Dimensionality Reduction via The Johnson and Lindenstrauss Lemma: Mathematical and Computational Improvements

by

John Fedoruk

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

In an increasingly data-driven society, there is a growing need to simplify high-dimensional data sets. Over the course of the past three decades, the Johnson and Lindenstrauss (JL) lemma has evolved from a highly abstract mathematical result into a useful tool for dealing with data sets of immense dimensionality. The lemma asserts that a set of high-dimensional points can be projected into lower dimensions while approximately preserving the pairwise distance structure. The JL lemma has been revisited many times, with improvements to both its sharpness (i.e., bound on the reduced dimensionality) and its simplicity (i.e., mathematical derivation). In 2008 Matoušek [36] provided generalizations of the JL lemma that lacked the sharpness of earlier approaches. The current investigation seeks to strengthen Matoušek's results by maintaining generality while improving sharpness. First, Matoušek's results are reproved with more detailed mathematics and, second, computational solutions are obtained on simulated data in Matlab. The reproofs result in a more specific bound than suggested by Matoušek while maintaining his level of generality. However, the reproofs lack the sharpness suggested by earlier, less general approaches to the JL lemma. The computational solutions suggest the existence of a result that maintains Matoušek's generality while attaining the sharpness suggested by his predecessors. The collective results of the current investigation support the notion that computational solutions play a critical role in the development of mathematical theory.

Acknowledgements

I thank my supervisor, Dr. Giseon Heo, for her patience and continued support. I also thank my co-supervisor, Dr. Byron Schmuland, whose passion for mathematics and statistics remains my greatest inspiration. I thank the Department of Mathematical and Statistical Sciences at the University of Alberta for providing me with a sound education. Finally, I thank my family.

Table of Contents

List of Tables

Notation

The following notation is used throughout this thesis.

- Random variables are denoted by capital letters, such as X, with the exception of the letter C, which is used to denote a constant value of particular importance.

- A sequence of n random variables is denoted as X_1, X_2, $\cdots, X_n$

- $X \sim F$ is the notation that is used to say that the random variable X follows the probability distribution F.

- $X \overset{D}{=} Y$ is the notation that is used to say that the random variables X and Y share the same probability distribution. That is, $X \overset{D}{=} Y$ means that there is a particular probability distribution F, such that $X \sim F$ and $Y \sim F$.

- $\|x\|$ is used to denote the L^2 norm of the point x. That is, if x is a d-dimensional vector, then $\|x\|= \sqrt{\sum_{i=1}^{d} x_i^2}$

- $\|x\|_\infty$ is used to denote the L^∞ norm of the point x. That is, if x is a d-dimensional vector, then $\|x\|_\infty= \max\{|x_i|: 1 \leq i \leq d\}$

Chapter 1

Introduction and Overview

Human decision making depends upon the analysis of data, although methods of data analysis have changed dramatically over time. Data analysis has evolved from basic cognitive processing of sensory input to statistical analysis of data sets that consist of a wide array of information. Classical statistical analysis requires data sets wherein the number of dimensions correspond to a small number of carefully chosen variables. In recent years, advances in computer technology have allowed for the collection of massive amounts of data which often include a large number of irrelevant and redundant variables. Accordingly, classical statistical methods are limited in their capacity to deal effectively with contemporary data sets. This has led to a new branch of statistics referred to as *high-dimensional data analysis* [21, 41]. To appreciate the scope of the current investigation, it is necessary to review the concepts of *dimensionality* and *dimensionality reduction.*

1.A Dimensionality

Statistics is a branch of mathematics focused primarily on the analysis of data. A data set is an $n \times d$ matrix X, consisting of n observations, where each observation is characterized by d covariates. From a purely mathematical standpoint, X is viewed a collection of n points, $x \in \mathbb{R}^d$. The *dimensionality* of x refers to the number of dimensions to which x belongs; in this case, x is said to be d-dimensional, expressed as $x = (x_1, x_2, \cdots, x_d)$, where $x_i \in \mathbb{R}$ is said to be the i^{th} *coordinate* of x, for $i = 1, 2, \cdots, d$. From a statistical standpoint, x is viewed as an observation whose dimensions correspond to different variables, where each coordinate x_i of x corresponds to a measurement of the i^{th} variable of interest.

Although each row of X is represented by a point x in d-dimensional space, it is often the case that most of the structure in X can be expressed through a lower dimensional representation. The *extrinsic dimensionality* of X refers to the dimensionality in which its data points are recorded; in this case, X has d extrinsic dimensions. There is an alternative, and arguably more important, type of dimensionality known as *intrinsic dimensionality*. The intrinsic dimensionality of X refers to the number of dimensions that are needed in order to answer a particular query of interest. For example, in d-dimensional signal processing, the intrinsic dimensionality is the number $k \leq d$ of variables that are required to effectively represent the signal [13].

1.B Dimensionality Reduction

The main goal of inferential statistics is to collect sample data in order to develop models that may be used to make claims about a population of interest. Classical statistical methods are not always viable when dealing with high-dimensional data. Thus, the first step in the analysis of a high-dimensional data set is to reduce its dimensionality. That is, given some $X_{n \times d}$, where $d >> n$, find a lower dimensional representation $Y_{n \times k}$, with $k < d$, so that much of the information contained in X can be obtained from Y. There are a number of statistical approaches to dimensionality reduction including model selection methods in regression and classification, regularization methods such as Lasso and support vector machines, principal component analysis, multidimensional scaling, and isometric mappings [13, 21, 24, 43]. Some of these approaches aim to directly identify the intrinsic dimensionality of the data while others aim to reduce the extrinsic dimensionality by transforming the data into an alternative, low-dimensional representation.

An issue common to many classical statistical approaches is that they require matrix operations that are computationally expensive when dealing with high-dimensional data. For example, principal component analysis and multidimensional scaling rely on some form of eigendecomposition while regression requires matrix inversion. Consequently, there is a growing need for techniques in dimensionality reduction that efficiently reduce the extrinsic dimensionality of high-dimensional data sets so that classical methods can be performed. The Johnson and Lindenstrauss lemma is essential to modern techniques in dimensionality reduction.

1.C Johnson and Lindenstrauss Lemma

In 1984 Johnson and Lindenstrauss [31] introduced a mathematical result that came to be known as the *JL lemma*. The lemma asserts that a set of high-dimensional points can be projected into lower dimensions while approximately preserving the pairwise distance structure between points. Specifically, the lemma guarantees the existence of a mapping $T : \mathbb{R}^d \to \mathbb{R}^k$, where $k = O(\epsilon^{-2} \log n)$, such that pairwise distances are maintained to within a multiplicative factor of $1 \pm \epsilon$. Johnson and Lindenstrauss use the lemma as a tool to prove extensions of Lipschitz mappings into a Hilbert space. The JL lemma has since evolved into an effective tool in high-dimensional data analysis.

Since its inception, the JL lemma has been subject to numerous improvements which have contributed to its evolution into an essential tool in dimensionality reduction. The evolution of the JL lemma is characterized by two phases. The first phase of the evolution reflects a number of improvements to the lower bound on the reduced dimensionality k. These improvements are the result of a series of probabilistic refinements which have both simplified the proof of the lemma and provided a tighter lower bound on k. The second phase of the evolution of the JL lemma involves improvements to the efficiency of the transformation T. Improvements to efficiency occur in two ways: 1) reduction of the number of operations needed to compute the transformation and 2) reduction of the amount of space needed to compute the transformation. Currently, the JL lemma is among the leading approaches to dimensionality reduction.

The JL lemma appears in a wide array of applications. In some cases, the lemma is used as an alternative to classical methods in dimensionality reduction; in other cases, it is used as a preprocessing step prior to use of classical methods. However, there are certain situations in which the JL lemma is the ideal method of dimensionality reduction. Applications of the JL lemma to dimensionality reduction include image retrieval, genetic algorithms, classification in machine learning, data streaming, nearest neighbor search, and compressed sensing. Despite successful applications of the JL lemma, recent treatments have lost the clarity and precision that characterized earlier stages of its evolution.

1.D The Problem and the Approach

In 2008 Matoušek [36] made a significant improvement to the JL lemma. Particularly, he proves several generalized statements of the JL lemma, two of which are relevant to this thesis: one ties together a number of earlier treatments of the lemma; the other stimulates subsequent treatments of the lemma. The following gives a brief summary of Matoušek's results while Chapter 2 contains a detailed discussion. Matoušek provides two families of mappings $T : \mathbb{R}^d \to \mathbb{R}^k$ such that each $x \in \mathbb{R}^d$ satisfies

$$\mathbb{P}((1-\epsilon)\|x\| \leq \|T(x)\| \leq (1+\epsilon)\|x\|) \geq 1-\delta.$$

Here, $d \in \mathbb{N}$ is the initial dimensionality of x, $\epsilon \in (0, 1/2]$ is the distortion parameter that controls the degree to which the length of x is distorted under T, and $\delta \in (0, 1)$ is the probability parameter that controls the probability that this length is distorted by no more than $1 \pm \epsilon$. Matoušek provides two mappings that satisfy the above conditions: one gives reduced dimensionality of $k = \frac{C \log(2/\delta)}{\epsilon^2}$ and the other gives reduced dimensionality of $k = \frac{C \log(4/\delta)}{\epsilon^2}$, for a constant C.

The generality of Matoušek's results leads to bounds on the reduced dimensionality k that are weaker than those given by previous mathematicians. In particular, Matoušek does not provide clear formulation for the constant C; he merely asserts that it must be "sufficiently large".

My approach to improving the JL lemma involves the development of mathematically refined theorems that are tested and further refined with a computational approach. First, I develop theorems based on Matoušek's treatment of the JL lemma [36]. Through detailed analysis, I obtain specific bounds on k while maintaining Matoušek's level of generality. Next, I measure the efficacy of these results using a computational approach with simulated data. Specifically, several data sets are simulated in the Matlab environment and each of these data sets are projected into k dimensions using the results of my refined theorems. The results of this computational approach imply a lower bound on k than that suggested by my theorems, since the proportion of lengths distorted by no more than $1 \pm \epsilon$ is much larger than $1 - \delta$. Accordingly, the computational approach is repeated several times, with repetitions projecting points into sequentially decreasing dimensions. The true value of k is approximated with the reduced dimensionality of the data points as the proportion of lengths distorted by no more than $1 \pm \epsilon$ approaches $1 - \delta$.

1.E Improvements to the Johnson and Lindenstrauss Lemma

My approach to addressing limitations of the JL lemma results in numerous improvements to the lemma. My mathematical results are obtained by following the same line of reasoning as Matoušek [36]. In particular, my proofs are built upon the construction of six lemmas which I prove in a sequential manner similar to Matoušek. However, each of my proofs contain more detail, clarity and precision. Accordingly, my results contain specific bounds, whereas Matoušek relies exclusively on asymptotic notation. More importantly, my detailed treatment clearly indicates dependencies between the constants and parameters of interest. This provides valuable insight that may motivate further research.

My computational results are of equal importance to my mathematical results. I first test the results of my refined theorems with simulated data and such tests indicate weakness in the bounds suggested by my theorems. I then utilize a computational approach to empirically estimate the bound on the reduced dimensionality. This approach provides evidence of a much smaller bound, although only on certain sets of simulated data. However, the computational results also contain patterns that may guide future refinement of my theorems.

My hybrid mathematical-computational approach results in significant improvements to the JL lemma. These improvements have far reaching implications for continued research in dimensionality reduction.

Chapter 2

Evolution of the Johnson and Lindenstrauss Lemma

This chapter provides a review of the literature relevant to a comprehensive understanding of the Johnson and Lindenstrauss lemma (JL lemma). First, the main features of the JL lemma are presented. Next, the key improvements to the JL lemma are discussed. Lastly, current applications of the JL lemma are reviewed.

2.A Inception of the Johnson and Lindenstrauss Lemma

In 1984 Johnson and Lindenstrauss [31] assert the existence of a mapping T that gives an orthogonal projection of n points from $\mathbb{R}^d$ onto a random k-dimensional subspace with dimensionality $O(\epsilon^{-2} \log n)$, such that pairwise distances are maintained to within a factor of $1 \pm \epsilon$. Johnson and Lindenstrauss propose following result:

Theorem 2.1: ***Johnson and Lindenstrauss (1984):*** *Given a set P of n points in $\mathbb{R}^d$, for some $n, d \in \mathbb{N}$, and given $\epsilon \in (0,1)$, there exists $k_0 = O(\epsilon^{-2} \log n)$ such that, if $k \geq \lceil k_0 \rceil$, there exists a linear mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ such that for any two points $u, v \in P$,*

$$(1-\epsilon)\|u-v\|^2 \leq \|T(u)-T(v)\|^2 \leq (1+\epsilon)\|u-v\|^2.$$

Since T is a linear mapping, there is no loss of generality in replacing the quantities $u-v$ and $T(u)-T(v)$ with x and $T(x)$, for a unit vector $x \in \mathbb{R}^d$, so that the above equation can be re-expressed in the following, more convenient form:

$$(1-\epsilon) \leq \|T(x)\|^2 \leq (1+\epsilon). \tag{2.1}$$

Johnson and Lindenstrauss provide a lengthy, technical proof using geometric approximation. In their proof, Johnson and Lindenstrauss choose the mapping T to be an orthogonal projection onto a random k-dimensional subspace of $\mathbb{R}^d$, multiplied by the scaling factor of $\sqrt{d/k}$. The main idea is as follows:

- Project a collection of points from d-dimensions into a random k-dimensional subspace.

- The pairwise distances between each set of points, both before and after projection, correspond to a vector starting at the origin.

- On average, the length of each k-dimensional vector is $\sqrt{k/d}$ times the length of the corresponding initial vector in d-dimensions and most of these lengths are closely concentrated about this expectation.

- Hence, multiplying each projection by the scaling factor of $\sqrt{d/k}$ yields a set of k-dimensional vectors, each similar in lengths to their d-dimensional counterpart.

- Lastly, choose some prespecified level of tolerance for distortion of length. Then, with nonzero probability, each length is preserved to within this level of tolerance.

A mapping T that satisfies (2.1) is said to preserve ϵ-distortion of the length of x. Thus, the JL lemma states that an arbitrary set of points can be projected into lower dimensions, under a mapping that preserves ϵ-distortion of pairwise distances. The image of a set P under the mapping T is referred to as a *JL embedding* [1]

Since its inception in 1984, the JL lemma has been subject to considerable scrutiny. The JL lemma has been reproved many times, with each new proof providing a sharpened (i.e., reduced bound) and/or simplified result. However, there is one particular feature that is common to all JL embeddings: the mapping T projects a vector into lower dimension, and the length of this projection is sharply concentrated around its expectation [1]. The existence of each such mapping is established through the probabilistic method: the random mapping T is shown to have nonzero probability of being sufficiently concentrated about its expectation. Each proof of the JL lemma relies on the construction of a random linear map $T : \mathbb{R}^d$ to $\mathbb{R}^k$ of the form

$$T = XR,$$

where R=$R_{d\times k}$ is a random projection matrix acting on the data structure

$X{=}X_{n\times d}$ (with row vectors corresponding to the points in P) and $T{=}T_{n\times k}$ is the resulting, transformed data structure in k-dimensions.

The proof then follows by establishing that the random mapping T satisfies a probability statement akin to the following: if x is a unit vector in $\mathbb{R}^d$, then

$$\mathbb{P}((1-\epsilon) \leq \|T(x)\|^2 \leq (1+\epsilon)) \geq 1 - \frac{1}{n^2}. \qquad (2.2)$$

There exists a wide array of mappings T that satisfy a statement similar to (2.2). However, there are two features of such a map that are of particular importance:

1. The transformation T leads to a reduced dimensionality k that is as small as possible for any fixed ϵ.
2. The transformation T is as efficient as possible so as to minimize runtime (which can be very expensive when dealing with data sets of immense dimensionality).

The next two sections contain a detailed summary of the evolution of the JL lemma with respect to the above two features of the transformation T. Section 2.B covers the early development of the JL lemma, wherein particular focus is placed on improving the lower bound on the reduced dimensionality k. Section 2.C covers some of the more recent treatments of the JL lemma which aim to improve the efficiency of JL embeddings. Section 2.D focuses on practical issues regarding applications of the JL lemma.

2.B Improving the Lower Bound of JL Embeddings

This section provides a detailed summary of the significant improvements on the lower bound of k, the reduced dimensionality of JL embeddings. Recall that Johnson and Lindenstrauss [31] provide the first statement of the lemma, which asserts the existence of a mapping T that gives an orthogonal projection of n points from $\mathbb{R}^d$ onto a random k-dimensional subspace with dimensionality $O(\epsilon^{-2} \log n)$, such that pairwise distances are maintained to within a factor of $1 \pm \epsilon$, i.e., they preserve ϵ-distortion. Although the JL lemma was an impactful result that stimulated considerable research, there was room for improvement. In particular, as Rojo observes [39], Johnson and Lindenstrauss do not actually provide a clear construction of the orthogonal mapping T but rather, they merely assert its existence. Moreover, the lower bound $O(\epsilon^{-2} \log n)$ begged further investigation.

Frankl and Meahara [23] provide the first significant improvement to the JL lemma. They tighten the lower bound on the reduced dimensionality k. Additionally, they provide an explicit formulation for a mapping that yields a JL embedding. The key to their improvement involves replacing the random k-dimensional subspace with a collection of k random, orthonormal vectors. Such an approach allows for a simpler proof than that given by Johnson and Lindenstrauss and, at the same time, it attains a sharper bound on k. The following is the result provided by Frankl and Meahara.

Theorem 2.2: ***Frankl and Meahara (1988):*** *Given a set P of n points in $\mathbb{R}^d$, for some $n, d \in \mathbb{N}$, and given $\epsilon \in (0, 1/2)$, choose $k \geq \lceil 9(\epsilon^2 - 2\epsilon^3/3)^{-1} \log(n) \rceil + 1$. If $n > k^2$, then there exists a linear mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ such that for any two points $u, v \in P$,*

$$(1-\epsilon)\|u-v\|^2 \leq \|T(u) - T(v)\|^2 \leq (1+\epsilon)\|u-v\|^2.$$

Frankl and Meahara establish that the mapping is of the form $T = \sqrt{\frac{d}{k}} XR$, where X=$X_{n\times d}$ is the data structure corresponding to the points in P, and R=$R_{d\times k}$ is the projection matrix consisting of random orthonormal column vectors.

Indyk and Motwani [28] provide the next improvement to the JL lemma by simplifying Frankl and Meahara's proof through relaxation of the conditions of orthogonality and unit length among the column vectors of the projection matrix R. More specifically, Indyk and Motwani reprove Theorem 2.2 using a projection matrix R that consist of independent, Gaussian random vectors, with each coordinate following $\mathcal{N}(0, 1/d)$.

The approach taken by Indyk and Motwani leads to a result that is almost equivalent to that of Frankl and Meahara because, in high dimensions, independent random vectors have high probability of being nearly orthogonal [11] and also, the length of each column vector R_i of R has high probability of being close to the expected length of 1. The latter result follows from *Gaussian 2-stability*: if $\alpha_i \in \mathbb{R}$ and $Z_i \sim \mathcal{N}(0,1)$ for $i = 1, 2, \cdots, d$, then $\sum_{i=1}^d \alpha_i Z_i \sim \mathcal{N}(0, \sum_{j=1}^d \alpha_j^2)$. Hence, choose projection coefficients r_{ij} that are independent, identically distributed random variables, $r_{ij} \sim \mathcal{N}(0, 1/d) \overset{D}{=} \frac{1}{\sqrt{d}} Z_i$, and let $R_i = (r_{1i}, r_{2i}, \cdots, r_{di})$ denote the i^{th} column of R so that

$$\begin{aligned}\|R_i\|^2 &\overset{D}{=} \sum_{j=1}^{d} r_{j,i}^2 \\ &\overset{D}{=} \frac{1}{d}\sum_{j=1}^{d} Z_i^2 \\ &\overset{D}{=} \frac{1}{d}\chi_d^2.\end{aligned}$$

Thus, the squared length of each column of R has mean 1 and variance $2/d$ which, for large d, is very close to 0. Gaussian 2-stability also plays a role in simplifying the proof of Theorem 2.2. In particular, an argument similar to that above shows that each projected point has squared length equal to

$$\begin{aligned}\|T(x)\|^2 &= \|\sqrt{\frac{d}{k}}xR\|^2 \\ &= \sum_{i=1}^{k}\Big(\sum_{j=1}^{d}\sqrt{\frac{d}{k}}x_j r_{j,i}\Big)^2 \\ &\overset{D}{=} \frac{1}{k}\sum_{i=1}^{k}\Big(\sum_{j=1}^{d} x_j Z_i\Big)^2 \\ &\overset{D}{=} \frac{1}{k}\sum_{i=1}^{k}\Big(\|x\|Z_i\Big)^2 \\ &\overset{D}{=} \frac{\|x\|^2}{k}\sum_{i=1}^{k} Z_i^2 \\ &\overset{D}{=} \frac{\|x\|^2}{k}\chi_k^2.\end{aligned}$$

Therefore,

$$\frac{k\|T(x)\|^2}{\|(x)\|^2} \sim \chi_k^2. \tag{2.3}$$

In order to verify that (2.2) holds true when using projection coefficients that are $\mathcal{N}(0, 1/d)$, use (2.3) and the well established concentration bounds for the Chi-square distribution.

Dasgupta and Gupta [17] provide the next significant improvement to the JL lemma. Like Indyk and Motwani, Dasgupta and Gupta rely on projection coefficients that are spherically symmetric, (in the form of normal random variables that are scaled so that the expected length of each column of R is 1, i.e., $r_{ij} \sim \mathcal{N}(0, 1/d)$). However, Dasgupta and Gupta take spherical symmetry one step further by making the following observation: the projection of a fixed unit vector onto a random hyperplane through the origin follows the same distribution as the projection of a uniformly random unit vector projected onto a fixed, k-dimensional subspace. A uniformly random unit vector is a Gaussian vector scaled to unit length and, for simplicity, the fixed k-dimensional subspace is taken to be the first k coordinates of this scaled Gaussian vector. Dasgupta and Gupta's proof reduces to analysis of a scaled Gaussian random variable, leading to the following result:

Theorem 2.3: ***Dasgupta and Gupta (1999):*** *Consider a set P of n points in $\mathbb{R}^d$, for some $n, d \in \mathbb{N}$. Choose $\epsilon \in (0, 1)$ and $k \geq \lceil 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \log(n) \rceil$. Then there exists a linear mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ such that for any two points $u, v \in P$,*

$$(1 - \epsilon)\|u - v\|^2 \leq \|T(u) - T(v)\|^2 \leq (1 + \epsilon)\|u - v\|^2.$$

Theorem 2.3 gives the tightest bound on k thus far discussed and, surprisingly, the proof is much simpler than the proofs of previously discussed approaches. In fact, Dasgupta and Gupta provide a bound on k that is so tight that nearly a decade would elapse before its next improvement.

Matoušek [36] makes the next contribution by generalizing and simplifying many of the previously discussed treatments of the JL lemma. In particular, Matoušek provides a generalized statement of the JL lemma using the language of subgaussian tails. Although Matoušek's results do not actually lead to a tighter JL embedding, his treatment of the JL lemma deserves mention due to its clever marriage of generality and simplicity.

Matoušek's first main result follows from the observation that many of the previous treatments of the JL lemma rely on projection coefficients that have a subgaussian tail. A random variable X is said to have a *subgaussian upper tail* if $\exists\ a > 0$ so that $\mathbb{P}(X > \lambda) \leq \exp(-a\lambda^2)$, for all $\lambda > 0$; if this inequality holds for all $\lambda \leq \lambda_0$, then X is said to have a *subgaussian upper tail up to* λ_0. Furthermore, if $-X$ also has a subgaussian upper tail, then X is said to have a *subgaussian tail.* Lastly, suppose that $X_1, X_2, \cdots$ is a sequence of random variables, each with subgaussian tail. If the constant a in the subgaussian tail inequality is the same for each X_i, then the X_is are said to have a *uniform subgaussian tail.* Matoušek's first result follows.

