((1-\varepsilon)\|x_j-x_i\|_2, (1+\varepsilon)\|x_j-x_i\|_2)\big\} \\
&> 1 - \binom{M}{2}\delta > 0.
\end{aligned}$$

□

4.3 Past results on tightness of JL-Lemma

4.3.1 Introduction

The problem of tightness of an estimation, by lower or upper bounds, refers to whether these bounds are attained by some of the variables being estimated. Regarding the JL-Lemma, in its first appearance in [JL1984], W. B. Johnson and J. Lindenstrauss proved that $m = \mathcal{O}\left(\varepsilon^{-2} \log M\right)$. We will say that the Lemma is tight, sharp or that its result is optimal if there is some M-size set $X \subset \mathbb{R}^N$ such that any map JL map, $f : \ell_2^N \to \ell_2^m$ must have $m = \Omega(\varepsilon^{-2} \log M)$.

The historical approach toward solving JL-Lemma's tightness problem has been done by exhibiting examples of sets yielding higher lower bounds on m. In the present section, we shall exhibit some of these results. However, our exposition is not comprehensive, excluding, for brevity, some important investigations such as the sharpness when the JL-embedding is a matrix satisfying the *Restricted Isometry Property (RIP)* (see [KW2011]).

4.3.2 Johnson and Lindenstrauss – 1984

The first step towards answering the optimality problem of the JL-Lemma was given in its very inception. In [JL1984], Johnson and Lindenstrauss argued that the projected dimension cannot have a dependence that is smaller than logarithmic in M, or, more precisely, such that $m = \Omega(\log M)$.

For the sake of brevity, we shall just exhibit an overview to the proof of this fact and direct the interested reader to [JL1984]. Namely, they used a result about ε-nets to conclude that in a ball with radius 2 in ℓ_2^m, there are at most 4^m vectors whose pairwise distance is at least 1. Consequently, for a sufficiently small distortion ε, there is no ε-isometry that projects an orthonormal set with more than 4^m vectors in a m-dimensional subspace of ℓ_2^N.

4.3.3 Noga Alon – 2003

Later, in 2003 [Alo2003], Noga Alon exhibited an example that almost attains a dimension $m = \mathcal{O}\left(\varepsilon^{-2} \log M\right)$. More precisely, he proved the following:

Theorem 4.3.1 (Alon's bound)**.** *If $X \subset \mathbb{R}^N$ is the N-simplex, i.e., $X = \{\mathbf{0}_N, e_1, \ldots, e_N\}$, with $M = N + 1$ and $\varepsilon \in (0, 1/2)$, then any JL-map, $f : X \to \ell_2^m$, must satisfy*

$$m = \Omega\left(\min\left\{N, \varepsilon^{-2} \frac{log\, N}{log(1/\varepsilon)}\right\}\right).$$

Even though the bound $m = \mathcal{O}\left(\varepsilon^{-2} \log M\right)$ is almost attained in Theorem 4.3.1, the term $\log(1/\varepsilon)$ is still not optimal. Furthermore, Alon [Alo2003] proved that this term cannot be removed for this particular set X, undermining completely the hope of a simplex yielding an optimal m (see [NNW12] for details).

Alon's work raises the question if the JL-Lemma is suboptimal for any set, i.e., if there is no $X \subset \mathbb{R}^N$ that can be projected with $m = \mathcal{O}\left(\varepsilon^{-2} \log M\right)$. This was a major open question in the area of dimensionality reduction that was still unsolved for 30 years until the work of Larsen and Nelson in [LN2016].

4.3.4 Larsen & Nelson – 2014

In 2014 [LN2016][1], Larsen and Nelson exhibited an example of a set $X \subset \mathbb{R}^N$ yielding $m = \mathcal{O}\left(\varepsilon^{-2} \log M\right)$ when the JL-embedding $f : \ell_2^N \to \ell_2^m$ is linear. More precisely, they proved the following:

Theorem 4.3.2 (Larsen & Nelson – 2014)**.** *For any $N > 1$ and $\varepsilon \in (0, 1/2)$, there is an $N^{\mathcal{O}(1)}$-point $X \subset \mathbb{R}^N$ such that any linear map $f : (X, \ell_2) \mapsto \ell_2^m$ satisfying the JL-guarantee must have:*

$$m = \Omega(min\{N, \varepsilon^{-2} \log N\}).$$

Notice this result improves Alon's [Alo2003] lower bound by getting rid of the $\log(1/\varepsilon)$ factor. This result is sharp since the identity map achieves the first term in the minimum and the JL-Lemma provides

$$m = \mathcal{O}\left(\varepsilon^{-2} \log M\right) = \mathcal{O}\left(\varepsilon^{-2} \log N^{\mathcal{O}(1)}\right) = \mathcal{O}\left(\varepsilon^{-2} \log N\right).$$

However, the optimality problem for the JL-Lemma is still unsolved since it may be suboptimal for non-linear embeddings f.

4.3.5 Larsen & Nelson – 2016

In 2016 [LN2017][2], the problem of JL-Lemma optimality was finally settled for nearly the full range of ε of interest. The same authors of [LN2016] exhibited an example attaining $m = \mathcal{O}\left(\varepsilon^{-2} \log M\right)$. More precisely, Larsen and Nelson proved the following:

Theorem 4.3.3. *For any integers $M, N \geq 2$ and $\varepsilon \in \left(\frac{\log^{0.5001} M}{\sqrt{\min\{M,N\}}}, 1\right)$, there is a set of points $X \in \mathbb{R}^N$ of size M, such that any map $f : (X, \ell_2) \to \ell_2^m$ providing the JL-guarantee must have*

$$m = \Omega(\varepsilon^{-2} \log(\varepsilon^2 M)).$$

4.4 Proof of Larsen & Nelson's Theorem – 2014

4.4.1 Overview of the proof of Larsen & Nelson's Theorem – 2014

The main result from Larsen and Nelson presented in [LN2016] claims the existence of a M-set $X \subset \mathbb{R}^N$ such that the JL-Lemma yields a linear projection, here denoted as a matrix $\boldsymbol{A} \in \mathcal{M}_{m \times N}$, of this set in a subspace whose dimension attains the lower bound $m = \mathcal{O}\left(\varepsilon^{-2} \log M\right)$. The construction of such set is done via a probabilistic argument. More precisely, X will be the union of the N vectors $\{e_1, \ldots, e_N\}$ of the canonical basis of $\mathbb{R}^N$ together with several independent Gaussian vectors.

At first, we make $X = \{e_1, \ldots, e_N\}$. Thus, if $\boldsymbol{A} \in M_{m \times N}$, $m \leq N$ is a ε-isometry in X, i.e.,

$$(1 - \varepsilon)\|x\|_2^2 \leq \|\boldsymbol{A}\, x\|_2^2 \leq (1 - \varepsilon)\|x\|_2^2, \quad \forall x \in X, \tag{4.4.1}$$

[1] This paper of 2014 was later published as a conference paper in 2016.
[2] This paper of 2016 was later published as a conference paper in 2017.

then its columns must heve nearly unit norm. Then, the results from covering and packing numbers will imply the existence of a family of matrices

$$\Phi \subset \bigcup_{t=1}^{N} \mathcal{M}_{t\times N}, \qquad \text{with} \qquad \#\Phi = \exp\{\mathcal{O}\left(N^2 \log N\right)\}$$

such that

$$\inf_{\hat{\boldsymbol{A}}\in\Phi\cap\mathcal{M}_{m\times N}} \|\boldsymbol{A} - \hat{\boldsymbol{A}}\|_{\mathcal{F}} \leq \frac{1}{N^C}, \tag{4.4.2}$$

with $C > 0$ being as large as desired. Also, by a Theorem of Latała [Lat1999], for any $\hat{\boldsymbol{A}} \in \Phi$ and a random Gaussian vector $g \in \mathbb{R}^N$, we have that

$$\mathbb{P}_g\left\{ \left| \|\hat{\boldsymbol{A}} g\|_2^2 - \mathrm{tr}(\hat{\boldsymbol{A}}^t \hat{\boldsymbol{A}}) \right| \geq \Omega\{\sqrt{\log(1/\delta)}\, \|\hat{\boldsymbol{A}}^t \hat{\boldsymbol{A}}\|_{\mathcal{F}}\} \right\} \geq \delta, \tag{4.4.3}$$

for any $0 < \delta < 1/2$.

Next, if follows from the standard Gaussian concentration of measure that a random Gaussian vector satisfies

$$\mathbb{P}_g\left\{ |\|g\|_2^2 - N| > C\sqrt{N\log(1/\delta)} \right\} < \delta/2. \tag{4.4.4}$$

Thus, by applying a union bound, we can conclude that the events in Equations 4.4.3 and 4.4.4 happen simultaneously with probability $\Omega(\delta)$. Consequently, if we take N independent random Gaussian vectors, the probability that the events in Equations 4.4.3 and 4.4.4 never happen simultaneously for any of these N vectors is at most

$$\{1 - \Omega(\delta)\}^N \approx \exp\{-\Omega(\delta N)\}.$$

Now, by taking a sufficiently large N and $\delta = 1/\mathrm{poly}(N)$, it is possible to show through a union bound over Φ that for every $\hat{\boldsymbol{A}} \in \Phi$, one of the N Gaussian vectors satisfies the events in Equations 4.4.3 and 4.4.4 simultaneously. More precisely, there exist $M = \mathcal{O}\left(N^3\right)$ vectors $\{v_1, \ldots, v_M\} := V \subset \mathbb{R}^N$ such that

$$\|v\|_2^2 = N \pm \mathcal{O}\left(\sqrt{N\log N}\right), \quad \forall v \in V; \tag{4.4.5}$$

and, for any $\hat{\boldsymbol{A}} \in \Phi$, there exists some $v \in V$ such that

$$\left| \|\hat{\boldsymbol{A}} v\|_2^2 - \mathrm{tr}(\hat{\boldsymbol{A}}^t \hat{\boldsymbol{A}}) \right| = \Omega(\sqrt{\log N}\, \|\hat{\boldsymbol{A}}\|_{\mathcal{F}}). \tag{4.4.6}$$

The final definition of X is $\{e_1, \ldots, e_N\} \cup V$. Then, using Equations 4.4.1 and 4.4.2, we show that Equation 4.4.6 implies

$$\mathrm{tr}(\hat{\boldsymbol{A}}^t \hat{\boldsymbol{A}}) = N \pm \mathcal{O}\left(\varepsilon N\right), \tag{4.4.7}$$

with $\pm B$ representing a value in $[-B, B]$; and

$$\left| \|\boldsymbol{A}\, v\|_2^2 - N \right| = \Omega(\sqrt{\log N}\, \|\hat{\boldsymbol{A}}^t \hat{\boldsymbol{A}}\|_{\mathcal{F}}) - \mathcal{O}\left(\varepsilon N\right). \tag{4.4.8}$$

Therefore, by the triangle inequality, Equations 4.4.5 and 4.4.1 imply

$$\left| \|\boldsymbol{A}\, v\|_2^2 - N \right| \leq \left| \|\boldsymbol{A}\, v\|_2^2 - \|v\|_2^2 \right| - \left| \|v\|_2^2 - N \right| = \mathcal{O}\left(\varepsilon N + \sqrt{N \log N}\right).$$

Finally, combining Equations 4.4.7 and 4.4.8, we obtain

$$\mathrm{tr}(\hat{\boldsymbol{A}}^t \hat{\boldsymbol{A}}) = \sum_{i=1}^{N} \hat{\lambda}_i \geq \{1 - \mathcal{O}(\varepsilon)\} N;$$

and

$$\|\hat{\boldsymbol{A}}^t \hat{\boldsymbol{A}}\|_{\mathcal{F}}^2 = \sum_{i=1}^{N} \hat{\lambda}_i^2 = \mathcal{O}\left(\frac{\varepsilon^2 N^2}{\log N} + N\right),$$

with $\{\hat{\lambda}_i\}_{i\in[N]}$ being the eigenvalues of $\hat{\boldsymbol{A}}^t \hat{\boldsymbol{A}}$. With bounds on $\sum_i \hat{\lambda}_i$ and $\sum_i \hat{\lambda}_i^2$ in hand, a lower bound on $\mathrm{rank}(\hat{\boldsymbol{A}}^t \hat{\boldsymbol{A}}) \leq m$ follows by the Cauchy-Schwarz-Bunyakovsky inequality.

Next, we shall exhibit this proof with more details. Namely, we will present some preliminary results in the Section 4.4.2 and then the proof itself in the Section 4.4.3.

