8-2017

Johnson-Lindenstrauss Transformations

Fiona Knoll
Clemson University, fknol309@gmail.com

Johnson-Lindenstrauss Transformations

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematical Sciences

by
Fiona Knoll
August 2017

Accepted by:
Dr. Shuhong Gao, Committee Chair
Dr. Gretchen Matthews
Dr. Felice Manganiello
Dr. June Luo

Abstract

With the quick progression of technology and the increasing need to process large data, there has been an increased interest in data-dependent and data-independent dimension reduction techniques such as principle component analysis (PCA) and Johnson-Lindenstrauss (JL) transformations, respectively. In 1984, Johnson and Lindenstrauss proved that any finite set of data in a high-dimensional space can be projected into a low-dimensional space while preserving the pairwise Euclidean distance within any desired accuracy, provided the projected dimension is sufficiently large; however, if the desired projected dimension is too small, Woodruff and Jayram, and Kane, Nelson, and Meka in 2011 separately proved such a projection does not exist. In this thesis, we answer an open problem by providing a precise threshold for the projected dimension, above which, there exists a projection approximately preserving the Euclidean distance, but below which, there does not exist such a projection. We, also, give a brief survey of JL constructions, covering the initial constructions and those based on fast-Fourier transforms and codes, and discuss applications in which JL transformations have been implemented.

Acknowledgments

I would like to thank my advisor Shuhong Gao for his guidance and patience these past few years. He has taught me more than he can ever imagine, for which I am very thankful.

I would also like to thank Dr. Burr who assisted in my understanding of the material and took time to provide me with insightful ideas.

In addition, I would like to express my gratitude to my committee: Dr. Gretchen Matthews, Dr. Felice Manganiello, and Dr. June Luo for their support in this process.

Finally, words cannot express my thankfulness for my family and friends who have encouraged me along the way.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Brief Overview of Projection Methods

High-dimensional data poses challenges to tasks such as data mining and pattern recognition. Data with a large amount of attributes, i.e., data in a high-dimensional space, is prone to noise and irrelevant information. In addition, the computation of running tasks on high-dimensional data is costly in comparison to low-dimensional data. Another challenge is the curse of dimensionality: as the dimension grows, the distance between the data points becomes indistinguishable.

Traditionally, to overcome these challenges, the data is projected to a lower dimension dependent on the inherent dimension of the data set; whereas, some modern projection methods, such as random projections, do not take into consideration the structure of the underlying data. Ideally, the projection, whether dependent or independent of the data, retains as much of the variation as possible. There are two basic techniques of data-dependent projections: linear and nonlinear. Linear techniques, such as Principal Component Analysis (PCA), assume the data lies in the proximity of a linear subspace; whereas, nonlinear techniques, such as multidimensional scaling (MDS), do not place any assumption of linearity on the data.

The PCA method seeks to preserve as much variance of the data's variance as possible by choosing dimensions along which the points are maximally spread out. It accomplishes this by finding linear combinations of the entries with maximal variance.

The MDS method, a method to visually compare data, seeks to preserve the predefined distances between points. The distances are either based on a metric or nonmetric. Typically, MDS is used to project the data into a 2-dimensional subspace to allow the user to visualize the relationships between the data. Due to the dependence on the data's underlying structure of both PCA and MDS, a disadvantage of both PCA and MDS is that the projected dimension relies on the inherent dimension of the data. A description of PCA and MDS can be found in Section 1.4 and Section 1.5, respectively.

Random projections, e.g., Johnson-Lindenstrauss transformations, are projections not dependent on the data. Random projections have been shown, both heuristically and theoretically, to approximately preserve the similarity of data, and are advantageous over the prior methods in terms of speed. This thesis predominantly focuses on Johnson-Lindenstrauss transformations.

1.2 Johnson-Lindenstrauss Transformation

Contrary to projections such as PCA and MDS, random projections do not take into consideration the data to be projected. In a random projection, a subspace onto which the data is to be projected, is randomly selected independent of the data. Random projections have been shown to preserve pairwise Euclidean distance with high probability [28]. In particular, Johnson and Lindenstrauss, in [28], showed for any set of n points in a d-dimensional space, there is a linear transformation which projects the points to a k-dimensional subspace, independent of d, such that pairwise Euclidean distance is approximately preserved. For a d-dimensional vector $x \in \mathbb{R}^d$, let $\|x\|_2 = \left(\sum_{i=1}^d x_i^2\right)^{1/2}$ denote the Euclidean norm. The following theorem is attributed to Johnson and Lindenstrauss.

Theorem 1 ([28]). For any $0 < \epsilon < \frac{1}{2}$ and $0 < \delta < \frac{1}{2}$, if $k \geq C \cdot \epsilon^{-2} \log \frac{1}{\delta}$ for some absolute constant $C > 0$, then there exists a probability distribution $\mathcal{D}$ on $k \times d$ real matrices such that, for any $x \in \mathbb{R}^d$,

$$\text{Prob}_{A\sim\mathcal{D}}\left[(1-\epsilon)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\epsilon)\|x\|_2^2\right] > 1-\delta, \tag{1.1}$$

where $A \sim \mathcal{D}$ means matrix A is chosen uniform and at random from distribution $\mathcal{D}$.

In this thesis, we write log for ln unless otherwise stated.

Definition 1. A distribution on $k \times d$ matrices such that for any $0 < \epsilon, \delta < \frac{1}{2}$ and $x \in \mathbb{R}^d$ Inequality (1.1) holds is called a Johnson-Lindenstrauss (JL) distribution. Any matrix from a JL distribution is called a JL transformation.

Observe that the projected dimension of JL transformations is not dependent on the data; whereas, in both PCA and MDS, the projected dimension depends on the underlying dimension of the data. Since JL transformations are independent of the data, the time required to both construct the projection and project the data is smaller than that of data dependent methods.

1.2.1 Computational Problems

With the quick progression of technology and the increasing need to process large amounts of data, there has been an increased interest in applying JL transformations. The ability to project a vector to a smaller dimension, independent of the original dimension, while approximately preserving the Euclidean norm is highly desirable and has applications in similarity tests, streaming algorithms, data mining, compressive sensing, machine learning, etc.

For instance, consider the bag of words scenario where each document is represented by a vector of occurrences. That is, each entry of a vector represents the number of occurrences of a word or phrase in the document. According to the Oxford dictionary, there are $171,476$ words currently in use and $47,156$ obsolete words. This implies each vector, when only considering words, will be at least $170,000$ in length. Adding in phrases, the length will be much larger. To store vectors of occurrences for n documents, $n \cdot d$ storage space is required, where d is the length of the vector; however, if only similarity is to be preserved, the vectors can be stored using a random projection, such as a JL transformation, requiring only $\mathcal{O}(n \log n)$ space. Other instances of application include, but are not limited to, rate matrices, images, genotypes, and network connections.

For a JL distribution $\mathcal{D}$ to be useful in practice, there are a few desirable traits. More precisely, we consider the following computation problems in this thesis.

Problem 1: Creating a JL distribution $\mathcal{D}$ such that the transforms from $\mathcal{D}$ are simple to construct. That is, the construction involves a few random bits or the transform has a simple form.

For instance, consider two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ on $k \times d$ matrices. Let $\mathcal{D}_1$ be a distribution on matrices such that each entry of the matrix is chosen at random from the normal distribution with mean 0 and variance 1. Let $\mathcal{D}_2$ be a distribution on matrices such that each entry is chosen uniformly and at random from the set $\{\pm 1\}$. In both cases, there are $k \cdot d$ random entries; however, matrices from $\mathcal{D}_2$ can be constructed faster.

Problem 2: Constructing a JL distribution $\mathcal{D}$ such that the computation of the projection is fast.

For instance, consider two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ on $k \times d$ matrices. Let $\mathcal{D}_1$ be a distribution on matrices such that each entry is chosen uniformly and at random from the set $\{\pm 1\}$. Let $\mathcal{D}_2$ be a distribution on matrices such that two-thirds of the entries of each matrix are 0 and otherwise chosen uniformly and at random from $\{\pm 1\}$. In comparison to $\mathcal{D}_1$, the distribution $\mathcal{D}_2$ gives a three-fold speed up in the computation of Ax for $A \sim \mathcal{D}_2$.

Problem 3: For a fixed d, k, and ϵ, find a distribution so that the probability of failure δ can be as small as possible.

Suppose we want to project a set of $n = 2^{10}$ vectors, e.g., images, to a smaller dimension. Suppose the probability of failure for each vector is $\delta = 2^{-20}$. Then, the probability of preserving pairwise distance will be $1-\delta'$, where

$$\delta' = \frac{\delta n^2}{2} = \frac{1}{2}.$$

Observe that in order to keep δ' fixed, as the number of vectors to project increases, the probability of failure δ must decrease. Hence, we desire to obtain a small probability of failure for a single vector for a fixed k, d, and ϵ.

Problem 4: Obtaining the smallest possible projected dimension for a fixed ϵ and δ. The question "What is the smallest possible projected dimension for an error factor of ϵ and probability of failure of δ?" arises with this problem.

1.3 Summary of Thesis

In Chapter 2, we present major progress on problems 1, 2, and 3. We first give an overview on the development of JL distributions. Then, we will expand on the main known constructions in literature, focusing on speed of both the projection and the construction of a transform.

In Chapter 3, we provide our main result, which addresses problem 4. For any ϵ, δ, let $k_0(\epsilon, \delta)$ denote the minimum k such that there is a JL distribution on $\mathbb{R}^{k\times d}$. For ease of notation, we write k_0 for $k_0(\epsilon, \delta)$. In 1988, Frankl and Maehara [20] showed there exists a JL distribution for $k \geq 9\epsilon^{-2}\log\frac{1}{\delta}$, resulting in an upper bound for k_0:

$$k_0 \leq 9\epsilon^{-2}\log\frac{1}{\delta}.$$

In 1998, Indyk and Motwani [26] provided a JL distribution for

$$k > 2\log\frac{2}{\delta}\left(\frac{\epsilon^2}{2}-\frac{\epsilon^3}{3}\right)^{-1},$$

resulting in an even tighter upper bound:

$$k_0 \leq 2\log\frac{2}{\delta}\left(\frac{\epsilon^2}{2}-\frac{\epsilon^3}{3}\right)^{-1} = 4\epsilon^{-2}\log\frac{1}{\delta}\left[1+\frac{\epsilon}{3-\epsilon}\left(1+\frac{\log 2}{\log\frac{1}{\delta}}\right)+\frac{\log 2}{\log\frac{1}{\delta}}\right].$$

It was not until 2003, when a lower bound for k_0 was given. Alon [4] proved for $k \leq C \cdot \epsilon^{-2}\log\frac{1}{\delta}/\log\frac{1}{\epsilon}$ for some global constant $C > 0$, there is no JL distribution. Woodruff and Jayram [27] and later Kane, Nelson, and Meka [37] improved Alon's result by eliminating the factor $\log\frac{1}{\epsilon}$. They showed for $k \leq C \cdot \epsilon^{-2}\log\frac{1}{\delta}$ for some global constant $C > 0$, there is no JL distribution. Hence, we have a lower bound

$$k_0 \geq C\epsilon^{-2}\log\frac{1}{\delta}$$

for some $C > 0$.

In Chapter 3, we provide a precise threshold for k_0. Our main result is the following theorem:

Theorem 2. For ϵ and δ sufficiently small, $k_0 \approx 4\epsilon^{-2}\log\frac{1}{\delta}$. That is,

$$\frac{k_0}{4\epsilon^{-2}\log\frac{1}{\delta}} \to 1 \text{ as } \epsilon, \delta \to 0.$$

Hence, $4\epsilon^{-2}\log\frac{1}{\delta}$ is a threshold for k_0.

In Chapter 4, we describe the applications of JL distributions in approximate nearest neighbors, linear algebra, machine learning, compressed sensing, and differential privacy.

For the purpose of comparison, we describe PCA and MDS in more depth, but they will not be referenced in the thesis.

1.4 Principal Component Analysis

Principal component analysis (PCA) is a linear projection technique exploiting the underlying structure of the data. Given a sample of n observations on d variables

$\{x_1, \dots, x_n\} \subset \mathbb{R}^d$, PCA retains a large proportion of the data's variance by finding linear combinations of the d variables with large variance and keeping the largest combinations. The linear combinations are called principal components.