Theorem 2.4: **Matoušek** ***(2008):*** *Consider a collection $\{R_{ij}\}_{i,j}$ of independent random variables, where $\mathbb{E}(R_{ij}) = 0$ and $\mathbb{V}(R_{ij}) = 1$ for each R_{ij} and also, suppose that $\{R_{ij}\}_{i,j}$ has a uniform subgaussian tail. Next, for fixed $d \in \mathbb{N}$, $\epsilon \in (0, 1/2]$, $\delta \in (0, 1)$, set $k = \frac{C \log(2/\delta)}{\epsilon^2}$, for a constant C which depends on the constant a in the subgaussian tail inequality for R_{ij}. Finally, define the random linear map $T : \mathbb{R}^d \to \mathbb{R}^k$ as follows:*

$$T(x)_i = \frac{1}{\sqrt{k}} \sum_{j=1}^{d} R_{ij} x_j, \ \textit{for } i = 1, 2, \cdots, k,$$

where $T(x)_i$ is the i^{th} coordinate of $T(x) \in \mathbb{R}^k$, and x_j is the j^{th} coordinate *of $x \in \mathbb{R}^d$. Then every $x \in \mathbb{R}^d$, satisfies*

$$\mathbb{P}((1-\epsilon)\|x\| \leq \|T(x)\| \leq (1+\epsilon)\|x\|) \geq 1-\delta.$$

In the proof of Theorem 2.4, Matoušek defines the random mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ as $T(x) = \frac{1}{\sqrt{k}}xR$, where $R = R_{d\times k}$ is the projection matrix consisting of projection coefficients with a uniform subgaussian tail. He then shows that $\mathbb{P}(\|T(x)\| \leq 1-\epsilon) \leq \delta/2$ and that $\mathbb{P}(\|T(x)\| \geq 1+\epsilon) \leq \delta/2$, so that $\mathbb{P}(1-\epsilon \leq \|T(x)\| \leq 1+\epsilon) \geq 1-\delta$. The proof relies on Markov's inequality (defined in Chapter 4), and the properties of subgaussian tails, together with the fact that $\|T(x)\|^2-1$ has a subgaussian tail up to $\sqrt{k}$.

Rojo and Nguyen [39] provide an alternative approach to improving the lower bound of JL embeddings which involves the use of numerical methods. Indeed, each of the previously discussed approaches generally involve the use of Markov's inequality and often resort to the use of moment generating functions. Rojo and Ngueyn's result is given below in Theorem 2.5.

Theorem 2.5: ***Rojo and Nguyen (2010):*** *For any $\epsilon \in (0,1)$, $n \in \mathbb{N}$, let k be the smallest even integer satisfying: $(\frac{1+\epsilon}{\epsilon})g(k,\epsilon) \leq \frac{1}{n^2}$, where $g(k,\epsilon) = \frac{\frac{k(1+\epsilon)}{2}^{\frac{k}{2}-1}}{(\frac{k}{2}-1)!}e^{-\frac{k(1+\epsilon)}{2}}$ is a decreasing function in k. Then for any set P of n points in $\mathbb{R}^d$, there is a linear map $T : \mathbb{R}^d \to \mathbb{R}^k$ such that for any $u, v \in P$,*

$$P[(1-\epsilon)\|u-v\|^2 \leq \|T(u)-T(v)\|^2 \leq (1+\epsilon)\|u-v\|^2] \geq 1-2/n^2.$$

The bound for k can be obtained numerically by finding the smallest even integer k satisfying the inequality $(\frac{1+\epsilon}{\epsilon})g(k,\epsilon) \leq \frac{1}{n^2}$.

Rojo and Nguyen present results from a variety of simulations which suggest that their approach can lead to significantly smaller k (in some instances up to 40 % smaller than what has been obtained using previously discussed methods). Their treatment of the JL lemma is the most recent improvement on the lower bound for the reduced dimensionality k.

The improvements to the JL lemma discussed in this section have led to a lower bound on k that is essentially optimal. However, each approach fails to address one major issue, namely, the immense runtime associated with the computation of JL embeddings. The next section shifts to improvements to the JL lemma regarding the efficiency of JL embeddings.

2.C Improving Efficiency of JL Embeddings

This section provides a detailed summary of significant improvements regarding the efficiency of JL embeddings. Although the JL lemma is improved upon by each of the results discussed in Section 2.B, none of these approaches address the immense runtime associated with the computation of the corresponding JL embedding. A JL embedding can be performed to transform high-dimensional data into lower dimensions so that computationally expensive operations can be performed with less runtime. The embedding itself can be computationally expensive, so much so that, in some situations, there is little to gain by performing the embedding. For this reason, the JL lemma has been subjected to a new sequence of improvements, each of which improve the efficiency of the JL embedding, either through reduction in the number of operations required or through reduction in the amount of space required.

2.C.1 Improving Efficiency via Sparsification of the Projection Matrix

One approach to improving the efficiency of the JL lemma is through increasing the speed of JL embeddings via sparsification of the projection matrices. In this context, *sparsification* refers to replacing a large number of matrix entries with zero entries. This leads to a significant reduction in the number of operations needed to perform an embedding. The methods discussed in the previous section involve matrix multiplication between the data structure X and a projection matrix R populated with real numbers; as n and d increase, the computational expense of such matrix multiplication grows rapidly. This issue is overcome by the use of the sparse projection matrices.

Achlioptas [1] (2003) provides the first development of a faster JL embedding through the use of projection coefficients r_{ij} that are independent random variables, identically distributed according to either of the following distributions:

$$r_{ij} = \begin{cases} 1 & \text{with probability } 1/2, \\ -1 & \text{with probability } 1/2, \end{cases} \tag{2.4}$$

or

$$r_{ij} = \begin{cases} \sqrt{3} & \text{with probability } 1/6, \\ -\sqrt{3} & \text{with probability } 1/6, \\ 0 & \text{with probability } 2/3. \end{cases} \tag{2.5}$$

Achlioptas essentially reproves Theorem 2.3, (the result of Dasgupta and Gupta discussed in section 2.B) using projection coefficients distributed according to either (2.4) or (2.5). The main advantage of using projection coefficients that are distributed according to (2.4) is that each coordinate $T(x)_i$ of the embedding is computed using only addition and subtraction of the original coordinates of a data point x while no multiplication is necessary. More specifically, $T(x)_i$ is calculated as follows: partition the coordinates of x randomly into two groups, compute the sum of each group, and set $T(x)_i$ to be the difference of these two sums. This approach significantly improves runtime when obtaining a JL embedding, since it is not necessary to perform repeated matrix multiplication (as is the case when projection coordinates are independent, identically distributed Gaussian random variables). Furthermore, a JL embedding can be found roughly 3 times faster when projection coefficients are distributed according to (2.5) instead of (2.4). Regardless of whether projection coefficients are distributed according to (2.4) or (2.5), computation of each coordinate $T(x)_i$ involves addition and subtraction of the original coordinates. However, when (2.5) is used, only about 1/3 of the original coordinates are considered while the remaining coordinates are set to 0 and therefore, roughly 1/3 as many operations are required.

Achlioptas' approach results in much faster computation of JL embeddings and, perhaps more importantly, this improvement in efficiency does not penalize the quality of the embedding. First, he shows that spherical symmetry of the projection coefficients is not necessary in order to obtain a JL embedding but rather, concentration of the projected points is sufficient. He then shows that the even moments of his random projections are dominated by those of the spherically symmetric case, so that a JL embedding can be found with probability at least as large as that in the spherical case. The question becomes: can efficiency be further improved while maintaining tightness of the embedding through use an even sparser projection matrix? Achlioptas addresses this question, claiming that the projection matrix cannot consist of much more than two thirds zero entries without sacrificing tightness of the embedding.

Ailon and Chazelle [2] (2006) extend Achlioptas' result by showing that highly sparse projection matrices can be used, but only on data points that are well-spread[1]. Recall that Achlioptas [1] shows that roughly two thirds of the projection matrix should be zero entries in order to guarantee that the optimal lower bound on k is attained; as the sparsity of the projection matrix increases beyond this threshold, the bound on k begins to suffer. In particular, sparse projection matrices are ineffective when dealing with data of low intrinsic dimensionality, since a sparse projection matrix tends to cause large distortion of a sparse vector. However, Ailon and Chazelle avoid this issue by considering data points that are well-spread across the dimensions in which they are observed; they show that, in this case, sparser projection matrices, more than two thirds of which are zero entries, can be used to obtain a faster JL embedding.

Ailon and Chazelle not only prove that sparse projection matrices may be used on data points that are well-spread, they also provide a clever construction that allows for the use of a sparse projection matrix on any data. The key to their approach involves preconditioning the projection with a randomized Fourier transform that isometrically increases the support of a sparse vector

[1] A unit vector is well-spread if it is close to $\frac{1}{\sqrt{d}}(\pm 1, \pm 1, \cdots, \pm 1)$ while something close to $(1, 0, \cdots, 0)$ is not well-spread since most of its mass lies in its first dimension.

[2]; such a preconditoning is achieved by defining the projection matrix R to be the product of three matrices: $R = (MHD)^T$, where M and D are random matrices, and H is deterministic[2]. More specifically:

- $M = M_{k \times d}$ is a sparse matrix consisting of roughly $1-q$ zero entries, and roughly q entries that are i.i.d $\mathcal{N}(0, 1/q)$, where $q = \min \left\{ \Theta(d^{-1} \log^2 n), 1 \right\}$.
- $H = H_{d \times d}$ is a normalized Walsh matrix[3].
- $D = D_{d \times d}$ is a diagonal matrix whose entries are independent, identically distributed Uniform$\{0, 1\}$.

Ailon and Chazelle then prove the following variant of the JL lemma using a projection matrix $R = (MHD)^T$:

Theorem 2.6: ***Ailon and Chazelle (2006):*** *Given a set P of n points in $\mathbb{R}^d$, for some $n, d \in \mathbb{N}$, and given $\epsilon \in (0, 1)$, choose $k = C\epsilon^{-2} \log(n)$, for some suitably large constant C. Then there exists a random linear mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ of the form $T(x) = x(MHD)^T$, such that, with probability of at least 2/3, the following two events occur:*

1. $(1 - \epsilon)k\|x\| \leq \|T(x)\| \leq (1 + \epsilon)k\|x\|$, *and*
2. *The mapping T requires $O(d \log(d) + \min\{d\epsilon^{-2} \log(n), \epsilon^{-2} \log^3(n)\})$ operations*

[2] Actually, Ailon and Chazelle consider x to be a $d \times 1$ column vector and they define $T(x) = MHDx$. In the current discussion, however, x is regarded as a $1 \times d$ row vector which leads to the alternative expression $T(x) = x(MHD)^T$

[3] A Walsh matrix consists of entries that are equal to ±1 such that any two row vectors are orthogonal, and any two column vectors are orthogonal; normalized simply means the matrix is multiplied by $d^{-1/2}$.

Ailon and Chazelle introduce the term *Fast Johnson and Lindenstrauss Transform*, or simply FJLT, to describe JL embeddings of the form discussed in Theorem 2.6. The FJLT has itself sparked a great deal of research focused on improving the speed of the transform. These improvements are subsequently discussed in Section 2.C.2

Matoušek [36] (2008) makes the next contribution by providing a simplified version of Ailon and Chazelle's result, wherein he introduces a computationally simpler alternative to the sparse matrix M from Theorem 2.6. In fact, Matoušek improves upon Ailon and Chazelle's result in a manner analogous to Achlioptas' improvement on the Indyk and Motwani result. In particular, Matoušek replaces the Gaussian projection coefficients of the matrix M with scaled coefficients distributed over ±1 such that the projection matrix R has entries that are distributed according to the following:

$$r_{ij} = \begin{cases} 1/\sqrt{q} \quad \text{with probability } q/2, \\ -1/\sqrt{q} \quad \text{with probability } q/2, \\ 0 \quad \text{with probability } 1-q, \end{cases} \tag{2.6}$$

where, q controls the sparsity of the projection matrix.

Before moving on, it is useful to note that Theorem 2.4 (Matoušek's first result, discussed in section 2.B) can be applied when projection coefficients are distributed according to (2.6), since such coefficients have a zero mean, unit variance, and a uniform subgaussian tail with coefficient $a = q^2/2$. However, recall that the reduced space has dimension $k = \frac{C \log(2/\delta)}{\epsilon^2}$, where C depends on the constant a in the subgaussian tail inequality of the elements of R. It can be shown that $C = O(a^{-2})$, so that $q \to 0$ implies $a \to 0$, which further implies

$k \to \infty$. Therefore, Theorem 2.4 is not practical when dealing with highly sparse projection matrices distributed according to (2.6). In order to deal with the issues that arise from using highly sparse projection matrices, Matoušek further considers data points that are well-spread; in particular, he defines the sparsity parameter q to be inversely proportional to the sparsity of the input vector. Matoušek's construction is such that the reduced dimensionality k no longer depends on the constant a_S, provided x is sufficiently well spread. More formally,

Theorem 2.7: **Matoušek** ***(2008):*** *Let each of $d \in \mathbb{N}^+$, $\epsilon \in (0, 1/2)$, $\delta \in (0, 1)$, and $\alpha \in [d^{-1/2}, 1]$ be parameters, and define the sparsity parameter*

$$q = C_0 \alpha^2 \log(d/\epsilon\delta),$$

where C_0 is a sufficiently large constant, and where q is assumed to be in $[0, 1]$. Next, define the independent, identically distributed random variables

$$S_{ij} = \begin{cases} q^{-1/2} & \text{with probability } q/2, \\ -q^{-1/2} & \text{with probability } q/2, \\ 0 & \text{with probability } 1 - q, \end{cases}$$

for $i = 1, \cdots, k$, $j = 1, \cdots, d$. Next, set $k = C\epsilon^{-2} \log(4/\delta)$, where C is a sufficiently large constant, and define the random linear mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ as follows:

$$T(x)_i = \frac{1}{\sqrt{k}} \sum_{j=1}^{d} S_{ij} x_j,$$

for $i = 1, \cdots, k$. Then if $x \in \mathbb{R}^d$ such that $\|x\|_\infty \leq \alpha \|x\|$, it follows that

$$\mathbb{P}((1-\epsilon)\|x\| \leq \|T(x)\| \leq (1+\epsilon)\|x\|) \geq 1-\delta.$$

In the proof of Theorem 2.7, Matoušek defines the random mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ as $T(x) = \frac{1}{\sqrt{k}} xS$, where $S = S_{dxk}$ is the sparse projection matrix consisting of projection coefficients following the distribution S defined above. Similar to proof of Theorem 2.4, Matoušek shows that $\mathbb{P}(\|T(x)\| \leq 1-\epsilon) \leq \delta/2$ and that $\mathbb{P}(\|T(x)\| \geq 1+\epsilon) \leq \delta/2$, so that $\mathbb{P}(1-\epsilon \leq \|T(x)\| \leq 1+\epsilon) \geq 1-\delta$. The proof also relies on Markov's inequality.

The main difference between Theorems 2.4 and 2.7 is that the proof of the former relies on projection coefficients with a subgaussian tail while the proof of the latter relies on the sparsity of the input vector. More specifically, Matoušek shows that $\|T(x)\|^2 - 1$ has a subgaussian tail, but this subgaussian tail is only guaranteed up to a constant that depends on the sparsity of both S and x. Recall that x is assumed to be a fixed data point, or input vector, and so the sparsity of x can be measured prior to constructing S. The sparsity of S is chosen in such a way that the subgaussian tail of $\|T(x)\|^2 - 1$ is guaranteed up to $\sqrt{k}$ and, from this point, the remainder of the proof follows in a manner similar to that of Theorem 2.4.

In summary, Achlioptas proves a variant of the JL lemma using slightly sparse projection matrices. His approach preserves the nearly optimal tightness of Dasgupta and Gupta's version of the JL lemma. Ailon and Chazelle as well as Matoušek prove variants of the JL lemma using highly sparse projection matrices. However, neither of their results are particularly tight, since the bound on the reduced dimensionality k depends on loosely defined constants

C and C_0. Such loose bounds are acceptable since an even tighter bound on k can be obtained using a data-driven approach, as illustrated by feature hashing.

2.C.2 Improving Efficiency via Feature Hashing

Computer scientists have also contributed to increased efficiency of the JL lemma by improving methods of data storage and retrieval; faster access to data leads to a faster JL embedding. The aim is efficiency in the amount of required computing resources: time and space. There is a time/space trade-off in implementation; a gain in time efficiency leads to a loss in space efficiency and vice versa. Since computing cycles are already fast, efficiency is gained by trading time for space [34].

A computer science approach to improving the JL lemma is *feature hashing*. Feature hashing is a space-efficient way to convert a feature vector into an index vector via the use of a *hash function*. A hash function is a storage organization strategy that speeds up retrieval by mapping data of arbitrary size to data of fixed size. Feature hashing applies a hash function to each feature in the feature vector; the features are then identified with the hash values in the index vector. Hashing methods offer a gain in efficiency of JL embeddings because the hash function computes the indices, reducing the size of the index vector needed to represent the input feature vector [34, 35, 45].

Dasgupta, Kumar and Sarlos [18] begin a series of enhancements that follow FJLT. They improve the speed of JL embeddings by suppressing the use of projection coefficients that are independent random variables. Instead, they rely upon the construction of a hash function that entails dependencies within feature vectors. For a subclass of matrices, their hashing scheme leads to fewer non-zero entries per column being needed to guarantee that the resulting matrix can lead to a JL embedding that maintains ϵ-distortion.

Each of the previously discussed treatments of the JL lemma have contributed to its evolution in terms of improvements to the lower bound of the reduced dimensionality k as well as the efficiency of JL embeddings via sparse projection matrices and feature hashing. The JL lemma has evolved into an essential tool in dimensionality reduction. Numerous practical applications of the JL lemma have resulted.

2.D Applications of the Johnson and Lindenstrauss Lemma

An application that includes an $n \times d$ matrix may benefit from a JL embedding, especially if d is large. Although there are other methods of dimensionality reduction [24], such methods typically involve algorithms, wherein the number of steps grows exponentially with the number of dimensions. Hence, a reduction in d prior to execution of such algorithms means that the corresponding problem can be solved more efficiently. One such application is image retrieval, wherein an image is retrieved from the internet or any large image database by scanning its feature vector. Images are characterized by a large number of features which must be distinguished and therefore, an image feature vector is of huge dimensionality. An efficient feature reduction technique is provided by the Fast Johnson and Lindenstrauss transform (FJLT) image hashing algorithm as described in Section 2.C.2 and improved upon by using a parameter estimation algorithm for FJLT introduced in [22].

Genetic algorithms (GAs) are another class of applications that involve high-dimensional data. In GAs, a population of individuals is viewed as a matrix such that each row is a bitmap that encodes a possible solution to a given problem. The goal is to find the best solution among possible solutions, each of which can be viewed as an organism with the matrix entry representing the organism's chromosome. Mutation occurs by flipping a random bit in the string that represents a solution and crossover by mingling bits from different solutions. The genome of an organism is typically huge as is the number of bits representing a solution, particularly with the occurrence of mutation and crossover. Bertoni and Valentini [7] use JL embeddings to reduce dimensionality in GAs.

Classification is an approach in statistics and machine learning that may benefit from dimensionality reduction. *Classification* refers to the development of models that predict class membership of new observations on the basis of a training set of examples with known membership [40, 27]. A specific application is the diagnosis of a new patient based on observed characteristics of previous patients. The symptoms of previous patients are many and varied, leading to feature vectors of high-dimensionality. A feature reduction technique would be helpful because the learning algorithm converges more quickly on a training set of reduced dimensionality [25].

Paul, Athithan and Murty [38] investigate the use of random projections (akin to JL embeddings) as a preprocessing step to data analysis. Specifically, they compare the efficacy of principal component analysis with random projections as a preliminary step that reduces the dimensionality of data prior to further analysis. Their results suggest that random projections provide a more efficient preprocessing step than principal component analysis, as illustrated by increased speed of the algorithm they consider.

The applications reviewed above do not necessitate the use of JL embeddings for dimensionality reduction but rather, a JL embedding is one choice for dimensionality reduction and/or it may be used as a preprocessing step prior to use of other approaches. However, there are applications where a JL embedding is the most suitable choice, which include data streaming, nearest neighbor searches and signal compression.

2.D.1 Data Streaming

Streaming applications [14, 15, 30, 33] are characterized by huge amounts of data that need to be processed with limited storage. In such situations, dimensionality reduction based on the JL lemma is particularly useful since not all data is needed prior to computation. Section 2.B discussed several improvements to the JL lemma and, in each, the reduced dimensionality k is of the order $\log(n)$. Thus, a data stream consisting of n points can be represented with only $O(\log(n))$ dimensions. Moreover, this bound on the reduced dimensionality is determined prior to streaming, and so dimensionality reduction may begin prior to receiving the n^{th} data point in the stream.

Theorem 2.3 in Section 2.B provides the tight bound of $k \geq \lceil 4(\epsilon^2/2 - \epsilon^3/3)^{-1}\log(n)\rceil$ given by Dasgupta and Gupta [17]. For some prespecified ϵ, the bound on k can be determined without any information other than the value of n. The bound on k can be used to determine the dimensionality of the projection matrix prior to receiving any data. Therefore, if a high-dimensional data set consisting of n points is streamed one point at a time, pointwise JL embeddings can be performed. This application illustrates a situation where a JL embedding is superior to other methods of dimensionality reduction, such as principal component analysis and multidimensional scaling, each of which require all n data points prior to dimensionality reduction.

2.D.2 Nearest Neighbor Search

Nearest neighbor searches can also benefit from application of the JL lemma, since the nearest neighbour problem is computationally expensive when dealing with high-dimensional data. The *nearest neighbor problem* states the following:

Given a set P consisting of n points in $\mathbb{R}^d$, preprocess P in such a way that for any query point, $q \in P$, its nearest neighbor p can be found quickly, where

$$p = \operatorname*{argmin}_{p_i \in P} \|p_i - q\|.$$

Since there are n points in d-dimensions, each such query requires $O(nd)$ steps, and this grows very fast with d. A solution to the nearest neighbor problem for high-dimensional data allows the use of randomization and approximation [2, 19, 20, 28]. In particular, the *ϵ-approximate nearest neighbor problem* is the following:

Given a set P consisting of n points in $\mathbb{R}^d$, preprocess P in such a way that for any query point, $q \in P$, the point p can be found quickly, where p is the point such that, for every $p' \in P$,

$$\|p - q\| \leq (1 + \epsilon)\|p' - q\|.$$

ϵ-nearest neighbor queries can be solved in as little as $O(\epsilon^{-2} d \log(n))$ time by preprocessing P with a JL embedding [2].

2.D.3 Compressed Sensing

JL embeddings are also a central tool in compressed sensing. *Compressed sensing* is a class of problems that aim to recover signals from few measurements with small error. Many signals are sparse (only a few non-zero coefficients) and they can be reconstructed in a known basis. However, there is a theoretical limit on the amount of information needed to provide a digital reconstruction of an analog signal, and this limit is larger than current technologies can support for applications such as video and medical imaging. Fortunately, the JL lemma provides a guarantee of the existence of a compressed sensing matrix that satisfies the restricted isometry property; such a matrix is an essential part of the reconstruction of a signal from limited information [4, 5, 6, 12, 16, 26, 44].

This chapter provided a review of the literature spanning 4 different topics regarding the JL lemma. Section 2.A provided an introduction to JL embeddings, which refer to transformations into lower dimensions using the result of the JL lemma. Section 2.B discussed the first phase of improvements to the JL lemma, which focus on improving the lower bound on the reduced dimensionality k. Section 2.C reviewed methods for dimensionality reduction based on the JL lemma, where such methods are efficient in the amount of computing resources required (time and space). Section 2.C also demonstrated improvements to the speed of JL embeddings by reducing the computational expense of matrix multiplication and by providing faster access to data via feature hashing. Section 2.D provided a review of applications of the JL lemma, together with a discussion of the lemma's suitability for such applications:

some applications may benefit from the JL lemma (image retrieval, genetic algorithms, and classification in machine learning) while the JL lemma is an ideal tool for other applications (data streaming, nearest neighbor search, and compressed sensing).