4.4.2 Preliminaries

At first, we exhibit the following result without proving it:

Theorem 4.4.1 (Latała – 1999 [Lat1999])**.** *There is a universal constant $c > 0$ such that for $g \in \mathbb{R}^N$ standard Gaussian random vector and a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ with vanishing diagonal elements, we have:*

$$\forall t \geq 1, \quad \mathbb{P}_g\big\{|g^T \boldsymbol{A} g| > c(\sqrt{t} \cdot \|\boldsymbol{A}\|_{\mathcal{F}} + t \cdot \|\boldsymbol{A}\|_{2\to 2})\big\} \geq \min\{c, e^{-t}\}$$

Also, this Theorem implies the following corollary:

Corollary 4.4.1. *Let g and $\boldsymbol{A}$ be as in Theorem 4.4.1, except for $\boldsymbol{A}$ being no longer restricted to have zero diagonal. Then there is a universal constant $c > 0$ such that*

$$\forall t \geq 1, \quad \mathbb{P}_g\{|g^T \boldsymbol{A} g - tr(\boldsymbol{A})| > c(\sqrt{t} \cdot \|\boldsymbol{A}\|_{\mathcal{F}} + t \cdot \|\boldsymbol{A}\|_{2\to 2})\} \geq \min\{c, e^{-t}\}.$$

Proof of the Corollary. As said before, this Corollary follows from Theorem 4.4.1. However, we will not apply it directly to matrix $\boldsymbol{A}$. We will instead construct an auxiliary matrix and conclude the proof by applying Theorem 4.4.1 and the *Law of Large Numbers*. The construction is as follows.

Let M be a positive integer and define

$$\tilde{g} = (\tilde{g}_{1,1}, \tilde{g}_{1,2}, \ldots, \tilde{g}_{1,M}, \ldots, \tilde{g}_{N,1}, \tilde{g}_{N,2}, \ldots, \tilde{g}_{N,M}),$$

an MN-standard Gaussian random vector. Then g_i is equal in distribution to $M^{-1/2} \sum_{j=1}^{M} \tilde{g}_{i,j}$.

Next, define $\tilde{\boldsymbol{A}}_{\boldsymbol{M}}$ as the $NM \times NM$ matrix formed by converting each entry a_{ij} of $\boldsymbol{A}$ into an $M \times M$ block with each entry being a_{ij}/M. We now claim that

$$g^T \boldsymbol{A} g - \operatorname{tr}(\boldsymbol{A}) \quad \text{and} \quad \tilde{g}^T \tilde{\boldsymbol{A}}_{\boldsymbol{M}} \tilde{g} - \operatorname{tr}(\tilde{\boldsymbol{A}}_{\boldsymbol{M}})$$

are equal in distribution.

In fact, $\operatorname{tr}(\boldsymbol{A}) = \operatorname{tr}(\tilde{\boldsymbol{A}}_{\boldsymbol{M}})$. Moreover,

$$\begin{aligned} g^T \boldsymbol{A} g - \operatorname{tr}(\boldsymbol{A}) &= \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij} g_i g_j - \operatorname{tr}(\boldsymbol{A}) \\ &\overset{d}{=} \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{r=1}^{M}\sum_{s=1}^{M} \frac{a_{ij}}{M} \tilde{g}_{i,r} \tilde{g}_{j,s} - \operatorname{tr}(\boldsymbol{A}) \\ &= \tilde{g}^T \tilde{\boldsymbol{A}}_{\boldsymbol{M}} \tilde{g} - \operatorname{tr}(\tilde{\boldsymbol{A}}_{\boldsymbol{M}}), \end{aligned}$$

with $\overset{d}{=}$ denoting the equality in distribution.

Consequently, for any $\forall t \geq 1$, the probability of

$$|g^T \boldsymbol{A} g - \operatorname{tr}(\boldsymbol{A})| > c(\sqrt{t} \cdot \|\boldsymbol{A}\|_{\mathcal{F}} + t \cdot \|\boldsymbol{A}\|_{2\to 2})$$

equals the probability of

$$|\tilde{g}^T \tilde{\boldsymbol{A}}_{\boldsymbol{M}} \tilde{g} - \operatorname{tr}(\tilde{\boldsymbol{A}}_{\boldsymbol{M}})| > c(\sqrt{t} \cdot \|\boldsymbol{A}\|_{\mathcal{F}} + t \cdot \|\boldsymbol{A}\|_{2\to 2}). \tag{4.4.9}$$

This is the first step towards substituting $\boldsymbol{A}$ by $\tilde{\boldsymbol{A}}_{\boldsymbol{M}}$ in the statement of the Corollary. We still have to substitute the terms in $\boldsymbol{A}$ in the right hand side of Equation 4.4.9, but we will first substitute its left hand side by a simpler expression in $\tilde{\boldsymbol{A}}_{\boldsymbol{M}}$. Namely, we will show, by the Weak Law of Large Numbers, that

$$\forall \lambda > 0, \lim_{M\to\infty} \mathbb{P}_g\left(|\tilde{g}^T \tilde{\boldsymbol{A}}_{\boldsymbol{M}} \tilde{g} - \operatorname{tr}(\tilde{\boldsymbol{A}}_{\boldsymbol{M}})| > \lambda\right) = \lim_{M\to\infty} \mathbb{P}_g\left(|\tilde{g}^T (\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}}) \tilde{g}| > \lambda\right), \tag{4.4.10}$$

with $\tilde{\boldsymbol{D}}_{\boldsymbol{M}}$ being a diagonal matrix containing the diagonal elements of $\tilde{\boldsymbol{A}}_{\boldsymbol{M}}$. Indeed, the Law of Large Numbers yields

$$\begin{aligned} \sum_{j=1}^{M} \frac{\tilde{g}_{ij}^2}{M} &\xrightarrow{p} \mathbb{E}\left\{\sum_{j=1}^{M} \frac{\tilde{g}_{ij}^2}{M}\right\} \\ &= \sum_{j=1}^{M} \frac{\mathbb{E}\tilde{g}_{ij}^2}{M} \\ &= \sum_{j=1}^{M} \frac{\operatorname{Var}(\tilde{g}_{ij})}{M} = 1, \end{aligned}$$

with this penultimate equality being justified by the fact that

$$\mathbb{E}\,\tilde{g}_{ij} = 0,\ \forall i \in [N] \ \text{ and } \ \forall j \in [M].$$

Finally, note that

$$\begin{aligned}\tilde{g}^T \tilde{\boldsymbol{D}}_{\boldsymbol{M}} \tilde{g} &= \sum_{i=1}^{N} \left(\sum_{j=1}^{M} \tilde{g}_{ij}^2 \right) \frac{a_{ii}}{M} \\ &= \sum_{i=1}^{N} \left(\sum_{j=1}^{M} \frac{\tilde{g}_{ij}^2}{M} \right) a_{ii} \xrightarrow{p} \sum_{i=1}^{N} a_{ii} = \operatorname{tr}(\boldsymbol{A}),\end{aligned}$$

which results in Equation 4.4.10. Consequently, for any $\forall t \geq 1$,

$$\mathbb{P}_g\{|g^T \boldsymbol{A} g - \operatorname{tr}(\boldsymbol{A})| > c(\sqrt{t} \cdot \|\boldsymbol{A}\|_{\mathcal{F}} + t \cdot \|\boldsymbol{A}\|_{2\to 2})\}$$

equals

$$\lim_{M\to\infty} \mathbb{P}_g\{|\tilde{g}^T \tilde{\boldsymbol{A}}_{\boldsymbol{M}} \tilde{g} - \operatorname{tr}(\tilde{\boldsymbol{A}}_{\boldsymbol{M}})| > c(\sqrt{t} \cdot \|\boldsymbol{A}\|_{\mathcal{F}} + t \cdot \|\boldsymbol{A}\|_{2\to 2})\}. \tag{4.4.11}$$

The next step now is to substitute $\boldsymbol{A}$ by $\tilde{\boldsymbol{A}}_{\boldsymbol{M}}$ in the right hand side of Equation 4.4.11. In order to do so, we prove that

$$\lim_{M\to\infty} \|\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{2\to 2} = \|\boldsymbol{A}\|_{2\to 2}. \tag{4.4.12}$$

In fact, since $\tilde{\boldsymbol{D}}_{\boldsymbol{M}}$ is diagonal, its non-zero elements are just its singular values, with the larger one being $\|\tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{2\to 2}$. This yields

$$|\,\|\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{2\to 2} - \|\boldsymbol{A}\|_{2\to 2}\,| \leq \|\tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{2\to 2} = \max_i |a_{ii}|/M$$

by the triangle inequality. Therefore

$$\lim_{M\to\infty} \|\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{2\to 2} = \|\boldsymbol{A}\|_{2\to 2}.$$

Also, by the equivalence of these matrix norms, the result is also valid for $\|\cdot\|_{\mathcal{F}}$. More precisely,

$$\lim_{M\to\infty} \|\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{\mathcal{F}} = \|\boldsymbol{A}\|_{\mathcal{F}}.$$

Consequently, for all $t \geq 1$, we may rewrite the right hand side of Equation 4.4.11 as follows:

$$c(\sqrt{t} \cdot \|\boldsymbol{A}\|_{\mathcal{F}} + t \cdot \|\boldsymbol{A}\|_{2\to 2}) = c\left(\sqrt{t} \cdot \lim_{M\to\infty} \|\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{\mathcal{F}} + t \cdot \lim_{M\to\infty} \|\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{2\to 2}\right).$$

Finally, for all $t \geq 1$, we have

$$\begin{aligned}
&\mathbb{P}_g\left\{|g^T\boldsymbol{A}g - \operatorname{tr}(\boldsymbol{A})| > c(\sqrt{t}\cdot\|\boldsymbol{A}\|_{\mathcal{F}} + t\cdot\|\boldsymbol{A}\|_{2\to 2})\right\}\\
&= \mathbb{P}_{\tilde{g}}\left\{\left|\,\tilde{g}^T\tilde{\boldsymbol{A}}_{\boldsymbol{M}}\tilde{g} - \operatorname{tr}(\tilde{\boldsymbol{A}}_{\boldsymbol{M}})\,\right| > c\left(\sqrt{t}\cdot\lim_{M\to\infty}\|\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{\mathcal{F}} + t\cdot\lim_{M\to\infty}\|\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{2\to 2}\right)\right\}\\
&= \lim_{M\to\infty}\mathbb{P}_{\tilde{g}}\left\{\left|\,\tilde{g}^T\tilde{\boldsymbol{A}}_{\boldsymbol{M}}\tilde{g} - \operatorname{tr}(\tilde{\boldsymbol{A}}_{\boldsymbol{M}})\,\right| > c\left(\sqrt{t}\cdot\|\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{\mathcal{F}} + t\cdot\|\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{2\to 2}\right)\right\}\\
&= \lim_{M\to\infty}\mathbb{P}_{\tilde{g}}\left\{\left|\,\tilde{g}^T(\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}})\tilde{g}\,\right| > c\left(\sqrt{t}\cdot\|\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{\mathcal{F}} + t\cdot\|\tilde{\boldsymbol{A}}_{\boldsymbol{M}} - \tilde{\boldsymbol{D}}_{\boldsymbol{M}}\|_{2\to 2}\right)\right\},
\end{aligned}$$

with the interchange of the limit being justified by the fact that $\tilde{g}^T\tilde{\boldsymbol{A}}_{\boldsymbol{M}}\tilde{g} - \operatorname{tr}(\tilde{\boldsymbol{A}}_{\boldsymbol{M}})$ has the same distribution of $g^T\boldsymbol{A}g$ for any natural M. Consequently, the claim in this Corollary is reduced to the zero diagonal case and we can apply Theorem 4.4.1 to prove the desired result. □

Lemma 4.4.1. *For some universal constant $c > 0$ and $g \in \mathbb{R}^N$ a standard Gaussian random vector,*

$$\forall t > 0, \quad \mathbb{P}\left\{\left|\,\|g\|_2^2 - N\,\right| > c\sqrt{Nt}\right\} < e^{-t}.$$

We now state and prove a Corollary of this Lemma that will be useful later.

Corollary 4.4.2. *For $\boldsymbol{A} \in \mathbb{R}^{d\times N}$, let $\lambda_1 \geq \ldots \geq \lambda_N \geq 0$ be the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$ and also let $g^{(1)}, \ldots, g^{(M)} \in \mathbb{R}^N$ be independent standard Gaussian random vectors. For some universal constants $c_1, c_2, \delta_0 > 0$ and any $0 < \delta < \delta_0$*

$$\begin{aligned}
\mathbb{P}\Bigg(\nexists j \in [M] : \Bigg\{\Big|\|\boldsymbol{A}g^{(j)}\|_2^2 - \sum_{i=1}^N \lambda_i\Big| \geq c_1\sqrt{\log(1/\delta)}\Big(\sum_{i=1}^N \lambda_i^2\Big)^{1/2}\Bigg\}\\
\wedge\left\{\,\left|\,\|g^{(j)}\|_2^2 - N\,\right| \leq c_2\sqrt{N\,\log(1/\delta)}\right\}\Bigg) \leq e^{-M\delta}. \qquad (4.4.13)
\end{aligned}$$

Before starting the proof, we state the following Lemma.