The first principal component for each point is the linear combination of the entries with maximal variance. The second principal component is the linear combination orthogonal to the first component with maximal variance. The other principal components are similarly chosen. In essence, PCA finds the dimensions along which the points are maximally spread out.

Let X be the data matrix whose rows consist of the n points

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n \times d}.$$

Let $x_p^{(i)}$ be the i^{th} entry (variable) of data point $x_p \in \mathbb{R}^d$. Let $\bar{x}^{(i)} := \frac{1}{n} \sum_{j=1}^{n} x_j^{(i)}$ be the mean of the i^{th} variable and $\bar{x} = (\bar{x}^{(1)}, \dots, \bar{x}^{(d)}) \in \mathbb{R}^d$ be the mean vector. Geometrically, PCA translates the origin of the points $x_1, \dots, x_n$ to the mean $\bar{x}$ and then rotates the axes to the natural axes to capture the maximum amount of variance. Let $\hat{x}_p = x_p - \bar{x}$ for $1 \leq p \leq n$ be the points centered about the mean and

$$\hat{X} = X - \frac{1}{n} J_n X = \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \hat{x}_n \end{pmatrix},$$

where J_n is the $n \times n$ matrix of ones, be the data matrix of those points. The rotation is accomplished by applying an orthogonal linear transformation $A \in \mathbb{R}^{k \times d}$ to each shifted point $\hat{x}_p$. Let $z_p = A\hat{x}_p$ for $1 \leq p \leq n$. Each entry $z_p^{(i)} = (A\hat{x}_p)^{(i)}$ is called a principal component of x_p. By design of A, the first principal component $z_p^{(1)}$ is the linear combination of the original entries with maximal variance.

To construct the orthogonal transform A, we must first consider the sample covariance matrix S of $x_1, \ldots, x_n$ and its spectral decomposition. The sample covariance matrix can be defined in terms of the shifted data matrix $\hat{X}$:

$$S = \frac{1}{n-1} \hat{X}^\top \hat{X}.$$

That is, the covariance between two variables r and t, $1 \leq r, t \leq d$ is

$$s_{rt} = \frac{1}{n-1} \sum_{p=1}^{n} \hat{x}_p^{(r)} \hat{x}_p^{(t)} = \frac{1}{n-1} \sum_{p=1}^{n} \left(x_p^{(r)} - \bar{x}^{(r)} \right) \left(x_p^{(t)} - \bar{x}^{(t)} \right)$$

and the variance of a variable t is

$$s_t = s_{tt} = \frac{1}{n-1} \sum_{p=1}^{n} \left(x_p^{(t)} - \bar{x}^{(t)} \right)^2.$$

The variance indicates how spread out the data is and the covariance indicates how closely related the variables are. More specifically, the sample covariance measures the linear relationship between the variables. For instance, if variables r and t are independent of each other, the covariance is zero. Since S is a real symmetric matrix, it is diagonalizable and has spectral decomposition

$$S = CDC^\top,$$

where $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ is the diagonal matrix consisting of the eigenvalues of S and C is an orthogonal matrix whose columns are the normalized eigenvectors of S. Let $A = C^\top$, i.e., the rows of A consist of the normalized eigenvectors of the sample covariance matrix S. Observe A can also be constructed using the singular value decomposition $U \Sigma V^\top$ of the shifted data matrix $\hat{X}$ since

$$S = \frac{1}{n-1}\hat{X}^\top \hat{X} = \frac{1}{n-1} V\Sigma^2 V^\top.$$

That is, $A = V^\top$ and $Ax = \begin{pmatrix} v_1 \cdot x \\ \vdots \\ v_d \cdot x \end{pmatrix}$, where v_i is the i^{th} column of V.

Since the covariance matrix of $z_1, \ldots, z_n$ is

$$S_z = ASA^\top$$

and $A = C^\top$, the covariance matrix S_z is a diagonal matrix consisting of the eigenvalues of S:

$$S_z = C^\top SC = \mathrm{diag}(\lambda_1, \ldots, \lambda_d).$$

Hence, the new variables are uncorrelated, implying that any two principal components are orthogonal and uncorrelated as desired. In addition, the variance of the first principal component $z^{(1)}$ corresponds to the largest eigenvalue, implying the linear combination has maximal variance.

1.4.1 Reducing the Dimension

To reduce the dimension of each point x_i for $1 \leq i \leq n$, the principal components with least variance are dropped. That is, entries $v_{k+1} \cdot x, \ldots, v_d \cdot x$ are dropped and $A = (v_1, \ldots, v_k)^\top$. In practice, there are four methods to determine the number of the principal components to retain.

1. The first k principal components which account for a specified proportion of variance are kept.

2. The principal components whose corresponding eigenvalues are greater than the average eigenvalue are kept.

3. Using the scree graph, a plot of the eigenvalues λ_i versus i, a natural break between the eigenvalues is determined and the principal components corresponding to the eigenvalues before the break are kept.

4. A hypothesis test with the hypothesis that the last $d - k$ eigenvalues are small and equal can be used.

Example 1. Consider the following sample covariance matrix

$$S = \begin{bmatrix} 2.69976 & 3.42433 & -.336648 & 3.0133 & 9.5361 & 1.93129 & 4.19314 \\ 3.42433 & 8.89447 & -.690373 & 5.1144 & 19.6511 & 4.56293 & 4.00574 \\ -.336648 & -.690373 & .190764 & -1.32775 & -4.00663 & 1.39379 & -.822954 \\ 3.0133 & 5.1144 & -1.32775 & 171.105 & 226.614 & 234.935 & 114.466 \\ 9.5361 & 19.6511 & -4.00663 & 226.614 & 457.672 & 222.548 & 178.559 \\ 1.93129 & 4.56293 & 1.39379 & 234.935 & 222.548 & 779.481 & 172.498 \\ 4.19314 & 4.00574 & -.822954 & 114.466 & 178.559 & 172.498 & 160.559 \end{bmatrix}$$

for samples on 7 variables. Then, the eigenvalues of S with the corresponding proportion of variance is

Eigenvalues	Proportion
1087.235	0.6879
383.487	0.2426
70.245	0.0444
29.731	0.0188
8.671	0.0055
1.129	$7.3199e^{-4}$
0.104	$6.5867e^{-5}$

The corresponding eigenvector matrix C, rounded to three decimal places, is

$$C = \begin{bmatrix} -0.008 & -0.019 & 0.002 & -0.070 & 0.404 & 0.911 & -0.047 \\ -0.015 & -0.036 & -0.060 & -0.167 & 0.892 & -0.411 & -0.046 \\ 0.001 & 0.011 & 0.008 & 0.006 & -0.060 & -0.023 & -0.998 \\ -0.347 & -0.203 & -0.041 & 0.902 & 0.150 & -0.005 & -0.006 \\ -0.473 & -0.727 & -0.341 & -0.341 & -0.120 & 0.008 & -0.006 \\ -0.761 & 0.620 & -0.105 & -0.154 & -0.021 & 0.005 & 0.005 \\ -0.276 & -0.208 & 0.931 & -0.113 & 0.017 & -0.025 & 0.003 \end{bmatrix}.$$

Method 1: Suppose we want to keep 80% of the the variance. Observe that the variance of λ_1 and λ_2 makes up approximately 94% of the data. As a result, we want to keep the eigenvectors corresponding to the first two eigenvalues.

Method 2: The average of the eigenvalues is 225.8. Observe that only λ_1 and λ_2 are greater than the average. As a result, we want to keep the eigenvectors corresponding to the first two eigenvalues.

Method 3: The scree graph of the eigenvalues of S is given in Figure 1.1. Observe that there is a natural break after the second eigenvalue. Again, we choose to keep the eigenvectors corresponding to the first two eigenvalues.

As method 4 is not relevant to this thesis, we do not provide an example using it.

1.5 Multidimensional Scaling

A nonlinear projection technique that exploits the structure of the data is multidimensional scaling (MDS) and is typically used to visually compare data. MDS seeks to project data to a smaller dimension k, usually $k = 2$, while preserving the pairwise distance between the points. If the distance is based on a metric, the technique is called metric multidimensional scaling. If the distance is a similarity measure based on judgment, the technique is called nonmetric multidimensional scaling.

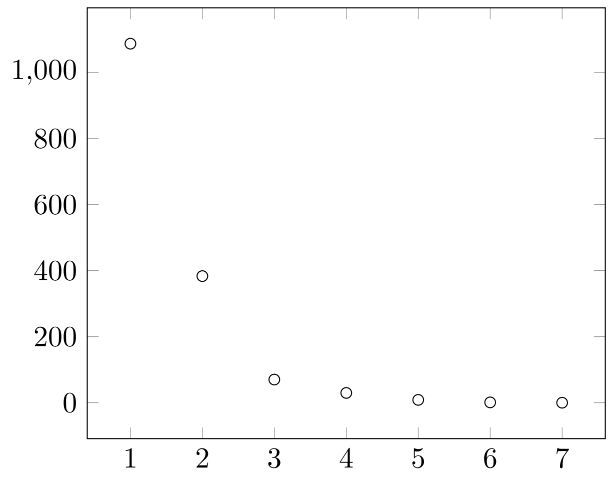


Figure 1.1: The scree graph is a plot of the eigenvalues, and is used in PCA to determine which eigenvalues to retain.

Given a set of points $x_1, \ldots, x_n$ and distances $d_{i,j} = d(x_i, x_j)$, MDS constructs a distance matrix $D = (d_{i,j})_{1 \leq i,j \leq n}$ and scales D by $-\frac{1}{2}$ to create matrix A:

$$A = -\frac{1}{2} (d_{i,j})_{1 \leq i,j \leq n} .$$

Let J_n be the $n \times n$ matrix of all 1's. Then, matrix B, from which we obtain the new set of points to which to project, is defined as

$$B = \left(I_n - \frac{1}{n} J_n \right) A \left(I_n - \frac{1}{n} J_n \right) .$$

Since B is a symmetric real matrix, it can be decomposed into

$$B = V \Lambda V^\top ,$$

where V consists of the eigenvectors of B and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ such that λ_i is an eigenvalue of B. Let

$$\Lambda_q = \operatorname{diag}(\lambda_1, \ldots, \lambda_q)$$

be the submatrix of Λ consisting of the first q eigenvalues. Then, the set of points to which to project is given by

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = V\Lambda_q^{1/2}.$$

Observe that $d(x_i, x_j) = d(y_i, y_j)$ if and only if B is positive semidefinite of rank q, that is, $\lambda_1, \ldots, \lambda_q > 0$ and $\lambda_{q+1}, \ldots, \lambda_n = 0$. In practice, $\text{rank}(B)$ is too large or B is not positive semidefinite. In either case, if the first k eigenvalues are positive and relatively large, then the rest of the eigenvalues can be dropped, giving k-dimensional points.

Chapter 2

Survey of JL Constructions

In 1984, Johnson and Lindenstrauss [28], in proving a bound on the Lipschitz constant, showed that any finite set of data in a high-dimensional space can be projected into a low-dimensional space while preserving the pairwise Euclidean distance within any desired accuracy. In particular, for any finite set of vectors $x_1, \ldots, x_n$ and for any $0 < \epsilon < \frac{1}{2}$, if $k \geq C \cdot \epsilon^{-2} \log n$ for some constant C, then there exists a linear map $A : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ such that

$$(1-\epsilon)\|x_i - x_j\|_2 \leq \|Ax_i - Ax_j\|_2 \leq (1+\epsilon)\|x_i - x_j\|_2$$

for all pairs $1 \leq i, j \leq n$. The following theorem is attributed to Johnson and Lindenstrauss and can be seen to be equivalent to the above by setting $\delta = \frac{1}{n^2}$ and taking the union bound, defined in (2.2).

Theorem 3 ([28])**.** For any $0 < \epsilon, \delta < \frac{1}{2}$, if $k \geq C \cdot \epsilon^{-2} \log \frac{1}{\delta}$ for some absolute constant $C > 0$, then there exists a probability distribution $\mathcal{D}$ on $k \times d$ real matrices such that, for any $x \in \mathbb{R}^d$,

$$\text{Prob}_{A \sim \mathcal{D}} \left[(1-\epsilon)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\epsilon)\|x\|_2^2\right] > 1 - \delta, \tag{2.1}$$

where $A \sim \mathcal{D}$ denotes that matrix A is chosen randomly from distribution $\mathcal{D}$.