In summary, the JL lemma first appeared in the literature 1984 [31]. Since then, the lemma has evolved in a number of ways. The first phase of its evolution is characterized by a number of theoretical improvements to the lower bound on reduced dimensionality [23, 28, 17, 36, 39]. The second phase of its evolution shifted to computational improvements in efficiency through reduction of the number of operations required [1, 2, 36] and reduction in the amount of space required [18, 34, 35, 45]. Such evolutionary processes have rendered the JL lemma an essential tool in high dimensional data analysis [2, 4, 5, 6, 7, 12, 14, 15, 16, 26, 22, 25, 27, 30, 33, 38, 40, 44]. Despite the successful applications which result from data-driven computational solutions, there is a pressing need to revisit recent treatments of the JL lemma with the mathematical rigor of earlier approaches. Matoušek [36] rests at the junction between mathematical and computational approaches.

Section 2.D provided a review of applications of the JL lemma, together with a discussion of the lemma's suitability for such applications: some applications may benefit from the JL lemma (image retrieval, genetic algorithms, and classification in machine learning) while the JL lemma is an ideal tool for other applications (data streaming, nearest neighbor search, and compressed sensing).

Chapter 3

Statement of Problem and Methodology

The JL lemma is an essential tool in high dimensional data analysis. The current investigation seeks to improve the mathematical foundations of the JL lemma through improving upon Matoušek's results [36]. First, Matoušek provides a generalized statement of the JL lemma using subgaussian tails and this statement subsumes previous treatments of the JL lemma. Second, Matoušek provides a variant of the JL lemma that contains a simple and efficient construction using sparse projection matrices that are included in current algorithms in dimensionality reduction. However, Matoušek's treatment of the JL lemma leads to results that lack the specificity of earlier approaches. In particular, Matoušek's results contain a lower bound on k that depends on constants C and C_0 which are not well-defined. This thesis aims to obtain specific values for C and C_0 using a blend of mathematical analysis and computational solutions.

3.A Statement of Problem

Recall from sections 2.B and 2.C that Matoušek's results are fairly general but this generality comes at the cost of weakened bounds on the reduced dimensionality k. More specifically, Theorems 2.4 and 2.7 respectively consider two families of projection coefficients: those with a subgaussian tail and those that are sparsely distributed. When the projection coefficients have a subgaussian tail, a JL embedding can be obtained for any input vector. Although this approach can reduce the dimensionality of a high-dimensional data set, the embedding itself can be computationally expensive and so this approach is not always useful. On the other hand, sparse projection matrices can be used in order to obtain a JL embedding that requires much less runtime. However, sparse projection matrices are only suitable when the input vectors are sufficiently dense: denser input vectors allow for sparser projection matrices; sparser input vectors require denser projection matrices. One problem with Matoušek's results, pertaining to these two situations, is that he does not give specific bounds for the reduced dimensionality k but rather, he relies on asymptotic notation.

Inspired by Matoušek's results, the first goal of this thesis is to obtain specific bounds for k. The approach is based largely on that of Matoušek, using nearly identical arguments but more detailed analysis and in some proofs, an altogether different approach is taken. The second goal of this thesis is to test and improve the accuracy of these new results on simulated data.

3.B Methodology

Two approaches are employed in an attempt to improve the JL lemma. First variants of Matoušek's results are proved using similar but more detailed analysis. Second, the results of these variants are tested and improved with simulated data.

3.B.1 Proving Variants of Matoušek's Results

Matoušek [36] provides results that are essential to the methodology of this thesis. Recall from Section 2.B and 2.C the following two theorems:

Theorem 2.4: **Matoušek** : *Consider a collection $\{R_{ij}\}_{i,j}$ of independent random variables, where $\mathbb{E}(R_{ij}) = 0$ and $\mathbb{V}(R_{ij}) = 1$ for each R_{ij} and also, suppose that $\{R_{ij}\}_{i,j}$ has a uniform subgaussian tail. Next, for fixed $d \in \mathbb{N}$, $\epsilon \in (0, 1/2]$, $\delta \in (0, 1)$, let us set $k = \frac{C \log(2/\delta)}{\epsilon^2}$, for a constant C which depends on the constant a in the subgaussian tail inequality for R_{ij}. Finally, let us define the random linear map $T : \mathbb{R}^d \to \mathbb{R}^k$ as follows:*

$$T(x)_i = \frac{1}{\sqrt{k}} \sum_{j=1}^{d} R_{ij} x_j, \; \textit{for } i = 1, 2, \cdots, k,$$

where $T(x)_i$ is the i^{th} coordinate of $T(x) \in \mathbb{R}^k$, and x_j is the j^{th} coordinate of $x \in \mathbb{R}^d$. Then every $x \in \mathbb{R}^d$, satisfies

$$\mathbb{P}((1 - \epsilon)\|x\| \leq \|T(x)\| \leq (1 + \epsilon)\|x\|) \geq 1 - \delta.$$

Theorem 2.7: **Matoušek** : *Let each of $d \in \mathbb{N}^+$, $\epsilon \in (0, 1/2)$, $\delta \in (0, 1)$, and $\alpha \in [d^{-1/2}, 1]$ be parameters, and define the sparsity parameter*

$$q = C_0 \alpha^2 \log(d/\epsilon\delta),$$

where C_0 is a sufficiently large constant, and where q is assumed to be in $[0, 1]$. Next, define the independent, identically distributed random variables

$$S_{ij} = \begin{cases} q^{-1/2} & \text{with probability } q/2, \\ -q^{-1/2} & \text{with probability } q/2, \\ 0 & \text{with probability } 1 - q, \end{cases}$$

for $i = 1, \cdots, k$, $j = 1, \cdots, d$. Next, set $k = C\epsilon^{-2} \log(4/\delta)$, where C is a sufficiently large constant, and define the random linear mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ as follows:

$$T(x)_i = \frac{1}{\sqrt{k}} \sum_{j=1}^{d} S_{ij} x_j,$$

for $i = 1, \cdots, k$. Then if $x \in \mathbb{R}^d$ such that $\|x\|_\infty \leq \alpha \|x\|$, it follows that

$$\mathbb{P}((1 - \epsilon)\|x\| \leq \|T(x)\| \leq (1 + \epsilon)\|x\|) \geq 1 - \delta.$$

Matoušek proves each of Theorems 2.4 and 2.7 with the use of standard techniques from analysis and probability theory; his proofs are outlined in Sections 2.B and 2.C. In his proofs, Matoušek does not provide specific bounds on the constants C or C_0. In the next chapter, I provide variants of Theorems 2.4 and 2.7, which provide specific bounds on the constants C and C_0. My proofs are structured in the same way as Matoušek's, but they contain detailed analysis that lead to the specific bounds.

3.B.2 Testing and Improving Variants of Matoušek's Results on Simulated Data

Having completed the proofs of my more detailed variants of Theorems 2.4 and 2.7, the next step is to determine the efficacy of these results on simulated data. JL embeddings are performed on four simulated data sets in order to empirically estimate the bound on C. For several choices of ϵ and δ, distortion is measured after each embedding and the proportion of lengths that maintain ϵ-distortion is compared to the theoretical probability $1 - \delta$. This process is repeated with decreasing values of C until the empirical probability approaches the theoretical probability of $1 - \delta$.

The simulations are performed using Matlab with the default random number generator, i.e., the random seed automatically generated by Matlab. Data sets are simulated using four different probability distributions: Uniform, Non-Central Cauchy, Exponential and Mixed Beta. To test the result of my variant of Theorem 2.4, projection coefficients are chosen from a variety of spherically symmetric distributions including the Gaussian and Uniform distributions as well as the distributions given in equations (2.4) and (2.5). To test the result of my variant of Theorem 2.7, the projection coefficients are chosen to be distributed over $\{-q^{-1/2}, 0, q^{-1/2}\}$, where 0 has probability $1 - q$, $\pm q^{-1/2}$ each have probability $q/2$, and where q is proportional to the L^∞ norm of the simulated data points in accordance with Theorem 2.7. Refer to Appendices A and B for the Matlab code.

In the next chapter, results are presented in two sections that correspond to the mathematical and computational approaches used to improve the JL lemma. It is important to note that considerable credit is given to Matoušek because he developed Theorems 2.4 and 2.7. However, it is also important that my own significant and extensive contribution is clearly credited. In essence, I borrow Matoušek blueprint to build a stronger house.

Chapter 4

Mathematical and Computational Results

I develop three theorems, two of which are refinements of Theorems 2.4 and 2.7 that give more specific bounds on C than given by Matoušek [36]. I then provide detailed proofs of my refined theorems using a similar, but more detailed, approach than that of Matoušek. Next, I test the results of my refined theorems by obtaining JL embeddings on a variety of simulated data sets. The results of these tests imply a lower dimensional embedding than that suggested by my refined theorems. Accordingly, a computational approach is taken in order to improve upon the results of my refined theorems by repeating these tests with sequentially decreasing values of C. This approach leads a tighter estimate of the bound on the reduced dimensionality than suggested by my refined theorems. The results of a representative selection of embeddings are then summarized.

4.A Mathematical Results

This section introduces three theorems, the proofs of which are presented in Section 4.B. Theorem 4.1 provides a simple and practical result that can be used in applications of the JL lemma. Theorems 4.2 and 4.3 are refinements of Theorems 2.4 and 2.7; refined in the sense that they provide more specific values for the bound on the reduced dimensionality. These theorems are based on Matoušek's treatment of the JL lemma [36].

Theorem 4.1: *Consider a set P of n points in $\mathbb{R}^d$, for some $n, d \in \mathbb{N}$. Given $\epsilon \in (0, 1/2)$, let $k = O(\epsilon^{-2} \log n)$. Then there is a mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ such that*

$$\mathbb{P}((1-\epsilon)\|u-v\| \leq \|T(u)-T(v)\| \leq (1+\epsilon)\|u-v\|, \forall u, v \in P) \geq 1/2.$$

The proof of Theorem 4.1[1] relies on the existence of a random linear map, $T : \mathbb{R}^d \to \mathbb{R}^k$ that satisfies the following condition: if $x \in \mathbb{R}^d$, then

$$\mathbb{P}((1-\epsilon)\|x\| \leq \|T(x)\| \leq (1+\epsilon)\|x\|) \geq 1 - \frac{1}{n^2}. \tag{4.1}$$

The next two theorems provide two families from which the mapping T can be taken, and any such T can be used to obtain the result of Theorem 4.1. In both theorems, the mapping T is of the form $T(X) = \frac{1}{\sqrt{k}} XR$, where R=$R_{d \times k}$ is the projection matrix, and X=$X_{n \times d}$ is the data structure. In each theorem, T is defined as in Theorems 2.4 and 2.7, with the former relying on subgaussian projection coefficients and the latter relying on sparse projection

[1] In fact, all known proofs of the JL lemma rely on statements akin to (4.1).

matrices.

Theorem 4.2: *Consider a collection $\{R_{ij}\}_{i,j}$ of independent random variables, where $\mathbb{E}(R_{ij}) = 0$ and $\mathbb{V}(R_{ij}) = 1$ for each R_{ij} and also, suppose that $\{R_{ij}\}_{i,j}$ has a uniform subgaussian tail. Next, for fixed $d \in \mathbb{N}$, $\epsilon \in (0, 1/2]$, $\delta \in (0, 1)$, let us set $k = \frac{C \log(2/\delta)}{\epsilon^2}$, for $C \geq 384(1 + 8/a_R)^2$, where a_R is the constant in the subgaussian upper tail of the R_{ij}s. Finally, let us define the random linear map $T : \mathbb{R}^d \to \mathbb{R}^k$ as follows:*

$$T(x)_i = \frac{1}{\sqrt{k}} \sum_{j=1}^{d} R_{ij} x_j, \ \textit{for } i = 1, 2, \cdots, k,$$

where $T(x)_i$ is the i^{th} coordinate of $T(x) \in \mathbb{R}^k$, and x_j is the j^{th} coordinate of $x \in \mathbb{R}^d$. For every $x \in \mathbb{R}^d$, it turns out that

$$\mathbb{P}((1 - \epsilon)\|x\| \leq \|T(x)\| \leq (1 + \epsilon)\|x\|) \geq 1 - \delta.$$

Theorem 4.2 can be improved upon by further requiring that the projection matrix is sparse, as discussed in Section 2.C.1. That is, define the mapping $T = \frac{1}{k} XS$, where elements of S are independent, identically distributed according to the following distribution

$$S_{ij} = \begin{cases} q^{-1/2} & \text{with probability } q/2, \\ -q^{-1/2} & \text{with probability } q/2, \\ 0 & \text{with probability } 1 - q. \end{cases}$$

In this case, the mapping T can be used to find a JL embedding provided the data points in X are sufficiently well-spread[2]. More specifically, the choice

[2] A unit vector is well-spread if it is close to $\frac{1}{\sqrt{d}}(\pm 1, \pm 1, \cdots, \pm 1)$ while something close

of the sparsity parameter q depends on the sparsity of data; sparse data vectors require large q while dense data vectors allow for small q (smaller q implies a sparser projection matrix).

Before moving on, it is useful to note that Theorem 4.2 can be applied when projection coefficients are independent, identically distributed according to S, since S is mean 0, unit variance, and S has a subgaussian tail with coefficient $a_S = q^2/2$ (a simple exercise involving inequality (4.2)) . However, recall that the reduced space has dimension $k = \frac{C\log(2/\delta)}{\epsilon^2}$, where $C \geq 384(1 + 8/a_S)^2$, so that $q \to 0$ implies $a_S \to 0$, which further implies $k \to \infty$. Therefore, Theorem 4.2 is not practical when dealing with highly sparse projection matrices distributed according to S while the following theorem is because it gives a formulation for the reduced dimensionality k that no longer depends on the constant a_S, so long as x is sufficiently well spread.

Theorem 4.3: *Let each of* $d \in \mathbb{N}^+$, $\epsilon \in (0, 1/2)$, $\delta \in (0, 1)$, *and* $\alpha \in [d^{-1/2}, 1]$*be parameters, and define the sparsity parameter*

$$q = C_0\alpha^2\log(d/\epsilon\delta),$$

where $C_0 \geq 1$ *and all parameters are chosen in such a way that* $q \in [0, 1]$. *Next, define the independent, identically distributed random variables*

$$S_{ij} = \begin{cases} q^{-1/2} & \textit{with probability } q/2, \\ -q^{-1/2} & \textit{with probability } q/2, \\ 0 & \textit{with probability } 1-q, \end{cases}$$

for $i = 1, \cdots, k,\ j = 1, \cdots, d$. *Next, set* $k = C\epsilon^{-2}\log(4/\delta)$, *where* $C \geq 768$,

to $(1, 0, \cdots, 0)$ is not well-spread since most of its mass lies in its first dimension.

and define the random linear mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ as follows:

$$T(x)_i = \frac{1}{\sqrt{k}} \sum_{j=1}^{d} S_{ij} x_j,$$

for $i = 1, \cdots, k$. Then if $x \in \mathbb{R}^d$ such that $\|x\|_\infty \leq \alpha \|x\|$, it follows that

$$\mathbb{P}((1-\epsilon)\|x\| \leq \|T(x)\| \leq (1+\epsilon)\|x\|) \geq 1 - \delta.$$

This section introduced Theorems 4.1, 4.2 and 4.3, each of which are based on Matoušek's discussion of the JL Lemma [36]. Indeed, Theorems 4.2 and 4.3 are simply refinements of Theorems 2.4 and 2.7 that provide more specific values for the bound on the reduced dimensionality. The next section contains detailed proofs of Theorems 4.1 through 4.3. The proofs contained in the next section are similar to Matoušek's proofs of 2.4 and 2.7, but they contain more detailed and specific arguments that lead to more specific results.

4.B Proofs of Theorems 4.1, 4.2 and 4.3

Much preliminary work is required in order to prove Theorems 4.1 through 4.3. This section provides proofs of Theorems 4.1, 4.2 and 4.3. First, relevant inequalities are stated, followed by necessary facts, lemmas and their respective proofs, and this section is concluded with proofs of Theorems 4.1 through 4.3. The proofs of Theorems 4.2 and 4.3 are structured in a manner that is essentially identical to that of Matoušek's proofs of Theorems 2.4 and 2.7, as found in [36]. The key difference between my proofs and those of Matoušek is that I provide more detailed arguments and I avoid the use of asymptotic notation in my proofs in order to obtain a more specific bound on the reduced dimensionality.

4.B.1 Inequalities

The proofs of my refined theorems make use of the following inequalities involving the exponential function:

$$1 + x \leq e^x, \ \forall x \in \mathbb{R}, \tag{4.2}$$

$$e^x \leq 1 + 2x \ \forall x \in [0, 1], \tag{4.3}$$

$$e^x \leq 1 + x + x^2, \ \forall x \leq 1, \tag{4.4}$$

$$\frac{1}{2}(e^x + e^{-x}) \leq e^{x^2/2} \ \forall x \in \mathbb{R}, \tag{4.5}$$

$$e^{-1/x} \leq x^k, \ \forall x > 0, k = 1, 2 \tag{4.6}$$

as well as *Markov's Inequality* which states the following: for any random variable $X \geq 0$, and for all $\lambda > 0$,

$$\mathbb{P}(X \geq \lambda) \leq \mathbb{E}(X)/\lambda. \tag{4.7}$$

4.B.2 Facts

In order to provide detailed proofs of my refined theorems, I introduce and prove the following facts.

Fact 1: The following equality holds for all constants a, t, provided $a > 0$.

$$\int_{-\infty}^{\infty} e^{tx} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}} e^{t^2/4a}.$$

Fact 2: If X has a subgaussian upper tail, and t is a constant, then

$$\lim_{x \to \infty} e^{tx} \mathbb{P}(X > x) = 0.$$

Proof of Fact 1: Let $X \sim \mathcal{N}(0, 1/2a)$. Then X has the following moment generating function.

$$\begin{aligned} M_X(t) &= \mathbb{E}(e^{tX}) \\ &= \sqrt{\frac{a}{\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-ax^2} dx \\ &= e^{t^2/4a} \sqrt{\frac{a}{\pi}} \int_{-\infty}^{\infty} e^{-a(x-t/2a)^2} dx \\ &= e^{t^2/4a}, \end{aligned}$$

by noting that $\frac{e^{-a(x-t/2a)^2}}{\sqrt{\pi/a}}$ is the density function of $X + t/2a$. In particular,

$$e^{t^2/4a} = \sqrt{\frac{a}{\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-ax^2} dx,$$

so that

$$\int_{-\infty}^{\infty} e^{tx} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}} e^{t^2/4a}$$

as claimed. □

Proof of Fact 2: Since X has a subgaussian upper tail, it's easy to see that

$$0 \leq e^{tx} \mathbb{P}(X > x) \leq e^{tx} e^{-ax^2}. \tag{4.8}$$

for every $x > 0$. Moreover, since $a > 0$ it follows that $tx - ax^2 \to -\infty$, as $x \to \infty$. Therefore, $\lim_{x \to \infty} e^{tx - ax^2} = 0$. □

4.B.3 Lemmas

The following six lemmas will be used to prove the Theorems 4.2 and 4.3. The lemmas are variants of five lemmas and one proposition that Matoušek establishes in order to prove Theorems 2.4 and 2.7. However, the lemmas in this section are treated in a more rigorous and detailed manner.

Lemma 4.1: *Let X be a mean 0 random variable. If there exists a constant c such that $\mathbb{E}[e^{tX}] \le e^{ct^2}$ for $t > 0$, then X has a subgaussian upper tail, with constant $a = 1/4c$. If, instead, $\mathbb{E}[e^{tX}] \le e^{ct^2}$ holds only for $t \in (0, t_0]$, then X has a subgaussian upper tail up to $2ct_0$.*

Proof of Lemma 4.1: Let $\lambda > 0$ be some constant. Then, using Markov's inequality,

$$\begin{aligned}
\mathbb{P}(X \ge \lambda) &= \mathbb{P}(e^{tX} \ge e^{t\lambda}),\ \forall\, t > 0 \\
&\le \frac{\mathbb{E}(e^{tX})}{e^{t\lambda}} \\
&\le \frac{e^{ct^2}}{e^{t\lambda}} \\
&= e^{ct^2 - \lambda t} \\
&= e^{-\frac{\lambda^2}{4c}},
\end{aligned}$$

by setting $t = \lambda/2c$. This shows that X has a subgaussian upper tail, with constant $a = 1/4c$. Note that if $\mathbb{E}[e^{tX}] \le e^{ct^2}$ holds only for $t \in (0, t_0]$ then, $t = \lambda/2c$ implies $\lambda \le 2ct_0$, so that X has a subgaussian upper tail up to $2ct_0$ as claimed. □.

The next lemma is the partial converse to Lemma 4.1. The proof requires the additional assumption that $\mathbb{V}(X) = 1$, and also, note the relationship between c and a is different than that in the previous case.

Lemma 4.2: *Let X be a random variable with $\mathbb{E}(X) = 0$, $\mathbb{V}(X) = 1$, and further suppose that X has a subgaussian upper tail. Then for all $t > 0$,*

$$\mathbb{E}[e^{tX}] \leq e^{ct^2}$$

where $c = 1 + 8/a$, with a being the constant in the subgaussian tail of X.

Proof of Lemma 4.2:

Lemma 2 is proved separately for the two different cases, where $t \leq \sqrt{a}/2$ and $t > \sqrt{a}/2$.

Case 1: $\mathbf{t \leq \sqrt{a}/2}$:

Let F be the distribution function of X, so that

$$\mathbb{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} dF(x) = \int_{-\infty}^{1/t} e^{tx} dF(x) + \int_{1/t}^{\infty} e^{tx} dF(x).$$

The above integrals are analyzed separately, starting with the left hand integral. Using inequality (4.4), and the fact that X is a mean zero, unit variance random variable, it's not hard to see that

$$\begin{aligned}\int_{-\infty}^{1/t} e^{tx} dF(x) &\leq \int_{-\infty}^{1/t} (1 + tx + (tx)^2) dF(x) \\ &\leq \int_{-\infty}^{\infty} (1 + tx + (tx)^2) dF(x) \\ &= 1 + t\mathbb{E}(X) + t^2\mathbb{E}(X^2) \\ &= 1 + t^2. \end{aligned} \tag{4.9}$$

Equation (4.9) contributes to the upper bound of $\mathbb{E}(e^{tX})$ when $X \leq 1/t$. Next, consider the second part of the expectation, where $X > 1/t$. Recall that

X has a subgaussian upper tail and the assumption that $t \leq \sqrt{a}/2$. Then, setting $j = tx$ gives

$$\begin{aligned}\int_{1/t}^{\infty} e^{tx} dF(x) = \int_{1}^{\infty} e^{j} dF(j/t) &\leq \sum_{i=1}^{\infty} \int_{i}^{i+1} e^{j} dF(j/t)\\ &\leq \sum_{i=1}^{\infty} e^{i+1} \int_{i}^{i+1} dF(j/t) \leq \sum_{i=1}^{\infty} e^{2i} \mathbb{P}(X \geq i/t)\\ &\leq \sum_{i=1}^{\infty} e^{2i-ai^2/t^2} = \sum_{i=1}^{\infty} e^{i(2-ia/t^2)}\\ &\leq \sum_{i=1}^{\infty} e^{-ia/2t^2} = \sum_{i=1}^{\infty} (e^{-a/2t^2})^i,\end{aligned}$$

so that

$$\int_{1/u}^{\infty} e^{tx} dF(x) \leq \sum_{i=1}^{\infty} (e^{-a/2t^2})^i. \tag{4.10}$$

Now, notice that the sum in (4.10) is a geometric series, with both the first term and the ratio equal to $e^{-a/2t^2}$. Hence,

$$\sum_{i=1}^{\infty} (e^{-a/2t^2})^i = \frac{e^{-a/2t^2}}{1 - e^{-a/2t^2}}. \tag{4.11}$$

Thus, since $t \leq \sqrt{a}/2$ implies $e^{-a/2t^2} \leq 1/2$, it follows from equation (4.11) that

$$\sum_{i=1}^{\infty} e^{(-a/2t^2)^i} \leq 2e^{-a/2t^2}. \tag{4.12}$$

Combining equations (4.10) and (4.12) yields

$$\int_{1/t}^{\infty} e^{tx} dF(x) \leq 2e^{-a/2t^2}. \tag{4.13}$$

Finally, (4.9) and (4.13) are combined to obtain

$$\begin{aligned}\mathbb{E}(e^{tX}) &= \int_{-\infty}^{1/t} e^{tx} dF(x) + \int_{1/t}^{\infty} e^{tx} dF(x) \\ &\leq 1 + t^2 + 2e^{-a/2t^2} \leq 1 + t^2 + 2(2t^2/a) \\ &= 1 + (1 + 4/a)t^2 \leq e^{(1+4/a)t^2},\end{aligned}$$

with the last 2 inequalities following from (4.6) and (4.2), respectively. This shows that

$$\mathbb{E}(e^{tX}) \leq e^{(1+4/a)t^2}, \tag{4.14}$$

for $t \leq \sqrt{a}/2$. Next, consider the case where $t > \sqrt{a}/2$.