Lemma 4.4.2. *For any $0 < \delta < 1$ and $M \in \mathbb{N}$, $(1-\delta)^M < e^{-\delta M}$.*

Proof of the Lemma. We argue by induction on M. At first, consider the case $M = 1$ and take the functions $f(\delta) = 1 - \delta$ and $g(\delta) = e^{-\delta}$. Clearly, $f(0) = g(0) = 1$. Moreover,

$$g'(\delta) = (-1)\,e^{-\delta} > -1 = f'(\delta), \quad \forall\delta \in (0,1)$$

and the result follows.

Finally, we take $M \in \mathbb{N}$ such that:

$$(1-\delta)^M < e^{-\delta M}, \quad \forall\delta \in (0,1).$$

Multiplying both sides of this inequality by $1-\delta$ and using the induction basis, yield the result:

$$(1-\delta)^{M+1} < (1-\delta)\,e^{-\delta M} < e^{-\delta}e^{-\delta M} = e^{-\delta(M+1)}.$$

□

Proof of the Corollary. We will show that, for any fixed $j \in [M]$, it holds that

$$\mathbb{P}\Bigg(\Big\{\big|\|\boldsymbol{A}g^{(j)}\|_2^2 - \sum_{i=1}^{N}\lambda_i\big| \le c_1\sqrt{\log(1/\delta)}\big(\sum_{i=1}^{N}\lambda_i^2\big)^{1/2}\Big\} \\ \vee\Big\{\ \big|\ \|g^{(j)}\|_2^2 - N\ \big| \ge c_2\sqrt{N\ \log(1/\delta)}\Big\}\Bigg) < 1-\delta. \tag{4.4.14}$$

Note that δ have to be less than 1 in order to Equation 4.4.14 make sense. As a consequence of Equation 4.4.14 and since the g_j are independent, we have that

$$\mathbb{P}\Bigg(\Big\{\Big|\|\boldsymbol{A}g^{(j)}\|_2^2 - \sum_{i=1}^{N}\lambda_i\Big| \le c_1\sqrt{\log(1/\delta)}\Big(\sum_{i=1}^{N}\lambda_i^2\Big)^{1/2}\Big\} \\ \vee\Big\{\ \big|\ \|g^{(j)}\|_2^2 - N\ \big| \ge c_2\sqrt{N\ \log(1/\delta)}\Big\}, \forall j \in [M]\Bigg) < (1-\delta)^M < e^{-\delta M}, \tag{4.4.15}$$

with this last inequality being resulted from Lemma 4.4.2. Moreover, note that Equation 4.4.15 is equivalent to 4.4.13.

Now, we must prove Equation 4.4.14. It suffices to show that

$$\mathbb{P}\left\{\left|\ \|\boldsymbol{A}\,g^{(j)}\|_2^2 - \sum_{i=1}^{N}\lambda_i\ \right| \ge c_1\sqrt{\log(1/\delta)}\ \left(\sum_{i=1}^{N}\lambda_i^2\right)^{1/2}\right\} > \frac{3}{2}\delta, \quad (\text{with } \delta < 2/3) \tag{4.4.16}$$

and

$$\mathbb{P}\left(\big|\ \|g^{(j)}\|_2^2 - N\ \big| \le c_2\sqrt{N\ \log(1/\delta)}\right) > 1 - \frac{1}{2}\delta, \tag{4.4.17}$$

since Equation 4.4.14 would then follow from a union bound.

For Equation 4.4.16, note that

1. $\|\boldsymbol{A}\,g^{(j)}\|_2^2 = {g^{(j)}}^T\boldsymbol{A}^T\boldsymbol{A}\ g^{(j)}$;
2. $\displaystyle\sum_{i=1}^{N}\lambda_i = \operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A})$;
3. $\displaystyle\left(\sum_{i=1}^{N}\lambda_i^2\right)^{1/2} = \|\boldsymbol{A}^T\boldsymbol{A}\|_{\mathcal{F}}$.

Then, Equation 4.4.16 follows from Corollary 4.4.1 for δ smaller than some sufficiently small constant δ_0, say $\delta_0 < 2/3$.

Indeed, by applying Corollary 4.4.1 to $\boldsymbol{A}^T\boldsymbol{A}$, we have, for all $t \geq 1$,

$$\mathbb{P}_g\left\{|g^{(j)^T}\boldsymbol{A}^T\boldsymbol{A}\ g^{(j)} - \operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A})| > c_1(\sqrt{t}\cdot\|\boldsymbol{A}^T\boldsymbol{A}\|_{\mathcal{F}} + t\cdot\|\boldsymbol{A}^T\boldsymbol{A}\|_{2\to 2})\right\} \geq \min\{c_1, e^{-t}\},$$

with $c_1 > 0$ being a universal constant. Since, for any fixed $t \geq 1$, the event above is a subset of

$$\left\{\left|\ g^{(j)^T}\boldsymbol{A}^T\boldsymbol{A}\ g^{(j)} - \operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A})\ \right| > c_1\sqrt{t}\cdot\|\boldsymbol{A}^T\boldsymbol{A}\|_{\mathcal{F}}\right\},$$

we have that

$$\mathbb{P}_g\left\{\left|\ g^{(j)^T}\boldsymbol{A}^T\boldsymbol{A}\ g^{(j)} - \operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A})\ \right| > c_1\sqrt{t}\cdot\|\boldsymbol{A}^T\boldsymbol{A}\|_{\mathcal{F}}\right\} \geq \min\{c, e^{-t}\}, \quad \forall t \geq 1.$$

Furthermore, we use the substitution made at the beginning of the present proof:

$$\forall t \geq 1, \quad \mathbb{P}_g\left\{\left|\ \|\boldsymbol{A}g^{(j)}\|_2^2 - \sum_i \lambda_i\ \right| > c_1\sqrt{t}\cdot\left(\sum_i \lambda_i^2\right)^{1/2}\right\} \geq \min\{c_1, e^{-t}\}.$$

In particular, for $t = \log(1/\delta^{\frac{3}{2}})$, we have

$$\begin{aligned}\mathbb{P}_g\left\{\left|\ \|\boldsymbol{A}g^{(j)}\|_2^2 - \sum_i \lambda_i\ \right| > c_1\sqrt{3/2}\sqrt{\log(1/\delta)}\cdot\left(\sum_i \lambda_i^2\right)^{1/2}\right\} &\geq \min\{c_1, e^{-\log(1/\delta^{\frac{3}{2}})}\} \\ &\geq \min\{c_1, e^{\frac{3}{2}\log(\delta)}\} \\ &= \min\{c_1, \delta^{\frac{3}{2}}\}.\end{aligned}$$

Finally, to make this probability larger than $3\,\delta/2$, requires

$$\min\{c_1, \delta^{\frac{3}{2}}\} > \frac{3\delta}{2}.$$

Note that

$$\delta^{\frac{3}{2}} < \frac{3\delta}{2}, \quad \forall \delta \in (0, 1).$$

Consequently, we have to choose the constant c_1 such that $c_1 > 3\delta/2$. This concludes the proof for

$$0 < \delta < \delta_0 < \min\left\{\frac{2}{3}c_1, \frac{2}{3}\right\}.$$

For a sufficiently large chosen c_2, Equation 4.4.17 follows from Lemma 4.4.1. Namely, such Lemma yields

$$\forall t > 0, \quad \mathbb{P}\left\{\left|\ \|g^{(j)}\|_2^2 - N\ \right| \leq c_2\sqrt{Nt}\right\} > 1 - e^{-t}.$$

For example, for $t = \log(2/\delta)$

$$\begin{aligned}\mathbb{P}\left\{\left|\ \|g^{(j)}\|_2^2 - N\ \right| \leq c_2\sqrt{N\ \log(2/\delta)}\right\} &> 1 - e^{-\log(2/\delta)} \\ &> 1 - e^{\log(\delta/2)} \\ &= 1 - \delta/2.\end{aligned}$$

□

Lemma 4.4.3. *For any parameter $0 < \alpha < 1$, there is a finite family $\Phi_\alpha \subset \bigcup_{m=1}^{N} \mathcal{M}_{m\times N}$ of matrices with the following properties:*

1. *For any matrix $\boldsymbol{A} \in \bigcup_{m=1}^{N} \mathcal{M}_{m\times N}$ with all entries bounded in absolute value by 2, there is a matrix $\hat{\boldsymbol{A}} \in \Phi_\alpha$ such that $\boldsymbol{A}$ and $\hat{\boldsymbol{A}}$ have the same number of rows and $\boldsymbol{B} = \boldsymbol{A} - \hat{\boldsymbol{A}}$ satisfies $tr(\boldsymbol{B}^T\boldsymbol{B}) \le \alpha/100$.*
2. *$\#\Phi_\alpha = e^{\mathcal{O}\left(N^2 \log(N/\alpha)\right)}$.*

Proof. We construct Φ_α as follows: for each integer $1 \le m \le N$, consider all $m \times N$-matrices whose entries are of the form $k\frac{\sqrt{\alpha}}{10\,N}$ for integers $k \in \left[\frac{-20N}{\sqrt{\alpha}}, \frac{20N}{\sqrt{\alpha}}\right] \cap \mathbb{Z}$. Note that adding to such set of entries the set $\{-2, 2\}$ we obtain partition of $[-2, 2]$ by intervals of size $\frac{\sqrt{\alpha}}{10N}$.

Now, let $\hat{\boldsymbol{A}} \in \Phi_\alpha$ be a $(m \times N)$-matrix with entries $[\hat{\boldsymbol{A}}]_{ij} = k_{ij}\frac{\sqrt{\alpha}}{10N}$. For all $i \in [m]$ and $j \in [N]$, we can choose k_{ij} such that

$$\left| [\boldsymbol{A}]_{ij} - [\hat{\boldsymbol{A}}]_{ij} \right| \le \frac{\sqrt{\alpha}}{10N}.$$

Then, for any matrix $\boldsymbol{A} \in \bigcup_{m=1}^{N} \mathcal{M}_{m\times N}$ with all entries bounded in absolute value by 2, there is $\hat{\boldsymbol{A}} \in \Phi_\alpha$ with the same number of rows and such that every entry of $\boldsymbol{B} = \boldsymbol{A} - \hat{\boldsymbol{A}}$ is bounded in absolute value by $\frac{\sqrt{\alpha}}{10N}$.

Consequently, every diagonal entry of $\boldsymbol{B}^T\boldsymbol{B}$ is bounded by $N\alpha/(100N^2)$ since

$$[\boldsymbol{B}^T\boldsymbol{B}]_{ii} = \sum_{t=1}^{N} [\boldsymbol{B}]_{tt}^2 \le N\left(\frac{\sqrt{\alpha}}{10N}\right)^2.$$

Thus, $\operatorname{tr}(\boldsymbol{B}^T\boldsymbol{B}) \le \alpha/100$.

Finally, we claim that the size of Φ_α is bounded by $N(1 + 40N/\sqrt{\alpha})^{N^2} = e^{\mathcal{O}\{N^2 \log(N/\alpha)\}}$. To see this, notice that, for any $x > 0$,

$$\#\{[-x, x] \cap \mathbb{Z}\} = 2\ \text{floor}(x) + 1 \le 2x + 1.$$

It follows that

$$\#\left\{\left[\frac{-20N}{\sqrt{\alpha}}, \frac{20N}{\sqrt{\alpha}}\right] \cap \mathbb{Z}\right\} \le \frac{40N}{\sqrt{\alpha}} + 1.$$

Now, we prove that the quantity of such $(m \times N)$-matrices (for a fixed m) is bounded by $(\frac{40N}{\sqrt{\alpha}} + 1)^{N^2}$. Indeed, there are less that $\frac{40N}{\sqrt{\alpha}} + 1$ ways to choose any of the mN entries of $\boldsymbol{A}$. Then, the number of such matrices is bounded above by

$$\left(\frac{40N}{\sqrt{\alpha}} + 1\right)^{mN} \le \left(\frac{40N}{\sqrt{\alpha}} + 1\right)^{N^2}.$$

We, therefore, finish the proof by taking the union bound over all m. □

4.4.3 Proof of the main Theorem

Lemma 4.4.4. *Let Φ_α be as in Lemma 4.4.3 with $0 \le \alpha < 1$. There is a set of $M = N^3$ vectors $v_1, \dots, v_M \in \mathbb{R}^N$ such that for every matrix $\boldsymbol{A} \in \Phi_\alpha$, there is an index $j \in [M]$ such that*

1. $\left| \|\boldsymbol{A}\, v_j\|_2^2 - \sum_{i=1}^{N} \lambda_i \right| = \Omega\left(\sqrt{(\log\ N) \sum_{i=1}^{N} \lambda_i^2} \right)$
   (here $\{\lambda_i\}_i$ are the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$),

2. $\left|\, \|v_j\|_2^2 - N \,\right| = \mathcal{O}\left(\sqrt{N \log N}\right).$

Proof. Let $g^{(1)}, \dots, g^{(M)} \in \mathbb{R}^N$ be independent standard Gaussian random vectors. We claim that these vectors satisfy both relations with probability exponentially close to 1.