There has been an ample amount of literature on explicit constructions of JL distributions. The construction provided by Johnson and Lindenstrauss [28] has the following constraints : the rows are both orthogonal and normal, and the transformation A is spherically symmetric, that is, A and BA have the same distribution for any orthogonal matrix B. Frankl and Maehara [20] simplified the proof of Johnson and Lindenstrauss and provided a JL distribution where the transformations are randomly chosen from the Stiefel manifold $V_d(\mathbb{R}^k) = \{B \in \mathbb{R}^{k\times d} : B^\top B = I_d\}$. Indyk and Motwani [26] relaxed the orthogonal and normal constraints and constructed a JL distribution by randomly and uniformly choosing each entry a_{ij} from the normal distribution $N(0, \frac{1}{\sqrt{k}})$. However, these constructions are dense and hence, are computationally inefficient. In 2003, Achlioptas [1] constructed a somewhat sparse distribution:

$$a_{ij} = \sqrt{\frac{3}{k}} \cdot \begin{cases} 1 & \text{with probability } 1/6 \\ 0 & \text{with probability } 2/3 \\ -1 & \text{with probability } 1/6 \end{cases}$$

speeding up the process three-fold. To further speed up the computation, Ailon and Chazelle [2] in 2009 implemented fast Fourier analysis in their construction Fast JL Transform (FJLT). In FJLT, a Fourier transform is applied to the vector x prior to projecting x with a sparse matrix whose nonzero entries are from the Gaussian distribution. In [36], Matousek improved the work of Ailon and Chazelle by choosing the nonzero entries of the sparse matrix from the set $\{\pm 1\}$ uniformly and at random. The construction of Dasgupta, Kumar, and Sarlós [12] further simplified the complexity by utilizing hash functions to create a sparser matrix. Furthering this idea, Kane and Nelson [29] provided an even sparser construction by utilizing codes and graph theory to specify the placement of the nonzero entries.

We look at these constructions in more detail in the subsequent sections.

2.1 Prerequisites

2.1.1 Results in Probability

The expected value of variable X, denoted $E[X]$, is the weighted average of all possible values that X can assume. In the countable discrete case,

$$E[X] = \sum_{i=1}^{\infty} x_i p_i,$$

where p_i is the likelihood of X taking on value x_i. When the expected value of an estimator is equal to the true value of the parameter being estimated, the estimator is called unbiased; otherwise, biased.

We refer to the following two inequalities throughout the paper:

Theorem 4 (Markov's Inequality). The probability of a nonnegative random variable X being at least a is bounded above by its expected value scaled by the factor $\frac{1}{a}$. That is,

$$\text{Prob}\left[X \geq a\right] \leq \frac{1}{a} E[X].$$

Theorem 5 (Union Bound - Boole's Inequality). For a countable set of events $A_1, A_2, \ldots$, the probability of at least one of the events occurring is bounded above by the sum of the probabilities of each event occurring:

$$\text{Prob}\left[\cup_i A_i\right] \leq \sum_i \text{Prob}\left[A_i\right]. \tag{2.2}$$

Definition 2. A set of random variables $X_1, \ldots, X_n$ is k-wise independent, $k \geq 2$, if for any $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ and events $a_{i_1} \in A_{i_1}, \ldots, a_{i_k} \in A_{i_k}$,

$$\text{Prob}\left[X_{i_1} = a_{i_1} \cap \cdots \cap X_{i_k} = a_{i_k}\right] = \prod_{j=1}^{k} \text{Prob}\left[X_{i_j} = a_{i_j}\right]. \tag{2.3}$$

Let $x \overset{iid}{\sim} \mathcal{D}$ denote when variable x is independent and identically distributed (i.i.d.)

from distribution $\mathcal{D}$. The main distribution employed in this paper is the normal distribution with mean 0 and variance 1, denoted $N(0,1)$. The likelihood of $x \overset{iid}{\sim} N(0,1)$ taking on value $t \in \mathbb{R}$ is

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2},$$

which is the probability density function of $N(0,1)$. A random variable X is said to be subgaussian if there is a constant $b > 0$ such that for every $t \in \mathbb{R}$,

$$E[e^{tx}] \leq e^{b^2t^2/2}.$$

An instance of a subgaussian random variable is a random variable from the centered normal distribution $N(0,\sigma^2)$. By the central limit theorem, the sum of random variables, regardless of the underlying distribution, tends to the normal distribution.

2.1.2 Notation

Let $\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{\frac{1}{p}}$, where $x \in \mathbb{R}^d$, denote the ℓ_p norm for $1 \leq p < \infty$. When $p = 2$, the ℓ_p norm is referred to as the Euclidean norm. We write $\|\cdot\|$ to denote the Euclidean norm unless otherwise specified. The set of vectors in $\mathbb{R}^d$ with Euclidean norm 1 is the $(d-1)$-dimensional unit sphere S^{d-1}. Let $\|x\|_\infty = \max_i\{|x_i|\}$ denote the ℓ_∞ norm and $\|x\|_0$ denote the number of nonzero entries of vector x.

Given matrix $A \in \mathbb{R}^{m\times n}$, the Frobenius norm, denoted $\|A\|_F$, is defined as

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2},$$

and the spectral norm, $\|A\|_2$, is defined as the largest singular value of A. It will be clear from context whether $\|\cdot\|_2$ refers to the Euclidean or the spectral norm.

In the construction of JL transformations, two well-known transforms will be employed: normalized Walsh-Hadamard and Fourier. A $d \times d$ normalized Walsh-Hadamard matrix, denoted H, has entries

$$h_{ij} = \frac{1}{\sqrt{d}}(-1)^{\langle i-1,j-1\rangle}, \tag{2.4}$$

where $\langle i, j\rangle$ is the inner product (mod 2) of the binary representation of i and j. A $d \times d$ Fourier transform, denoted F, has entries

$$f_{jk} = e^{-2\pi ikj/d}$$

for $k = 0, \ldots, d-1$ and $j = 0, \ldots, d-1$, where $i = \sqrt{-1}$. That is, for $x \in \mathbb{R}^d$,

$$F(x) = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{-2\pi i/d} & e^{-4\pi i/d} & \cdots & e^{-2(d-1)\pi i/d} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & e^{-2(d-1)\pi i/d} & e^{-4(d-1)\pi i/d} & \cdots & e^{-2(d-1)^2\pi i/d} \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{d-1} \end{pmatrix}.$$

Let $f(x)$ and $g(x)$ be functions on $\mathbb{R}$. Then,

- $f(x) = \mathcal{O}(g(x))$ if and only if for some constants x_0 and $m > 0$, $|f(x)| \geq m|g(x)|$ for $x \geq x_0$.

- $f(x) = \Omega(g(x))$ if and only if for some constants x_0 and $M > 0$, $|f(x)| \leq M|g(x)|$ for $x \geq x_0$.

- $f(x) = \Theta(g(x))$ if and only if $f(x) = \mathcal{O}(g(x))$ and $f(x) = \Omega(g(x))$. That is, there are constants $m, M > 0$ and x_0 such that

  $$m|g(x)| \leq |f(x)| \leq M|g(x)|$$

  for $x \geq x_0$.

- $f(x) = o(g(x))$ if and only if $f(x) = \mathcal{O}(g(x))$, but $f(x) \neq \Theta(g(x))$.

- $f(x) = \omega(g(x))$ if and only if $f(x) = \Omega(g(x))$ and $f(x) \neq \Theta(g(x))$.

2.2 Dense Constructions

2.2.1 Johnson and Lindenstrauss (1984)

In the Lipschitz extension problem, one desires to find the smallest constant L such that for a function f mapping any n-element subset X' of a metric space X to ℓ_2, f can be extended to a function $\tilde{f}$ that maps X to ℓ_2 and for all $x \in X'$

$$\|\tilde{f}(x)\| \leq L\|f(x)\|.$$

Johnson and Lindenstrauss [28], upon giving a bound to the Lipschitz constant L, proved that any finite set of data in a high-dimensional space can be projected into a low-dimensional space while preserving the pairwise Euclidean distance within any desired accuracy.

Their distribution consists of transformations, which are orthogonal projections to a random k-dimensional subspace of $\mathbb{R}^d$. In particular, let

$$Q = \left(\begin{array}{c|c} I_k & 0 \\ 0 & 0 \end{array} \right)_{d\times d} . \tag{2.5}$$

Consider the Euclidean norm $\|Qx\|_2$ for $x \in S^{d-1}$. Observe that $\|Qx\|_2^2$ is simply the sum of the first k entries of x squared. Consider the constant $M \in \mathbb{R}$ such that

$$\text{Prob}\left[\sqrt{d}\|Qx\|_2 \geq M\right] = \frac{1}{2} = \text{Prob}\left[\sqrt{d}\|Qx\|_2 \leq M\right].$$

The constant M is called the Levy median of $\sqrt{d}\|Qx\|_2$. Combining Q and M with an orthonormal matrix, we obtain a transformation from the JL distribution.

Distribution: Let $M > 0$ be a fixed constant as described. Let $Q \in \mathbb{R}^{d\times d}$, $d \geq n$, be defined as in Equation (2.5). Let $U \in \mathbb{R}^{d\times d}$ be a random orthonormal matrix. Then, let $\mathcal{D}$

be a distribution on $d \times d$ matrices such that, for $A \sim \mathcal{D}$,

$$A = \sqrt{\frac{d}{M}} QU.$$

To show $\mathcal{D}$ is a JL distribution, it must be shown that, for $x \in S^{d-1}$, the sum of the first k entries squared is sharply concentrated around $\sqrt{\frac{M}{d}}$ with high probability. We will give a brief overview of this proof. More details can be found in [28] and [35].

Lemma 1 (Levy's Lemma)**.** Let $f : S^{n-1} \to \mathbb{R}$ be 1-Lipschitz, that is $|f(a)-f(b)| \leq |a-b|$. Then for all $t \in [0,1]$,

$$\text{Prob}\left[f \geq \text{med}(f) + t\right] \leq 2e^{-t^2 n/2} \quad \text{and} \quad \text{Prob}\left[f \leq \text{med}(f) - t\right] \leq 2e^{-t^2 n/2},$$

where $\text{med}(f) = \sup\{\ell \in \mathbb{R} : \text{Prob}\left[f \leq \ell\right] \leq \frac{1}{2}\}$.

Observe $f(x) = \|Qx\|_2$ is 1-Lipschitz. Let $m = \text{med}(f)$ and $t = \epsilon m$, then combining the probabilities in Levy's lemma, we obtain

$$\text{Prob}\left[|\|Qx\|_2 - m| > \epsilon m\right] \leq 4e^{-\epsilon^2 m^2 d/2}. \tag{2.6}$$

It was shown in [35], that $m \geq \frac{1}{2}\sqrt{\frac{k}{d}}$. Substituting this inequality into Expression (2.6) and noting that $m < 1$, we obtain

$$\text{Prob}\left[|\|Qx\|_2 - m| > \epsilon\right] \leq 4e^{-\epsilon^2 k/8}.$$

As a result,

Theorem 6 ([28])**.** Let $Q \in \mathbb{R}^{d \times d}$ be defined as in Equation (2.5). Let $M > 0$ be the Levy median of $\sqrt{d}\|Qx\|_2$. Let U be a random orthonormal matrix. Let $\mathcal{D}$ be a distribution on $d \times d$ matrices such that for $A \sim \mathcal{D}$,

$$A = \sqrt{\frac{d}{M}} QU.$$

Then, $\mathcal{D}$ is a JL distribution for $k > 8\epsilon^{-2}\log\frac{4}{\delta}$.

2.2.2 Frankl and Maehara (1988)

Frankl and Maehara [20] were the first to formally state the results given by Johnson and Lindenstrauss. In addition to formally stating it, Frankl and Maehara improved it by providing an explicit lower bound of the projected dimension.