Case 2: $\mathbf{t} > \sqrt{\mathbf{a}}/\mathbf{2}$:

First, split the integral as follows:

$$\mathbb{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} dF(x) = \int_{-\infty}^{0} e^{tx} dF(x) + \int_{0}^{\infty} e^{tx} dF(x).$$

Since $x \leq 0$ implies $e^{tx} \leq 1$, the above can be estimated with

$$\begin{aligned}\mathbb{E}(e^{tX}) &\leq \int_{-\infty}^{0} dF(x) + \int_{0}^{\infty} e^{tx} dF(x), \\ &= F(0) + \int_{0}^{\infty} e^{tx} dF(x).\end{aligned} \tag{4.15}$$

Next, apply integration by parts to the right hand integral of equation

(4.15), and apply the result of Fact 2 in order to obtain

$$\begin{aligned}\int_0^\infty e^{tx}dF(x) &= \int_0^\infty (1-F(x))te^{tx}dx - e^{tx}(1-F(x))\Big|_0^\infty \\ &= \int_0^\infty te^{tx}\mathbb{P}(X \geq x)dx - \Big(\lim_{x\to\infty} e^{tx}(1-F(x)) - 1 + F(0)\Big) \\ &\leq 1 - F(0) + \int_0^\infty te^{tx}\mathbb{P}(X \geq x)dx,\end{aligned}$$

with the inequality following from the fact that $-e^{tx}(1-F(x)) \leq 0$ for all $x > 0$. Thus,

$$\int_0^\infty e^{tx}dF(x) \leq 1 - F(0) + t\int_0^\infty e^{tx}\mathbb{P}(X \geq x)dx,$$

and since X has a subgaussian upper tail,

$$\begin{aligned}\int_0^\infty e^{tx}dF(x) &\leq 1 - F(0) + t\int_0^\infty e^{tx}e^{-ax^2}dx \\ &\leq 1 - F(0) + t\int_{-\infty}^\infty e^{tx}e^{-ax^2}dx.\end{aligned}$$

Therefore Fact 1 implies

$$\int_0^\infty e^{tx}dF(x) \leq 1 - F(0) + t\sqrt{\frac{\pi}{a}}e^{t^2/4a}. \tag{4.16}$$

Note that $t > \sqrt{a}/2$ implies

$$e^{2t^2\sqrt{\pi}/a} > e^{t\sqrt{\pi/a}}. \tag{4.17}$$

Therefore, applying inequality (4.6) to equation (4.16) and then using (4.17), leads to

$$\begin{aligned}\int_0^\infty e^{tx} dF(x) &\le 1 - F(0) + e^{t\sqrt{\frac{\pi}{a}}} e^{\frac{t^2}{4a}} \\ &\le 1 - F(0) + e^{\frac{2t^2\sqrt{\pi}}{a}} e^{\frac{t^2}{4a}} \\ &= 1 - F(0) + e^{\frac{8\sqrt{\pi}+1}{4a}t^2} \\ &\le 1 - F(0) + e^{\frac{4}{a}t^2}.\end{aligned}$$

Thus,

$$\int_0^\infty e^{tx} dF(x) \le 1 - F(0) + e^{\frac{4}{a}t^2}. \tag{4.18}$$

Next, note that $t > \sqrt{a}/2$ implies $4t^2/a > 1$, so that

$$2 < e^1 < e^{4t^2/a}. \tag{4.19}$$

Plugging (4.18) into (4.15), and then using (4.19), gives

$$\mathbb{E}(e^{tX}) \le 1 + e^{\frac{4}{a}t^2} \le 2e^{\frac{4}{a}t^2} \le e^{\frac{8}{a}t^2}.$$

In summary, when $t \le \sqrt{a}/2$

$$\mathbb{E}(e^{tX}) \le e^{(1+4/a)t^2},$$

and when $t > \sqrt{a}/2$

$$\mathbb{E}(e^{tX}) \le e^{(8/a)t^2}.$$

Finally, setting $c = 1 + 8/a$, gives $\mathbb{E}[e^{tX}] \le e^{ct^2}$ for any $t > 0$, which proves the claim. □

Lemma 4.3: *Let $X_1, X_2, \cdots, X_d$ be independent random variables with uniform subgaussian tail with constant a_X, and assume that $\mathbb{E}(X_i) = 0$ and that $\mathbb{V}(X_i) = 1$, for each i. Next, let $\alpha \in \mathbb{R}^d$ be any unit vector, and define $Y = \sum_{i=1}^{d} \alpha_i X_i$. Then $\mathbb{E}(Y) = 0,\ \mathbb{V}(Y) = 1$, and Y has a subgaussian tail with constant $a = \frac{a_X}{4(a_X+8)}$.*

Proof of Lemma 4.3:

To show that Y has mean 0, simply use the linearity of the expectation and that fact that each $\mathbb{E}(X_i) = 0$, so that

$$\mathbb{E}(Y) = \mathbb{E}(\sum_{i=1}^{d} \alpha_i X_i) = \sum_{i=1}^{d} \alpha_i \mathbb{E}(X_i) = 0.$$

Next, show that Y has unit variance by using the basic properties of the variance operator, and that fact each $\mathbb{V}(X_i) = 1$. That is,

$$\mathbb{V}(Y) = \mathbb{V}(\sum_{i=1}^{d} \alpha_i X_i) = \sum_{i=1}^{d} \alpha_i^2 \mathbb{V}(X_i) = \sum_{i=1}^{d} \alpha_i^2 = 1,$$

with the last equality following from the assumption that α is a unit vector. Then, show that Y has a subgaussian tail. Let $t \geq 0$ and apply Lemma 4.2 to obtain

$$\begin{aligned}
\mathbb{E}(e^{tY}) &= \mathbb{E}(e^{\sum_i t\alpha_i X_i}) \\
&= \prod_{i=1}^{d} \mathbb{E}(e^{t\alpha_i X_i}) \\
&\leq \prod_{i=1}^{d} e^{(1+8/a_X)(t\alpha_i)^2} \\
&= e^{(1+8/a_X)t^2 \sum_i \alpha_i^2} \\
&= e^{(1+8/a_X)t^2}.
\end{aligned}$$

Therefore, since $\mathbb{E}(e^{tY}) \leq e^{(1+8/a_X)t^2}$, it follows from Lemma 4.1 that Y has a subgaussian upper tail, with constant $a = \frac{1}{4(1+\frac{8}{a_X})} = \frac{a_X}{4(a_X+8)}$. Moreover, since the X_is have a subgaussian tail, an identical argument shows that $-Y$ also has a subgaussian upper tail. This completes the proof. □

Lemma 4.4: *Let Y be a random variable with a subgaussian tail with constant a, $\mathbb{E}(Y) = 0$, and $\mathbb{V}(Y) = 1$. Then for all $t \in [0, a/4]$,*

$$\mathbb{E}(e^{t(Y^2-1)}) \leq e^{(12/a^2)t^2}, \tag{i}$$

and

$$\mathbb{E}(e^{t(1-Y^2)}) \leq e^{(4/a^2)t^2}. \tag{ii}$$

Proof of Lemma 4.4 (i)

First, let F be the distribution function of Y^2 so that, for $t \geq 0$,

$$\mathbb{E}(e^{tY^2}) = \int_0^\infty e^{tx} dF(x) = \int_0^{1/t} e^{tx} dF(x) + \int_{1/t}^\infty e^{tx} dF(x) \tag{4.20}$$

Once again, the two integrals are analyzed separately. The left integral of equation (4.20) is estimated using inequality (4.4), in order to obtain

$$\begin{aligned}\int_0^{1/t} e^{tx} dF(x) &\leq \int_0^{1/t} (1 + tx + t^2x^2) dF(x)\\ &\leq \int_0^\infty (1 + tx) dF(x) + t^2 \int_0^{1/t} x^2 dF(x)\\ &= 1 + t\mathbb{E}(Y^2) + t^2 \int_0^{1/t} x^2 dF(x)\\ &= 1 + t + t^2 \int_0^{1/t} x^2 dF(x). \end{aligned} \tag{4.21}$$

The remaining integral in (4.21) is then analyzed by two applications of integration by parts:

$$\begin{aligned}
t^2\int_0^{1/t} x^2 dF(x) &= t^2[-x^2(1-F(x))|_0^{1/t}+2\int_0^{1/t} x(1-F(x))dx] \\
&= -\mathbb{P}(Y^2 > 1/t) + 2t^2\int_0^{1/t} x\mathbb{P}(Y^2 > x)dx \\
&\leq 2t^2\int_0^{1/t} x\mathbb{P}(Y^2 > x)dx \\
&= 2t^2\int_0^{1/t} x\mathbb{P}(|Y|> \sqrt{x})dx \\
&\leq 2t^2\int_0^{1/t} x2e^{-ax}dx, \quad \text{since } Y \text{ has a subgaussian tail,} \\
&= 4t^2\int_0^{1/t} xe^{-ax}dx \\
&= \frac{4t^2}{a}[-xe^{-ax}|_0^{1/t}+\int_0^{1/t} e^{-ax}dx] \\
&= \frac{4t^2}{a}[-\frac{1}{t}e^{-a/t} - \frac{1}{a}e^{-ax}|_0^{1/t}] \\
&= \frac{-4t}{a}e^{-a/t} - \frac{4t^2}{a^2}(e^{-a/t}-1) \\
&\leq \frac{4t^2}{a^2}. \qquad (4.22)
\end{aligned}$$

The left integral of (4.20) is then estimated by combining (4.21) and (4.22):

$$\int_0^{1/t} e^{tx}dF(x) \leq 1 + t + \frac{4t^2}{a^2}. \qquad (4.23)$$

Before analyzing the right integral of equation (4.20), note that $t \in [0, a/4]$ implies

$$t - a \leq -3a/4 < 0. \qquad (4.24)$$

With this in mind, apply integration by parts to the right hand integral of equation (4.20) to obtain

$$\begin{aligned}\int_{1/t}^{\infty} e^{tx} dF(x) &= -e^{tx}(1-F(x))|_{1/t}^{\infty} + t\int_{1/t}^{\infty} e^{tx}(1-F(x))dx \\ &\leq e^{1}(1-F(1/t)) + t\int_{1/t}^{\infty} e^{tx}(1-F(x))dx \\ &= e^{1}\mathbb{P}(Y^2 \geq 1/t) + t\int_{1/t}^{\infty} e^{tx}\mathbb{P}(Y^2 \geq x)dx \\ &\leq 2e^{1}e^{-a/t} + 2t\int_{1/t}^{\infty} e^{tx}e^{-ax}dx \\ &= 2e^{\frac{t-a}{t}} + 2t\int_{1/t}^{\infty} e^{tx-ax}dx \\ &= 2e^{\frac{t-a}{t}} + \frac{2t}{a-t}e^{\frac{t-a}{t}}. \end{aligned} \tag{4.25}$$

Applying inequalities (4.5) and (4.24) to (4.25) yields

$$\begin{aligned}\int_{1/t}^{\infty} e^{tx} dF(x) &\leq 2e^{\frac{t-a}{t}} + \frac{2t}{a-t}e^{\frac{t-a}{t}} \\ &\leq 2\left(\frac{t}{a-t}\right)^2 + \frac{2t}{a-t}\left(\frac{t}{a-t}\right) \\ &= \frac{4t^2}{(a-t)^2} \\ &\leq \frac{64t^2}{9a^2} \leq \frac{8}{a^2}t^2. \end{aligned} \tag{4.26}$$

Combining (4.23) with (4.26) gives

$$\begin{aligned}\mathbb{E}(e^{tY^2}) &= \int_{0}^{1/t} e^{tx}dF(x) + \int_{1/t}^{\infty} e^{tx}dF(x) \\ &\leq 1 + t + \frac{4}{a^2}t^2 + \frac{8}{a^2}t^2 \\ &= 1 + t + \frac{12}{a^2}t^2 \\ &\leq e^{t+\frac{12}{a^2}t^2}. \end{aligned}$$

Finally, we conclude that

$$\begin{aligned}\mathbb{E}(e^{t(Y^2-1)}) &= \mathbb{E}(e^{tY^2})e^{-t} \\ &\leq e^{t+\frac{12}{a^2}t^2}e^{-t} \\ &= e^{\frac{12}{a^2}t^2},\end{aligned}$$

which proves our first claim.

Proof of Lemma 4.4 (ii)

To prove part (ii) of this lemma, use inequalities (4.4) and (4.2), and integration by parts:

$$\begin{aligned}\mathbb{E}(e^{-tY^2}) &= \int_0^\infty e^{-tx}dF(x) \\ &\leq \int_0^\infty (1-tx+(tx)^2)dF(x) \\ &= 1-t\mathbb{E}(Y^2)+t^2\int_0^\infty x^2dF(x) \\ &= 1-t+t^2[-x^2(1-F(x))\Big|_0^\infty + 2\int_0^\infty x(1-F(x))dx] \\ &= 1-t+t^2[-x^2(\mathbb{P}(|Y|\geq\sqrt{x})\Big|_0^\infty + 2\int_0^\infty x\mathbb{P}(|Y|\geq\sqrt{x})dx] \\ &\leq 1-t+2t^2\int_0^\infty x2e^{-ax}dx \\ &= 1-t+\frac{4t^2}{a}[-xe^{-ax}\Big|_0^\infty + \int_0^\infty e^{-ax}dx] \\ &\leq 1-t+\frac{4t^2}{a}\int_0^\infty e^{-ax}dx \\ &= 1-t+\frac{4t^2}{a^2} \\ &\leq e^{-t+(4/a^2)t^2},\end{aligned}$$

so that

$$\mathbb{E}(e^{t(1-Y^2)}) = \mathbb{E}(e^{-tY^2})e^t \leq e^{-t+\frac{4}{a^2}t^2}e^t = e^{\frac{4}{a^2}t^2},$$

which proves the second claim. □

Lemma 4.5: *Given some $k \in \mathbb{N}^+$, let $Y_1, \cdots, Y_k$ be independent random variables with uniform subgaussian tail with constant a, and moreover, suppose that each $\mathbb{E}(Y_i) = 0$ and $\mathbb{V}(Y_i) = 1$. Next, define the random variable $Z = \frac{1}{\sqrt{k}}\sum_{i=1}^k (Y_i^2 - 1)$. Then, Z has subgaussian tail up to $\sqrt{k}$, with subgaussian tail coefficient $a_Z = a^2/48$.*

Proof of Lemma 4.5: Suppose that $t \in (0, \frac{a}{4}\sqrt{k}]$. Since the Y_is are independent with uniform subgaussian tail, it follows from Lemma 4.4 that

$$\begin{aligned}\mathbb{E}(e^{tZ}) &= \mathbb{E}(e^{(t/\sqrt{k})(Y_1^2+\cdots+Y_k^2-k)}) \\ &= \mathbb{E}(e^{(t/\sqrt{k})(Y_1^2-1)})^k \\ &\leq (e^{(12/a^2)t^2/k})^k \\ &= e^{(12/a^2)t^2},\end{aligned}$$

and that

$$\begin{aligned}\mathbb{E}(e^{-tZ}) &= \mathbb{E}(e^{(-t/\sqrt{k})(Y_1^2+\cdots+Y_k^2-k)}) \\ &= \mathbb{E}(e^{(t/\sqrt{k})(1-Y_1^2)})^k \\ &\leq (e^{(4/a^2)t^2/k})^k \\ &= e^{(4/a^2)t^2}.\end{aligned}$$

Thus, Lemma 4.1 implies Z has a subgaussian upper tail up to $6\sqrt{k}/a$, with constant $a_{Z^+} = a^2/48$, and that $-Z$ has a subgaussian upper tail up to $2\sqrt{k}/a$, with constant $a_{Z^-} = a^2/16$. Hence, setting $a_Z = \min(a_{Z^-}, a_{Z^+}) = a^2/48$, and noting that $6\sqrt{k}/a \geq 2\sqrt{k}/a \geq \sqrt{k}$, [3] it follows that Z has a subgaussian tail up to $\sqrt{k}$, with constant $a_Z = a^2/48$ which completes the proof. □

Lemmas 4.1 through 4.5 are sufficient for proving Theorem 4.2. The proof of Theorem 4.3 require one more lemma.

Lemma 4.6: *Let $\alpha^2 \leq q \leq 1$ and, for notational simplicity, let $Y = \sum_{j=1}^{d} S_{1j}x_j = \sqrt{k}T(x)_1$. Then Y has a subgaussian tail up to $2\sqrt{2q}/\alpha$, with coefficient $a = 1/4$.*

Proof of Lemma 4.6

First, note that

$$\begin{aligned}\mathbb{E}(e^{tY}) &= \prod_{j=1}^{d} \mathbb{E}(e^{tx_jS_{1j}}) \\ &= \prod_{j=1}^{d} (\frac{q}{2}(e^{tx_j/\sqrt{q}} + e^{-tx_j/\sqrt{q}}) + 1 - q) \\ &\leq \prod_{j=1}^{d} (qe^{t^2x_j^2/2q} + 1 - q), \qquad (4.27)\end{aligned}$$

with the inequality following from (4.5). Next, let $t \in (0, \sqrt{2q}/\alpha]$, so that $t^2\alpha^2/2q \leq 1$. Then, since $\|x\|_\infty \leq \alpha$, it follows that $t^2x_j^2/2q \leq 1$. Thus, using inequalities (4.3) and (4.2),

$$\begin{aligned}qe^{t^2x_j^2/2q} + 1 - q &\leq q(1 + 2(t^2x_j^2/2q)) + 1 - q \\ &= 1 + t^2x_j^2 \\ &\leq e^{t^2x_j^2}. \qquad (4.28)\end{aligned}$$

[3] Lemmas 4.1 and 4.2 imply $a = a_R/(4a_R + 32)$, where $a_R > 0$ is the constant in the subgaussian tail of R. This further implies that $1/a > 4$, and the inequality follows.

Next, apply (4.28) to (4.27) in order to obtain

$$\begin{aligned}\mathbb{E}(e^{tY}) &\leq \prod_{j=1}^{d} (qe^{t^2x_j^2/2q} + 1 - q)\\ &\leq \prod_{j=1}^{d} e^{t^2x_j^2}\\ &= e^{t^2\|x\|^2}\\ &= e^{t^2}.\end{aligned}$$

Thus, since the above holds for $t \in (0, \sqrt{2q}/\alpha]$, Lemma 4.1 may be applied to show that Y has a subgaussian upper tail up to $2\sqrt{2q}/\alpha$ and also, that the coefficient in the subgaussian upper tail of Y is $a = 1/4$.

Finally, since Y is symmetric about 0, an identical approach verifies the subgaussian upper tail of $-Y$, which completes the proof. □

The above 6 lemmas allow for the following proofs of Theorems 4.2 and 4.3.

4.B.4 Proofs of Theorems

The following are proofs of Theorems 4.1 through 4.3. The proof of Theorem 4.1 relies on the results of the Theorems 4.2 and 4.3, and so it is proved last. The proofs Theorems 4.2 and 4.3 follow the same overall structure as Matoušek's proofs of Theorems 2.4 and 2.7 [36]. However, the provided proofs contain more detail, as was the case with the proofs of Lemmas 4.1 through 4.6. Moreover, the proofs depend on those of Lemmas 4.1 through 4.6, and so the increased specificity of these lemmas is reflected in the results of Theorems 4.2 and 4.3.