Indeed, take $\boldsymbol{A} \in \Phi_\alpha$ and apply Corollary 4.4.2 with $\delta = N^{-1/4} = M^{-1/12}$. Thus, with probability at most $e^{-N^{3-1/4}}$ no $g^{(j)}$ satisfies (1) and (2) for $\boldsymbol{A}$. Equivalently, the probability of one of the $g^{(j)}$ satisfying (1) and (2) for $\boldsymbol{A}$ is at least $1 - e^{-N^{3-1/4}}$.

Since $\#\Phi_\alpha = e^{\mathcal{O}\{N^2 \log(N/\alpha)\}}$, the claim follows by a union bound over all matrices in Φ_α. More precisely,

$$\begin{aligned}
\mathbb{P}\{(1) \wedge (2),\, \forall \boldsymbol{A} \in \Phi_\alpha\} &= 1 - \mathbb{P}\{\sim (1) \vee \sim (2) \text{ for some } \boldsymbol{A} \in \Phi_\alpha\} \\
&\ge 1 - \sum_{\boldsymbol{A} \in \Phi_\alpha} \mathbb{P}\{\sim (1) \vee \sim (2) \text{ for } \boldsymbol{A} \text{ fixed}\} \\
&= 1 - \sum_{\boldsymbol{A} \in \Phi_\alpha} (1 - \mathbb{P}\{(1) \wedge (2) \text{ for } \boldsymbol{A} \text{ fixed}\}) \\
&= 1 - e^{\mathcal{O}\{N^2 \log(N/\alpha)\}} + \sum_{\boldsymbol{A} \in \Phi_\alpha} \mathbb{P}\{(1) \wedge (2) \text{ for } \boldsymbol{A} \text{ fixed}\} \\
&\ge 1 - e^{\mathcal{O}\{N^2 \log(N/\alpha)\}} + e^{\mathcal{O}\{N^2 \log(N/\alpha)\}}(1 - e^{-N^{3-1/4}}) \\
&= 1 - e^{\mathcal{O}\{N^2 \log(N/\alpha)\}} e^{-N^{3-1/4}}.
\end{aligned}$$

Finally, we claim that this last term will be larger than zero; or more precisely, that

$$\mathcal{O}\{N^2 \log(N/\alpha)\} - N^{3-1/4} < 0.$$

Indeed, by design,

$$\frac{1}{\text{poly}(1)} < \alpha < 1.$$

Consequently,

$$N^2 \log(N/\alpha) \le N^2 \log\{N^2\, \text{poly}(N)\} = \mathcal{O}\left(N^2 \log N\right), \tag{4.4.18}$$

from which

$$N^2 \log(N/\alpha) \le \mathcal{O}\left(N^2 \log N\right).$$

On the other hand,

$$N^{3-1/4} > \mathcal{O}\left(N^2 \log N\right)$$

and the result follows. □

The proof of Theorem 4.3.2 follows from the next result.

Theorem 4.4.2. *For any* $0 < \varepsilon < 1/2$, *there is a set* $V \subset \mathbb{R}^N (N > 4), \#V = M = N^3 + N$, *such that if* $\boldsymbol{A}$ *is a matrix in* $\mathcal{M}_{m\times N}$ *satisfies*

$$(1-\varepsilon)\|v_i\|_2^2 \leq \|\boldsymbol{A}\, v_i\|_2^2 \leq (1+\varepsilon)\|v_i\|_2^2, \quad v_i \in V,$$

then $m = \Omega(min\{N, \varepsilon^{-2} \log M\}) = \Omega(min\{N, \varepsilon^{-2} \log N\})$.

Proof. Since $N > 4$, we can assume that $1/\sqrt{N} < \varepsilon < 1/2$. In fact, the case $\varepsilon \leq 1/\sqrt{N}$ implies the lower bound $m = \Omega(N)$ by [Alo2003, Lemma 9.1]. To construct V, we first invoke Lemma 4.4.4 with $\alpha = \varepsilon^2/N^2$ to find N^3 vectors $\omega_1, \ldots, \omega_{N^3}$ such that for all matrices $\tilde{\boldsymbol{A}} \in \Phi_{\varepsilon^2/N^2}$, there is an index $j \in [N^3]$ for which:

1. $\left| \|\tilde{\boldsymbol{A}}\, \omega_j\|_2^2 - \sum_{i=1}^{N} \tilde{\lambda}_i \right| = \Omega\left(\sqrt{(\log\, N) \sum_{i=1}^{N} \tilde{\lambda}_i^2} \right)$;
2. $\left|\, \|\omega_j\|_2^2 - N \,\right| = \mathcal{O}\left(\sqrt{N\,(\log\, N)} \right)$,

with $\tilde{\lambda}_1 \geq \ldots \geq \tilde{\lambda}_N \geq 0$ being the eigenvalues of $\tilde{\boldsymbol{A}}^{\boldsymbol{T}}\, \tilde{\boldsymbol{A}}$. We let $V = \{e_1, \ldots, e_N, \omega_1, \ldots, \omega_{N^3}\}$ and claim this set of $N^3 + N$ vectors satisfies the Theorem.

In order to prove it, let $\boldsymbol{A} \in \mathbb{R}^{m\times N}$ be the matrix with $m \leq N$ satisfying

$$(1-\varepsilon)\|v\|_2^2 \leq \|\boldsymbol{A}\, v\|_2^2 \leq (1+\varepsilon)\|v\|_2^2, \quad v \in V.$$

Now observe that, since $e_1, \ldots, e_N \in V$, the matrix $\boldsymbol{A}$ satisfies

$$\|\boldsymbol{A}\, e_i\|_2^2 \in ((1-\varepsilon)\|e_i\|_2^2, (1+\varepsilon)\|e_i\|_2^2) = (1-\varepsilon, 1+\varepsilon), \quad \forall i \in [N]. \tag{4.4.19}$$

Hence, for all entries a_{ij} of $\boldsymbol{A}$ we must have $a_{ij}^2 \leq 1 + \varepsilon < 2$ (and, in fact, all columns of $\boldsymbol{A}$ have ℓ_2 norm at most $\sqrt{2}$). This implies that there is an $m \times N$ matrix $\hat{\boldsymbol{A}} \in \Phi_{\varepsilon^2/N^2}$ such that $\boldsymbol{B} = \boldsymbol{A} - \hat{\boldsymbol{A}} = (b_{ij})$ satisfies $\operatorname{tr}(\boldsymbol{B}^T \boldsymbol{B}) \leq \varepsilon^2/(100\, N^2)$, by Lemma 4.4.3. Since $\operatorname{tr}(\boldsymbol{B}^T \boldsymbol{B}) = \|\boldsymbol{B}\|_{\mathcal{F}}^2$, this also implies $\|\boldsymbol{B}\|_{\mathcal{F}} \leq \varepsilon/(10N)$. Then,

$$\begin{aligned} \sum_{i=1}^{N} \hat{\lambda}_i &= \operatorname{tr}(\hat{\boldsymbol{A}}^T \hat{\boldsymbol{A}}) && (4.4.20)\\ &= \operatorname{tr}\{(\boldsymbol{A}-\boldsymbol{B})^T(\boldsymbol{A}-\boldsymbol{B})\} && (4.4.21)\\ &= \operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B}^T\boldsymbol{B}) - \operatorname{tr}(\boldsymbol{A}^T\boldsymbol{B}) - \operatorname{tr}(\boldsymbol{B}^T\boldsymbol{A}). && (4.4.22) \end{aligned}$$

At first, we have from Equation 4.4.19 that

$$\operatorname{tr}\left(\boldsymbol{A}^T\boldsymbol{A}\right) = \sum_{i,j=1}^{N} a_{ij}^2 = \sum_{i=1}^{N} \|\boldsymbol{A}\, e_i\|_2^2 \in (N \pm \varepsilon N).$$

Thus,

$$\operatorname{tr}\left(\boldsymbol{A}^T\boldsymbol{A}\right) + \operatorname{tr}\left(\boldsymbol{B}^T\boldsymbol{B}\right) \in (N \pm \mathcal{O}\left(\varepsilon\, N\right)).$$

Also, from *Cauchy-Schwarz-Bunyakovsky inequality*,

$$\left|\operatorname{tr}\left(\boldsymbol{A}^T\boldsymbol{B}\right)\right| = \left|\operatorname{tr}\left(\boldsymbol{B}^T\boldsymbol{A}\right)\right| \leq \sqrt{\operatorname{tr}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\, \operatorname{tr}\left(\boldsymbol{B}^T\boldsymbol{B}\right)} = \left(\sum_{i,j=1}^{N} a_{ij}^2\right)^{1/2} \left(\sum_{i,j=1}^{N} b_{ij}^2\right)^{1/2}.$$

Consequently,

$$\begin{aligned}
\left|\operatorname{tr}\left(\boldsymbol{A}^T\boldsymbol{B}\right) + \operatorname{tr}\left(\boldsymbol{B}^T\boldsymbol{A}\right)\right| &\leq \left|\operatorname{tr}\left(\boldsymbol{A}^T\boldsymbol{B}\right)\right| + \left|\operatorname{tr}\left(\boldsymbol{B}^T\boldsymbol{A}\right)\right| \\
&\leq 2\left(\sum_{i,j=1}^{N} a_{ij}^2\right)^{1/2} \left(\sum_{i,j=1}^{N} b_{ij}^2\right)^{1/2} \\
&\leq 2\left(N \cdot \max_k \sum_{i=1}^{N} a_{ik}^2\right)^{1/2} \left(N \cdot \max_j \sum_{i=1}^{N} b_{ij}^2\right)^{1/2}.
\end{aligned}$$

Therefore,

$$-\operatorname{tr}\left(\boldsymbol{A}^T\boldsymbol{B}\right) - \operatorname{tr}\left(\boldsymbol{B}^T\boldsymbol{A}\right) \in \left(\pm\, 2N \cdot \max_j \left(\sum_{i=1}^{N} b_{ij}^2\right)^{1/2} \cdot \max_k \left(\sum_{i=1}^{N} a_{ik}^2\right)^{1/2}\right).$$

Finally, by substituting the equations above in Equation 4.4.22, we conclude

$$\begin{aligned}
\sum_{i=1}^{N} \hat{\lambda}_i &\in \left[N \pm \left\{\mathcal{O}\left(\varepsilon N\right) + 2N \cdot \max_j \left(\sum_i b_{ij}^2\right)^{1/2} \cdot \max_k \left(\sum_i a_{ik}^2\right)^{1/2}\right\}\right] \\
&\subset \left[N \pm \left\{\mathcal{O}\left(\varepsilon N\right) + 2N \cdot \|\boldsymbol{B}\|_{\mathcal{F}} \cdot \sqrt{2}\right\}\right] = \left[N \pm \mathcal{O}\left(\varepsilon N\right)\right], \qquad (4.4.23)
\end{aligned}$$

with the penultimate step being justified by the fact that

$$\max_j \left(\sum_i b_{ij}^2\right)^{1/2} \leq \|\boldsymbol{B}\|_{\mathcal{F}}.$$

Thus, from Lemma 4.4.4, our choice of V yields the existence of a vector $v^* \in V$ such that

$$\Big|\ \|\hat{\boldsymbol{A}}\, v^*\|_2^2 - N\ \Big| \geq \Omega\left(\sqrt{(\log\ N)\sum_{i=1}^{N}\hat{\lambda}_i^2}\right) - \mathcal{O}\left(\varepsilon N\right), \tag{4.4.24}$$

$$\big|\ \|v^*\|_2^2 - N\ \big| = \mathcal{O}\left(\sqrt{N\,(\log\ N)}\right). \tag{4.4.25}$$

Note that

$$\|\boldsymbol{B}\|^2 \leq \|\boldsymbol{B}\|_{\mathcal{F}}^2 = \operatorname{tr}\left(\boldsymbol{B}^T\boldsymbol{B}\right) \leq \varepsilon^2/(100\,N^2)$$

and that

$$\|\hat{\boldsymbol{A}}\|_2^2 \leq \|\hat{\boldsymbol{A}}\|_{\mathcal{F}}^2 \leq (\|\boldsymbol{A}\|_{\mathcal{F}} + \|\boldsymbol{B}\|_{\mathcal{F}})^2 = \mathcal{O}\left(N\right),$$

with the second inequality above being due to the Cauchy-Schwarz-Bunyakovsky inequality.