Lemma 2 ([20])**.** For $0 < \epsilon < \frac{1}{2}$ and $n \in \mathbb{Z}^+$, let $k(n,\epsilon) = \left\lceil 9\left(\epsilon^2 - \frac{2\epsilon^3}{3}\right)^{-1}\log n\right\rceil + 1$. If $n > k(n,\epsilon)^2$, then for any n-point set S in $\mathbb{R}^d$, $d \geq n$, there exists a function $f : S \to \mathbb{R}^{k(n,\epsilon)}$ such that

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

for all $u, v \in S$.

To determine the parameter $k(n,\epsilon)$, they constructed a distribution similar to that given in [28] and showed it is a JL distribution.

Distribution: Let B be chosen uniform at random from the Stiefel manifold $V_k(\mathbb{R}^d) = \{B \in \mathbb{R}^{k\times d} : B^\top B = I_d\}$. Then, let $\mathcal{D}$ be a distribution on $k \times d$ matrices such that for $A \sim \mathcal{D}$,

$$A = \sqrt{\frac{d}{k}}B.$$

To prove $\mathcal{D}$ is a JL distribution, Frankl and Maehara considered the k-dimensional space, on which x is to be projected, to be fixed and the vector $x \in S^{d-1}$ to be uniform random on S^{d-1}. With this perspective, they bounded the surface area of S^{d-1} such that inequality

$$\left|\|Ax\|_2^2 - \frac{k}{d}\right| \geq \frac{k}{d}\epsilon$$

holds. To bound the surface area, they considered θ to be the angle between x and the space of projection, i.e., $\|Ax\|_2^2 = \cos^2\theta$, and found the surface area when

$$\theta \leq \arccos\sqrt{(1+\epsilon)\frac{k}{d}} \quad \text{and} \quad \theta \geq \arccos\sqrt{(1-\epsilon)\frac{k}{d}}.$$

With the surface area, they give an upper bound to Prob $\left[|X - \frac{k}{d}| > \epsilon \cdot \frac{k}{d}\right]$:

$$\text{Prob}\left[|X - \frac{k}{d}| > \epsilon \cdot \frac{k}{d}\right] < 2\sqrt{k}\exp\left(-(k-1)\left(\frac{\epsilon^2}{4} - \frac{\epsilon^3}{6}\right)\right),$$

and as a result, provide the parameter

$$k(n,\epsilon) = \left\lceil 9\left(\epsilon^2 - \frac{2\epsilon^3}{3}\right)^{-1}\log n\right\rceil + 1.$$

2.2.3 Indyk and Motwani (1998)

Indyk and Motwani [26] simplified the construction of a JL distribution by making two observations. First, rather than requiring the rows to be normal, a sufficient condition is for the Euclidean norm of each of the rows of the transformation to have expected value one. Second, it is sufficient for the projected space of k rotationally invariant vectors to be random rather than random and orthogonal. Observe that random vectors become more orthogonal as d approaches infinity. Exploiting these observations, they constructed the following distribution:

Distribution: Let $\mathcal{D}$ be a distribution on $k \times d$ matrices such that, for $A \sim \mathcal{D}$, $R_{ij} \overset{iid}{\sim} N(0,1)$ for $1 \leq i \leq k$ and $1 \leq j \leq d$, and $A = \frac{1}{\sqrt{k}}R$.

One motivation for such a construction is it is easier to generate than the previous constructions and in addition, due to the 2-stability of the normal distribution, we have a concentration of measure of $\|Ax\|_2^2$, $x \in S^{d-1}$. In particular, $\|Ax\|_2^2 \sim \frac{1}{k}\chi^2(k)$, where $\chi^2(k)$ is the chi-square distribution with k degrees of freedom. To show $\mathcal{D}$ is a JL distribution, we turn to the lemma provided by Achlioptas in [1].

Lemma 3 (Lemma 4.1, [1])**.** For any $\epsilon > 0$ and any unit vector $x \in \mathbb{R}^d$,

$$\text{Prob}_{A\sim D}\left[\|Ax\|^2 > 1+\epsilon\right] < \exp\left(-\frac{k}{2}\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right)\right) \quad \text{and}$$
$$\text{Prob}_{A\sim D}\left[\|Ax\|^2 < 1-\epsilon\right] < \exp\left(-\frac{k}{2}\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right)\right).$$

Combining these two probabilities, we obtain

$$\mathrm{Prob}_{A\sim D}\left[\|Ax\|^2 < 1-\epsilon \text{ or } \|Ax\|^2 > 1+\epsilon\right] < 2\exp\left(-\frac{k}{2}\left(\frac{\epsilon^2}{2}-\frac{\epsilon^3}{3}\right)\right). \qquad (2.7)$$

When $k > 2\log\frac{2}{\delta}\left(\frac{\epsilon^2}{2}-\frac{\epsilon^3}{3}\right)^{-1}$, the right hand side of Inequality (2.7) is bounded above by δ. Hence, $\mathcal{D}$ is a JL distribution for $k > 2\log\frac{2}{\delta}\left(\frac{\epsilon^2}{2}-\frac{\epsilon^3}{3}\right)^{-1}$.

2.2.4 Dasgupta and Gupta (2003)

Dasgupta and Gupta [13] further simplified the proof of Theorem 3 by making the following two observations.

1. The Euclidean norm of a unit vector projected onto a random k-dimensional subspace has the same distribution as a random unit vector projected onto a fixed k-dimensional subspace.

2. Choosing a point uniformly at random from the $(d-1)$-dimensional unit sphere S^{d-1} is equivalent to choosing each entry of a point uniformly at random from the distribution $N(0,1)$ and normalizing it.

Hence to prove Theorem 3, let the k-dimensional projected subspace be fixed and let $x \in \mathbb{R}^d$ be random such that $x_i \overset{iid}{\sim} N(0,1)$. Then,

$$E(e^{sx_i^2}) = \frac{1}{\sqrt{1-2s}} \qquad (2.8)$$

for $-\infty < s < \frac{1}{2}$ and $1 \leq i \leq d$. Combining (2.8) with Markov's inequality, the sum of the first k entries squared, $\sum_{i=1}^k x_i^2$, is shown in [13] to be concentrated around its expectation k. In particular, let $L = \|Z\|^2$ where Z is the projection of x onto its first k coordinates.

Then by Lemma 2.2 in [13],

$$\begin{aligned}\text{Prob}\left[L \leq \frac{k}{d}(1-\epsilon)\right] &\leq \exp\left(\frac{k}{2}(\epsilon + \log(1-\epsilon)\right) \text{ and} \qquad (2.9)\\ \text{Prob}\left[L \geq \frac{k}{d}(1+\epsilon)\right] &\leq \exp\left(\frac{k}{2}(-\epsilon + \log(1+\epsilon)\right). \qquad (2.10)\end{aligned}$$

Using inequalities $\log(1-x) \leq -x - \frac{x^2}{2}$ and $\log(1+x) \leq x - \frac{x^2}{2}$ derived from Taylor's expansion, expressions (2.9) and (2.10) simplify to

$$\begin{aligned}\text{Prob}\left[L \leq \frac{k}{d}(1-\epsilon)\right] &\leq \exp\left(-\frac{k\epsilon^2}{4}\right) \text{ and}\\ \text{Prob}\left[L \geq \frac{k}{d}(1+\epsilon)\right] &\leq \exp\left(-\frac{k}{2}\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right)\right).\end{aligned}$$

Combining these two probabilities, we obtain

$$\text{Prob}\left[|L - \frac{k}{d}| > \epsilon \cdot \frac{k}{d}\right] \leq 2\exp\left(-\frac{k}{2}\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right)\right). \qquad (2.11)$$

The right hand side of (2.11) is bounded above by δ when

$$k \geq 2\log\left(\frac{2}{\delta}\right)\left[\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right]^{-1} = 4\epsilon^{-2}\log\frac{1}{\delta}\left[1 + \frac{\epsilon}{3-\epsilon}\left(1 + \frac{\log 2}{\log\frac{1}{\delta}}\right) + \frac{\log 2}{\log\frac{1}{\delta}}\right].$$

As a result, the distribution provided in Section 2.2.1 is a JL distribution for $k \geq 2\log\left(\frac{2}{\delta}\right)\left[\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right]^{-1}$.

2.2.5 Achlioptas (2003)

Achlioptas [1] constructed a JL distribution that is both easy to generate and is relatively quick to compute. He deviated from previous constructions by observing that a JL transformation does not need to be rotationally invariant, but rather it is sufficient for the projection of a unit vector to be tightly concentrated about $\frac{1}{d}$. This observation is derived from the central limit theorem, which states that one obtains a "good estimate of the original length" with a sufficient number of unbiased estimators with bounded variance.

Hence, one just needs unbiased estimators with bounded variance. Achlioptas pointed out for any independent set of variables $\{a_{ij}\}$, where $E(a_{ij}) = 0$ and $\text{Var}(a_{ij}) = 1$, we obtain such estimators. That is,

$$E(\|Ax\|_2^2) = \|x\|_2^2,$$

where $E(a_{ij}) = 0$ and $\text{Var}(a_{ij}) = 1$. Exploiting these observations, Achlioptas provided two constructions of JL distributions, both of which are easy to generate and the second of which gives a three-fold speed up for computing Ax.

Construction 1:

In his first construction, Achlioptas lets each entry of the transform be chosen uniformly at random from $\{\pm 1/\sqrt{k}\}$.

Distribution: Let $\mathcal{D}_1$ be a distribution on $k \times d$ matrices such that, for $A \sim \mathcal{D}_1$, $a_{ij} \overset{iid}{\sim} \left\{\pm\frac{1}{\sqrt{k}}\right\}$ for $1 \leq i \leq k$ and $1 \leq j \leq d$.

In [1], it was shown for $\epsilon, \beta > 0$, if $k \geq \frac{2+\beta}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}} \log\frac{1}{\delta}$, then

$$\text{Prob}\left[(1-\epsilon)\|x\|^2 \leq \|Ax\|^2 \leq (1+\epsilon)\|x\|^2\right] \geq 1 - \delta^{\beta/2}.$$

For sake of comparison, we let $\beta = 2$ and find for $k \geq 8\epsilon^{-2}\log\frac{1}{\delta}\left[1 + \frac{2\epsilon}{3-2\epsilon}\right]$,

$$\text{Prob}\left[(1-\epsilon)\|x\|^2 \leq \|Ax\|^2 \leq (1+\epsilon)\|x\|^2\right] \geq 1 - \delta.$$

Construction 2:

Unlike his first construction, the second construction is somewhat sparse and as a result, gives a three-fold speed up in the projection of x.

Distribution: Let $\mathcal{D}_2$ be a distribution on $k \times d$ matrices such that, for $A \sim \mathcal{D}_2$,

$$a_{ij} = \sqrt{\frac{3}{k}} \cdot \begin{cases} +1 & \text{with probability } \frac{1}{6} \\ 0 & \text{with probability } \frac{2}{3} \\ -1 & \text{with probability } \frac{1}{6} \end{cases}$$

for $1 \leq i \leq k$ and $1 \leq j \leq d$.

Then for $\epsilon, \beta > 0$,

$$\text{Prob}\left[(1-\epsilon)\|x\|^2 \leq \|Ax\|^2 \leq (1+\epsilon)\|x\|^2\right] \geq 1 - \delta^{\beta/2},$$

if $k \geq \frac{2+\beta}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}} \log \frac{1}{\delta}$. Again, we are interested when $\beta = 2$ and find for
$k \geq 8\epsilon^{-2} \log \frac{1}{\delta}\left[1 + \frac{2\epsilon}{3-2\epsilon}\right]$,

$$\text{Prob}\left[(1-\epsilon)\|x\|^2 \leq \|Ax\|^2 \leq (1+\epsilon)\|x\|^2\right] \geq 1 - \delta.$$

To prove both constructions are JL distributions, Achlioptas applies Markov's inequality to moment generating functions. We will not go into the details of the proof, but rather we will present a sketch of a proof provided by Matousek in [36] that envelops the constructions of both Achlioptas and Indyk and Motwani.