Proof of Theorem 4.2

Consider a fixed unit vector $x \in \mathbb{R}^d$, and let $Y_i = \sum_{j=1}^{d} R_{ij}x_j$, for $i = 1, \cdots, k$. It follows from Lemma 4.3 that $\mathbb{E}(Y_i) = 0$ and $\mathbb{V}(Y_i) = 1$, for $i = 1, \cdots, k$, and that the collection of Y_is has a uniform subgaussian tail, with constant $a = \frac{a_R}{4(a_R+8)}$. Therefore, Lemma 4.5 shows that $Z = \frac{1}{\sqrt{k}}((\sum_{i=1}^{k} Y_i^2) - k)$ has subgaussian tail up to $\sqrt{k}$, with constant $a_Z = \frac{a_R^2}{768(a_R+8)^2}$. Next, observe that

$$\|T(x)\|^2 - 1 = \frac{1}{\sqrt{k}} Z. \tag{4.29}$$

Thus, using equation (4.29), the fact that Z has a subgaussian tail up to $\sqrt{k}$, and recalling that $k = \frac{C\log(2/\delta)}{\epsilon^2}$,

$$\begin{aligned}
\mathbb{P}(\|T(x)\| \geq 1+\epsilon) &\leq \mathbb{P}(\|T(x)\|^2 \geq 1+2\epsilon) \\
&= \mathbb{P}(\|T(x)\|^2 - 1 \geq 2\epsilon) \\
&= \mathbb{P}(Z \geq 2\epsilon\sqrt{k}) \\
&\leq e^{-a_Z(2\epsilon\sqrt{k})^2} \\
&= e^{-4a_Z\epsilon^2 k} \\
&= e^{-4a_Z C\log(2/\delta)}.
\end{aligned}$$

Hence ,

$$\mathbb{P}(\|T(x)\| \geq 1+\epsilon) \leq e^{-4a_Z C\log(2/\delta)}. \tag{4.30}$$

Choose $C \geq \frac{1}{2a_Z}$ $(= 384(1+8/a_R)^2$ by Lemma 4.5) so that equation (4.30) becomes

$$\mathbb{P}(\|T(x)\| \geq 1+\epsilon) \leq e^{-2\log(2/\delta)} = \frac{1}{(2/\delta)^2} \leq \frac{\delta}{2},$$

which shows that

$$\mathbb{P}(\|T(x)\|\geq 1+\epsilon) \leq \delta/2. \tag{4.31}$$

A similar argument shows that

$$\begin{aligned}\mathbb{P}(\|T(x)\|\leq 1-\epsilon) &= \mathbb{P}(\|T(x)\|^2\leq 1-2\epsilon+\epsilon^2)\\ &= \mathbb{P}(\|T(x)\|^2-1 \leq \epsilon^2-2\epsilon)\\ &= \mathbb{P}(Z \leq (\epsilon^2-2\epsilon)\sqrt{k})\\ &= \mathbb{P}(-Z \geq (2\epsilon-\epsilon^2)\sqrt{k})\\ &\leq e^{-a_Z((2\epsilon-\epsilon^2)\sqrt{k})^2}\\ &= e^{-a_Z\epsilon^2 k(4-4\epsilon+\epsilon^2)}\\ &\leq e^{-2a_Z\epsilon^2 k}\\ &= e^{-2a_Z C\log(2/\delta)}.\end{aligned}$$

Hence ,

$$\mathbb{P}(\|T(x)\|\leq 1-\epsilon) \leq e^{-2a_Z C\log(2/\delta)}. \tag{4.32}$$

The choice of $C \geq \frac{1}{2a_Z}$, together with (4.32), leads to

$$\mathbb{P}(\|T(x)\|\leq 1-\epsilon) \leq e^{-\log(2/\delta)} = \frac{\delta}{2},$$

which shows that

$$\mathbb{P}(\|T(x)\|\leq 1-\epsilon) \leq \delta/2 \tag{4.33}$$

Thus, combine equations (4.31) and (4.33) in order to obtain

$$\mathbb{P}(\{\|T(x)\|\leq 1-\epsilon\} \cup \{\|T(x)\|\geq 1+\epsilon\}) \leq \delta/2+\delta/2 = \delta,$$

which shows that

$$\mathbb{P}((1-\epsilon) \leq \|T(x)\| \leq (1+\epsilon)) \geq 1-\delta. \tag{4.34}$$

Lastly, observe that equation (4.34) is true for any fixed unit vector $x \in \mathbb{R}^d$. Thus, since T is a linear mapping, it follows that for any $y \in \mathbb{R}^d$ x can be replaced with $y/\|y\|$ to obtain

$$\mathbb{P}((1-\epsilon)\|y\| \leq \|T(y)\| \leq (1+\epsilon)\|y\|) \geq 1-\delta,$$

which completes the proof. □

Proof of Theorem 4.3

Let $Y_i = \sum_{j=1}^{d} S_{ij}x_j$, for $i = 1, 2, \cdots, k$. Then, since $\mathbb{E}(S_{ij}) = 0$ and $\mathbb{V}(S_{ij}) = 1$, it's easy to see that $\mathbb{E}(Y_i) = 0$ and $\mathbb{V}(Y_i) = 1$. Next, set $Z = \frac{1}{\sqrt{k}}(Y_1^2 + \cdots + Y_k^2 - k)$, so that $\frac{1}{\sqrt{k}}Z = \|T(x)\|^2 - 1$. Hence, the proof requires establishing that

$$\mathbb{P}(\{\|T(x)\| \geq 1+\epsilon\} \cup \{\|T(x)\| \leq 1-\epsilon\}) \leq \delta. \tag{4.35}$$

First, note that

$$\begin{aligned}\mathbb{P}(\|T(x)\| \geq 1+\epsilon) &= \mathbb{P}(\|T(x)\|^2 \geq 1+2\epsilon+\epsilon^2) \\ &\leq \mathbb{P}(Z \geq 2\epsilon\sqrt{k}). \qquad (4.36)\end{aligned}$$

Next, observe that

$$\begin{aligned}\mathbb{P}(\|T(x)\|\leq 1-\epsilon) &= \mathbb{P}(\|T(x)\|^2\leq 1-2\epsilon+\epsilon^2)\\ &= \mathbb{P}(Z\leq(\epsilon^2-2\epsilon)\sqrt{k}). \qquad (4.37)\end{aligned}$$

Equation (4.35) is verified by considering equations (4.36) and (4.37) separately. However, Lemma 4.5 cannot be used to prove (4.36) or (4.37), since the subgaussian tail of each Y_i is only guaranteed up to the threshold $2\frac{\sqrt{2q}}{\alpha}$. This problem addressed by truncating Y at the level $2\sqrt{2q}/\alpha$. That is, define the random variables $\tilde{Y_i}$ as follows

$$\tilde{Y_i} = \begin{cases} Y_i & \text{if } |Y_i|\leq 2\sqrt{2q}/\alpha, \\ 0 & \text{otherwise,} \end{cases}$$

for $i=1,\cdots,k$. Accordingly, define $\tilde{Z}$ using each $\tilde{Y_i}$ (in the same manner as Z was defined using each Y_i in the proof of Theorem 4.2). That is, set $\tilde{Z}=\frac{1}{\sqrt{k}}(\tilde{Y_1}^2+\cdots+\tilde{Y_k}^2-k)$, so that

$$\begin{aligned}&\mathbb{P}(Z\geq 2\epsilon\sqrt{k})\\ &= \mathbb{P}(Z\geq 2\epsilon\sqrt{k}\cap_i \tilde{Y_i}=Y_i)+\mathbb{P}(Z\geq 2\epsilon\sqrt{k}\cap\exists i:\tilde{Y_i}\neq Y_i)\\ &= \mathbb{P}(\tilde{Z}\geq 2\epsilon\sqrt{k}\cap_i \tilde{Y_i}=Y_i)+\mathbb{P}(Z\geq 2\epsilon\sqrt{k}\cap\exists i:\tilde{Y_i}\neq Y_i)\\ &\leq \mathbb{P}(\tilde{Z}\geq 2\epsilon\sqrt{k})+\mathbb{P}(Z\geq 2\epsilon\sqrt{k}\cap\exists i:\tilde{Y_i}\neq Y_i). \qquad (4.38)\end{aligned}$$

Similarly,

$$\begin{aligned}&\mathbb{P}(Z\leq(\epsilon^2-2\epsilon)\sqrt{k})\\ &\leq \mathbb{P}(\tilde{Z}\leq(\epsilon^2-2\epsilon)\sqrt{k})+\mathbb{P}(Z\leq(\epsilon^2-2\epsilon)\sqrt{k}\cap\exists i:\tilde{Y_i}\neq Y_i). \qquad (4.39)\end{aligned}$$

Next, combine equations (4.36) through (4.39) , in order to obtain

$$\begin{aligned}&\mathbb{P}(\{\|T(x)\|\geq 1+\epsilon\}\cup\{\|T(x)\|\leq 1-\epsilon\})\\ &\leq \mathbb{P}(\tilde{Z}\geq 2\epsilon\sqrt{k})+\mathbb{P}(\tilde{Z}\leq(\epsilon^2-2\epsilon)\sqrt{k})+\mathbb{P}(\exists i:\tilde{Y_i}\neq Y_i). \qquad (4.40)\end{aligned}$$

The next step of the proof is to analyze the 3 terms to the right of the inequality in equation (4.40). First, consider the rightmost term of equation (4.40). Note that Lemma 4.6 implies

$$\mathbb{P}(\tilde{Y_i}\neq Y_i)=\mathbb{P}(|Y_i|\geq 2\sqrt{2q}/\alpha)\leq 2e^{-2q/\alpha^2},$$

so that

$$\begin{aligned}\mathbb{P}(\exists i:\tilde{Y_i}\neq Y_i)&=\mathbb{P}(\cup_i\tilde{Y_i}\neq Y_i)\\ &=\sum_{i=1}^{k}\mathbb{P}(\tilde{Y_i}\neq Y_i)\\ &\leq 2ke^{-2q/\alpha^2}. \qquad (4.41)\end{aligned}$$

Next, choose $C_0\geq 1$. Then it follows from (4.41) and the choice of the parameters that

$$\begin{aligned}\mathbb{P}(\exists i:\tilde{Y_i}\neq Y_i)&\leq 2ke^{-2q/\alpha^2}\\ &=2ke^{-2C_0\log(d/\epsilon\delta)}\\ &\leq 2ke^{-2\log(d/\epsilon\delta)}\\ &\leq 2d^2e^{-2\log(d/\epsilon\delta)}\\ &\leq \delta/2, \qquad (4.42)\end{aligned}$$

provided $k \leq d^2$ is satisfied[4]. Next, consider the two remaining terms on the right hand side of equation (4.40). Note that the construction of $\tilde{Y}$ implies $\tilde{Y}$ has a subgaussian tail, $\mathbb{E}(\tilde{Y}) = 0$ (recalling that Y is symmetric about 0) and $\mathbb{V}(\tilde{Y}) \leq 1$. However, application of Lemma 4.5 requires $\mathbb{V}(\tilde{Y}) = 1$. Hence, $\tilde{Y}$ requires further manipulation as follows. First, note that

$$\begin{aligned}
|Y_i| &= \Big|\sum_{j=1}^{d} S_{ij}x_j\Big| \leq \sum_{j=1}^{d} |S_{ij}x_j| \\
&= \sum_{j=1}^{d} |S_{ij}||x_j| \leq \sum_{j=1}^{d} \frac{\alpha}{\sqrt{q}} \\
&= \frac{d\alpha}{\sqrt{q}},
\end{aligned}$$

so that $\max_i |Y_i| \leq \frac{d\alpha}{\sqrt{q}}$, and therefore

$$\begin{aligned}
1 &= \mathbb{E}(Y^2) \\
&= \mathbb{E}(Y^2 \,|\, |Y| \leq 2\sqrt{2q}/\alpha)\mathbb{P}(|Y| \leq 2\sqrt{2q}/\alpha) + \mathbb{E}(Y^2 \,|\, |Y| > 2\sqrt{2q}/\alpha)\mathbb{P}(|Y| > 2\sqrt{2q}/\alpha) \\
&\leq \mathbb{E}(\tilde{Y}^2) + \max(Y^2)\mathbb{P}(|Y| > 2\sqrt{2q}/\alpha) \\
&\leq \mathbb{E}(\tilde{Y}^2) + \frac{d^2\alpha^2}{q} 2e^{-2q/\alpha^2} \\
&\leq \mathbb{E}(\tilde{Y}^2) + \frac{d^2\alpha^2}{q} 2e^{-2\log(d/\epsilon\delta)} \\
&\leq \mathbb{E}(\tilde{Y}^2) + \epsilon,
\end{aligned}$$

which implies $\mathbb{V}(\tilde{Y}) \geq 1 - \epsilon$. So, let $1 - v = \mathbb{V}(\tilde{Y})$ and define $\hat{Y} = \frac{1}{\sqrt{1-v}}\tilde{Y}$. Then, $\mathbb{E}(\hat{Y}) = 0$, $\mathbb{V}(\hat{Y}) = 1$, and $\hat{Y}$ has a subgaussian tail with coefficient $a = 1/4$.

[4] In fact, this theorem is only applicable when d is sufficiently large so that $d > k$ holds for some fixed ϵ and δ; $d < k$, implies projection into higher dimensions which is clearly not the objective of this theorem.

Next, define $\hat{Z}$ in a manner analogous to Z and $\tilde{Z}$, so that Lemma 4.5 implies $\hat{Z}$ has subgaussian tail up to $\sqrt{k}$, with $a_{\hat{Z}} = a^2/48 = 1/768$. Then, for all $t \in (0, \sqrt{k}]$,

$$\mathbb{P}(\hat{Y_1}^2 + \cdots + \hat{Y_k}^2 \geq k + t\sqrt{k}) \leq e^{-a_{\hat{Z}}t^2}, \tag{4.43}$$

and

$$\mathbb{P}(\hat{Y_1}^2 + \cdots + \hat{Y_k}^2 \leq k - t\sqrt{k}) \leq e^{-a_{\hat{Z}}t^2}. \tag{4.44}$$

Thus, choose $C \geq 1/a_{\hat{Z}} = 768$, so that (4.43), gives

$$\begin{aligned}
\mathbb{P}(\tilde{Z} \geq 2\epsilon\sqrt{k}) &= \mathbb{P}(\tilde{Y_1}^2 + \cdots + \tilde{Y_k}^2 \geq (1+2\epsilon)k) \\
&= \mathbb{P}(\hat{Y_1}^2 + \cdots + \hat{Y_k}^2 \geq \frac{1+2\epsilon}{1-v}k) \\
&\leq \mathbb{P}(\hat{Y_1}^2 + \cdots + \hat{Y_k}^2 \geq k + 2\epsilon k) \\
&\leq e^{-a_{\hat{Z}}(2\epsilon\sqrt{k})^2} \\
&= e^{-4a_{\hat{Z}}\epsilon^2 k} \\
&= e^{-4a_{\hat{Z}}C\log(4/\delta)} \\
&\leq e^{-4\log(4/\delta)} \\
&= (\delta/4)^4 \\
&\leq \delta/4.
\end{aligned} \tag{4.45}$$

Similarly, apply the result of (4.44), in order to obtain

$$\begin{aligned}
\mathbb{P}(\tilde{Z} \leq (\epsilon^2 - 2\epsilon)\sqrt{k}) &= \mathbb{P}(\tilde{Y_1}^2 + \cdots + \tilde{Y_k}^2 \leq (1 - 2\epsilon + \epsilon^2)k) \\
&= \mathbb{P}(\tilde{Y_1}^2 + \cdots + \tilde{Y_k}^2 \leq (1-\epsilon)^2 k) \\
&= \mathbb{P}(\hat{Y_1}^2 + \cdots + \hat{Y_k}^2 \leq \frac{(1-\epsilon)^2}{1-v} k) \\
&\leq \mathbb{P}(\hat{Y_1}^2 + \cdots + \hat{Y_k}^2 \leq (1-\epsilon)k) \\
&= \mathbb{P}(\hat{Y_1}^2 + \cdots + \hat{Y_k}^2 \leq k - \epsilon k) \\
&\leq e^{-a_{\hat{Z}}\epsilon^2 k} \\
&= e^{-a_{\hat{Z}} C \log(4/\delta)} \\
&\leq e^{-\log(4/\delta)} \\
&= \delta/4. \qquad (4.46)
\end{aligned}$$

Finally, plug (4.42), (4.45) and (4.46) into (4.40) to obtain

$$\begin{aligned}
&\mathbb{P}(\{\|T(x)\| \geq 1 + \epsilon\} \cup \{\|T(x)\| \leq 1 - \epsilon\}) \leq \cdots \\
&\leq \mathbb{P}(\tilde{Z} \geq 2\epsilon\sqrt{k}) + \mathbb{P}(\tilde{Z} \leq (\epsilon^2 - 2\epsilon)\sqrt{k}) + \mathbb{P}(\exists i : \tilde{Y}_i \neq Y_i). \\
&\leq \delta/4 + \delta/4 + \delta/2 = \delta.
\end{aligned}$$

This completes the proof. □

Proof of Theorem 4.1: This proof follows quite simply by applying the result of either of the above Theorems, and by choosing $\delta = \frac{1}{n^2}$:

$$\begin{aligned}
&\mathbb{P}((1-\epsilon)\|u-v\| \leq \|T(u)-T(v)\| \leq (1+\epsilon)\|u-v\|, \forall u,v \in P) \\
&= \mathbb{P}(\cap_{u,v\in P}\{(1-\epsilon)\|u-v\| \leq \|T(u)-T(v)\| \leq (1+\epsilon)\|u-v\|\}) \\
&= 1-\mathbb{P}(\cup_{u,v\in P}\{(1-\epsilon)\|u-v\| \leq \|T(u)-T(v)\| \leq (1+\epsilon)\|u-v\|\}^C) \\
&\geq 1-\binom{n}{2}\mathbb{P}\Big(\Big\{(1-\epsilon)) \leq \|T\Big(\frac{u-v}{\|u-v\|}\Big)\| \leq (1+\epsilon)\Big\}^C\Big) \\
&\geq 1-\binom{n}{2}\delta \\
&\geq 1-\frac{n^2}{2n^2} \\
&= 1/2.
\end{aligned}$$

□

This section provided detailed proofs of Theorems 4.1, 4.2 and 4.3, the last two of which were based on Matoušek's proofs of Theorems 2.4 and 2.7 [36]. The proof of Theorem 4.1 was short and simple; it required the choice of $\delta = \frac{1}{n^2}$, application of the simple union bound, along with the result of either of Theorems 4.2 or 4.3. The proof of Theorems 4.2 and 4.3 required Lemmas 4.1 through 4.5 while the proof of Theorem 4.3 further required Lemma 4.6.

The key difference between Theorems 4.2 and 4.3 is that the former relies upon projection coefficients with a uniform subgaussian tail while the latter relies upon projection coefficients that follow the sparse distribution S, given in Theorem 4.3. Note, however, that Theorem 4.2 can be applied with such projection coefficients, since they are mean 0, unit variance, and have a subgaussian tail with coefficient $a_S = q^2/2$ (a simple exercise involving inequality (4.2)). The reason Theorem 4.2 should not be applied when

S is sparse is because it gives reduced dimensionality of $k = \frac{C\log(2/\delta)}{\epsilon^2}$, where $C \geq 384(1 + 8/a_S)^2$. Thus, $q \to 0$ implies $a_S \to 0$, which further implies $k \to \infty$. For this reason, Theorem 4.2 is not practical when dealing with highly sparse projection matrices distributed according to S (i.e., when q is very small).

Theorem 4.3 is very practical when dealing with sparse projection matrices since it gives a formulation for the reduced dimensionality k that no longer depends on the constant a_S, provided x is sufficiently well spread. In particular, Lemma 4.6, provides an additional constraint that ensures that $a = 1/4$, which does not depend on the coefficient a_S in the subgaussian tail of S. This is key to obtaining the improved value of the constant k, since the constant C no longer depends on a_S. The idea is as follows. Let $Y = \frac{1}{\sqrt{k}}xS$ be the embedding of a point x as in Theorem 4.3. The result of Lemma 4.6 only guarantees the subgaussian tail of Y up to $2\sqrt{2q}/\alpha$. On the other hand, the proof of Theorem 4.2 requires elements of the projection matrix to have subgaussian tail up to $\sqrt{k}$ which is likely much larger than $2\sqrt{2q}/\alpha$. Thus, the proof of Theorem 4.3, truncates Y and then defines Z on the truncated version of Y in order to obtain a mean zero random variable with subgaussian tail. Next, Z is transformed so that it also has unit variance and then Lemma 4.5 is applied in a manner similar to the proof of Theorem 4.2. Finally, the proof follows since truncation sets Y to 0 with probability $\leq \delta/2$ while that the transformed Z maintains ϵ-distortion with probability $\delta/4 + \delta/4$ (similar to the proof of Theorem 4.2 where we show that Z maintains ϵ-distortion with probability $\delta/2 + \delta/2$).

To summarize, refinements of Theorems 2.4 and 2.7 were presented as Theorems 4.2 and 4.3. My proofs established the specific bounds on C and C_0 that are given in the statement of Theorems 4.2 and 4.3. I next test the results of my refined theorems by obtaining JL embeddings on a variety of simulated data sets.

4.C Computational Results

This section summarizes the results of a representative selection of JL embeddings obtained from simulated data. First, the results of Theorems 4.2 and 4.3 are tested on data sets simulated using three different probability distributions. The efficacy of each theorem is measured by comparing the theoretical and empirical probability of obtaining a JL embedding that maintains ϵ-distortion. Here, the theoretical probabilities refer to $1-\delta$, for various choices of δ, and empirical probabilities refer to the relative frequencies of pairs of points that preserve ϵ-distortion, i.e., $\frac{\text{\#pairs preserving distortion}}{\text{\#pairs}} = \frac{\text{\#pairs preserving distortion}}{\binom{n}{2}}$. For all simulated data sets, for all choices of the distortion parameter ϵ, and for all choices of the probability parameter δ, the empirical probabilities are equal to 1. The discrepancy between the theoretical and empirical probabilities suggests an inflated bound on the constant C. Hence, the tests are repeated with smaller and smaller values of C, until the empirical probabilities match those suggested by Theorems 4.2 and 4.3, for fixed values of ϵ and δ. More specifically, as the empirical probabilities approach the corresponding theoretical probabilities of $1-\delta$, a neighborhood of typical values of C begins to emerge (refer to Appendices A and B for Matlab code).

4.C.1 Improving the Results of Theorems 4.2 and 4.3 with Computation

In order to test the results of Theorems 4.2 and 4.3, several data sets are simulated and then projected into lower dimensions using the transformations outlined in each theorem. Each data set consists of $n = 1000$ points, each of which are simulated as random, 10 000-dimensional vectors. The simulations are performed using Matlab and using the default random number generator, i.e., the random seed automatically generated by Matlab. Data sets are simulated using four different probability distributions: Uniform over (0,1), Exponential with mean 1, Non-Central Cauchy with non centrality parameter of 5, and Mixed Beta, where points are selected from Beta (1,3) with probability 0.25 and from Beta(4,1) with probability 0.75. These distributions are available through the built-in Matlab functions: "unifrnd, exprnd, nctrnd, and betarnd"; in order to construct the mixed distribution, points are randomly selected from two different distributions, which further required use of the built-in function "rand." refer to Appendices A and B for Matlab code.

The result of each theorem is tested by estimating the probability of obtaining a JL embedding that maintains ϵ-distortion of pairwise distances. That is, for each simulated data set, every simulated data point x is transformed into the embedded point $T(x)$, and for each x, y in the data set, the distortion of distance is computed as $\|T(x) - T(y)\|/\|x - y\|$; the transformation is said to preserve ϵ-distortion of the distance between x and y if $1 - \epsilon \leq \|T(x) - T(y)\|/\|x - y\| \leq 1 + \epsilon$. The probability that the embedding preserves ϵ-distortion is then estimated with the relative frequency of embeddings that maintain ϵ-distortion of pairwise distances between the points within each simulated data set. That is, for a particular data set, let f denote the frequency of pairs of points whose distance is distorted by no more than $1 \pm \epsilon$. Then the empirical probability that the particular JL embedding preserves ϵ-distortion is:

$$\hat{p} = \frac{\#\text{pairs preserving } \epsilon\text{-distortion}}{\#\text{pairs}} = \frac{f}{\binom{1000}{2}} = \frac{f}{499500}$$

In all cases, 100% of the embeddings maintain ϵ-distortion and this number is far too high, especially for the situations when the probability parameter δ is chosen to be rather large; for example, if δ is chosen to be 0.5, at least 50% of the pairwise distances are expected to preserve ϵ-distortion; it seems very unlikely, however, that 100% of these distances preserve ϵ-distortion. The discrepancy between the theoretical and empirical probabilities suggests an inflated bound on the constant C. Hence, in order to empirically estimate these values, the constant C is first chosen to be C=384 (when strengthening Theorem 4.2), and C=768 (when strengthening Theorem 4.3). The simulations are repeated with smaller and smaller values of C until the empirical probabilities begin to converge upon those suggested by Theorems 4.2 and 4.3. This process is repeated for each combination of ϵ and δ equal to 0.1, 0.3, and 0.5. The computation results are summarized in the next two sections.

4.C.2 Computational Approach to Theorem 4.2

To simulate the result of Theorem 4.2, the projection coefficients are chosen from a variety of spherically symmetric distributions, including the Gaussian and Uniform distributions as well as the discrete discrete distributions given by Achlioptas [1] which were discussed in section 2.C. Moreover, the parameters for each of these distributions are chosen in such a way that the projection coefficients are mean zero with unit variance. The coefficients are simulated using the built-in Matlab functions "unifrnd, normrnd", as well as the "discretesample" function that may be downloaded from the internet. Three of the four data sets are considered [5] , those following the Uniform, Cauchy, and Mixed Beta distributions outlined in Section 4.C.1. Refer to Appendix A for Matlab code.

Each simulated data set is projected into k dimensions, where $k = \frac{C}{\epsilon^2}\log(2/\delta)$, and this is repeated for each combination of ϵ and δ being equal to 0.1, 0.3 and 0.5. The constant C is first chosen to be $C = 384$ (which is smaller than that suggested by Theorem 4.2, being equal to the value of C if $a_R = \infty$). However, this value is too large since 100% of the projections preserve ϵ-distortion of pairwise distances. Thus, the simulations are repeated with sequentially decreasing values of C until the empirical probabilities begin to converge upon $1 - \delta$, as suggested by Theorem 4.2.

For each choice of projection coefficients, for each of the three data sets, nine combinations of ϵ and δ are considered for a number of choices of C. This results in a large number of embeddings. Similar results are obtained for each choice of projection coefficients and for each data set. For this reason, only the embeddings using Gaussian projection coefficients are reported; further reporting would be redundant.

The following tables summarize the results of the embeddings that rely upon projection coefficients that are scaled Gaussian random variables, following $\frac{1}{\sqrt{k}}\mathcal{N}(0, 1)$. In the tables, the columns correspond to each of the three

[5] Initially, only three data sets were simulated. However, the embeddings discussed in the next section required the use of an additional data set

simulated data sets, the rows correspond to each combination of the considered values of ϵ and δ, and the intersecting cells give the relative frequency of pairwise distances that preserve ϵ-distortion.