Then, we have

$$\begin{aligned}
\big|\ \|\boldsymbol{A}\, v^*\|_2^2 - N\ \big| &= \Big|\ \|\hat{\boldsymbol{A}}\, v^*\|_2^2 + \|\boldsymbol{B}\, v^*\|_2^2 + 2\left\langle \hat{\boldsymbol{A}}v^* : \boldsymbol{B}v^* \right\rangle - N\ \Big| \\
&\geq |\ \|\hat{\boldsymbol{A}}\, v^*\|_2^2 - N\ | - \|\boldsymbol{B}\, v^*\|_2^2 - 2\left|\left\langle \hat{\boldsymbol{A}}v^* : \boldsymbol{B}v^* \right\rangle\right| \\
&\geq \Omega\left(\sqrt{(\log\ N)\sum_{i=1}^{N}\hat{\lambda}_i^2}\right) - \mathcal{O}\left(\varepsilon N\right) - \|\boldsymbol{B}\, v^*\|_2^2 - 2\left|\left\langle \hat{\boldsymbol{A}}v^* : \boldsymbol{B}v^* \right\rangle\right| \\
&\geq \Omega\left(\sqrt{(\log\ N)\sum_{i=1}^{N}\hat{\lambda}_i^2}\right) - \mathcal{O}\left(\varepsilon N\right) - \|\boldsymbol{B}\|^2 \cdot \|v^*\|_2^2 - 2\|\boldsymbol{B}\| \cdot \|\boldsymbol{A}\| \cdot \|v^*\|_2^2 \\
&= \Omega\left(\sqrt{(\log\ N)\sum_{i=1}^{N}\hat{\lambda}_i^2}\right) - \mathcal{O}\left(\varepsilon N\right),
\end{aligned} \tag{4.4.26}$$

with the second inequality above being resulted from Equation 4.4.24.

Also, since $\boldsymbol{A}$ is a JL-embedding, we have $|\ \|\boldsymbol{A}\, v^*\|_2^2 - \|v^*\|_2^2\ | = \mathcal{O}\left(\varepsilon\|v^*\|_2^2\right) = \mathcal{O}\left(\varepsilon N\right)$. Therefore by Equation 4.4.25,

$$\begin{aligned}
|\ \|\boldsymbol{A}\, v^*\|_2^2 - N| &\leq |\ \|\boldsymbol{A}\, v^*\|_2^2 - \|v^*\|_2^2\ | + |\ \|v^*\|_2^2 - N\ | \\
&= \mathcal{O}\left(\varepsilon N + \sqrt{N\,\log N}\right),
\end{aligned} \tag{4.4.27}$$

which when combined with Equation 4.4.26 implies

$$\sum_{i=1}^{N}\hat{\lambda}_i^2 = \mathcal{O}\left(\frac{\varepsilon^2\, N^2}{\log\ N} + N\right). \tag{4.4.28}$$

To prove the statement above, note that Equations 4.4.26 and 4.4.27 above imply the existence of $c_1, c_2, c_3 \geq 0$ such that

$$c_1\sqrt{(\log\ N)\sum_{i=1}^{N}\hat{\lambda}_i^2} - c_2\varepsilon N \leq \left|\ \|\boldsymbol{A}v^*\|_2^2 - N\ \right| \leq c_3(\varepsilon N + \sqrt{N\log N}).$$

Consequently,

$$\sqrt{\sum_{i=1}^{N}\hat{\lambda}_i^2} \leq \frac{\tilde{c}_1\varepsilon N}{\sqrt{\log N}} + \tilde{c}_2\sqrt{N},$$

with $\tilde{c}_1, \tilde{c}_2 \geq 0$. By squaring both sides of this inequality, we obtain

$$\begin{aligned}\sum_{i=1}^{N}\hat{\lambda}_i^2 &\leq \frac{\tilde{c}_1^2\varepsilon^2 N^2}{\log N} + \tilde{c}_2^2 N + 2\tilde{c}_1\tilde{c}_2\frac{\varepsilon N^{3/2}}{\sqrt{\log N}} \\ &\leq (\tilde{c}_1^2 + \tilde{c}_1\tilde{c}_2)\frac{\varepsilon^2 N^2}{\log N} + \tilde{c}_2^2 N,\end{aligned}$$

since $\log N \leq N$ and $\varepsilon < 1/2$, and Equation 4.4.28 follows.

Now we claim that

$$\frac{N^2}{2} \leq \left(\sum_{i=1}^{N}\hat{\lambda}_i\right)^2.$$

In fact, from Equation 4.4.23, there is $c \geq 0$ such that, for sufficiently large N,

$$\sum_{i=1}^{N}\hat{\lambda}_i \geq (1 - c\,\varepsilon)N > \left(1 - \frac{c}{2}\right)N.$$

Note that we can choose any constant $\tilde{c} \geq c$ at the expense of tightness in the inequalities above. Consequently,

$$\left(\sum_{i=1}^{N}\hat{\lambda}_i\right)^2 > \left(1 - c + \frac{c^2}{4}\right)N^2 \geq \frac{N^2}{2},$$

since we substitute c for $\tilde{c} \geq \min\{c,\, 2+\sqrt{2}\}$.

Also, note that the amount of non-zero $\hat{\lambda}_i$ is exactly $\operatorname{rank}(\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}})$. Furthermore, consider the vectors $\vec{a} := (\hat{\lambda}_1, \ldots, \hat{\lambda}_N)$ and $\vec{b} = (b_i)_{i\in[N]}$ such that $b_i = 0$ if $\hat{\lambda}_i$ vanishes and $b_i = 1$ otherwise. Now, by the *Cauchy-Schwarz-Bunyakovsky inequality*,

$$\begin{aligned}\left(\sum_{i=1}^{N}\hat{\lambda}_i\right)^2 = |\langle\vec{a} : \vec{b}\rangle|^2 &\leq \|\vec{a}\|_2^2 \cdot \|\vec{b}\|_2^2 \\ &= \left(\sum_{i=1}^{N}\hat{\lambda}_i^2\right) \cdot\ \operatorname{rank}(\hat{\boldsymbol{A}}^T\boldsymbol{A}).\end{aligned}$$

Finally, the fact $\operatorname{rank}(\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}})$ is at most m since $\boldsymbol{A}$ in an $(m \times N)$-matrix and Equation 4.4.28 together imply

$$\left(\sum_{i=1}^{N}\hat{\lambda}_i^2\right) \cdot \operatorname{rank}(\hat{\boldsymbol{A}}^T\boldsymbol{A}) \leq \mathcal{O}\left(\frac{\varepsilon^2 N^2}{\log N} + N\right) \cdot m.$$

Additionally, the assembly of the inequalities we have just proven leads to

$$\frac{N^2}{2} \leq \left(\sum_{i=1}^{N}\hat{\lambda}_i\right)^2 \leq \operatorname{rank}(\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}}) \left(\sum_{i=1}^{N}\hat{\lambda}_i^2\right) \leq m\,\mathcal{O}\left(\frac{\varepsilon^2 N^2}{\log N} + N\right).$$

We claim that rearranging the terms will give

$$m = \Omega(\min\{N, \varepsilon^{-2}\log N\}) = \Omega(\min\{N, \varepsilon^{-2}\log M\})$$

as desired. Indeed, we have concluded that

$$\frac{N^2}{2} \leq m\,\mathcal{O}\left(\frac{\varepsilon^2 N^2}{\log N} + N\right).$$

Consequently, there is $c \geq 0$ such that, for a sufficiently large N,

$$\frac{N^2}{2} \leq m\,c\left(\frac{\varepsilon^2 N^2}{\log N} + N\right).$$

Equivalently,

$$m \geq \frac{1}{2c}\varepsilon^{-2}\log N\left(1 + \frac{\varepsilon^{-2}}{N}\log N\right)^{-1}.$$

Now, we have to get rid of that last factor. In order to do so, note that

$$\lim_{N\to\infty}\frac{\log N}{N} = 0.$$

Thus, there is $N_0 \in \mathbb{N}$ such that

$$\forall N \geq N_0, \quad \frac{\log N}{N} \leq \frac{1}{\varepsilon^2}$$

that implies

$$\forall N \geq N_0, \quad \frac{1}{1 + \frac{\varepsilon^{-2}}{N}\log N} \geq 1/2$$

and the result follows.

□

4.5 Overview of the proof of Larsen & Nelson's Theorem – 2016

Let us recall the statement of Larsen and Nelson's result.

Theorem 4.5.1. *For any integers $M, N \geq 2$ and $\varepsilon \in \left(\frac{\log^{0.5001} M}{\sqrt{\min\{M,N\}}}, 1\right)$, there is a set of points $X \in \mathbb{R}^N$ of size M, such that any map $f : (X, \ell_2) \to \ell_2^m$ providing the JL-guarantee must have*

$$m = \Omega(\varepsilon^{-2}\log(\varepsilon^2 M)).$$

4.5.1 Counting argument

Intuitively, one might ask what prevents our dimension reduction results from embedding almost isometrically any set of high dimensional vectors in an Euclidean vector space with arbitrarily small dimension. A *naïve* answer to this question can be elaborated by noting that not any set of vectors are equal. Some are said to be very "different" from each other. That being said, our argument is based on the fact that there are not too many "different" sets with a big amount of elements in a tiny dimensional space, say $\mathbb{R}^m$.

More precisely, Larsen and Nelson construct a family of very "different" sets of M vectors $\mathcal{P} = \{P_1, P_2 \ldots\}$ in a high dimensional vector space, say $\mathbb{R}^N$. It is then assumed that all point sets in $\mathcal{P}$ can be embedded (not necessarily with the same map) into $\mathbb{R}^m$ while preserving all pairwise distances within $1 \pm \varepsilon$.

Letting $f_1(P_1), f_2(P_2), \ldots$ denote the embedded point sets, then it is argued that such choice of $\mathcal{P}$ ensures that any two $f_i(P_i)$ and $f_j(P_j)$ must be very "different". Therefore, if m is too low, this is impossible as there are not enough sufficiently "different" point sets in $\mathbb{R}^m$.

The construction of $\mathcal{P}$ is as follows. Let $\{e_1, \ldots, e_N\}$ denote the canonical basis of $\mathbb{R}^N$. Assume $N = \frac{M}{\log(1/\varepsilon)}$ and take $\varepsilon \in (0, 1)$. For any index set $S \subset [N]$ of $k = \frac{1}{\varepsilon^2 c_0^2}$ elements (with $c_0 > 0$ a sufficiently large constant), define the vector y_S as

$$y_S := \sum_{j \in S} \frac{e_j}{\sqrt{k}}.$$

Note that these vectors are such that

$$\langle y_S : e_j \rangle = \begin{cases} 0, & j \notin S \\ c_0 \varepsilon, & j \in S \end{cases}.$$

The construction of such vectors starts to shed light on our intuition of what "different set" means, since vectors y_S will be used to construct such sets and there is a gap of $c_0\varepsilon$ between the products $\langle y_S : e_j \rangle$ for j being or not an index in S. That is, different choices of the index set S leads to different vectors y_S, following this gap criteria.

Now let $f : \mathbb{R}^N \to \mathbb{R}^m$ be a JL-embedding for $P = \{\mathbf{0}_N, e_1, \ldots, e_N, y_S\}$. By assuming that $f(\mathbf{0}_N) = \mathbf{0}_m$ (what can be made w.l.o.g. since pairwise distances are translation invariant), one concludes that f must preserve norms of vectors in P with at most $1 \pm \varepsilon$ distortion. Indeed,

$$\begin{aligned} (1-\varepsilon)\|x\|_2^2 &= (1-\varepsilon)\|x - \mathbf{0}_N\|_2^2 \\ &\leq \|f(x) - f(\mathbf{0}_N)\|_2^2 \quad (= \|f(x) - \mathbf{0}_m\|_2^2 = \|f(x)\|_2^2) \\ &\leq (1+\varepsilon)\|x - \mathbf{0}_N\|_2^2 \\ &= (1+\varepsilon)\|x\|_2^2. \end{aligned}$$

As a consequence of such norm distortion, we claim that f must preserve inner products

$\langle e_j : y_S \rangle$ up to an $\mathcal{O}(\varepsilon)$ summand. In fact, for $x \in P$:

$$\begin{aligned} \|f(e_j) - f(y_s)\|_2^2 &= \|f(e_j)\|_2^2 + \|f(y_S)\|_2^2 - 2\langle f(e_j) : f(y_S)\rangle \implies \\ 2\langle f(e_j) : f(y_S)\rangle &\in [1 \pm \varepsilon]\|e_j\|_2^2 + [1 \pm \varepsilon]\|y_S\|_2^2 - [1 \pm \varepsilon]\|e_j - y_S\|_2^2 \implies \\ 2\langle f(e_j) : f(y_S)\rangle &\in 2\langle e_j : y_S\rangle \pm \varepsilon(\|e_j\|_2^2 + \|y_S\|_2^2 + \|e_j - y_S\|_2^2) \implies \\ \langle f(e_j) : f(y_S)\rangle &\in \langle e_j : y_S\rangle \pm 4\varepsilon. \end{aligned}$$

Also, note that

$$\langle e_j : y_S\rangle \pm 4\varepsilon \subset [-4\varepsilon,\, c_0\varepsilon + 4\varepsilon].$$

This means that after applying f, there is a gap of $(c_0 + 8)\varepsilon = \Omega(\varepsilon)$ between $\langle f(e_j) : f(y_S)\rangle$ for j being or not an index in S.