Theorem 7 (Theorem 3.1, [36])**.** Let d be an integer, $0 < \epsilon \leq \frac{1}{2}$ and $0 < \delta < 1$. Let $k = C\epsilon^{-2} \log \frac{2}{\delta}$ for some constant $C > 0$. Let $T : \mathbb{R}^d \to \mathbb{R}^k$ be a random linear map defined by

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

where R_{ij} are independent random variables with $E(R_{ij}) = 0$, $\text{Var}(R_{ij}) = 1$, and a uniform subgaussian tail. That is, there exists a constant $a > 0$ such that for all $\lambda > 0$,

$$\text{Prob}\left[X > \lambda\right] \leq e^{-a\lambda^2}$$

for each R_{ij}. Then for every $x \in \mathbb{R}^d$,

$$\text{Prob}\left[(1-\epsilon)\|x\| \leq \|T(x)\| \leq (1+\epsilon)\|x\|\right] \geq 1 - \delta.$$

To prove Theorem 7, Matousek implements the following proposition.

Proposition 1. Let $k \in \mathbb{Z}_{\geq 1}$. Let $Y_1, \ldots, Y_k$ be independent random variables with $E(Y_i) = 0$, $\mathrm{Var}(Y_i) = 1$ and uniform subgaussian tail, that is

$$\mathrm{Prob}\left[Y_i < \lambda\right] \leq e^{-a\lambda^2}$$

for same a. Then

$$Z = \frac{1}{\sqrt{k}}\left[\sum Y_i^2 - k\right]$$

has a subgaussian tail up to $\sqrt{k}$, that is

$$\mathrm{Prob}\left[Z > \lambda\right] \leq e^{-a\lambda^2}$$

for $0 < \lambda < \sqrt{k}$ and some constant $a > 0$.

Let $Y_i = \sum_{j=1}^{d} R_{ij}x_j = \sqrt{k}T(x)_i$. By construction, $E(R_{ij}) = 0$ and $\mathrm{Var}(R_{ij}) = 1$, and each x_i has subgaussian tail for $1 \leq i \leq d$. It follows that $E(Y_i) = 0$, $\mathrm{Var}(Y_i) = 1$, and each Y_i has a subgaussian tail for $1 \leq i \leq d$. Then by Proposition 1, for $Z = \frac{1}{\sqrt{k}}(Y_1^2 + \cdots + Y_k^2 - k)$

$$\mathrm{Prob}\left[Z > \lambda\right] \leq e^{-a\lambda^2} \tag{2.12}$$

for $0 < \lambda \leq \sqrt{k}$ and some constant $a > 0$. We observe that

$$\mathrm{Prob}\left[\|T(x)\| \geq 1 + \epsilon\right] \leq \mathrm{Prob}\left[\|T(x)\|^2 \geq 1 + 2\epsilon\right] = \mathrm{Prob}\left[Z \geq 2\epsilon\sqrt{k}\right].$$

By Inequality (2.12), the above is bounded in the following way

$$\mathrm{Prob}\left[\|T(x)\| \geq 1 + \epsilon\right] \leq e^{-4a\epsilon^2 k}. \tag{2.13}$$

We note $e^{-4a\epsilon^2 k} < \delta$ when $k > \frac{1}{4a}\epsilon^{-2}\log\frac{2}{\delta}$, where a depends on (2.13). Hence, T is a JL transformation when $k > \frac{1}{4a}\epsilon^{-2}\log\frac{2}{\delta}$.

2.3 Fast Fourier Transform-Based Constructions

To quicken the computation, one may turn to sparsifying the projection as done by Achlioptas. However, a problem arises when the vector to be projected is itself sparse. When this is the case, the projected norm may be greatly distorted by sparse projections. For instance, when $x = (0, \ldots, 0, 1)$, x will be projected to the zero vector under the projection which takes the first k entries of x. To navigate around this problem, the vector is preconditioned. Preconditioning of a vector consists of increasing the support of x, i.e., it increases the number of nonzero entries with high probability. Observe that for $x \in S^{d-1}$,

$$\frac{1}{\sqrt{d}} \leq \|x\|_\infty \leq 1.$$

A sparse vector has ℓ_∞ norm closer to 1 while a dense vector has ℓ_∞ norm closer to the lower bound $\frac{1}{\sqrt{d}}$.

2.3.1 Ailon and Chazelle (2006)

To ensure $\|x\|_\infty$ lies close to the lower bound $\frac{1}{\sqrt{d}}$ for $x \in S^{d-1}$, Ailon and Chazelle [2] implemented randomized Fourier transforms before applying a sparse projection. To avoid sparsifying an already dense vector, the Fourier transform is randomized. Let H be a $d \times d$ normalized Walsh-Hadamard matrix as defined in (2.4). To randomize the transform, H is multiplied by a random $d \times d$ diagonal matrix D consisting of diagonal entries chosen uniformly and at random from the set $\{\pm 1\}$. In [2], Ailon and Chazelle showed for a set X of n vectors,

$$\max_{x \in X} \|HDx\|_\infty \leq \sqrt{\frac{\log n}{d}}$$

with high probability. Observe as HD is an orthogonal transformation, it is Euclidean norm invariant. That is, $\|HDx\|_2 = \|x\|_2$. Let P be a $k \times d$ matrix such that

$$p_{ij} = \begin{cases} p_{ij} \overset{iid}{\sim} N\left(0, \frac{1}{q}\right) & \text{with probability } q \\ 0 & \text{with probability } 1-q \end{cases}, \tag{2.14}$$

where $q = \min\left\{\Theta\left(\frac{\log^2 n}{d}\right), 1\right\}$.

Distribution: Let H be a $d \times d$ Walsh-Hadamard matrix. Let D be a random diagonal matrix as described above. Let $P \in \mathbb{R}^{k \times d}$ be random as defined in (2.14). Then, let $\mathcal{D}$ be a distribution on $k \times d$ matrices such that, for $A \sim \mathcal{D}$,

$$A = \frac{1}{\sqrt{k}} PHD.$$

Theorem 8. Let distribution $\mathcal{D}$ be as described. Let $0 < \epsilon < 1$. Then, for all $x \in \mathbb{R}^d$,

$$\text{Prob}_{A \sim \mathcal{D}}\left[(1-\epsilon)\|x\|_2 \leq \|Ax\|_2 \leq (1+\epsilon)\|x\|_2\right] \geq \frac{2}{3} = 1 - \frac{1}{3}.$$

Choosing a random transformation from distribution $\mathcal{D}$ $\log \frac{1}{\delta}$ times and picking the best one will decrease the probability of failure to a desired constant δ.

The proof, given in [2], to show $\mathcal{D}$ is a JL distribution employs Markov's inequality and the bounding of moments.

2.3.2 Matousek (2008)

Matousek [36] improved Ailon and Chazelle's construction by simplifying the nonzero entries of the sparse matrix, and as a result, made it easier to generate the sparse projection. In addition, he showed the number of nonzero entries per column of the transform must be at least $\Omega\left(\frac{\alpha^2}{\epsilon^2}\right)$, where $\|x\|_\infty \leq \alpha$.

Distribution: Let H be a $d \times d$ Walsh-Hadamard matrix as defined in (2.4). Let D be a random diagonal matrix with diagonal entries chosen uniformly and at random from the set

$\{\pm 1\}$. Let P be a $k \times d$ random matrix such that

$$p_{ij} = \begin{cases} 1 & \text{with probability } \frac{q}{2} \\ 0 & \text{with probability } 1-q \\ -1 & \text{with probability } \frac{q}{2} \end{cases},$$

where $q = \Omega\left(\alpha^2 \log \frac{d}{\epsilon\delta}\right)$. Then, let $\mathcal{D}$ be a distribution on $k \times d$ matrices such that, for $A \sim \mathcal{D}$,

$$A = \frac{1}{\sqrt{qk}} PHD$$

Theorem 9. Let distribution $\mathcal{D}$ be as described. Let $0 < \epsilon < 1$. Then, if $k = \Omega\left(\epsilon^{-2} \log \frac{4}{\delta}\right)$, for $x \in \mathbb{R}^d$,

$$\text{Prob}_{A \sim \mathcal{D}}\left[(1-\epsilon)\|x\|_2 \le \|Ax\|_2 \le (1+\epsilon)\|x\|_2\right] \ge 1-\delta.$$

The proof in [36] implements the subgaussian tail and is similar to his proof for the constructions of Achlioptas and Indyk and Motwani, see Section 2.2.5.

2.3.3 Dasgupta, Kumar, and Sarlós (2010)

To allow for a sparser matrix than that provided by the previous two constructions, Dasgupta, Kumar, and Sarlós [12] placed dependence among the entries of the sparse matrix P. The dependence is a direct consequence of the use of hash functions to determine the location of the nonzero entries. They present two different constructions.

Construction 1:

Let $0 < \epsilon, \delta < \frac{1}{2}$. Let $c = \frac{16}{\epsilon} \log \frac{1}{\delta} \log^2 \frac{k}{\delta}$. Let $G \in \mathbb{R}^{cd \times d}$ be a fixed matrix such that

$$g_{ij} = \begin{cases} \frac{1}{\sqrt{c}} & (j-1)c + 1 \le i \le jc \\ 0 & \text{otherwise} \end{cases}$$

for $1 \leq i \leq cd$ and $1 \leq j \leq d$. For example, if $c = 2$,

$$G = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \cdots 0 & \\ \frac{1}{\sqrt{2}} & 0 & \cdots 0 & \\ 0 & \frac{1}{\sqrt{2}} & \cdots & 0 \\ 0 & \frac{1}{\sqrt{2}} & \cdots 0 & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Let $\mathcal{H}$ be the set of all hash functions from $[cd]$ to $[k]$, where $[n]$ denotes the set $\{1, \ldots, n\}$. Let $h \in \mathcal{H}$ be a random function for $1 \leq j \leq cd$. This implies for each $j \in [cd]$, $h(j)$ is uniform random in $[k]$. Let $r_j \overset{iid}{\sim} \{\pm 1\}$ for $1 \leq j \leq cd$. Let $M \in \mathbb{R}^{k \times cd}$ such that $m_{ij} = \delta(i, h(j))r_j$, where $\delta(i, j) = 1$ if $i = j$ and 0 otherwise.

Distribution: Let $G \in \mathbb{R}^{cd \times d}$ be a fixed matrix as described. Let $M \in \mathbb{R}^{k \times cd}$ be a random matrix as described. Then, let $\mathcal{D}_1$ be a distribution on $k \times d$ matrices such that, for $A \sim \mathcal{D}$, $A = MG$.

Theorem 10 ([12])**.** Let $\mathcal{D}_1$ be the distribution described. If $k \geq \frac{12}{\epsilon^2} \log \frac{1}{\delta}$, then for all $x \in \mathbb{R}^d$,

$$\text{Prob}_{A \sim D_1} \left[(1 - \epsilon)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \epsilon)\|x\|_2^2 \right] \geq 1 - 4\delta.$$

Projection Ax takes time $\mathcal{O}\left(\frac{1}{\epsilon} \log^2 \frac{k}{\delta} \log \frac{1}{\delta}\right) \cdot \|x\|_0$.

Construction 2: Let $0 < \delta < 1$ and $c = \frac{16}{\epsilon} \log \frac{1}{\delta} \log^2 \frac{k}{\delta} \geq 1$. Let $b := 6c \log \frac{3c}{\delta}$ and assume $b \leq d$. Let H be a $b \times b$ Walsh-Hadamard matrix as defined in (2.4). Let $G \in \mathbb{R}^{d \times d}$ be a random block diagonal matrix, where each of the $\frac{d}{b}$ diagonal blocks of G is a $b \times b$ randomized Hadamard matrix HD, where the matrices D are independent random diagonal matrices such that the diagonal entries are chosen uniformly at random from the set $\{\pm 1\}$. Let $\mathcal{H}$ be the set of all hash functions from $[d]$ to $[k]$. Let $h \in \mathcal{H}$ be a random function for $1 \leq j \leq d$

as in Construction 1. Let $r_{ij} \overset{iid}{\sim} \{\pm 1\}$. Let $P \in \mathbb{R}^{k \times d}$ be a random matrix such that

$$p_{ij} = \delta(i, h(j)) r_{ij}.$$

Note that P has one nonzero entry per column.

Distribution: Let $\mathcal{D}_2$ be a distribution on $k \times d$ matrices such that, for $A \sim \mathcal{D}$, $A = PG$.

To prove the latter construction is a JL distribution, the following two statements must hold.