	Uniform	**N.C. Cauchy**	**Mixed Beta**
ϵ=0.1,δ=0.1	1.00	1.00	1.00
ϵ=0.1,δ=0.3	1.00	1.00	1.00
ϵ=0.1,δ=0.5	1.00	1.00	1.00
ϵ=0.3,δ=0.1	1.00	1.00	1.00
ϵ=0.3,δ=0.3	1.00	1.00	1.00
ϵ=0.3,δ=0.5	1.00	1.00	1.00
ϵ=0.5,δ=0.1	1.00	1.00	1.00
ϵ=0.5,δ=0.3	1.00	1.00	1.00
ϵ=0.5,δ=0.5	1.00	1.00	1.00

Table 4.1: Estimated probability of preserving ϵ-distortion using Theorem 4.2 with C=384, so that k ranges from 2130 (when $\delta = \epsilon = 0.5$) to 115037 (when $\delta = \epsilon = 0.1$).

	Uniform	**N.C. Cauchy**	**Mixed Beta**
ϵ=0.1,δ=0.1	1.00	1.00	1.00
ϵ=0.1,δ=0.3	1.00	1.00	1.00
ϵ=0.1,δ=0.5	1.00	1.00	1.00
ϵ=0.3,δ=0.1	1.00	1.00	1.00
ϵ=0.3,δ=0.3	1.00	1.00	1.00
ϵ=0.3,δ=0.5	1.00	1.00	1.00
ϵ=0.5,δ=0.1	1.00	1.00	1.00
ϵ=0.5,δ=0.3	1.00	1.00	1.00
ϵ=0.5,δ=0.5	1.00	1.00	1.00

Table 4.2: Estimated probability of preserving ϵ-distortion using Theorem 4.2 with C=10, so that k ranges from 56 (when $\delta = \epsilon = 0.5$) to 2996 (when $\delta = \epsilon = 0.1$).

	Uniform	**N.C. Cauchy**	**Mixed Beta**
$\boldsymbol{\epsilon}$**=0.1,**$\boldsymbol{\delta}$**=0.1**	0.99	0.99	0.99
$\boldsymbol{\epsilon}$**=0.1,**$\boldsymbol{\delta}$**=0.3**	0.94	0.94	0.95
$\boldsymbol{\epsilon}$**=0.1,**$\boldsymbol{\delta}$**=0.5**	0.90	0.90	0.90
$\boldsymbol{\epsilon}$**=0.3,**$\boldsymbol{\delta}$**=0.1**	0.99	0.99	0.99
$\boldsymbol{\epsilon}$**=0.3,**$\boldsymbol{\delta}$**=0.3**	0.95	0.96	0.95
$\boldsymbol{\epsilon}$**=0.3,**$\boldsymbol{\delta}$**=0.5**	0.90	0.91	0.90
$\boldsymbol{\epsilon}$**=0.5,**$\boldsymbol{\delta}$**=0.1**	0.98	0.98	0.98
$\boldsymbol{\epsilon}$**=0.5,**$\boldsymbol{\delta}$**=0.3**	0.94	0.94	0.94
$\boldsymbol{\epsilon}$**=0.5,**$\boldsymbol{\delta}$**=0.5**	0.89	0.90	0.89

Table 4.3: Estimated probability of preserving ϵ-distortion using Theorem 4.2 with C=1, so that k ranges from 6 (when $\delta = \epsilon = 0.5$) to 300 (when $\delta = \epsilon = 0.1$).

	Uniform	**N.C. Cauchy**	**Mixed Beta**
$\boldsymbol{\epsilon}$**=0.1,**$\boldsymbol{\delta}$**=0.1**	0.92	0.92	0.91
$\boldsymbol{\epsilon}$**=0.1,**$\boldsymbol{\delta}$**=0.3**	0.83	0.83	0.83
$\boldsymbol{\epsilon}$**=0.1,**$\boldsymbol{\delta}$**=0.5**	0.76	0.75	0.76
$\boldsymbol{\epsilon}$**=0.3,**$\boldsymbol{\delta}$**=0.1**	0.92	0.92	0.91
$\boldsymbol{\epsilon}$**=0.3,**$\boldsymbol{\delta}$**=0.3**	0.82	0.84	0.83
$\boldsymbol{\epsilon}$**=0.3,**$\boldsymbol{\delta}$**=0.5**	0.73	0.75	0.73
$\boldsymbol{\epsilon}$**=0.5,**$\boldsymbol{\delta}$**=0.1**	0.89	0.90	0.89
$\boldsymbol{\epsilon}$**=0.5,**$\boldsymbol{\delta}$**=0.3**	0.78	0.77	0.78
$\boldsymbol{\epsilon}$**=0.5,**$\boldsymbol{\delta}$**=0.5**	0.67	0.68	0.67

Table 4.4: Estimated probability of preserving ϵ-distortion using Theorem 4.2 with C=0.5, so that k ranges from 3 (when $\delta = \epsilon = 0.5$) to 150 (when $\delta = \epsilon = 0.1$).

The above tables illustrate that, on these particular data sets, the constant C is actually less than 0.5. To see this, recall from Theorem 4.2 that if C is chosen to be greater than $384(1 + 8/a)^2 > 384$, then the probability of maintaining ϵ-distortion is bounded below by $1 - \delta$:

$$\mathbb{P}((1-\epsilon)\|x-y\| \leq \|T(x)-T(y)\| \leq (1+\epsilon)\|x-y\|) \geq 1-\delta.$$

However, each choice of $C \geq 1$ leads to a relative frequency that is far greater than $1-\delta$. Only when $C \simeq 0.5$ do the empirical probabilities begin to converge upon the desired probability of $1-\delta$. Moreover, the relative frequencies begin to fall below $1-\delta$ when C is chosen to be much smaller than 0.5.

4.C.3 Computational Approach to Theorem 4.3

To simulate the result of Theorem 4.3, the projection coefficients are chosen to be distributed over $\{-q^{-1/2}, 0, q^{-1/2}\}$, where 0 has probability $1-q$, $\pm q^{-1/2}$ each have probability $q/2$, and where q is proportional to the L^∞ norm of the simulated data points in accordance with Theorem 4.3. The coefficients are simulated using the "discretesample" function that may be downloaded from the internet. Three of the four data sets are considered, those following the Uniform, Exponential, and Mixed Beta distributions outlined in Section 4.C.1 (refer to Appendix B for Matlab code). Note that the current set of computations, the Exponential distribution replaces the Cauchy distribution for simulation of the third data set; this is due to the fact that Cauchy random variables have infinite variance and therefore, data simulated from this distribution (and scaled to have unit length) is too sparse to apply the result of Theorem 4.3.

Each simulated data set is projected into k dimensions, where $k = \frac{C}{\epsilon^2}\log(4/\delta)$, and this is repeated for each combination of ϵ and δ being equal to 0.1, 0.3 and 0.5. The constant C is first chosen to be $C = 768$ (as suggested by Theorem 4.3) but this value is too large since 100% of the projections preserve ϵ-distortion of pairwise distances. Thus, the simulations are repeated in the same way as those from the previous section, with sequentially decreasing values of C, until the empirical probabilities begin to converge upon $1-\delta$, as suggested by Theorem 4.3.

The following tables summarize the results of the embeddings on each of the simulated data sets in the same way as the tables from 4.C.2. That is, the columns correspond to each of the three simulated data sets, the rows correspond to each combination of the considered values of ϵ and δ, and the intersecting cells give the relative frequency of pairwise distances that preserve ϵ-distortion.

	Uniform	**Exponential**	**Mixed Beta**
ϵ=0.1,δ=0.1	1.00	1.00	1.00
ϵ=0.1,δ=0.3	1.00	1.00	1.00
ϵ=0.1,δ=0.5	1.00	1.00	1.00
ϵ=0.3,δ=0.1	1.00	1.00	1.00
ϵ=0.3,δ=0.3	1.00	1.00	1.00
ϵ=0.3,δ=0.5	1.00	1.00	1.00
ϵ=0.5,δ=0.1	1.00	1.00	1.00
ϵ=0.5,δ=0.3	1.00	1.00	1.00
ϵ=0.5,δ=0.5	1.00	1.00	1.00

Table 4.5: Estimated probability of preserving ϵ-distortion using Theorem 4.3 with C=768, so that k ranges from 6389 (when $\delta=\epsilon=0.5$) to 283306 (when $\delta=\epsilon=0.1$).

	Uniform	**Exponential**	**Mixed Beta**
ϵ=0.1,δ=0.1	1.00	1.00	1.00
ϵ=0.1,δ=0.3	1.00	1.00	1.00
ϵ=0.1,δ=0.5	1.00	1.00	1.00
ϵ=0.3,δ=0.1	1.00	1.00	1.00
ϵ=0.3,δ=0.3	1.00	1.00	1.00
ϵ=0.3,δ=0.5	1.00	1.00	1.00
ϵ=0.5,δ=0.1	1.00	1.00	1.00
ϵ=0.5,δ=0.3	1.00	1.00	1.00
ϵ=0.5,δ=0.5	1.00	1.00	1.00

Table 4.6: Estimated probability of preserving ϵ-distortion using Theorem 4.3 with C=10, so that k ranges from 84 (when $\delta=\epsilon=0.5$) to 3689 (when $\delta=\epsilon=0.1$).

	Uniform	**Exponential**	**Mixed Beta**
ϵ=0.1,δ=0.1	0.99	0.99	0.99
ϵ=0.1,δ=0.3	0.98	0.98	0.98
ϵ=0.1,δ=0.5	0.96	0.96	0.95
ϵ=0.3,δ=0.1	0.99	0.99	0.99
ϵ=0.3,δ=0.3	0.98	0.97	0.98
ϵ=0.3,δ=0.5	0.96	0.96	0.96
ϵ=0.5,δ=0.1	0.99	0.99	0.99
ϵ=0.5,δ=0.3	0.98	0.98	0.97
ϵ=0.5,δ=0.5	0.95	0.96	0.95

Table 4.7: Estimated probability of preserving ϵ-distortion using Theorem 4.3 with C=1, so that k ranges from 8 (when $\delta = \epsilon = 0.5$) to 369 (when $\delta = \epsilon = 0.1$).

	Uniform	**Exponential**	**Mixed Beta**
ϵ=0.1,δ=0.1	0.95	0.95	0.94
ϵ=0.1,δ=0.3	0.89	0.89	0.89
ϵ=0.1,δ=0.5	0.85	0.84	0.85
ϵ=0.3,δ=0.1	0.94	0.95	0.94
ϵ=0.3,δ=0.3	0.89	0.89	0.89
ϵ=0.3,δ=0.5	0.84	0.84	0.83
ϵ=0.5,δ=0.1	0.94	0.95	0.94
ϵ=0.5,δ=0.3	0.89	0.89	0.88
ϵ=0.5,δ=0.5	0.84	0.85	0.84

Table 4.8: Estimated probability of preserving ϵ-distortion using Theorem 4.3 with C=0.5, so that k ranges from 4 (when $\delta = \epsilon = 0.5$) to 184. (when $\delta = \epsilon = 0.1$).

The above tables illustrate that, on these particular data sets, the constant C is actually less than 0.5. To see this, recall from Theorem 4.3 that if C is chosen to be greater than 768, then the probability of maintaining ϵ-distortion is bounded below by $1 - \delta$:

$$\mathbb{P}((1-\epsilon)\|x-y\| \leq \|T(x)-T(y)\| \leq (1+\epsilon)\|x-y\|) \geq 1-\delta.$$

However, each choice of $C \geq 1$ leads to a relative frequency that is far greater than $1 - \delta$. Only when $C \simeq 0.5$ do the empirical probabilities begin to converge upon the desired probability of $1 - \delta$. Moreover, the relative frequencies begin to fall below $1-\delta$ when C is chosen to be much smaller than 0.5.

This chapter presented Theorems 4.1, 4.2 and 4.3; Theorems 4.2 and 4.3 are refinements of Theorems 2.4 and 2.7 given by Matoušek [36]. I provided detailed proofs that contained novel arguments that led to more specific bounds on C than given by Matoušek. The results were tested with a computational approach using simulated data. The computational approach implied a lower dimensional embedding than suggested by my refined theorems. Such results contribute to the continued evolution of the JL lemma.

Chapter 5

Summary and Discussion of Improvements to the Johnson and Lindenstrauss Lemma

In view of the mathematical detail and complexity of Chapter 4, the current chapter provides a succinct summary of the research results. A detailed review of the mathematical improvements is provided including a list of specific contributions to the work of Matoušek. The chapter concludes with a summary of the computational results that further contribute to improvements to the JL lemma.

5.A Summary of Results

My contribution to improving the JL lemma involved mathematical developments that were tested and improved with a computational approach. First, I developed theorems that are refinements of two of Matoušek's previous results

[36]. Next, I tested the results of my refined theorems with simulated data; these tests suggested further improvements to my refined theorems. As such, I utilized a computational approach that provided a strong indication of the magnitude of such improvements and also, it supported existing arguments that are of practical significance, but have not yet been well established.

Section 4.A presented Theorems 4.1, 4.2 and 4.3, all of which are based on Matoušek's treatment of the JL lemma [36]. Theorem 4.1 is a simple and practical result that can be used in applications of the JL lemma. Theorems 4.2 and 4.3, refinements of Theorems 2.4 and 2.7, give more specific bounds on the reduced dimensionality k than those given by Matoušek. I proved each of my theorems in Section 4.B using a similar, more detailed approach than Matoušek. It is through this detailed treatment that I obtained the specific bounds on k.

Section 4.C presented the results of tests of Theorems 4.2 and 4.3 by obtaining JL embeddings on a variety of simulated data sets. The results of these tests implied a lower dimensional embedding than that suggested by my refined theorems. Accordingly, a computational approach was taken in order to empirically estimate the bound on the reduced dimensionality by performing JL embeddings on simulated data. The results of a representative selection of embeddings were then summarized.

5.B Mathematical Improvements

The essence of my results is reflected in Theorems 4.1, 4.2 and 4.3, all of which are based on Matoušek's treatment of the JL lemma [36]. Theorem 4.2 establishes the existence of a JL embedding using projection coefficients with a subgaussian tail. Theorem 4.3 establishes the existence of a JL embedding using sparse projection matrices. Theorem 4.1 is a simple and practical result that can be used in applications of the JL lemma. The results of Theorems 4.2 or 4.3 provide two families of mappings that can be used in Theorem 4.1

Motivation for the inclusion of both Theorems 4.2 and 4.3 follows. Theorem 4.2 asserts the existence of a mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ of the form $T(X) = \frac{1}{\sqrt{k}} XR$, where elements of R are mean zero and unit variance with uniform subgaussian tail (with constant a in the subgaussian tail inequality), and where $k = C\epsilon^{-2} \log(\delta/2)$, with $C = 384(1 + 8/a)^2$. However, such a mapping can be computationally expensive since it involves multiplication of high dimensional matrices. This is overcome by considering sparse projection matrices which contain coefficients that are distributed according to

$$S = \begin{cases} q^{-1/2} & \text{with probability } q/2, \\ -q^{-1/2} & \text{with probability } q/2, \\ 0 & \text{with probability } 1 - q, \end{cases}$$

Such projection matrices greatly improve the speed of JL embeddings, since matrix multiplication reduces to aggregate evaluation of approximately q of the original coordinates of each data point. Moreover, Theorem 4.2 can be applied when projection coefficients are independent, identically distributed according to S, since S is mean 0, unit variance, and has a subgaussian tail with coefficient $a_S = q^2/2$ (a simple exercise involving inequality (4.2)). However, Theorem 4.2 is not practical when S is highly sparse, since $q \to 0$ implies $a_S \to 0$, which implies $C \to \infty$, which further implies $k \to \infty$. This limitation of Theorem 4.2 is remedied by the additional assumptions of Theorem 4.3. That is, Theorem 4.3 asserts the existence of a mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ of the form $T(X) = \frac{1}{\sqrt{k}}XS$, where elements of S follow the above distribution, and where $k = C\epsilon^{-2}\log(\delta/4)$, with $C = 768$. Theorem 4.3 is, therefore, applicable when S is highly sparse since the reduced dimensionality k no longer depends on the constant a_S, as long as x is sufficiently well spread.

As previously mentioned, Matoušek [36] deserves much credit since Theorems 4.2 and 4.3 are based on his treatment of the JL lemma. Indeed, Theorems 4.2 and 4.3 are refinements of Theorems 2.4 and 2.7, first given by Matoušek. Moreover, my proofs follow the same overall structure of those given by Matoušek. However, I prove each of my theorems using a more detailed approach which leads to more specific results. My contributions to the improvement of Matoušek's proofs include the following:

1. The proofs of Theorems 4.2 and 4.3 rely upon six results, which I refer to as Lemmas 4.1 through 4.6. These lemmas are based on six results[1] given by Matoušek, which involve loosely defined constants. My detailed approach leads to the specific formulations for each of the constants in Lemmas 4.1 through 4.6.

2. My proof of Lemma 4.2 is broken into two cases, as was Matoušek's proof. However, Matoušek does not prove the second case but rather, he claims it is true with two short, somewhat intuitive sentences. I provide a convincing, three page proof of the second case. Moreover, my proof of Lemma 4.2 requires the inclusion of Facts 1 and 2, which I introduce and prove in Section 4.B.2.

3. My proof of Lemma 4.4 is far more detailed than Matoušek's proof. In

[1] Matoušek refers to Lemmas 4.1 through 4.6, respectively, as Lemmas 2.3, 2.4, 2.2, 3.3, Proposition 3.2, and Lemma 4.2. To avoid confusion, I will refer to these results only as Lemmas 4.1 through 4.6 even when discussing Matoušek's approach.

particular, Matoušek's proof relies upon asymptotic statements that he does not prove. I avoid the asymptotic notation and I provide a lengthy step-by-step proof of the result.

4. My proof of Lemma 4.5 is more detailed than Matoušek's proof and, more importantly, his proof contains a rather questionable assumption about the constant C. In particular, Matoušek's proof requires the assumption that $C \geq 1/2$, but he provides no justification for this assumption; it is not possible for such a conclusion to be reached from the loosely defined constants in Matoušek's approach. In contrast, my detailed approach to Lemmas 4.1 to 4.4 allows for a proof of Lemma 4.5 that is free of ambiguous assumptions.

5. The increased specificity of Lemmas 4.1 to 4.5 leads to the more specific bounds on k that are given in Theorems 4.2 and 4.3.

6. I justify the necessity of Theorem 4.3 when dealing with highly sparse projection matrices. My detailed arguments make clear that Theorem 4.2 can be applied when the sparse projection matrix S is used, but that it is impractical to do so because the reduced dimensionality grows with the sparsity of S.

My contributions to improving the JL lemma are not limited to mathematical refinements but rather, extend to computational improvements that suggest a lower dimensional embedding than given by my theorems.

5.C Computational Improvements

A computational approach was employed to improve the results of Theorems 4.2 and 4.3. In particular, the theorems were tested on data sets simulated in the Matlab environment. The efficacy of each theorem was measured by comparing the theoretical and empirical probability of obtaining a JL embedding that maintains ϵ-distortion. Here, the theoretical probabilities refer to $1 - \delta$ and empirical probabilities refer to the relative frequencies of pairs of points that preserve ϵ-distortion, i.e., $\frac{\#\text{pairs preserving distortion}}{\#\text{pairs}} = \frac{\#\text{pairs preserving distortion}}{\binom{n}{2}}$.

Preliminary tests gave an empirical probability that was far too large, which suggested an inflated bound on the constant C. This idea of an inflated bound on C is supported by previous treatments of the JL lemma. Indeed, Chapter 2 discussed the results of Dasgupta and Gupta [17] as well as Achlioptas [1], both of which give the bound $k \geq \frac{4}{\epsilon^2/2-\epsilon^3/3}\log(n)$. Note that Dasgupta and Gupta and Achlioptas take different approaches than myself and Matoušek, since they provide a probability statements involving the squared length of the points before and after projection. Nevertheless, Dasgupta and Gupta use Gaussian projection coefficients (which fit into the subgaussian class) while Achlioptas uses slightly sparse projection matrices (which fit into the class of sparse projection matrices). Hence, the results of Dasgupta and Gupta and of Achlioptas can be compared to the results of Matoušek in order to approximate the value of C, when these specific projection coefficients are used. Such comparisons do not provide an exact value for C but rather, they provide a bound for C which may only apply to a given choice of ϵ, when $\delta = 1/n^2$. In any case, such comparisons seem to support the notion of an inflated bound on the constant C given in Theorems 4.2 and 4.3.

In light of the discrepancy between the empirical and theoretical probabilities, a computational approach was taken in order to obtain a tighter bound on the constant C. In particular, the tests were repeated using sequentially decreasing values of C until the empirical probabilities matched those suggested by Theorems 4.2 and 4.3. For all sets of simulated data, for all choices of projection coefficients, and for all choices of ϵ and δ, the empirical results suggested that the constant C is less than 1. Specifically, the empirical probabilities began to converge upon the theoretical probabilities of $1 - \delta$ when C approached 1, and the empirical probabilities began to fall below $1 - \delta$ when C fell below 0.5. Since these results depended on simulated data, it is possible that the true lower bound on C is greater than 1. Nonetheless, once any empirical probability falls below $1 - \delta$, the corresponding value of C provides an upper bound under which the true C cannot exist. For example, for each of the JL embeddings performed using subgaussian projection coefficients with $C = 0.5$, more than $1 - \delta$ of the pairwise distances maintained ϵ-distortion of distances. However, when $C = 0.4$ was selected, certain embedded data sets had less than $1-\delta$ of the pairwise distances maintain ϵ-distortion. This implies that the use of subgaussian projection coefficients has a true lower bound on C that is somewhere between 0.4 and $384(1+8/a)^2$. Indeed, it is possible that the true lower bound on C (for all data) could be larger than 1. However, the computational results suggested that the value of C is in the neighborhood of $C = 1$. This C is much smaller than those given in Theorems 4.2 and 4.3 suggesting the need for further refinement, as described in the section on future research in the next chapter.

The computational results led to another noteworthy observation: there seems to be a dependency between the constant C and the parameter δ. In particular, the empirical probabilities tended to first fall below $1-\delta$, for small δ while the empirical probabilities remain large relative to $1-\delta$, for large δ. The discrepancy between the theoretical and empirical results, together with the apparent dependency between C and δ indicates that the bounds on C can be improved in each of Theorems 4.2 and 4.3. Such observations compelled me to reinvestigate my proofs of Theorems 4.2 and 4.3 in an attempt to identify any weak statements relating C to δ. Reinvestigation led to the discovery of the following weak inequality used in my proof of Theorem 4.3:

$$\begin{aligned}
\mathbb{P}(\exists i : \tilde{Y}_i \neq Y_i) &\leq 2ke^{-2q/\alpha^2} \\
&= 2ke^{-2C_0 \log(d/\epsilon\delta)} \\
&\leq 2ke^{-2\log(d/\epsilon\delta)} \\
&\leq 2d^2e^{-2\log(d/\epsilon\delta)} \\
&\leq \delta/2.
\end{aligned}$$

This weak inequality is improved using the fact that $\epsilon \in (0, 1/2]$ and $\delta \in (0, 1)$. It is likely the first of many theoretical improvements based on insights from my computational results. The improvement follows:

$$\begin{aligned}
\mathbb{P}(\exists i : \tilde{Y}_i \neq Y_i) &\leq 2ke^{-2q/\alpha^2} \\
&= 2ke^{-2C_0 \log(d/\epsilon\delta)} \\
&\leq 2ke^{-2\log(d/\epsilon\delta)} \\
&\leq 2k\frac{(\epsilon\delta)^2}{d^2} \\
&\leq \frac{\delta^2}{2d}
\end{aligned}$$

A particular consequence of this tightened inequality is that each of the remaining probabilities in equation (4.40) can be bounded above by $\frac{1}{2}(1-\frac{\delta^2}{2d})$, instead of $\delta/4$, as in the proof of Theorem 4.3. This leads to the improved bound

$$C \geq \frac{-\log(\frac{1}{2}-\frac{\delta^2}{2d})}{\log(\frac{4}{\delta})a_Z}. \tag{5.1}$$

For example, in the situation where $d = 10000$ and $\delta = 0.5$ the improved bound on C reduces to

$$\begin{aligned} C &\geq \frac{-\log(\frac{1}{2}-\frac{1}{160000})}{\log(8)a_Z} \\ &\simeq \frac{0.337}{a_Z}, \end{aligned}$$

It was established in the proof of Theorem 4.3 that $a_Z = 1/768$. Consequently, the new bound on C decreases to roughly 258, given the particular choices of $d = 10000$ and $\delta = 0.5$. Moreover, it is apparent that the bound on C becomes larger for small choice of δ while the bound decreases as δ grows. The new bound on C given in equation (5.1) remains suspiciously large, but it is only the first improvement to my mathematical results that follows from my computational results.