Now, we are ready to describe the point sets in $\mathcal{P}$. Namely, let $Q = M - N - 1$. Then, for every choice of Q index sets, $S_1, \ldots, S_Q \subset [N]$, each with k elements, we add to $\mathcal{P}$ a new point set P such that

$$P = \{\mathbf{0}_N, e_1, \ldots, e_N, y_{S_1}, \ldots, y_{S_Q}\}.$$

This gives a family $\mathcal{P}$ of size $\binom{N}{k}^Q$.

Moreover, it is argued that the JL-embedded sets, $f_1(P_1), f_2(P_2), \ldots$ have to be quite "different". More precisely, it comes from the fact the sets $S_1, \ldots, S_Q$ and consequently $P_1, P_2, \ldots$ are "different". This intuition is formalized by proving that these maps uniquely determine which point set it embeds.

In fact, for a given map, f_i, $i \in \mathbb{N}$, if suffices to evaluate the inner products

$$\langle f_i(e_j) : f_i(y_{S_\ell})\rangle, \text{ with } j \in [N] \text{ and } \ell \in [Q].$$

Given the gaps, it is possible to determine for any j and ℓ if $j \in S_\ell$. Consequently, all S_ℓ can be reconstructed and then all P_i.

The problem now is that there are infinitely many sets of M points in $\mathbb{R}^m$ that one can embed to. Thus it is necessary to discretize $\mathbb{R}^m$ in a careful manner and argue that there are not enough M-sized sets of points in this discretization to uniquely embed each P_i when m is too low.

4.5.2 Encoding argument

To give a formal proof that there are not enough ways to embed the point sets in $\mathcal{P}$ into $\mathbb{R}^m$ when m is too low, it is given an encoding argument. More specifically, it is assumed that it is possible to embed every point set in $\mathcal{P}$ into $\mathbb{R}^m$ while preserving pairwise distances to within $1 \pm \varepsilon$. Larsen and Nelson then present an algorithm that can take any point set $P_i \in \mathcal{P}$ and encode it into a bit string of length $\mathcal{O}(Mm)$. Such encoding will be represented by the injective mapping $g : \mathcal{P} \to \{0,1\}^{\mathcal{O}(Mm)}$ so we can uniquely recover P_i from the bit string.

Since g is injective, one must must have $\#\mathcal{P} \leq 2^{\mathcal{O}(Mm)}$. But

$$\#P = \binom{N}{k}^Q \geq \left\{\frac{\varepsilon^2 M}{\log(1/\varepsilon)}\right\}^{\frac{1}{\varepsilon^2 c_0^2}\left(M - \frac{M}{\log(1/\varepsilon)} - 1\right)}$$

and it can be concluded that

$$m = \Omega\left\{\varepsilon^{-2}\log\left(\frac{\varepsilon^2 M}{\log(1/\varepsilon)}\right)\right\}.$$

Now, for $\varepsilon > 1/M^{0.4999}$, we get $m = \Omega(\varepsilon^{-2}\log M)$.

Chapter 5

Conclusion

The main goal of this work was the detalied discussion on the theoretical basis of a dimensionality reduction method, the Johnson-Lindenstrauss lemma. We motivated the need of such tools via the presentation of some common problems in the high-dimensional setting. Overall, the text was organized as follows: the original proof of the Lemma was presented in Chapter 2; an overview of the historical improvements was discussed in Chapter 3; and the recent sharpness results were discussed in Chapter 4. Furthermore, in this last Chapter, we presented in details the sharpness results due to Larsen and Nelson [LN2016, LN2017].

However, for the sake of brevity, some topics were not discussed and we will point them out now. In Chapter 3, we dealt just with theoretical improvements on the JL-Lemma and omitted the practical results that improve complexity and storage issues. We did not discuss, for example, the application of special matrices to construct the JL-embedding satisfying important properties, like sparsity, Restricted Isometry Property (RIP), circulant matrices, among others. The Fast Johnson-Lindenstrauss Transform (FJLT), that speeds up the projections using a Fast Fourier Transform was also ommited from the text. Finally, we could not discuss several important applications of the Lemma like the ones in Machine Learning and Compressed Sensing. Also, the interested reader can find further improvements on the result like its generalization for other norms than $\|\cdot\|_2$ and its application for infinite sets in Vershnyn's *High-Dimensional Probability: An Introduction with Applications in Data Science* [Vsh2018]. Finally, we have not discussed results obtained from Algorithmic and Coding Theory, as the proof of the main result in [LN2017], that proves the sharpness of the JL-Lemma for non-linear embeddings. All these points should be addressed in future studies about the subject.

Chapter A

Appendix A

A.1 What is an uniformly chosen matrix on $O(N)$?

In Chapter 2 of this text we presented the original proof to the Johnson-Lindenstrauss Lemma through an argument of *random projection*. More precisely, we selected a random matrix from an uniform distribution on $O(N)$. However, we still have to explain in more details what we mean by a uniform distribution on $O(N)$.

The first idea that comes up to our minds when talking about sampling random matrices is to select its entries from some fixed probability distribution. However, individual entries of elements of matrix groups such as $O(N)$ do not have a simple characterization, but the rows or columns of such matrices satisfy simple relations, and also it is possible to infer geometrical properties the respective linear transformation will satisfy. Consequently, our sampling problem will be simpler if we give a geometric interpretation to these matrix groups.

The keystone property here is the *rotation invariance*. For example, assume that our goal is to select uniformly a point on $\mathbb{S}^1 \subset \mathbb{R}^2$. We can think of a point on the circle as $z = x + iy$, with the condition that $x^2 + y^2 = 1$, but it does not lead us to any ideas about how to choose coordinates x and y such that the respective point is a "uniformly chosen random point" on the circle. However, it makes sense to think about a complex random variable taking values in $\mathbb{S}^1 \subset \mathbb{C}$, whose distribution is rotation invariant. More precisely, for a Borel set $A \subset \mathbb{S}^1$, the probability that our random point lies in A should be the same as the probability that it lies in $e^{i\theta}A := \{e^{i\theta}a : a \in A\}$.

The reasoning for matrix groups is analogous. In particular, let $\boldsymbol{M}$ be a fixed element in $O(N)$. If a matrix $\boldsymbol{U}$ is selected from a uniform distribution on $O(N)$, we must have the following equalities in distribution:

$$\boldsymbol{MU} \overset{d}{=} \boldsymbol{UM} \overset{d}{=} \boldsymbol{U}.$$

From another point of view, a uniform measure σ on $O(N)$ is such that for any subset $A \subset O(N)$ and any fixed $\boldsymbol{M} \in O(N)$, we have

$$\sigma(\boldsymbol{M}A) = \sigma(A\boldsymbol{M}) = \sigma(A),$$

with

$$\boldsymbol{M}A := \{\boldsymbol{MU} : \boldsymbol{U} \in A\} \text{ and } A\boldsymbol{M} := \{\boldsymbol{UM} : \boldsymbol{U} \in A\}.$$

Fortunately, there is a Theorem called *Haar Theorem* that guarantees the existence and uniqueness of such a probability measure on matrix groups, in particular, $O(N)$. This theorem will be discussed – without proof and in a more general setting – in more details in the next section.

A.2 Haar Theorem

As explained at the end of the previous section, we shall not prove the Haar Theorem for the sake of brevity. However, the interested reader may be directed to references [Foll1999] and [DiSp2014].

The present section goes as follows. We first motivate the Haar Theorem as a generalization of the Lebesgue measure on $\mathbb{R}^N$ for the left translations by a general group G. In order to do so, we also exhibit some definitions from Group Theory that allow us to reinterpret the vector translations of $\mathbb{R}^N$ as left translations of an additive group. We then introduce some concepts from Measure Theory that are necessary to understand the Haar Theorem (Theo. A.2.1) and state it in a very general context, without proof. Finally, since the matrix groups $O(N)$ and $SO(N)$ are very specific cases, it is possible to explicitly construct the desired measure, and this construction is briefly outlined.

A.2.1 Motivation from the Lebesgue measure

As it is widely known, the Lebesgue measure $\mathcal{L}^N$ in $\mathbb{R}^N$ has the *translation invariance property*. That is, given a Lebesgue-measurable set $E \subset \mathbb{R}^N$ and $c \in \mathbb{R}^N$, we have that

$$c + E := \{c + x : x \in E\}$$

is also Lebesgue-measurable and

$$\mathcal{L}^N(c + E) = \mathcal{L}^N(E).$$

Note that such property defines a "uniformity" of this measure on $\mathbb{R}^N$ since sets that differ only by translations have the same measure. That is, these "analogous sets" have the same measure independently of where in $\mathbb{R}^N$ they are.

That being said, we might ask if we can generalize this invariance of Lebesgue measure to other measurable spaces under another operation instead of translation. As we shall see, it is possible to generalize such invariance when our measurable space has the structure of a *group*.

A.2.2 Left Haar measures

Since a matrix $\boldsymbol{M} \in \mathcal{M}_{N\times N}$ usually multiplies a vector $v \in \mathbb{R}^N$ by the left, the action of matrix groups on subsets of $\mathbb{R}^N$ is done by the left. Thus, we shall focus our discussion on *left invariance* for the sake of formality.

Definition A.2.1 (Left translation)**.** *Let S be a subset of a group $(G, \cdot)$. For a fixed $g \in G$, we define the left translation $g \cdot S$ or gS of S by g as the set*

$$gS := \{g \cdot x : x \in G\}.$$

In this new context, it would be desirable to have a measure satisfying the following property:

Definition A.2.2 (Left translation invariant Borel measure)**.** *A Borel measure μ on a group G is called left translation invariant or G-invariant if, for any Borel subset $S \subset G$ and all $g \in G$, the left translation gS is μ-measurable and*

$$\mu(gS) = \mu(S).$$

Fortunately, in a very general setting, the existence and uniqueness of such measure can be guaranteed:

Theorem A.2.1 (Haar Theorem)**.** *Let $(G, \cdot)$ be a locally compact Hausdorff topological group. Then, there exists, up to a positive multiplicative constant, a unique left translation invariant measure. This measure is called the Haar measure over G.*

Indeed, matrix groups endowed with the Euclidean topology satisfy the required conditions, and in the particular case of G being compact this measure is such that

$$\mu(G) < \infty.$$

Now, as the left Haar measure is unique up to a positive multiplicative constant, a compact space G yields a *unique probability measure* via normalization, that we will henceforth denote as σ.

A.2.3 Explicit construction of the Haar measure on $O(N)$

The Haar Theorem is indeed very useful to our random projection scheme. However, its statement by itself does not tell us how to construct the Haar measure in specific cases. We shall discuss one explicit construction of this measure on $O(N)$, which also allows us to uniformly select a matrix from this group using a standard Gaussian matrix.

Firstly, we claim that the distribution of a random matrix whose entries were sorted from a standard Gaussian distribution is invariant under left-multiplication by an orthogonal matrix. Indeed, let $\boldsymbol{X} \in \mathcal{M}_{N\times N}$ be such matrix. The joint density of the N^2 entries of $\boldsymbol{X}$ is given by

$$\frac{1}{(2\pi)^{N^2}} \prod_{i,j=1}^{N} \exp\left\{-\frac{x_{i,j}^2}{2}\right\} = \frac{1}{(2\pi)^{N^2}} \exp\left\{-\frac{1}{2}\sum_{i,j=1}^{N} x_{i,j}^2\right\}.$$

By applying the following change of variables

$$y_{i,j} := [\boldsymbol{M}\boldsymbol{X}]_{i,j} = \sum_{k=1}^{N} [\boldsymbol{M}]_{i,k}[\boldsymbol{X}]_{k,j}$$

it follows that the joint density of the entries of $\boldsymbol{MX}$ is given by

$$\frac{\left|\det(\boldsymbol{M}^{-1})\right|}{(2\pi)^{N^2}} \exp\left\{-\frac{1}{2}\sum_{i,j}^{N}[\boldsymbol{M}^{-1}y]_{i,j}^2\right\} = \frac{1}{(2\pi)^{N^2}} \exp\left\{-\frac{1}{2}\sum_{i,j}^{N} y_{i,j}^2\right\},$$

since $\boldsymbol{M}^{-1}$ is an isometry.