1. Let $\epsilon < 1$ and $0 < \delta < \frac{1}{10}$. Then, for $x \in \mathbb{R}^d$ such that $\|x\|_\infty \leq \frac{1}{\sqrt{c}}$,

$$\text{Prob}\left[(1-\epsilon)\|x\|_2^2 \leq \|Px\|_2^2 \leq (1+\epsilon)\|x\|_2^2\right] \geq 1 - 3\delta.$$

2. The largest entry of Gx is at most $\frac{1}{\sqrt{c}}$, i.e., $\|Gx\|_\infty \leq \frac{1}{\sqrt{c}}$.

We will focus on step 2 and outline the proof as given in [12]. For the proof of step 1, see [12]. Ailon and Chazelle in [2] showed for $\|x\|_2 \leq 1$ and $\|Ax\|_2 = 1$,

$$\text{Prob}\left[\|Ax\|_\infty \geq s\right] \leq 2b \exp\left(-\frac{s^2 b}{2}\right). \tag{2.15}$$

Observe each $b \times b$ block of G, which will be denoted G_j for $1 \leq j \leq \frac{d}{b}$, is norm invariant. Let x_j denote the block of entries of x associated with block G_j. That is,

$$Gx = (G_1 x_1, \ldots, G_{d/b} x_{d/b})^\top.$$

Then, $\|G_j x_j\|_2 = \|x_j\|_2$ and since $\|x_j\|_2 \leq \|x\|_2 \leq 1$,

$$\text{Prob}\left[\|G_j x_j\|_\infty \geq s\right] \leq 2b \exp\left(-\frac{s^2 b}{2}\right)$$

for $1 \leq j \leq \frac{d}{b}$. To show $\|Gx\|_\infty \leq \frac{1}{\sqrt{c}}$, we consider two cases.

1. Suppose $\|x_j\|_2 \leq \frac{1}{\sqrt{c}}$ for $1 \leq j \leq \frac{d}{b}$. Then $\|G_j x_j\|_2 \leq \frac{1}{\sqrt{c}}$ for $1 \leq j \leq \frac{d}{b}$, and as a

result,

$$\text{Prob}\left[\|Gx\|_\infty \geq \frac{1}{\sqrt{c}}\right] = \text{Prob}\left[\max_j \|G_j x_j\|_\infty \geq \frac{1}{\sqrt{c}}\right] \leq \delta.$$

2. Suppose $\|x_j\|_2 \geq \frac{1}{\sqrt{c}}$ for some j. By substituting $\frac{1}{\sqrt{c}}$ for s, Expression (2.15) becomes

$$\text{Prob}\left[\|G_j x_j\|_\infty \geq \frac{1}{\sqrt{c}}\right] \leq 2b \exp\left(-\frac{b}{2c}\right) \quad \text{for } 1 \leq j \leq \frac{d}{b}.$$

Taking the union bound up to c blocks, we obtain

$$\begin{aligned}\text{Prob}\left[\|Gx\|_\infty \geq \frac{1}{\sqrt{c}}\right] &\leq c \cdot \text{Prob}\left[\|G_j x_j\|_\infty \geq \frac{1}{\sqrt{c}}\right] \\ &\leq 2bc \exp\left(-\frac{b}{2c}\right) \\ &= \frac{4}{9} \log \frac{3c}{\delta} \delta^3 < \delta.\end{aligned}$$

Theorem 11. Let $\mathcal{D}_2$ be the distribution described. Then, for any $x \in \mathbb{R}^d$,

$$ppP(1-\epsilon)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\epsilon)\|x\|_2^2 A \sim \mathcal{D}_2 \leq 1 - 4\delta.$$

The time required to compute Ax is

$$\mathcal{O}\left(\min\left(\frac{\|x\|_0}{\epsilon} \log^4\left(\frac{1}{\epsilon\delta}\right), d\right) \cdot \log\left(\frac{1}{\epsilon\delta}\right)\right).$$

2.3.4 Liberty, Ailon, and Singer (2008)

Liberty, Ailon, and Singer [33] sought to combine the steps of preconditioning and projection, and accomplished this for a subset of vectors in $\mathbb{R}^d$ with their Lean-Walsh construction. A transform from their distribution consists of a Lean-Walsh transform and a diagonal matrix whose entries are chosen uniformly at random from $\{\pm 1\}$.

A Lean-Walsh transform is constructed from a seed matrix, which will be denoted B. Let $B \in \mathbb{C}^{r \times c}$ such that $r < c$, $|B_{ij}| = \frac{1}{\sqrt{r}}$, and the rows are orthogonal. For instance, let B be a sub-Hadamard matrix or a sub-Fourier matrix:

$$\underbrace{B = \frac{1}{\sqrt{3}}\begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}}_{\text{sub-Hadamard}} \quad \text{or} \quad \underbrace{B = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 & 1 \\ 1 & e^{2\pi i/3} & e^{4\pi i/3} \end{pmatrix}}_{\text{sub-Fourier}}.$$

Matrix B is called a Lean-Walsh seed and the Lean-Walsh transform A_ℓ is constructed by recursively taking the Kronecker product of B and the previous output. That is,

$$A_\ell = B \otimes A_{\ell-1} = \begin{bmatrix} b_{11}A_{\ell-1} & \cdots & b_{1c}A_{\ell-1} \\ \vdots & & \vdots \\ b_{r1}A_{\ell-1} & \cdots & b_{rc}A_{\ell-1} \end{bmatrix},$$

where $A_1 = B$. The parameter ℓ is chosen according to the desired dimension.

Example 2. Let $B = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{pmatrix}$. Then, $A_1 = B$ and

$$A_2 = B \otimes B = \frac{1}{\sqrt{2}}\begin{pmatrix} B & B & -B & -B \\ B & -B & B & -B \end{pmatrix}.$$

If we choose $k = 4$, then we stop at $\ell = 2$.

Distribution: Let $A_\ell \in \mathbb{R}^{k \times d}$ be a fixed Lean-Walsh transform. Let D be a diagonal matrix with entries $d_{ii} \overset{iid}{\sim} \{\pm 1\}$. Then, let $\mathcal{D}$ be a distribution on $k \times d$ matrices such that for $A \sim \mathcal{D}$, $A = A_\ell D$.

Theorem 12. For a distribution $\mathcal{D}$ constructed from a Lean-Walsh seed $B \in \mathbb{R}^{r \times c}$ as described, $\mathcal{D}$ is a JL distribution for

$$\{x \in S^{d-1} : \|x\|_2 \leq k^{-1/2} d^{\frac{1-\alpha}{4}}\},$$

where $\alpha = \frac{\log r}{\log c}$.

2.4 Code-Based Constructions

2.4.1 Ailon and Liberty (2009)

Ailon and Liberty [3] abandoned the sparse matrix of Matousek's construction and replaced it with a dense matrix based on Rademacher variables and BCH codes. Let $B_k \in \mathbb{R}^{k \times f_{BCH}(k)}$ be a 4-wise independent code matrix, where $f_{BCH}(k) = \Theta(k^2)$. That is, every row of B_k is a row of the Walsh-Hadamard matrix, defined in (2.4), scaled by constant $\sqrt{\frac{f_{BCH}(k)}{k}}$. The dense matrix B is a result of concatenating multiple copies of the 4-wise independent code matrix B_k:

$$B = [B_k, \ldots, B_k].$$

Distribution: Let $B \in \mathbb{R}^{k \times d}$ be a fixed matrix as described. Let H be the Walsh-Hadamard matrix as defined in (2.4). Let D and $D^{(i)}$ for $1 \leq i \leq r = \lceil \frac{1}{2\delta} \rceil$ be independent random diagonal matrices such that the diagonal entries are chosen uniformly at random from the set $\{\pm 1\}$. Then, let $\mathcal{D}$ be a distribution on $k \times d$ matrices such that, for $A \sim \mathcal{D}$,

$$A = BD\Phi_d^{(r)},$$

where $\Phi_d^{(r)} = HD^{(r)} \cdots HD^{(1)}$.

Theorem 13 ([3])**.** Let distribution $\mathcal{D}$ be as described. Let $0 < \epsilon < 1$ and $0 < \delta < 1$. If $\frac{1}{c_2}\epsilon^{-2} \log \frac{c_1}{\delta} \leq k \leq d^{1/2-\delta}$, where c_1 and c_2 are some global constants, then

$$\text{Prob}_{A \sim \mathcal{D}} \left[(1-\epsilon)\|x\|_2 \leq \|Ax\|_2 \leq (1+\epsilon)\|x\|_2\right] \geq 1 - \delta,$$

and it takes $\mathcal{O}(d \log k)$ to compute Ax for $A \sim D$, and $\mathcal{O}(d)$ random bits to construct A.

To exhibit an instance of a 4-wise independent code matrix B_k, we briefly introduce BCH codes and their parity check matrices. We will not discuss the topic of BCH codes in depth, but direct those interested to [34].

Let $GF(2)$ be the finite field of two elements, e.g., $\{0, 1\}$. The field $GF(2^m)$ containing 2^m elements can be constructed by considering an irreducible polynomial $f(x)$ of degree m over the field $GF(2)$. Let α be the root of polynomial $f(x)$. Then, the field $GF(2^m)$ can be defined as

$$GF(2^m) = \{0, \alpha, \alpha^2, \ldots, \alpha^{2^m-1}\}.$$

Example 3. Let $m = 2$. Then, $f(x) = x^2 + x + 1$ is irreducible over $GF(2)$. Let α be a root of $f(x)$. That is, $\alpha^2 + \alpha + 1 = 0$. The field $GF(2^2)$ can then be defined as

$$GF(2^2) = \{0, \alpha^0 = 1, \alpha, \alpha^2 = \alpha + 1\}.$$

A binary BCH code C of length $n = 2^m - 1$ is a linear subspace of $GF(2)^n$. The elements of C can be defined by the kernel of the parity check matrix

$$W = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n-1} \\ 1 & \alpha^2 & (\alpha^2)^2 & \cdots & (\alpha^{n-1})^2 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \alpha^{2t} & (\alpha^2)^{2t} & \cdots & (\alpha^{n-1})^{2t} \end{bmatrix},$$

where $t < 2^{m-1}$ is an integer. That is, $C = \{c \in GF(2)^n | Wc = 0\}$, and $c = (c_0, c_1, \ldots, c_{n-1})$ is a codeword if and only if the polynomial

$$g(x) = c_0 + c_1 x + \cdots + c_{n-1} x^{n-1}$$

has $\alpha, \ldots, \alpha^{2t}$ as roots. Consequently, by design of matrix W, $2t$ of its columns are linearly independent giving us a $2t$-wise independent matrix.

Example 4. Let $m = 4$. Then, an irreducible polynomial over $GF(2)$ is

$$f(x) = x^4 + x + 1.$$

Let α be a root of $f(x)$. Then,

$$GF(2^4) = \{0, \alpha, \alpha^2, \ldots, \alpha^{15} = 1\}.$$

A parity check matrix W of a BCH code C with length $n = 2^4 - 1 = 15$ is

$$W = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{14} \\ 1 & \alpha^2 & (\alpha^2)^2 & \cdots & (\alpha^{14})^2 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \alpha^{2t} & (\alpha^2)^{2t} & \cdots & (\alpha^{14})^{2t} \end{bmatrix},$$

where $t < 8$. Observe that matrix W is a $2t$-wise independent matrix, and hence, we desire $t \geq 2$.

2.4.2 Kane and Nelson (2014)

Kane and Nelson in [29] showed the DKS construction in Section 2.3.3 requires at least $\Omega\left(\epsilon^{-1}\log^2\frac{1}{\delta}\right)$ nonzero entries per column in order for the distribution to be a JL distribution. To achieve a smaller sparsity, Kane and Nelson restricted the set $\mathcal{H}$ of hash functions from which to choose. In the DKS construction, the location of the nonzero entries are determined by a hash function chosen at random from the set of all hash functions. In [29], Kane and Nelson present two means to determine the locations: a linear code or hash functions chosen at random from a set $\mathcal{H}$ of hash functions without replacement. They provide two constructions in which both methods can be implemented. For both constructions, a transform has entries of the form

$$A_{ij} = \frac{1}{\sqrt{s}}\eta_{ij}\sigma_{ij},$$

where $\sigma_{ij} \overset{iid}{\sim} \{\pm 1\}$ and η_{ij} is an indicator function dependent on a code or a set of s hash functions. In both cases, each column contains exactly s nonzero entries.