This chapter reviewed and discussed the manner in which my results contributed to the continued evolution of the JL lemma. A hybrid mathematical-computational approach resulted in significant improvements to the JL lemma. These improvements have far reaching implications for continued research in dimensionality reduction. Although such continued research is beyond the scope of this thesis, it is briefly discussed in the next chapter.

Chapter 6

Conclusions, Limitations, and Future Research

The first version of the JL lemma was introduced over thirty years ago; it is a piece of mathematical beauty. Subsequent research collectively contributed to the evolution of the lemma and today it is an essential tool in dimensionality reduction. In fact, the lemma is of such practical significance that its applicability has evolved faster than its theoretical development. As such, contemporary variants of the JL lemma lack mathematical clarity.

In 2008 Matoušek [36] provided particularly important evolutionary improvements to the JL lemma in the form of Theorems 2.4 and 2.7. Unfortunately, due to his untimely death in 2015, continuation of his work is left to others. I am honored to follow the work of Matoušek by providing Theorems 4.2 and 4.3 in an effort to re-establish the mathematical clarity that once characterized the JL lemma. (Refer to Appendix C for my first contribution to the continued evolution of the JL lemma).

6.A Conclusions

This thesis focused on improving the JL lemma. A union of mathematical and computational techniques was applied in order to obtain specific bounds on the reduced dimensionality of JL embeddings constructed with the use of either subgaussian projection coefficients or sparse projection matrices. Subgaussian projection coefficients are particularly important because they guarantee the existence of a JL embedding for any input data. Moreover, the proof of the JL lemma with subgaussian projection coefficients leads into the proof of the JL lemma with sparse projection matrices. Sparse projection matrices allow for faster JL embeddings, but their applicability is restricted to non-sparse input data. Matoušek acknowledged such restriction, but failed to clearly articulate the conditions under which one approach is preferred over the other.

The contributions of this thesis to ongoing research include:

- indication that sparse projection matrices should be used whenever possible, since the coefficients are particular members of the subgaussian class that lead to a faster JL embedding
- specific values for the constants in the statements of refined theorems which led to specific bounds on the reduced dimensionality k
- step-by-step proofs of theoretical results, free of ambiguity
- "apparent" reduction to the bounds on k as indicated through computational solutions
- indication of relationships between constants and parameters that may guide subsequent improvements to the bound on k

- concrete evidence that theoretical development benefits from reciprocity between mathematical analysis and computational solutions

6.B Limitations

Despite the significant contributions of this thesis, it is not without limitations. Although mathematical results provided specific bounds on k, computational results suggested that this bound is inflated. This was apparent for at least two reasons: 1) unacceptably large empirical probabilities that JL embeddings maintain ϵ-distortion of pairwise distances and 2) significantly lower dimensional embeddings regarding squared distance are provided in the literature. Computational results further suggested a relationship between the constant C (in the formulation of k) and the probability parameter δ: small δ seems to require a larger choice of C while larger δ seems to permit a smaller choice of C. This relationship was not addressed in the my theoretical results. Thus, my improvements to Matoušek's results remain incomplete.

Another limitation of this thesis is its disregard for the L^1 norm. The current investigation focused exclusively on statements of the JL lemma that regard the L^2 norm, although the L^1 norm is an alternative, and sometimes more relevant, measure of distance. The literature includes a number of treatments of the JL lemma regarding the L^1 norm. Although such treatments are relevant, they are beyond the scope of this thesis.

This thesis is directed toward mathematical and computational improvements to the JL lemma to provide data analysts with improved tools for dimensionality reduction. However, my results demand further refinement prior to application to real-world data. This thesis addresses the increasingly apparent necessity to re-establish mathematical clarity that has been lost in recent years due to the pandemic of trial-and-error applications that currently characterize data science.

6.C Future Research

Limitations of the current thesis suggest areas of future research. It is important to continue revision of Matoušek's discussion of the JL lemma [36]. For example, Matoušek provides an alternative treatment of the JL lemma involving the L^1 norm; he provides loosely defined constants similar to those given in Theorems 2.4 and 2.7 (which involve the L^2 norm). Thus, there is room to improve Matoušek's L^1 results using a more detailed approach as I did in the L^2 case.

It is necessary to more clearly discuss the trade-off between subgaussian projection coefficients and sparse projection matrices. Recall that Matoušek fails to clearly articulate the conditions under which sparse projection matrices should be used instead of subgaussian projection coefficients and vice versa. Although I provided an explanation of why sparse projection matrices are the preferred choice, I did not provide an indication of when sparse projection matrices cannot be used. Comprehension of the applicability of sparse projection matrices requires more careful analysis of the sparsity parameter q (from Theorems 2.7 and 4.3). In particular, Theorem 4.3 defines $q = C_0\alpha^2 \log(d/\epsilon\delta)$, where $C_0 \geq 1$, $\alpha \in [d^{-1/2}, 1]$ and $\|x\|_\infty \leq \alpha\|x\|$. However, this further implies that $\alpha \leq \sqrt{\frac{1}{\log(\frac{d}{\epsilon\delta})}}$, which illustrates that Theorem 4.3 is only applicable when the input data are sufficiently well spread. I am not satisfied with this statement alone, since α is defined over a region of possible values, but minimal α

is desired in order to maximize sparsity. Moreover, the inclusion of the constant C_0 is questionable since $C_0 = 1$ seems sufficient to prove Theorem 4.3. A deeper investigation into the possible relationship between C_0 and α may provide valuable insight prerequisite to the development of the optimal choice of q.

My computational results indicate dependencies among the constants and parameters in my theorems; such dependencies are not included in my results. Future research may provide further improvements to the lower bound on k which are contingent upon reducing the lower bound on the constant C. My computational results suggest a relationship between C and δ that is not stated in my theorems. Brief reinvestigation of my proofs provides immediate and overwhelming evidence of a significant reduction to the bound on C. For example, in the proof of Theorem 4.3, the following reduced bound on C can be obtained

$$C \geq \frac{-\log(\frac{1}{2} - \frac{\delta^2}{2d})}{\log(\frac{4}{\delta})a_Z}. \tag{6.1}$$

However, it is not yet clear whether the relationship between C and δ given in (6.1) is generally true (i.e., it is not yet known whether the relationship holds in alternative contexts). Future research may examine this relationship more closely and subsequently provide a tighter bound on k.

Finally, further investigation is required of data-driven solutions to improving the JL lemma. Such improvements lack mathematical clarity and yet, they can lead to highly efficient JL embeddings. For example, improvements to the JL lemma have been made via hashing schemes which can lead to sparser projection matrices than guaranteed by previous treatments of the JL lemma regarding sparse projeciton matrices. Such data-driven approaches are fairly recent, highly computational in nature and as such, there is room for theoretical development. It may be of interest to applied statisticians to explore the practical utility of data-driven approaches to the JL lemma. Moreover, such applied approaches may provide insight that motivates subsequent theoretical improvements. Indeed, regardless of whether computation is performed on real-world or simulated data, a hybrid mathematical-computational approach is an effecive means of theoretical development.

Bibliography

[1] Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of computer and System Sciences,* 66(4), 671-687.

[2] Ailon, N., & Chazelle, B. (2006, May). Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing* (pp. 557-563). ACM.

[3] Allen-Zhu, Z., Gelashvili, R., Micali, S., & Shavit, N. (2014). Sparse sign-consistent Johnson-Lindenstrauss matrices: Compression with neuroscience-based constraints. *Proceedings of the National Academy of Sciences,* 111(47), 16872-16876.

[4] Baraniuk, R. G. (2007). Compressive Sensing. *IEEE signal processing magazine,* 24(4).

[5] Baraniuk, R. G., Cevher, V., Wakin, M. B. (2010). Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective *Proceedings of the IEEE,* 98(6), 959-971.

[6] Baraniuk, R., Davenport, M., Devore, R., & Wakin, M. (2006). The Johnson-Lindenstrauss lemma meets compressed sensing. *Submitted manuscript, June.*

[7] Bertoni, A., & Valentini, G. (2005). Random projections for assessing gene expression cluster stability. In *Proceedings. 2005 IEEE International Joint Conference on*, 2005. (Vol 1, pp. 149-154). IEEE.

[8] Blum, A. (2006). Random projection, margins, kernels, and feature-selection. In *Subspace, Latent Structure and Feature Selection*, pages 5268. Springer.

[9] Bourgain, J., Dirksen, S., & Nelson, J. (2015). Toward a unified theory of sparse dimensionality reduction in euclidean space. *Geometric and Functional Analysis,* 25(4), 1009-1088.

[10] Braverman, V., Ostrovsky, R., & Rabani, Y. (2010). Rademacher chaos, random Eulerian graphs and the sparse Johnson-Lindenstrauss transform. *arXiv preprint arXiv:1011.2590.*

[11] Cai, T. & Jiang, T. (2013) Distributions of angles in random packing on spheres. *The Journal of Machine Learning Research,* 14(1), 1837-1864.

[12] Candes, E. J., & Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory,* 52(12):54065425.

[13] Carter, K. M. (2009). *Dimensionality reduction on statistical manifolds.* ProQuest. PhD Thesis.

[14] Charikar, M., Chen, K., & Farach-Colton, M. (2004). Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3-15.

[15] Clarkson, K. L., & and Woodruff, D. P. (2009). Numerical linear algebra in the streaming model. In *Proceedings of the 41st ACM Symposium on Theory of Computing (STOC)*, pages 205-214.

[16] Haimi-Cohen, R., & Lai, Y. M. (2016). Compressive measurements generated by structurally random matrices: Asymptotic normality and quantization. *Signal Processing,* 120, 71-87.

[17] Dasgupta, S., & Gupta, A. (1999). An elementary proof of the Johnson-Lindenstrauss lemma. *International Computer Science Institute, Technical Report,* 99-006.

[18] Dasgupta, A., Kumar, R., & Sarls, T. (2010, June). A sparse Johnson-Lindenstrauss transform. In *Proceedings of the forty-second ACM symposium on Theory of computing* (pp. 341-350). ACM. Chicago

[19] Deegalla, S., & Bostrom, H. (2006). Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In *Machine Learning and Applications, 2006. ICMLA06. 5th International Conference*, pages 245-250. IEEE.

[20] Deegalla, S., & Bostrom, H. Classification of microarrays with knn: Comparison of dimensionality reduction methods. In *Intelligent Data Engineering and Automated Learning- IDEAL 2007*, pages 800-809. Springer.

[21] Donoho, D. L. (2000). Aide-memoire. high-dimensional data analysis: The curses and blessings of dimensionality. *American Math. Society Lecture-Math Challenges of the 21st Century.*

[22] Fatourechi, M., Lv, X., Wang, Z. J., & Ward, R.K. (2009). Towards automated image hashing based on the Fast Johnson-Lindenstrauss Transform (FJLT). *2009 First IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 121-125. IEEE

[23] Frankl, P., & Maehara, H. (1988). The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B,* 44(3), 355-362.

[24] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157-1182.

[25] Gyllensten, A. C., & Sahlgren, M. (2015). Navigating the Semantic Horizon using Relative Neighborhood Graphs. *arXiv preprint arXiv:1501.02670.*

[26] Haimi-Cohen, R., & Lai, Y. M. (2016). Compressive measurements generated by structurally random matrices: Asymptotic normality and quantization. *Signal Processing*, 120, 71-87.

[27] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580.*

[28] Indyk, P., & Motwani, R. (1998, May). Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing* (pp. 604-613). ACM.

[29] Jacques, L. (2015). A quantized Johnson Lindenstrauss lemma: The finding of Buffon's needle. *IEEE Transactions on Information Theory,* 66(9):5012-5027.

[30] Jayram, T. S., & Woodruff, D. P. (2013). Optimal bounds for johnson-lindenstrauss transforms and streaming problems with subconstant error. *ACM Transactions on Algorithms (TALG)*, 9(3):26.

[31] Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics, 26(189-206),* 1.

[32] Kane, D. M., & Nelson, J. (2014). Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM), 61(1),* 4.

[33] Kane, D. M., Nelson, J., Porat, E., & Woodruff, D. P. (2011). Fast moment estimation in data streams in optimal space. In *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC)*, pages 745-754.

[34] Konheim, A. G. (2010). Hashing in Computer Science: Fifty Years of Slicing and Dicing. John Wiley & Sons.

[35] Lv. X., & Wang. J. (2008). Fast Johnson-Lindenstrauss transform for robust and secure image hashing. In *Multimedia Signal Processing, 2008 IEEE 10th Workshop*, pages 725-729. IEEE.

[36] Matoušek, J. (2008). On variants of the Johnson-Lindenstrauss lemma. *Random Structures & Algorithms,* 33(2), 142-156.

[37] Meng, X., & Mahoney, M. (2013). Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the 45th ACM Symposium on Theory of Computing (STOC)*, pages 91-100. ACM.

[38] Paul, B., Athithan, G., & Murty, M. N. (2009). Speeding up AdaBoost classifier with random projection. *Advances in Pattern Recognition, 2009. ICAPR '09. Seventh International Conference on* (pp. 251-254). IEEE.

[39] Rojo, J., & Nguyen, T. (2010). Improving the Johnson-Lindenstrauss Lemma. *arXiv preprint arXiv:1005.1440.*

[40] Sharma, A., & Paliwal, K. K., (2007). Fast principal component analysis using fixed-point algorithm. *Pattern Recognition Letters*, 28(10):1151-1155.

[41] Tariyal, S., Narendra, N., & Chandra, M. G. (2015). JL Lemma Based Dimensionality Reduction: On Using CDS Based Partial Fourier Matrices. In *2015 IEEE 22nd International Conference on High Performance Computing Workshops* (pp. 44-47). IEEE.

[42] Thorup, M., & Zhang, Y. (2012). Tabulation-based 5-independent hashing with applications to linear probing and second moment estimation. *SIAM J. Comput.,* 41(2): 293-331.

[43] Wang, J. (2011). Classical multidimensional scaling. In *Geometric Structure of High-Dimensional Data and Dimensionality Reduction* (pp. 115-129). Springer Berlin Heidelberg.

[44] Ward, R. (2014). Cross validation in compressed sensing via the Johnson Lindenstrauss Lemma.
http://www.citebase.org/abstract?id=oai:arXiv.org:0803.1845

[45] Weinberger, K. Q., Dasgupta, A., Langford, J., Smola, A. J., & Attenberg, J. (2009). Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 1113-1120.

Appendix A

Matlab Code: Theorem 4.2

```
%       JL Simulation 1: Subguassian Projection Coefficients
%
%% Step 1: Initialize
%Clear workspace and reset random number generator
clear all clc
rng('default') %set random seed

% Choose Parameters
N=1000; %number of points

% Choose one of the following for the distortion parameter
e=0.1; %e=0.3; %e=0.5;

% Choose one of the following for the probability parameter
d=0.1; %d=0.3; %d=0.5;

% Choose one of the following constants
C=0.4; %C=0.5; %C=1;  %C=10; %C=384

% Set dimsensionality
D=10000; %initial dimsensionality
K=floor(C*e^(-2)*log(2/d)); %reduced dimensionality

%% Step 2: Simulate Data
p=0.75; %used when simulating from mixture distribution
u= rand(N,D)<=p; % u is true with probability p and false with probability 1-p

% Choose one of the following distributions for simulated data

P1=unifrnd(0,1,N,D); %P1=nctrnd(1,5,N,D); %P1=(1-u).*betarnd(1,3,N,D)+u.*betarnd(4,1,N,D);

%% Step 3: Chooseprojection matrix
%
% Gaussian(0,1) projection coefficients:
Basis=normrnd(0,1,K,D);

% Uniform(-sqrt(3),sqrt(3)) projection coefficients:
%Basis=unifrnd(-sqrt(3),sqrt(3),K,D);

% Discrete projection coefficient
% prob=[1/6,2/3,1/6]; %Prob. Dist'n for each element of projection matrix/
% Basis=zeros(K,D);
%   for j=1:K
%   Basis(j,:)=(discretesample(prob,D)-2);
%   end
%   Basis=(1/3)^(-1/2)*Basis;

%% Step 4: Project into K dimensions.

P2=sqrt(1/K)*P1*Basis'; %P2=Data point after projection into K dimension

%% Step 5: Compute prob of success
NUMPAIRS=nchoosek(N,2);
D1=pdist(P1);
D2=pdist(P2);
  RAT=bsxfun(@rdivide,D2,D1);

  freq=nnz(1-e<RAT1&RAT<1+e);
  prob=freq/NUMPAIRS
```

Appendix B

Matlab Code: Theorem 4.3

```matlab
%       JL Simulation 2: Sparse Projection Matrix

%% Step 1: Initialize

%Clear workspace and set random number generator
%
clear all clc
rng('default') %set random seed

% Choose Parameters
%
N=1000; %number of points

% Choose one of the following for the distortion parameter
e=0.1; %e=0.3; %e=0.5;

% Choose one of the following for the probability parameter
d=0.1; %d=0.3; %d=0.5;

% Choose one of the following constants
C=0.4;%C=0.5; %C=1; %C=10;%C=786;

% Set dimensionality

D=10000; %initial dimsensionality
K=floor(C*e^(-2)*log(4/d)); %reduced dimensionality

%% Step 2: Simulate Data
p=0.75; %used when simulating from mixture distribution
u= rand(N,D)<=p; % u is true with probability p and false with probability 1-p

%Choose one of the following distributions for simulated data
P1=unifrnd(0,1,N,D); %P1=exprnd(1,N,D); %P1=(1-u).*betarnd(1,3,N,D)+u.*betarnd(4,1,N,D);

%% Step 3: Compute sparsity parameter
rowNorms=sqrt(sum(abs(P1).^2,2));
normalizedP1=bsxfun(@rdivide,P1,rowNorms);
alpha=max(normalizedP1(:)); %Sparsity parameter
q=alpha^2*log(D/(e*d));
prob=[q/2,1-q,q/2]; %probability distribution for each element of the spare projection matrix

%% Step 4: Construct projection matrix
Basis=zeros(K,D);
  for j=1:K
  Basis(j,:)=(discretesample(prob,D)-2);
  end
  Basis=q^(-1/2)*Basis;

%% Step 5: Project into K dimensions
P2=sqrt(1/K)*P1*Basis'; %P2=Data point after projection into K dimension

%% Step 6: Compute prob of success
NUMPAIRS=nchoosek(N,2);
D1=pdist(P1);
D2=pdist(P2);
  RAT=bsxfun(@rdivide,D2,D1);

  freq=nnz(1-e<RAT&RAT<1+e);
  prob=freq/NUMPAIRS;
```

Appendix C

Paper Presented at the 3rd International Conference on Advances in Big Data Analytics, July 2016, Las Vegas, Nevada (published proceedings pending).

Dimensionality Reduction via the Johnson-Lindenstrauss Lemma

J. Fedoruk[1], **B. Schmuland**[1], **J. Johnson**[3], **and G. Heo**[1,2]
[1]Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Canada.
[2]Department of Dentistry, Faculty of Medicince and Dentistry, University of Alberta, Edmonton, Canada.
[3]Department of Mathematics and Computer Sciences, Laurentian University, Sudbury, Canada.

Abstract— *The Johnson-Lindenstrauss lemma is a famous result that has lead to the development of tools that may be used when dealing with datasets of immense dimensionality. The lemma asserts that a set of high dimensional points can be projected into lower dimensions, while approximately preserving the pairwise distance structure. Significant improvements of the JL-lemma are summarized, followed by a detailed treatment of the more recent approach taken by Matoušek [13]. Particular focus is placed on reproving Matoušek's versions of the lemma first using subgassian projection coefficients and then using sparse projection matrices. The results of the lemma are then tested using simulated data. The simulation suggests a projection that is more effective in terms of dimensionality reduction than that which is born out by the theory.*

Keywords: Johnson-Lindenstrauss, dimensionality reduction

1. Introduction

Statistics is a branch of mathematics focused largely on *data*. To the applied statistician, a dataset is an $n \times d$ matrix X, consisting of n observations, where each observation is characterized by d covariates. From a geometric standpoint, one can view X as a collection of n points, $x \in \mathbb{R}^d$. The *dimensionality* of x refers to the number of dimensions to which x belongs; in this case, x is said to be d-dimensional, and we can express x as $x = (x_1, x_2, \cdots, x_d)$, where $x_i \in \mathbb{R}$ is said to be the i^{th} *coordinate* of x, for $i = 1, 2, \cdots, d$. From a statistical standpoint, one can view x as an observation consisting of d measurements, with each coordinate x_i of x corresponding to measurement for the i^{th} variable.

The main objective of statistics is to collect sample data in order to develop models that may be used to make claims about a population of interest. However, methods of data collection and model development have evolved over the years. David Donoho [5] argues that traditional statistical analyses relied upon the collection of a large number of observations, each characterized by a few carefully chosen variables. Accordingly, the observations themselves correspond to points in relatively low-dimensional space. However, as Donoho goes on to claim, modern data are often represented by a number of dimensions that is too large for classical statistical approaches to be feasible. Indeed, thanks to advancement in computer power, our capacity to sense and record information has grown immensely; so much so that the dimensionality of modern datasets can be in the thousands or even in the millions. This has created a new challenge for statisticians: how does one begin to fit a model to a dataset consisting of significantly more variables than observations (when d is much larger than n)?

The difficulty in analyzing high-dimensional data is known as *The Curse of Dimensionality*. Issues revolving around the Curse of Dimensionality have become commonplace in data analysis, and this has lead us to an exciting area of research known as dimensionality reduction.

1.1 Dimensionality Reduction

The first step in the analysis of a high dimensional data set is to reduce its dimensionality. That is, given some dataset $X_{n \times d}$, where $d >> n$, we wish to find a lower dimensional representation $Y_{n \times k}$ of X, with $k < d$, so that much of the information contained in X can be obtained from Y. Techniques in dimensionality reduction are being used in a variety of fields, including research in dentistry and orthodontics. For example Heo et al. explore the use of dimensionality reduction techniques to landmark-based data [7], [8], [9]. In particular, they apply dimensionality reduction techniques to orthodontic data sets in order to compare two types of rapid maxillary expansion treatments. Their initial dataset consisted of high-dimensional landmark configuration data that were obtained from cone beam CT scans. Techniques in dimensionality reduction were applied to these data in order to allow for computation of between-subject variation. The next question to address is this: what are the different methods of dimensionality reduction, and when should one method be used instead of another?

There are a number of statistical approaches that may be used to reduce the dimensionality of a dataset, and such approaches can be classified as either *feature selection* or *feature extraction* techniques. Some of the well-known methods of feature selection include model selection methods in regression and classification, as well as regularization methods such as Lasso and support vector machines. Some of the well-known methods of feature extraction include clustering, principal component analysis, multidimensional scaling, and ISO maps. Most of the statistical approaches to dimensionality reduction are based on uncovering the *intrinsic dimensionality* of a data set, which is the number of dimensions (variables) that contribute to the majority of the observed structure in the data; on the other hand, the *extrinsic dimensionality* of a data set gives the number of dimensions in which the data are observed [14].

Although it is of interest for us to uncover the intrinsic dimensionality of a dataset, it is not always possible to do so. In particular, many of the above approaches rely matrix operations that are computationally expensive for high dimensional data. For example, regression requires matrix inversion, while MDS and PCA rely on eigendecomposition and such matrix operations require a great deal of memory when acting on high-dimensional matrices. As such, there is a growing need for methods of dimensionality reduction that enable us to significantly decrease the extrinsic dimensionality of the data while preserving its structure. Accordingly, new methods in dimensionality reduction are emerging, and such methods effectively reduce the extrinsic dimensionality of the data, without any consideration of the true intrinsic dimensionality. As a result, these new methods do not provide a clear picture of the intrinsic dimensionality of a dataset, nor do they provide us with the variables responsible for much of the structure in the data. Nevertheless, the new approaches to dimensionality reduction are becoming an integral part of various algorithms designed to deal with high-dimensional data. The following gives a brief summary of the Lemma that started this movement, and some of the key improvements it has seen since its inception.