Nevertheless, that Gaussian matrix is still not what we want since it is not orthogonal. We can solve this problem by applying the Gram-Schmidt algorithm. In fact, performing the Gram-Schmidt process commutes with multiplication by a fixed orthogonal matrix $\boldsymbol{M}$: applying the algorithm to $\boldsymbol{X}$ and multiplying the result by $\boldsymbol{M}$ yields the same matrix as applying it directly to $\boldsymbol{MX}$. Moreover, we recall that $\boldsymbol{MX}$ and $\boldsymbol{X}$ have the same probability distribution given the discussion made so far.

Now, let $\{X_1, \ldots, X_N\}$ be the columns of $\boldsymbol{X}$. Given a column, X_1 for example, the Gram-Schmidt processes will substitute X_2 by

$$X_2 - \frac{\langle X_1 : X_2\rangle}{\langle X_1 : X_1\rangle} X_1 \tag{A.2.1}$$

normalized. On the other hand, doing the same procedure to the matrix $\boldsymbol{MX}$, whose columns are

$$\{\boldsymbol{M}X_1, \ldots, \boldsymbol{M}X_N\},$$

produces, as the new second column, the vector

$$\boldsymbol{M}X_2 - \frac{\langle \boldsymbol{M}X_1 : \boldsymbol{M}X_2\rangle}{\langle \boldsymbol{M}X_1 : \boldsymbol{M}X_1\rangle} \boldsymbol{M}X_1$$

normalized. Note that it is equal to A.2.1 multiplied on the left side by $\boldsymbol{M}$, since it is an isometry.

Finally, the probability measure constructed this way results in a random orthogonal matrix whose distribution is invariant under left-multiplication by a fixed orthogonal matrix. In other words, we presented a way to construct the Haar measure on $O(N)$.

Chapter B

Appendix B

B.1 The sphere

As discussed in Chapter 2, the original proof of the JL-Lemma is done through a *concentration of measure* argument of its relative surface area. More precisely, in a high dimensional setting, most of the relative surface area of $\mathbb{S}^{N-1}$ will be concentrated around any of its equators. Furthermore, if the opening angle of this equatorial strip is $2\varepsilon \in (0, \pi/2)$, its surface area will be larger than

$$1 - 4\exp\{-N\varepsilon^2/2\}.$$

Such impressive result is one of the counterintuitive behaviors that motivates the study of *Asymptotic Geometric Analysis*.

A direct consequence of the concentration of measure around equatorial strips on the sphere is that any spherical cap containing a hemisphere concentrates exponentially the sphere's relative surface area. A key result, broadly discussed – without proof – in the Chapter 2 of this text, is the *isoperimetric inequality on the sphere*. This inequality allows us to relate the relative area of any such Borel subset with the relative area of a spherical cap. As a consequence, we may extend the concentration of measure inequality for the sphere to any Borel subset of $\mathbb{S}^{N-1}$ whose relative area is bigger than $1/2$.

The discussion made so far in this introduction motivates the use of the relative surface area as a probability measure on the sphere. Moreover, the Haar and Weil's Theorems guarantee the existence and uniqueness of the Haar measure on $\mathbb{S}^{N-1}$, but do not exhibit how to compute it while the surface area can be directly calculated using tools from *Multivariable Calculus*. However, we are still interested in the good properties, mainly the *rotation invariance*, of the Haar measure on $\mathbb{S}^{N-1}$. Consequently, the main goal of this section is to prove that these probability measures are equivalent, obtaining the best of all worlds.

In sum, this section goes as follows. In the first subsection, we exhibit some formulas for surface areas on $\mathbb{S}^{N-1}$ that will be useful to prove the concentration of measure on the sphere. Next, we define the *Lebesgue measure on the sphere*. In this presentation, such measure will play just the theoretical role to link the two measures we are really interested in. Namely, the Lebesgue measure on $\mathbb{S}^{N-1}$ will be resulted from the Lebesgue measure on $\mathbb{R}^N$, that is the Haar measure on $\mathbb{R}^N$. Consequently, it will inherit the regularity and invariance properties of that measure. Finally, we show that, for the sphere, the surface area is equal to the Lebesgue

measure up to a multiplicative constant, concluding the result.

B.1.1 Surface area of spherical caps

Given the central role of the relative surface area of spherical caps motivated in the introduction of this section, we will make some considerations on how to calculate it. We start by remarking the formulas for surface area and volume of the unitary sphere $\mathbb{S}^{N-1}$:

$$\text{Vol}\left\{\mathbb{S}^{N-1}\right\} = \frac{\pi^{N/2}}{\Gamma(1+N/2)} \quad \text{and} \quad \text{Area}\left\{\mathbb{S}^{N-1}\right\} = \frac{2\pi^{N/2}}{\Gamma(N/2)}, \tag{B.1.1}$$

with $\Gamma : \mathbb{C} \setminus (-\infty, 0] \to \mathbb{C}$ being the Gamma function, that is

$$\Gamma(z) := \int_0^\infty x^{z-1} e^{-x}\, dx, \quad \mathcal{R}(z) > 0.$$

Furthermore, in the present section, we shall discuss about the surface area of *spherical caps*. Such sets can be more easily described in terms of the following:

Definition B.1.1 (Geodesic distance on the sphere)**.** *Let $x, y \in \mathbb{S}^{N-1}$. We define the geodesic distance $\rho(x, y)$ between these points as the angle $\widehat{xOy} \in [0, \pi]$, i.e., the convex (shortest) angle between x and y.*

Provided with this Definition, we may write the spherical caps as balls of such metric. Namely, for $\phi \in [0, \pi]$, we define the ϕ-cap centered in $x \in \mathbb{S}^{N-1}$ as

$$K_{N-1}(x, \phi) := \{z \in \mathbb{S}^{N-1} : \rho(x, z) \leq \phi\}.$$

Henceforth, when we are not particularly interested in the center x of the spherical cap $K_{N-1}(x, \phi)$, we shall denote it simply as $K_{N-1}(\phi)$.

Now, from [Li2011], we have, for $N > 2$, that

$$\begin{aligned} \text{Area}\{K_{N-1}(\phi)\} &= \text{Area}\{\mathbb{S}^{N-2}\} \int_0^\phi \sin^{N-2}\theta\, d\theta \\ &= \frac{2\pi^{(N-1)/2}}{\Gamma\{(N-1)/2\}} \int_0^\phi \sin^{N-2}\theta\, d\theta, \end{aligned}$$

with $\text{Area}\{\mathbb{S}^1\} = 2\pi$. That is, we have an equation that relates the surface area of a subset of $\mathbb{S}^{N-1}$ with the surface area of $\mathbb{S}^{N-2}$. In particular, for $\phi = \pi$, we obtain, from the Equation above, the recursion below:

$$\begin{aligned} \text{Area}\{\mathbb{S}^{N-1}\} &= \text{Area}\{K_{N-1}(\pi)\} \\ &= \text{Area}\{\mathbb{S}^{N-2}\} \int_0^\pi \sin^{N-2}\theta\, d\theta \\ &= \frac{2\pi^{(N-1)/2}}{\Gamma\{(N-1)/2\}} \int_0^\pi \sin^{N-2}\theta\, d\theta, \end{aligned}$$

Finally, we define a probability measure $\mu_{N-1}(S)$ on $\mathbb{S}^{N-1}$ as the relative area of the Borel subset $S \subset \mathbb{S}^{N-1}$. That is,

$$\mu_{N-1}(S) := \frac{\text{Area}(S)}{\text{Area}(\mathbb{S}^{N-1})}.$$

In particular, for spherical caps, we have

$$\mu_{N-1}\{K_{N-1}(\phi)\} = \frac{\int_0^\phi \sin^{N-2}\theta\, d\theta}{\int_0^\pi \sin^{N-2}\theta\, d\theta} = \gamma_N \int_0^\phi \sin^{N-2}\theta\, d\theta \quad (\phi \in [0, \pi/2]),$$

naming

$$\gamma_N^{-1} = \int_0^\pi \sin^{N-2}\theta\, d\theta$$

to simplify the notation.

In order to prove that the surface area is equivalent to the Haar measure, we shall exhibit how it is related to the Lebesgue measure, $\mathcal{L}^N$, on ℓ_2^N. As we shall see in the next subsection, $\mathcal{L}^N$ is the Haar measure on $\mathbb{R}^N$. That being settled, writing the normalized surface area in terms of $\mathcal{L}^N$ yields a simple proof that it is indeed the Haar measure on $\mathbb{S}^{N-1}$.

B.1.2 Lebesgue measure on $\mathbb{S}^{N-1}$

We start the present subsection regarding the Lebesgue measure $\mathcal{L}^N$ on the Euclidean space $\mathbb{R}^N$. Such measure translates our geometrical idea of volume in $\mathbb{R}^N$. Moreover, it is a well known result that the $\mathcal{L}^N$ is *rotation invariant*. It give us a clue that Lebesgue measure is a good starting point in our search for a rotation invariant measure on the sphere. Thus, we shall briefly discuss of how to determine $\mathcal{L}^N$ on Borel subset of $\mathbb{R}^N$.

Remark. *The σ-algebra $\mathcal{R}^N$ over which the measure $\mathcal{L}^N$ is defined includes more sets than just the Borel subsets of $\mathbb{R}^N$. However, since we are interested only in defining a Lebesgue measure on Borel subsets of the sphere, we shall skip such deeper details.*

In order to define the Lebesgue measure space $(\mathbb{R}^N, \mathcal{R}^N, \mathcal{L}^N)$, we claim that the N-boxes, i.e., sets of of the form $B = \prod_{k=1}^N I_k$, with I_k, $k \in [N]$ being intervals of $\mathbb{R}$ are $\mathcal{L}^N$-measurable sets and its measure are given by

$$\mathcal{L}^N(B) := \prod_{k=1}^N \ell(I_k),$$

with $\ell(I_k)$ being the length of the interval $I_k \subset \mathbb{R}$. More generally, we claim that Borel sets are still elements of the σ-algebra $\mathcal{R}^N$ and also that the measure of a Borel set $S \subset \mathbb{R}^N$ is given by

$$\mathcal{L}^N(S) := \inf\left\{ \sum_{i\in\mathbb{N}} \mathcal{L}^N(B_i) : \{B_i\}_{i\in\mathbb{N}} \text{ being a covering by } N\text{-boxes of } S \right\}.$$

Now, we can define the measure space $(\mathbb{S}^{N-1}, \mathcal{S}^{N-1}, \Sigma_{N-1})$ whose measurable sets are given by the intersection of $\mathbb{S}^{N-1}$ with the elements of $\mathcal{R}^N$. In particular, the Borel sets of

the sphere are the intersection of $\mathbb{S}^{N-1}$ with the Borel sets of $\mathbb{R}^N$. Note that this definition is not arbitrary, it is based in the definition of induced topology on a subset of a topological space.

Additionally, in order to define the measure of a Borel subset $S \subset \mathbb{S}^{N-1}$, we shall relate it to the following Borel subset $C(S)$ of the Euclidean space $\mathbb{R}^N$:

$$C(S) := \bigcup_{r\in[0,1]} r \cdot S,$$

with

$$r \cdot S := \{rx \in \mathbb{R}^N : x \in S\},\ r \geq 0.$$

In particular, $C(\mathbb{S}^{N-1})$ is given by the unit closed ball $\overline{B(0,1)} \subset \mathbb{R}^N$. Also, we remark that the *Cavalieri's Principle* yields

$$\mathrm{Vol}\{C(S)\} = \int_0^1 \mathrm{Area}\{r \cdot S\}\, dr.$$

Moreover, for all $r \in [0,1]$,

$$\mathrm{Area}\{r \cdot S\} = r^{N-1}\mathrm{Area}(S),$$

since $S \mapsto r \cdot S$ is the transformation of a $(N-1)$-dimensional set by the linear map $x \mapsto rx$. Consequently,

$$\mathcal{L}^N\{C(S)\} = \int_0^1 r^{N-1}\mathrm{Area}(S)\, dr = \frac{1}{N}\,\mathrm{Area}(S).$$

That being said, we define the Lebesgue measure on the sphere as follows.