Construction 1 - Graph Construction:

In the graph construction, there are no restrictions placed on the location of the s nonzero entries in each column. In this case, the indicator function η_{ij} can be represented by a bipartite graph G with d left vertices, k right vertices, and left degree s. A bipartite graph is a graph with vertices which can be separated into two disjoint sets, left and right, such that there is no edge within either set. We say G has left degree s if each left vertex has exactly s edges.

Example 5. Let $d = 6$, $k = 4$, and $s = 2$. Consider the following bipartite graph G

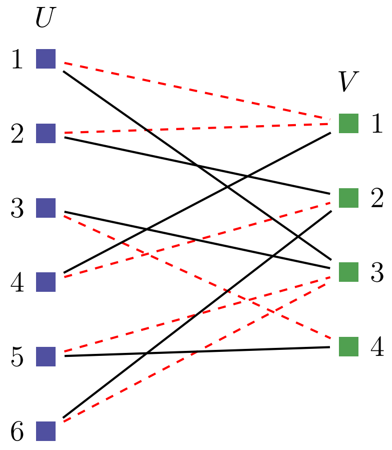


Figure 2.1: The bipartite graph G represents the graph construction of Kane and Nelson. Each vertex is hashed s times to the lower dimension.

Then,

$$A = \frac{1}{\sqrt{2}} \begin{bmatrix} \sigma_{11} & \sigma_{12} & 0 & \sigma_{14} & 0 & 0 \\ 0 & \sigma_{22} & 0 & \sigma_{24} & 0 & \sigma_{26} \\ \sigma_{31} & 0 & \sigma_{33} & 0 & \sigma_{35} & \sigma_{36} \\ 0 & 0 & \sigma_{43} & 0 & \sigma_{45} & 0 \end{bmatrix},$$

where $\sigma_{ij} \overset{iid}{\sim} \{\pm 1\}$.

The indicator η_{ij} can be determined by either a code or a set of s hash functions. For the graph construction, a weight s binary code with minimum distance $d = 2s - \mathcal{O}\left(\frac{s^2}{k}\right)$

may be used or the set of hash functions $\mathcal{H}$ where the s hash functions are drawn without replacement.

We define the indicator function η_{ij} by a set of hash functions $h_1, \ldots, h_s$ independently chosen from $\mathcal{H}$ (without replacement) in the following manner:

$$\eta_{ij} = \begin{cases} 1 & h_t(j) = i \quad \text{for some } t \in \{1, \ldots, s\} \\ 0 & \text{otherwise} \end{cases}.$$

Example 5 (continued)**.** Graph G in Figure 2.1 can be defined by hash functions

$$\begin{aligned} h_1 : &\{1,2\} \mapsto 1 & \qquad h_2 : &\{4\} \mapsto 1 \\ &\{4\} \mapsto 2 & &\{2,6\} \mapsto 2 \\ &\{5,6\} \mapsto 3 & &\{1,3\} \mapsto 3 \\ &\{3\} \mapsto 4 & &\{5\} \mapsto 4 \end{aligned}$$

Construction 2 - Block Construction:

Differing from the graph construction, the block construction places restrictions on the location of the s nonzero entries in each column. Each column is split into blocks of length $\frac{k}{s}$, and one nonzero entry must be located in each block. The indicator function for the block construction can also be represented by a bipartite graph. For the block construction, we split the right vertices of the bipartite graph into s blocks:

$$\underbrace{v_1, \ldots, v_{k/s}}_{B_1}, \ldots, \underbrace{v_{k-k/s+1}, \ldots, v_k}_{B_s}.$$

Then, each of the d vertices on the left side of the bipartite graph must be adjacent to one vertex from each of the blocks as defined, rather than being adjacent to random s of the k vertices on the right. This creates a higher dependence among the indicators η_{ij}, since if $\eta_{ij} = 1$ for one of the entries on the block, η_{ij} must be zero for the rest of that block.

Example 6. Let $d = 6$, $k = 4$, and $s = 2$. Consider the following bipartite graph G

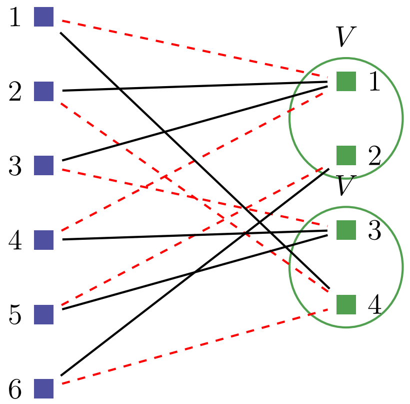


Figure 2.2: The bipartite graph G above represents the block construction of Kane and Nelson. Each vertex is hashed s times to the lower dimension. Each time the vertex is hashed, it must go to a different block.

Then,

$$A = \frac{1}{\sqrt{2}} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{25} & \sigma_{26} \\ 0 & 0 & \sigma_{33} & \sigma_{34} & \sigma_{35} & 0 \\ \sigma_{41} & \sigma_{42} & 0 & 0 & 0 & \sigma_{46} \end{bmatrix},$$

where $\sigma_{ij} \overset{iid}{\sim} \{\pm 1\}$.

For the block construction, an $[s, k, s - \mathcal{O}\left(\frac{s^2}{k}\right)]$ code in $\mathbb{F}_{k/s}$, field of k/s elements, may be used to determine the indicator function. The code construction is described in Section 2.4.2.1. Similar to the graph construction, an indicator function may also be defined by a set of s hash functions such that the entries are hashed $\mathcal{O}\left(\log \frac{1}{\delta}\right)$-wise independently. When the indicator is defined by the latter, sparsity $\Theta\left(\epsilon^{-1} \log \frac{1}{\delta}\right)$ is reached.

Theorem 14 ([29])**.** Let $\mathcal{D}$ be either the graph of block construction dependent on hashing functions as defined. Then, for $k = \Theta\left(\epsilon^{-2} \log \frac{1}{\delta}\right)$ and $s = \Theta\left(\epsilon k\right)$,

$$\text{Prob}_{A \sim \mathcal{D}} \left[(1-\epsilon)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\epsilon)\|x\|_2^2 \right] > \delta.$$

2.4.2.1 Block Construction via Codes:

We first define a code and its properties. We then explain the block construction via a code.

Definition 3. Let $\mathbb{F}_q$ be the field of q elements. For $x, y \in \mathbb{F}_q^n$, the *Hamming distance* is defined to be

$$d(x, y) := |\{i; x_i \neq y_i\}|$$

Definition 4. A $[d, m, d(C)]$ *code* C over $\mathbb{F}_q$ is a linear subspace of $\mathbb{F}_q^d$ with dimension m and minimum distance $d(C)$. The minimum distance $d(C)$ of a code C is defined as

$$d(C) := \min\{d(x, y) : x, y \in C \quad \text{and} \quad x \neq y\}.$$

The elements of C are called codewords.

Place an order on the elements of $\mathbb{F}_q$. Let $p : \mathbb{F}_q \to \mathbb{Z}$ be the order function defined by $p(a) = z$ where z is the location of a in the ordering of $\mathbb{F}_q$. Let $C \subseteq \mathbb{F}_q^s$ be an $[s, m, d(C)]$-code such that $|C| \geq d$. Choose d codewords $c_1, \ldots, c_d$ from C. Let c_{ij} represent the j^{th} entry of codeword c_i. The $\mathbb{F}_q$-element c_{ij} determines the position of the nonzero element of the j^{th} block of the i^{th} column. In particular, the nonzero entry occurs at entry $p(c_{ij})$ of block j.

Let $E : \mathbb{F}_q \to \{0, 1\}^q$ defined by $E(z) = e_{p(z)} \in \{0, 1\}^q$ such that e_i is the i^{th} standard basis of $\mathbb{R}^q$.

Example 7. Consider $\mathbb{F}_4$ with elements $\{0, 1, \alpha, \alpha^2\}$ ordered as listed, where α is a root of $x^2 + x + 1$ over $\mathbb{F}_2$. Then,

$$E(\alpha) = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

Distribution: Let $\mathbb{F}_q$ be a field of q elements with fixed order. Let C be a predetermined $[d, m, d(C)]$-code over F_q. Let $\sigma_{ij} \overset{iid}{\sim} \{\pm 1\}$ for $1 \leq i \leq s$ and $1 \leq j \leq d$. Then, let $\mathcal{D}$ be a distribution defined by code C such that, for $A \sim \mathcal{D}$,

$$A = \frac{1}{\sqrt{s}} \begin{pmatrix} \sigma_{11} E(c_{11}) & \cdots & \sigma_{1d} E(c_{d1}) \\ \vdots & \ddots & \vdots \\ \sigma_{s1} E(c_{1s}) & \cdots & \sigma_{sd} E(c_{ds}) \end{pmatrix}.$$

Theorem 15 ([29])**.** Let $\mathcal{D}$ be either the graph of block construction dependent on codes as defined. Then, for $k = \Theta\left(\epsilon^{-2} \log \frac{1}{\delta}\right)$ and $s \geq 2(2\epsilon - \epsilon^2)^{-1} \log \frac{1}{\delta}$,

$$\text{Prob}_{A \sim \mathcal{D}} \left[(1-\epsilon) \|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\epsilon) \|x\|_2^2 \right] > \delta.$$

2.4.2.2 Block Construction via Algebraic Geometry Codes

An explicit block construction defined by a code was given by Gao et al. in [21]. In [21] they showed that the distribution $\mathcal{D}$ defined by a code is a JL distribution, if the $[s, m, d(C)]$-code off which it is based holds the following inequality:

$$\frac{s^2}{s-d} \geq 4\epsilon^{-2} \left[(2\ell - 1)!!\right]^{1/\ell}, \tag{2.16}$$

where $\delta = 2^{-\ell}$. Consequently, in order to allow for a small ϵ and δ, codes with high ratio $\frac{s^2}{s-d}$ are desired. For such codes, Gao et al. turned to Algebraic Geometry (AG) codes.

An AG code is a good code if the ratio between the number N of rational points and the genus g of the curve, $\frac{N}{g}$, is large. An upper bound of the ratio was given by Drinfeld and Vladut in [18]:

$$\limsup_{g \to \infty} \frac{N}{g} \leq \sqrt{q} - 1.$$

An AG code, which attains the Drinfeld-Vladut (DV) bound, is called an asymptotically good code. Garcia and Stichtenoth constructed such a code from towers.

Definition 5 (Garcia-Stichtenoth Tower)**.** Let $\mathcal{F} := (F_0, F_1, \ldots)$ denote the tower of extensions of rational function field $F_0 = \mathbb{F}_{q^2}(x_0)$, where q is a prime power. For $i \geq 1$, let $F_i := F_{i-1}(x_i)$ where

$$x_i^q + x_i = \frac{x_{i-1}^q}{x_{i-1}^{q-1} + 1}.$$

They showed that a linear code such that the DV bound is attained can be explicitly constructed from the towers.

Theorem 16 ([22])**.** Let $u \geq 1$ be some integer and $q \geq 2$ be a prime power. Let

$$s = q^u(q^2 - q) \quad \text{and} \quad g = (q^{\lfloor \frac{u+1}{2} \rfloor} - 1)(q^{\lceil \frac{u+1}{2} \rceil} - 1).$$

Suppose $m < s$ is an integer and

$$d(C) = s - m - g.$$

Then one can explicitly construct a linear $[s, k, d]$ AG code over $\mathbb{F}_{q^2}$ with code length s, dimension m and minimum distance $d(C)$.