1.2 Johnson-Lindstrauss Lemma

The Johnson-Lindenstrauss Lemma is a famous result that has lead to the creation of a new class of techniques in dimensionality reduction. The approach is much more general than some of the classical, statistical methods in that it may be applied to *any* set of points in high dimensions (unlike statistical methods of dimensionality reduction, in which it is assumed that the intrinsic dimensionality is very small relative to the extrinsic dimensionality).

The Johnson Lindenstrauss Lemma asserts that a set of high dimensional points can be projected into lower dimensions, while approximately preserving the pairwise distance structure between points. More formally, the JL Lemma states the following:

Given a set P of n points in $\mathbb{R}^d$, for some $n, d \in \mathbb{N}$, there exists $k_0 = O(\epsilon^{-2} \log n)$ such that, if $k \geq \lceil k_0 \rceil$, there exists a linear mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ such that for any two points $u, v \in P$,

$$(1-\epsilon)\|u-v\| \leq \|T(u)-T(v)\| \leq (1+\epsilon)\|u-v\|.$$

Since T is a linear mapping, we can, without loss of generality, replace the quantities $u-v$ and $T(u)-T(v)$ with x and $T(x)$, for a unit vector $x \in \mathbb{R}^d$. That is, x represents the distance between two points in P, and $T(x)$ represents the distance between the two mapped points. The mapping T, is referred to as a *JL-embedding*.

The result of this theorem ensures that any set of points can be projected into $O(\epsilon^{-2} \log n)$ dimensions while maintaining ϵ-distortion of pairwise distances between points. Here, ϵ-distortion implies the ratio of distance after projection over that before projection is within $(1-\epsilon, 1+\epsilon)$.

2. Evolution of the JL Lemma

Over the years, the JL-Lemma has been reproved many times, with new proofs providing a sharpening and/or simplification of the result. However, there is one particular feature that is common to all JL-embeddings: the mapping T projects a vector into lower dimension, and the length of this projection is sharply concentrated around its expectation. Moreover, the existence of such mappings are typically established through the probabilistic method, i.e. one shows that the random mapping T has nonzero probability of being sufficiently concentrated about its expectation.

In the original paper that introduced the JL-lemma, Johnson and Lindenstrauss [12] assert the existence of a mapping T that gives an orthogonal projection of n points from $\mathbb{R}^d$ onto a random k-dimensional subspace with dimensionality $O(\log(n/\epsilon^2))$, such that pairwise distances are maintained to within a factor of $1\pm\epsilon$. Johnson and Lindenstrauss provide a lengthy, technical proof using geometric approximation, and reading through every detail of their proof is a challenging endeavor, even for an experienced mathematician.

The first significant improvement to the JL-lemma came from Frankl and Meahara [6], who replace the random k-dimensional subspace with a collection of k random, orthonormal vectors; this approach requires a much simpler proof that attains a sharper bound on the reduced dimensionality of $T(x)$. In particular, Frankl and Meahara show that n points from $\mathbb{R}^d$ can be projected into $k \geq \lceil 9(\epsilon^2 - \epsilon^3/3)^{-1} \log(n) \rceil$ dimensions while maintaining ϵ-distortion of pairwise distances. Moreover, Frankl and Meahara establish that the mapping is of the form $T = \sqrt{\frac{d}{k}} XR$, where X=$X_{n \times d}$ is the data structure corresponding to the points in P, and R=$R_{d \times k}$ is the projection matrix consisting of random orthonormal column vectors.

Indyk and Motwani [11] then provide the next improvement by relaxing the condition of orthogonality in the projection matrix. Instead, they show that a projection matrix need only consist of independent, Gaussian random vectors, with each coordinate following $\mathcal{N}(0, 1/d)$. This result greatly simplifies the proof of the JL-lemma since independent vectors are easier to deal with than orthogonal vectors and in high dimensions, independent Gaussian vectors are almost orthogonal.

Dasgupta and Gupta [4] then provide an alternative, much simpler proof of the result of Indyk and Motwani using moment generating functions. Moreover, they provide a tighter bound than all previous versions of the JL-lemma, wherein n points from $\mathbb{R}^d$ can be projected into $k \geq$

$\lceil 4(\epsilon^2/2 - \epsilon^3/3)^{-1}\log(n)\rceil$ dimensions while maintaining ϵ-distortion. The results of both Indyk and Motwani, and Dasgupta and Gupta rely on projection coefficients that are spherically symmetric.

Achlioptas [1] then shows that spherical symmetry of the projection coefficients is not necessary in order to obtain a JL-embedding that maintains ϵ-distortion. Instead, he shows that concentration of the projected points is sufficient. In particular, he chooses projection coefficients that are independent, identically distributed (i.i.d.) random variables, uniformly distributed over $\{-1, 1\}$ or, alternatively, distributed over $1/\sqrt{3}\{-1, 0, 1\}$, where ± 1 occur with probability 1/6 and 0 occurs with probability 2/3; he then shows that the even moments of such random projections are dominated by those of the spherically symmetric case, so that a JL-embedding can be found with probability at least as large as that in the spherical case (that is, when spherically symmetric projection coefficients are used).

Finally, Matoušek [13] improves upon the above results in two ways. First, he proves a generalized version of the JL-lemma using the language of subgaussian tails, and this approach contains many of the previously mentioned approaches, which involve spherical symmetry of the projection coefficients. In particular, Matoušek shows that a JL-embedding can be found by using i.i.d. projection coefficients that follow a distribution with a mean of 0, variance of 1, and with tails that are tighter than those of the standard normal distribution. Matoušek's next contribution is an extension of Achlioptas' result mentioned above. More specifically, Matoušek proves that highly sparse projection matrices can be used, but the sparsity of the projection matrix depends on the density of the input vectors: denser input vectors allow for sparser projection matrices which is desirable since sparse projection matrices lead to faster embeddings.

3. Two Approaches to the JL-Lemma: Subgaussian Projection Coefficients and Sparse Projection Matrices

The following three theorems are based largely on Matoušek's rendition of the JL Lemma [13].

Theorem 1: Consider a set P of n points in $\mathbb{R}^d$, for some $n, d \in \mathbb{N}$. Given $\epsilon \in (0, 1/2)$, let $k = O(\epsilon^{-2}\log n)$. Then there is a mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ such that

$$\mathbb{P}\big((1-\epsilon)\|u-v\| \leq \|T(u)-T(v)\| \leq (1+\epsilon)\|u-v\|, \forall u, v \in P\big) \geq 1/2.$$

The proof of Theorem 1[1] relies on the existence of a random linear map, $T : \mathbb{R}^d \to \mathbb{R}^k$ that satisfies the following condition: if $x \in \mathbb{R}^d$, then

$$\mathbb{P}\big((1-\epsilon)\|x\| \leq \|T(x)\| \leq (1+\epsilon)\|x\|\big) \geq 1 - \frac{1}{n^2}. \quad (1)$$

The proof then follows by choosing $\delta = 1/n^2$, and applying the result of either of the next two together with the union bound. The next two theorems provide two particular families of mappings T, that can be used in Theorem 1. In both theorems, the mapping T is of the form $T(x) = XR^T$, where R=$R_{k\times d}$ is the projection matrix, and X=$X_{n\times d}$ is the data structure. Theorem 2 requires that elements of R are i.i.d. random variables, with mean 0, unit variance, and uniform a subgaussian tail, while Theorem 3 uses a sparse projection matrix.

Definition 1: **Subgaussian Tails**
Let X be a real-valued random variable, with $\mathbb{E}(X) = 0$. X is said to have a *subgaussian upper tail* if $\exists\ a > 0$ so that

$$\mathbb{P}(X > \lambda) \leq \exp(-a\lambda^2), \quad (2)$$

for every $\lambda > 0$. If there is some λ_0 such that equation (2) holds only when $\lambda \in (0, \lambda_0)$, then we say that X has a subgaussian upper tail *up to* λ_0. Furthermore, we say that X has a *subgaussian tail* if both X and $-X$ have subgaussian upper tails. Lastly, suppose that $X_1, X_2, \cdots$ is a sequence of random variables, each with subgaussian tail. If the constant a in the subgaussian tail inequality is the same for each X_i, then we say that the X_is have a *uniform subgaussian tail.*

Theorem 2: Consider a collection $\{R_{ij}\}_{i,j}$ of independent random variables, where $\mathbb{E}(R_{ij}) = 0$ and $\mathbb{V}(R_{ij}) = 1$ for each R_{ij} and also, suppose that $\{R_{ij}\}_{i,j}$ has a uniform subgaussian tail. Next, for fixed $d \in \mathbb{N}$, $\epsilon \in (0, 1/2]$, $\delta \in (0, 1)$, let us set $k = \frac{C\log(2/\delta)}{\epsilon^2}$, for $C \geq 384(1+8/a_R)^2$, where a_R is the constant in the subgaussian upper tail of the R_{ij}s. Finally, let us define the random linear map $T : \mathbb{R}^d \to \mathbb{R}^k$ as follows:

$$T(x)_i = \frac{1}{\sqrt{k}}\sum_{j=1}^{d} R_{ij}x_j, \text{ for } i = 1, 2, \cdots, k,$$

where $T(x)_i$ is the i^{th} coordinate of $T(x) \in \mathbb{R}^k$, and x_j is the j^{th} coordinate of $x \in \mathbb{R}^d$. For every $x \in \mathbb{R}^d$, it turns out that

$$\mathbb{P}\big((1-\epsilon)\|x\| \leq \|T(x)\| \leq (1+\epsilon)\|x\|\big) \geq 1 - \delta.$$

Theorem 2, can be improved upon by further requiring that the projection matrix is sparse. That is, define the mapping $T = XS^T$, where elements of S are i.i.d. according to the following distribution

$$S_{ij} = \begin{cases} q^{-1/2} & \text{with probability } q/2, \\ -q^{-1/2} & \text{with probability } q/2, \\ 0 & \text{with probability } 1-q. \end{cases}$$

[1] In fact, all known proofs of the JL-Lemma rely on statements akin to (1).

In this case, the mapping T can be used to find a JL embedding provided the data points in X are sufficiently well-spread[2].

This idea was first introduced by Achlioptas [1], who considers the two specific cases where $q = 1$ and $q = 1/3$, and shows that $q = 1/3$ is nearly optimal. Ailon and Chazelle [2], then extend this idea by considering highly sparse matrices with $q \to 0$; they show that, as the sparsity of our projection matrix increases, so too does the need for our data points to be well-spread across the dimensions in which they are observed. That is, if our projection matrix consists largely of 0s, then each coordinate of a data point x should hold about the same mass as each other coordinate.

One advantage to using projection coefficients that are i.i.d uniform over $\{-1, 1\}$ is that each coordinate $T(x)_i$ of our projection involves only addition and subtraction of the original coordinates x_j. More specifically, $T(x)_i$ is calculated as follows: partition the coordinates of x randomly into two groups, compute the sum of each group, and set $T(x)_i$ to be the difference of these two sums. This greatly improves runtime when searching for a JL-embedding, since we need not perform repeated matrix multiplication (as is the case when our projection coordinates are i.i.d. gaussian random variables).

If we use i.i.d. projection coefficients with distribution equal to that of S, then we can obtain a JL-embedding about q times faster than when using projection coefficients that are uniform over $\{-1, 1\}$. This is because, in both cases, computation of each coordinate $T(x)_i$ involves addition and subtraction of the original coordinates, but when the projection coefficients are distributed as S, only about q of the original coordinates are considered, with the remaining coordinates sent to 0.

Before moving on, it is useful to note that Theorem 2 can be applied when projection coefficients are i.i.d. according to S, since S is mean 0, unit variance, and S has a subgaussian tail with coefficient $a_S = q^2/2$ (a simple exercise) . However, recall that the reduced space has dimension $k = \frac{C\log(2/\delta)}{\epsilon^2}$, where $C \geq 384(1 + 8/a_S)^2$, so that $q \to 0$ implies $a_S \to 0$, which further implies $k \to \infty$. Therefore, Theorem 2 is not practical when dealing with highly sparse projection matrices distributed according to S.

The following provides a formal discussion of JL-embeddings using sparse projection matrices, following closely the work present in [13]. The key difference between this theorem and Theorem 2 is that the reduced dimensionality k no longer depends on the constant a_S, so long as x is sufficiently well spread.

Theorem 3: Let each of $d \in \mathbb{N}^+$, $\epsilon \in (0, 1/2)$, $\delta \in (0, 1)$, and $\alpha \in [d^{-1/2}, 1]$ be parameters, and define the *sparsity parameter*

$$q = C_0\alpha^2\log(d/\epsilon\delta),$$

where $C_0 \geq 1$ and all parameters are chosen in such a way that $q \in [0, 1]$. Next, define the i.i.d. random variables

$$S_{ij} = \begin{cases} q^{-1/2} & \text{with probability } q/2, \\ -q^{-1/2} & \text{with probability } q/2, \\ 0 & \text{with probability } 1-q, \end{cases}$$

for $i = 1, \cdots, k$, $j = 1, \cdots, d$. Next, set $k = C\epsilon^{-2}\log(4/\delta)$, where $C \geq 768$, and define the random linear mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ as follows:

$$T(x)_i = \frac{1}{\sqrt{k}}\sum_{j=1}^{d} S_{ij}x_j,$$

for $i = 1, \cdots, k$. Then if $x \in \mathbb{R}^d$ such that $\|x\|_\infty \leq \alpha\|x\|$, it follows that

$$\mathbb{P}\big((1-\epsilon)\|x\| \leq \|T(x)\| \leq (1+\epsilon)\|x\|\big) \geq 1 - \delta.$$

3.1 Methodology

Our goal is provide a more specific bound on k than that given by Matoušek. Matoušek gives the same bounds as those given in Theorems 2 and 3, only in both cases, he does not give a specific bound on the constant C but rather, he simply asserts that C is "a sufficiently large constant". First, we reprove Matoušek's results in a more detailed manner in order to obtain specific lower bound on the constant C. Next, we perform a variety of simulations in order to empirically estimate the bound on C.

4. Theoretical Results

Through mathematical analyses similar to those used by Matoušek, we obtain the bounds on k given in Theorems 2 and 3. That is, when using subgaussian projection coefficients we obtain $k = C\log(2/\delta)/(\epsilon^2)$, where $C > 384(1 + 8/a_R)^2$, and where a_R is the coefficient in the subgaussian tail inequality of the projection coefficients R. On the other hand, when using sparse projection matrices we obtain $k = C\log(4/\delta)/(\epsilon^2)$, where $C > 768$.

5. Simulation Results and Discussion

The simulations were performed using Matlab R2013b and using the default random number generator, i.e. the random seed automatically generated by Matlab. To simulate the result of Theorem 2, the projection coefficients were chosen to be standard normal random variables (scaled so that the expected length of each row is equal to 1) using the built-in function normrnd. To simulate the result of Theorem 3, the projection coefficients were chosen to be multinomial distributed over $\{-q^{-1/2}, 0, q^{-1/2}\}$, where 0 has probability $1-q$, $\pm q^{-1/2}$ each have probability $q/2$, and where q is proportional to the L^∞ norm of the simulated

[2] A unit vector is well-spread if it is close to $\frac{1}{\sqrt{d}}(\pm 1, \pm 1, \cdots, \pm 1)$, while something close to $(1, 0, \cdots, 0)$ is not well-spread since most of its mass lies in its first dimension.

data points in accordance with Theorem 3. In an attempt to compare the results for different types of data, datasets were simulated using four different probability distributions: Uniform, Cauchy, Mixed Non-Central Cauchy, and Mixed Beta. These distributions are available through the built-in Matlab functions: unifrnd, trnd, nctrnd, and betarnd; in order to construct each of the mixed distributions, points were randomly selected from two different distributions, which further required use of the built-in function rand.

Each of the simulated datasets consist of $n = 1000$ 10000-dimensional points which are projected into lower dimensions using several different choices of the parameters ϵ and δ, and the values for C suggested by Theorems 2 and 3. The theory is then tested by using the relative frequency approach in order to estimate the probability that each JL embedding maintains ϵ-distortion. That is, for each simulated data set, and for each choice of ϵ and δ, we construct the mapping T using projection matrices outlined in Theorems 2 and 3. Then, for each simulated point x, we compute the ratio $\|T(x)\|/\|x\|$; if this ratio is within $(1-\epsilon, 1+\epsilon)$, then this particular embedding is considered to be a success. Finally, for each simulated data set, and each choice of ϵ and δ, the probability of success is estimated by the number of successful embeddings, over the number of points, $n = 1000$.

Now, according to Theorems 2 and 3, for each fixed ϵ and δ, each point should preserve ϵ-distortion with probability of at least $1-\delta$. However, for each simulated data set, 100% of the projected points preserve ϵ-distortion. This very high frequency of success seems unusual, especially for situations when δ is chosen to be rather large. This discrepancy between the theoretical and empirical results is likely due to an inflated bound on the constant C in each of Theorems 2 and 3. For this reason, the above simulations are repeated using smaller and smaller values of C until the probability bound appears to fall closer to the expected bound of $1-\delta$. Repeating the simulations in this way seems to suggest a significantly lower bound on the reduced dimensionality k than that suggested by the theory. In particular, the simulations consistently suggest that the constant C is between 0.5 and 2.

Concrete Example: The following example illustrates the above discrepancy between the theorized value for C and that which is suggested by simulations. Using $\delta = 0.2$, $\epsilon = 0.5$. and the sparse projection matrix given in Theorem 3, we project $n = 1000$ uniformly random, 10000-dimensional datapoints into k dimensions, where

$$k = C\log(4/\delta)/(\epsilon^2).$$

Thus, our choices of $\delta = 0.2$ and $\epsilon = 0.5$, together with the bound $C > 768$ imply

$$K > 768\log(4/0.2)/(0.5^2) = 9202.$$

Thus, the random mapping T sends a 10000-dimensional point x to the 9202-dimensional point $T(x)$ such that

$$P(1-\epsilon < ||T(x)||/||x|| < 1+\epsilon) > 1-\delta.$$

Due to our choices of $\delta = 0.2$ and $\epsilon = 0.5$, we should therefore expect

$$P(0.5 < ||T(x)||/||x|| < 1.5) > 0.8. \qquad (3)$$

Now, in order to check the validity of (3), we simply compute the ratio of norms $||T(x)||/||x||$ for each projected point and count the number of projections that are not distorted by more than 0.5. Finally, we estimate the probability of success with the relative frequency of such successful projections.

Using the value of $C = 768$, we obtain a success rate of 100%, which is quite large compared to the probability bound of 0.8 suggested by Theorem 3. Accordingly, the above was repeated using smaller and smaller values of C until a value was found that seems to have roughly 80% success rate. It turns out that for C as low as $C = 10$, we still have 100% success rate. Choosing $C = 1$, leads to 93.9% success; choosing $C = 0.75$ leads to 89.5% success, choosing $C = 0.5$ leads to 79.6% success probability. Thus, using the value $C = 768$, given in Theorem 3, leads to a reduced dimensionality of $k = 9202$, whereas the simulations suggested instead that we can use $C = 0.5$ which leads to $k = 6$.

There are a few questions that should follow from the result of this example:

1) Do these results change significantly if we use different data points? (In this particular example, the points were simulated by generating uniformly random 10000-dimensional vectors). The answer is that after generating various random data sets and repeating the above approach, it seems that the type of data point is not a major factor contributing to the huge discrepancy between the reduced dimensionality k obtained by the math vs that obtained by simulations (different data results in slightly different reduced dimensionality, maybe as high as 20 dimensions, but never anything close to 9202).
2) Do these results change significantly if we try different values for the parameters ϵ and δ? The answer is that it does not seem to matter. Changing the values of ϵ and δ leads to different values of k and different probabilities of success (according to the math) but once again, the probabilities are consistently far too high for any fixed k, and in order to make the simulated probability (relative frequency of successful projections) match with the theoretical probability of $1-\delta$ we need to make the constant C much smaller than the value of 768 given in the theorem.
3) Do these results change significantly if we use Theorem 2 instead of Theorem 3? Once again, it seems

that the observed discrepancy is not due to the choice of theorem, but agian due to an inflated value of C.

In summary, the mathematical bounds are far too large and not of much practical use. However, the simulated results seem to suggest that the value C can simply be estimated and tweaked to the particular dataset. Moreover, the simulated results suggest a much more practical result. In the above, for example, the math says that we can go from 10000 dimensions to 9202 (not very helpful), while the simulated results suggest that we can go from 10000 dimensions into only 6 dimensions (very useful indeed).

6. Conclusions

We have discussed a non-statistical method of dimensionality reduction, where any given set of points can be embedded into lower dimensions, although such embeddings are typically subject to some form of distortion. Regardless of the initial dimensionality, the JL-lemma guarantees the existence of a lower dimensional representation, the dimensionality of which depends on the number of points as well as the level of distortion one is willing to accept.

Mathematics gives a weaker bound on k than do our simulations. In particular, the simulations seem to suggest that C is generally around $C = 1$. This means that our mathematical result (in particular, the bound on C) is hundreds of times larger than the simulations suggest (or even thousands when using Theorem 2 , depending on the choice of subgaussin projection coefficients) and as such, our bound on k is hundreds (to thousands) times larger than that which is suggested by simulation.

References

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

[2] N. Ailon and B. Chazelle, *Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform*. Proceedings of the 38th ACM Symposium on the Theory of Computing, 2006, pp. 557-563.

[3] K.M. Carter , *Dimensionality reduction on statistical manifolds*, PhD thesis, The University of Michigan, Department of Engineering, 2009.

[4] S. Dasgupta and A. Gupta, *An elementary proof of the Johnson-Lindenstrauss lemma*. Technical Report 99-006, UC Berkely,March 1999.

[5] D. Donoho, *Aide-memoire. High-dimensional data analysis: the curses and blessings of dimensionality*, (2000), available at http://statweb.stanford.edu/ donoho/Lectures/AMS2000/Curses.pdf.

[6] P. Frankl and H. Maehara, *The Johnson-Lindenstrauss lemmas and the sphericity of some graphs*, Journal of Combinatorial Theory Series B 44(3):355-362, 1988.

[7] G. Heo, J. Gamble, P.T. Kim. *Topological analysis of variance and the maxillary complex*, J Am Stat Assoc. 2011

[8] H. Gao, W. Hong, J. Cui, Y. Zhao and H. Meng, *Pattern recognition of multivariate information based on non-statistical techniques*, International Conference on Information and Automation, 2008, pp.697–702.

[9] J. Gamble, H. Geo, *Exploring uses of persistent homology for statistical analysis of landmark-based shape data*, J Multivar Anal. 101:2184-2199, 2010.

[10] A.C. Gyllensten and M. Sahlgren , *Navigating the semantic horizon using relative neighborhood graphs*, 2015, CoRR, abs/1501.02670.

[11] P. Indyk and R. Motwani, *Approximate nearest neighbors: towards removing the curse of dimensionality*, 30th Annual ACM Symposium on Theory of Computing, Dallas, TX, ACM, New York, 1998, pp.604-613.

[12] W.B. Johnson and J. Lindenstrauss, *Extensions of Lipshitz mappings into a Hilbert space*, Conference in modern analysis and probability, New Haven, CI, 1982, American Mathematical Society, Providence, RI, 1984, pp.189-206

[13] J. Matoušek, *On variants of the Johnson-Lindenstrauss lemma*, Random structures and algorithms 33(2):142–156, 2008.

[14] J. Wang, *Classical multidimensional scaling. Geometric Structure of High-Dimensional Data and Dimensionality Reduction*, (2011) Springer Heidelberg Dordrecht, London New York, pp. 115-129.