Definition B.1.2 (Lebesgue measure on Borel subsets of the sphere)**.** *Let $\mathcal{L}^N$ be the Lebesgue measure on $\mathbb{R}^N$. For any Borel set $S \subset \mathbb{S}^{N-1}$, we can define its Lebesgue measure Σ_{N-1} on the sphere as follows:*

$$\Sigma_{N-1}(S) := N\mathcal{L}^N\{C(S)\}.$$

The measure Σ_{N-1} inherits from the Radon regularity, the countable summability and the finiteness on compact sets from $\mathcal{L}^N$. Consequently, to prove that it is the Haar measure on $\mathbb{S}^{N-1}$, from its uniqueness, it suffices to prove that Σ_{N-1} is *rotation invariant*. As we will, see this is a consequence of the rotation invariance of $\mathcal{L}^N$. Namely, recall that, for a Borel set $S \subset \mathbb{R}^N$,

$$\mathcal{L}^N\left\{\boldsymbol{T}(S)\right\} = |\det(\boldsymbol{T})|\,\mathcal{L}^N(S)$$

for any matrix $\boldsymbol{T} \in \mathcal{M}_{N\times N}$. As the determinants of matrices in $O(N)$ and in $SO(N)$ equals 1, the result follows. Finally, we have the following result:

Proposition B.1.1 (Rotation invariance of Σ_{N-1})**.** *The Lebesgue measure on $\mathbb{S}^{N-1}$ is invariant by the action of the orthogonal group $O(N)$.*

Proof. At first, we claim that, for any $\boldsymbol{T} \in O(N)$ and for a Borel set $S \subset \mathbb{S}^{N-1}$, we have

$$\boldsymbol{T}C(S) = C(\boldsymbol{T}S),$$

with the notation $\boldsymbol{T}X$ representing the set

$$\boldsymbol{T}X := \{\boldsymbol{T}x : x \in X\}.$$

In fact, from the notation remarked above,

$$\begin{aligned}\boldsymbol{T}C(S) &= \{\boldsymbol{T}z : z \in C(S)\} \\ &= \{\boldsymbol{T}rx : x \in S \text{ and } r \in [0,1]\} \\ &= \{ry : y \in \boldsymbol{T}S \text{ and } r \in [0,1]\} := C(\boldsymbol{T}S).\end{aligned}$$

Finally, we have,

$$\begin{aligned}\Sigma_{N-1}(\boldsymbol{T}S) :&= N\mathcal{L}^N\{C(\boldsymbol{T}S)\} \\ &= N\mathcal{L}^N\{\boldsymbol{T}C(S)\} \\ &= N\mathcal{L}^N\{C(S)\} \quad := \Sigma_{N-1}(S),\end{aligned}$$

concluding that Σ_{N-1} is rotation invariant and, together with the previously cited conditions, the Haar measure on $\mathbb{S}^{N-1}$. □

As we said before, the Lebesgue/ Haar measure on $\mathbb{R}^N$ gives us an intuition of *uniformity of the space* since it is invariant by translations. Provided with this motivation, we shall define a uniform probability measure on $\mathbb{S}^{N-1}$ as the one with the *normalized Lebesgue measure* μ_{N-1} on $\mathbb{S}^{N-1}$, that is,

$$\mu_{N-1}(S) := \frac{\Sigma_{N-1}(S)}{\Sigma_{N-1}(\mathbb{S}^{N-1})},$$

with $S \subset \mathbb{S}^{N-1}$ being a Borel subset of $\mathbb{S}^{N-1}$.

Finally, we finish this section by remarking that the normalized Haar measure μ, the normalized surface area, and the normalized Lebesgue measure on $\mathbb{S}^{N-1}$ yield the same probability distribution in $\mathbb{S}^{N-1}$.

B.1.3 The Haar measures on $\mathbb{S}^{N-1}$ and $O(N)$

We end this section by exhibiting how the uniqueness of the Haar measure provides us with an way to compute the Haar measure of a set in $O(N)$ (or $SO(N)$) from the Haar measure of sets in $\mathbb{S}^{N-1}$, More precisely, we have the following:

Theorem B.1.1. *Let σ and μ_{N-1} be the normalized Haar measures on $(O(N), \|\cdot\|_{\mathcal{F}})$ (or $(SO(N), \|\cdot\|_{\mathcal{F}})$) and $(\mathbb{S}^{N-1}, \|\cdot\|_2)$, respectively. Moreover, let*

$$\begin{aligned}\phi : O(N) \times \mathbb{S}^{N-1} &\to \mathbb{S}^{N-1} \\ (\boldsymbol{T}, x) &\mapsto \boldsymbol{T}x.\end{aligned}$$

be a (left) group action of $O(N)$ on $\mathbb{S}^{N-1}$ with the product topology in its domain. For any fixed Borel $S \subset \mathbb{S}^{N-1}$ and $x_0 \in \mathbb{S}^{N-1}$, we have that

$$\mu_{N-1}(S) = \{\phi_*^{(x_0)}(\sigma)\}(S).$$

That is, the orbit map $\phi^{(x_0)} : O(N) \to \mathbb{S}^{N-1}$ pushes the Haar measure on $O(N)$ to the one on $\mathbb{S}^{N-1}$.

Bibliography

[Ach2003] Achlioptas, Dimitris. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

[AilChz2006] Ailon, Nir and Chazelle, Bernard. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *STOC '06: Proceedings of the Thirty-eighth Annual ACM Symposium on Theory of Computing*, pages 557–563, Seattle, WA, USA, 5 2006. ACM - Association for Computing Machinery.

[AilChz2009] Ailon, Nir and Chazelle, Bernard. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal in Computing*, 39(1):302–322, 2009.

[AlCh2006] Ailon, Nir and Chazelle, Bernard. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *STOC '06: Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, page 557–563, Seattle, WA, USA, 5 2006. Association for Computing Machinery.

[Alo2003] Noga Alon. Problems and results in extremal combinatorics—I. *Discrete Mathematics*, 273(1-3):31 – 53, 2003.

[Asc1973] F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, 1973.

[BDDW2006] Baraniuk, Richard and Davenport, M and DeVore, R and Wakin, M. The Johnson-Lindenstrauss Lemma meets compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006.

[Bsp2006] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, USA, 1 edition, 2006.

[BV2005] A. Bertoni and G. Valentini. Random projections for assessing gene expression cluster stability. In *IJCNN '05: Proceedings of the 2005 IEEE International Joint Conference on Neural Networks*, pages 149–154, Montreal, Que., Canada, 7 2005. IEEE.

[CT2005] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[DasGup2003] Dasgupta, Sanjoy and Gupta, Anupam. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.

[DeBm2006] S. Deegalla and H. Bostrom. Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In *ICMLA'06: Proceedings of the 2006 5th International Conference on Machine Learning and Applications*, pages 245–250, Orlando, FL, USA, 12 2006. IEEE.

[DeBm2007] Deegalla, Sampath and Boström, Henrik. Classification of microarrays with kNN: Comparison of dimensionality reduction methods. In *IDEAL '07: Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning*, pages 800–809, Birmingham, UK, 12 2007. Springer, Berlin, Heidelberg.

[DiSp2014] Joe Diestel; Angela Spalsbury. *The joys of Haar measure.* American Mathematical Society, 2014.

[Don2000] David Donoho. Aide-memoire. high-dimensional data analysis: The curses and blessings of dimensionality. *Components*, 2000.

[Don2006] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[Fdk2016] John Fedoruk. Dimensionality reduction via the Johnson and Lindenstrauss lemma: Mathematical and computational improvements. M.Sc. dissertation, University of Alberta, Alberta, Canada, 2016.

[Fdr1996] Herbert Federer. *Geometric measure theory.* Springer, Berlim, Germany, 1 edition, 1996.

[Feller1991] William Feller. *An Introduction to Probability Theory and Its Applications, Volume II.* Wiley, New York, USA, 2 edition, 1991.

[FklMae1986] P. Frankl and H. Maehara. Embedding the n-cube in lower dimensions. *European Journal of Combinatorics*, 7(3):221 – 225, 1986.

[FklMae1988] P Frankl and H Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B*, 44(3):355 – 362, 1988.

[FLM1977] Figiel, T. and Lindenstrauss, J. and Milman, V. D. The dimension of almost spherical sections of convex bodies. *Acta Math.*, 139:53–94, 1977.

[Foll1999] Gerald B. Folland. *Real analysis: modern techniques and their applications.* Wiley, 2 edition, 1999.

[FSJH2018] Fedoruk, John and Schmuland, Byron and Johnson, J. and Heo, Giseon. Dimensionality reduction via the Johnson–Lindenstrauss lemma: theoretical and empirical bounds on embedding dimension. *The Journal of Supercomputing*, 74(8):3933–3949, 2018.

[Gir2014] Christophe Giraud. *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC, 1 edition, 2014.

[GoLo2012] Gene H. Golub, Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, fourth edition, 2012.

[Herb1987] Herbert Edelsbrunner. *Algorithms in Combinatorial Geometry*. Springer, Berlin, Germany, 1 edition, 1987.

[HMae1984] Maehara, Hiroshi. On the sphericity for the join of many graphs. *Discrete Math.*, 49(3):311–313, 1984.

[HMae1986] Hiroshi Maehara. Sphericity exceeds cubicity for almost all complete bipartite graphs. *Journal of Combinatorial Theory, Series B*, 40(2):231 – 235, 1986.

[Hrup1981] Haagerup, Uffe. The best constants in the Khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

[Htl1933] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417–441, 1933.

[Ind2001] P. Indyk. Algorithmic applications of low-distortion geometric embeddings. In *FOCS '01: Proceedings of the 42nd IEEE symposium on Foundations of Computer Science*, pages 10–33, Newport Beach, CA, USA, USA, 10 2001. IEEE.

[IndMot1998] Indyk, Piotr and Motwani, Rajeev. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *STOC '98: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, pages 604–613, Dallas, Texas, USA, 1998. Association for Computing Machinery.

[IndMot2012] Indyk, Piotr and Motwani, Rajeev. Approximate nearest neighbors: Towards removing the curse of dimensionality. *Theory of Computing*, 8:321–360, 2012.

[JL1984] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[JW2013] Jayram, T. S. and Woodruff, David P. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *ACM Trans. Algorithms*, 9(3), 2013.

[KP1901] Karl Pearson F.R.S. Liii. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[KW2011] Krahmer, Felix and Ward, Rachel. New and improved Johnson–Lindenstrauss embeddings via the Restricted Isometry Property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

[Lat1999] Latała, Rafał. Tail and moment estimates for some types of chaos. *Studia Mathematica*, 135(1):39–53, 1999.

[Li2011] Li S. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4(1):66–70, 2011.

[LN2016] Kasper Green Larsen and Jelani Nelson. The Johnson-Lindenstrauss lemma is optimal for linear dimensionality reduction. In *ICALP '16: Proceedings of the 43rd International Colloquium on Automata, Languages and Programming*, page 775, Rome, Italy, 7 2016. Sapienza, Universita di Roma.

[LN2017] Green Larsen, Kasper and Nelson, Jelani. Optimality of the Johnson-Lindenstrauss lemma. In *FOCS '17: Proceedings of the 2017 IEEE 58th Annual IEEE Symposium on Foundations of Computer Science*, pages 633–638, Berkeley, CA, USA, 10 2017. IEEE - Computer Society.

[LV2007] Lee, John A., Verleysen, Michel. *Nonlinear Dimensionality Reduction*. Springer-Verlag New York, New York, USA, 1 edition, 2007.

[Mae1984] Hiroshi Maehara. Space graphs and sphericity. *Discrete Applied Mathematics*, 7(1):55 – 64, 1984.

[Mae1986] Hiroshi Maehara. On the sphericity of the graphs of semiregular polyhedra. *Discrete Mathematics*, 58(3):311 – 315, 1986.

[Mat2008] Matoušek, Jiří. On variants of the Johnson–Lindenstrauss lemma. *Random Structures & Algorithms*, 33(2):142–156, 2008.

[MiApAv2015] Shiri Artstein-avidan, Apostolos Giannopoulos, Vitali D. Milman. *Asymptotic geometric analysis. Part I*. American Mathematical Society, 2015.

[Mrk1994] Robert J. Marks II. *Computational Intelligence: Imitating Life*. IEEE, 1 edition, 1994.

[NNW12] Nelson, Jelani and L. Nguyen, Huy and Woodruff, David. On deterministic sketching and streaming for sparse recovery and norm estimation. *Linear Algebra and its Applications*, 441:152 – 167, 2012.

[Pach1980] Pach, János. Decomposition of multiple packing and covering. *2. Kolloquium über Diskrete Geometrie*, pages 169–178, 1980.

[RojNg2010] Rojo, Javier and Nguyen, Tuan. Improving the Johnson-Lindenstrauss lemma. 2010.

[TBau1997] Lloyd N. Trefethen, David Bau III. *Numerical linear algebra*. Siam - Society for Industrial and Applied Mathematics, Philadelphia, USA, 1997.

[Vem2004] Santosh Vempala. *The random projection method*. AMS - American Mathematical Society, Rhode Island, USA, 2004.

[Vsh2018] Vershynin, Roman. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, 2018.

[Ward2008] Rachel Mizsei Ward. Cross validation in compressed sensing via the Johnson Lindenstrauss lemma. 2008.