The number of rational points $N(F_u)$ of tower F_u is bounded below by

$$q^u(q^2 - q).$$

The genus $g(F_u)$ of tower F_u is

$$g(F_u) = (q^{\lfloor \frac{u+1}{2} \rfloor} - 1)(q^{\lceil \frac{u+1}{2} \rceil} - 1).$$

From the parameters defined in Theorem 16, we observe $\frac{s^2}{s-d(C)} \approx \frac{n^2}{m+g}$. Hence, as the code is asymptotically optimal, $\frac{s^2}{s-d(C)}$ is large. Examples of parameters for which an AG code holds Inequality (2.16) are presented in Table 2.1. As Inequality (2.16) holds true, these examples can be used to define a JL transformation under the block construction.

q	u	m	$d(C)$	g	$s = q^u(q^2 - q)$	$k = s \cdot q^2$	$d = (q^2)^m$	ϵ	$\delta = 0.5^\ell$
2	8	16	31	465	512	2048	4.29×10^{09}	0.30	0.5^4
2	9	16	47	961	1024	4096	4.29×10^{09}	0.42	0.5^8
2	13	18	237	16129	16,384	65,536	6.87×10^{10}	0.4	0.5^{32}
2	15	19	492	65,025	65,536	262,144	2.75×10^{11}	0.2	0.5^{32}
2	15	19	492	65,025	65,536	262,144	2.75×10^{11}	0.4	0.5^{64}
3	4	11	267	208	486	4374	3.14×10^{10}	0.21	0.5^4
3	4	11	267	208	486	4374	3.14×10^{10}	0.42	0.5^8
3	6	12	2,282	2,080	4,374	39,366	2.82×10^{11}	0.3	0.5^{16}
3	7	12	6,710	6,400	13,122	118,098	2.82×10^{11}	0.3	0.5^{32}
3	8	13	19,993	19,360	39,366	354,294	2.54×10^{12}	0.4	0.5^{64}
4	2	8	139	45	192	3072	4.29×10^{09}	0.26	0.5^4
4	4	9	2,118	945	3,072	49,152	6.87×10^{10}	0.3	0.5^{16}
4	5	10	8,309	3,969	12,288	196,608	1.10×10^{12}	0.3	0.5^{32}

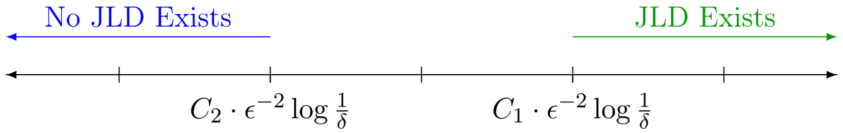
Table 2.1: The construction of A is based on an AG code from the u-th level GS-tower over $\mathbb{F}_{q^2}$ where q is a prime power. The parameters s, m, $d(C)$ and g correspond to the length, dimension, minimum distance and genus.

Those who are interested in AG codes are directed to [45] for a comprehensive study on the matter, and those interested in the block construction via AG codes are directed to [21].

Chapter 3

Optimal Bounds of the Projected Dimension

In this chapter, we focus on the problem of finding a precise threshold for k_0, the minimum dimension k such that there exist a JL distribution on $\mathbb{R}^{k\times d}$. Johnson and Lindenstrauss [28] showed the existence of a JL transformation for projected dimensions $k \geq C_1 \cdot \left(\epsilon^{-2}\log\frac{1}{\delta}\right)$ for some constant $C_1 > 0$. In 1988, Frankl and Maehara [20] provided an explicit construction for the projected dimensions $k \geq 9\epsilon^{-2}\log\frac{1}{\delta}$. In 2003, a threshold under which no JL distribution exists was given by Alon. Alon [4] proved for $k = \mathcal{O}\left(\epsilon^{-2}\log\frac{1}{\delta}/\log\frac{1}{\epsilon}\right)$ there is no JL transformation. Woodruff and Jayram [27] and later Kane, Nelson, and Meka [29] improved Alon's result to $k = \mathcal{O}\left(\epsilon^{-2}\log\frac{1}{\delta}\right)$. Hence, for a projected dimension less than $C_2 \cdot \left(\epsilon^{-2}\log\left(\frac{1}{\delta}\right)\right)$ for some constant C_2, there is no JL distribution; whereas, for a projected dimension greater than $C_1 \cdot \left(\epsilon^{-2}\log\left(\frac{1}{\delta}\right)\right)$ for some constant C_1, there exists a transformation. The smallest projected dimension k_0 lies between the values $C_2 \cdot \epsilon^{-2}\log\frac{1}{\delta}$ and $C_1 \cdot \left(\epsilon^{-2}\log\frac{1}{\delta}\right)$.


No JLD Exists
JLD Exists
$C_2 \cdot \epsilon^{-2}\log\frac{1}{\delta}$
$C_1 \cdot \epsilon^{-2}\log\frac{1}{\delta}$

In this paper, we provide explicit thresholds between which the smallest projected dimension k_0 lies and show these thresholds become one as $\epsilon, \delta \to 0$ and $\frac{k}{d} \to 0$. In particular, we have the following theorem:

Theorem 17. Let $0 < \epsilon, \delta < \frac{1}{2}$.

a.) If $k > 4\epsilon^{-2} \log\left(\frac{1}{\delta}\right)[1 + o(1)]$, then there exists a JL distribution.

The term $o(1)$ is dependent on the error factor ϵ and probability of failure δ, and approaches zero as ϵ and δ approach zero. The exact term $o(1)$ may be found in the proof of Theorem 21.

b.) If $k < 4\epsilon^{-2} \log\left(\frac{1}{\delta}\right)[1 - o(1)]$, then there does not exist a JL distribution.

The term $o(1)$ is dependent on the error factor ϵ, probability of failure δ, and the ratio $\frac{k}{d}$, and approaches zero as ϵ, δ and the ratio $\frac{k}{d}$ approach zero.

As a result, we obtain Theorem 2.

Theorem 2. For ϵ and δ sufficiently small, $k_0 \approx 4\epsilon^{-2} \log \frac{1}{\delta}$. That is,

$$\frac{k_0}{4\epsilon^{-2} \log \frac{1}{\delta}} \to 1 \text{ as } \epsilon, \delta \to 0.$$

In Section 3.1, we review uniform measure on unit spheres S^{d-1} and show how the metric on S^{d-1} relates to metrics on S^{d-k-1} and S^{k-1} with $k < d$. In Section 3.2.1, we provide a lower bound that guarantees the existence of an orthogonal projection from $\mathbb{R}^d$ to $\mathbb{R}^k$ such that Inequality (2.1) holds, and hence give an upper bound to k_0. In Section 3.2.2, we provide the lower bound of $\text{Prob}_{A\sim D}\left[|\|Ax\|^2 - 1| < \epsilon\right]$, and as a result provide a threshold for the projected dimension, below which there does not exist a JL distribution.

3.1 Uniform Measure of Unit Spheres in High Dimensions

In this section, we review uniform measure and provide a relation between the surface area measure on S^{d-1} and that on S^{d-k-1} and S^{k-1}, $k < d$. A uniform distribution on a unit sphere can be described as follows: let $x \overset{iid}{\sim} N(0,1)^d$, then $\frac{x}{\|x\|_2}$ is uniformly distributed on S^{d-1}. The majority of measure lies around the equator of the sphere S^{d-1}. We exploit this property in Section 3.2.1 and Section 3.2.2 to bound the measure of concentration.

3.1.1 Surface Area Measure

Let S^{d-1} denote the $(d-1)$-dimensional unit sphere and $d\Omega_{d-1}$ denote the $(d-1)$-dimensional surface area measure on the sphere S^{d-1}. Let

$$D^{d-1} := \left\{ x \in \mathbb{R}^{d-1} : \sum_{i=1}^{d-1} x_i^2 \leq 1 \right\}$$

be the $(d-1)$-dimensional disk.

Lemma 4. Let $(y_1, \ldots, y_d)$ with $y_d > 0$ be coordinates of points on the upper hemisphere of a $(d-1)$-dimensional unit sphere. Then, the surface area measure of the unit sphere S^{d-1} is

$$d\Omega_{d-1} = \frac{1}{y_d} dy_1 \cdots dy_{d-1}.$$

Proof. In 3-dimensional space, the surface area measure of a sphere can be found by dividing the domain into small rectangular regions, computing the sum of the areas of the parallelograms tangent to points on the surface above the rectangular regions, and decreasing the size of the regions to zero. In t-dimensional space, the surface area measure of a unit sphere S^{d-1} can be determined by computing the $(d-1)$-dimensional volume of parallelepipeds defined by the tangent vectors, and decreasing the size of the parallelepipeds to zero.

Let $(z_1, \ldots, z_{d-1})$ be coordinates of points on the $(d-1)$-dimensional disk D^{d-1}. Let $\phi : D^{d-1} \to \mathbb{R}_{\geq 0}$ be defined by

$$\phi(z_1, \ldots, z_{d-1}) = \sqrt{1 - \sum_{i=1}^{d-1} z_i^2}.$$

Observe that $y_d = \phi(y_1, \ldots, y_{d-1})$.

Let the parallelepiped P^{d-1} defined by the tangent vectors at a point in D^{d-1}. The tangent vectors $\Delta z_i e_i$ have length Δz_i and direction e_i, where e_i is the i^{th} standard basis of $\mathbb{R}^d$. Let P^d denote the parallelepiped defined by the tangent vectors $\Delta z_i e_i$ and the vector he_d. Then, as the vector he_d is perpendicular to the tangent vectors of the disk D^{d-1}, the t-dimensional volume of P^d can be computed in terms of the $(d-1)$-dimensional volume of P^{d-1} and height h, i.e.,

$$\mathrm{Vol}_d\left(P^d\right) = h \cdot \mathrm{Vol}_{d-1}\left(P^{d-1}\right). \tag{3.1}$$

Now, consider the map $\Phi : D^{d-1} \times \mathbb{R} \to \mathbb{R}^d$ defined by

$$(z_1, \dots, z_{d-1}, z_d) \mapsto ((1+z_d)z_1, (1+z_d)z_2, \dots, (1+z_d)z_{d-1}, (1+z_d)\phi(z_1, \dots, z_{d-1}))\,.$$

We first note that $\Phi\big|_{D^{d-1}\times\{0\}}$ maps the disk D^{d-1} surjectively onto the upper hemisphere of the unit sphere S^{d-1}, see Figure 3.1. The Jacobian of Φ when restricted to $D^{d-1} \times \{0\}$ is

$$(\mathrm{Jac}\Phi)\,|_{D^{d-1}\times\{0\}} = \left[\begin{array}{ccc|c} & & & z_1 \\ & I & & \vdots \\ & & & z_{d-1} \\ \hline -\frac{z_1}{\phi(z_1,\dots,z_{d-1})} & \cdots & -\frac{z_{d-1}}{\phi(z_1,\dots,z_{d-1})} & \phi(z_1, \dots, z_{d-1}) \end{array}\right].$$

Under the map of the Jacobian, the image of the tangent vectors $\Delta z_i e_i$ for $1 \leq i \leq d-1$, is

$$\Delta z_i \left(f_i - \frac{z_i}{\phi(z_1, \dots, z_{d-1})} f_d \right),$$

where f_i is the i^{th} standard basis of $\mathbb{R}^d$, and the image of the vector he_d is the vector

$$h\langle z_1, \dots, z_{d-1}, \phi(z_1, \dots, z_{d-1})\rangle.$$

Observe that the vectors $\Delta z_i \left(f_i - \frac{z_i}{\phi(z_1,\ldots,z_{d-1})} f_d \right)$ for $1 \le i \le d-1$ are tangent vectors to the sphere S^{d-1}, and that $h\langle z_1, \ldots, z_{d-1}, \phi(z_1, \ldots, z_{d-1})\rangle$ is the outward pointing surface normal with length h.

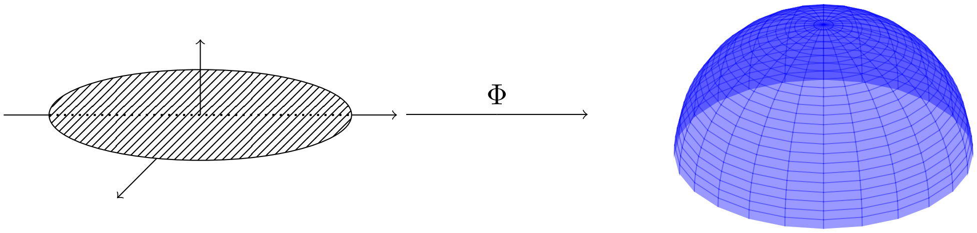


Figure 3.1: Φ maps the disk D^{d-1} surjectively onto the surface area on the upper hemisphere of the $d-1$ dimensional unit sphere S^{d-1}. We observe that $\Phi^{-1} = \Pi$ is the projection map onto the first $d-1$ coordinates.

Let Q^{d-1} denote the parallelepiped defined by the tangent vectors $\Delta z_i \left(f_i - \frac{z_i}{\phi(z_1,\ldots,z_{d-1})} f_d \right)$ for $1 \le i \le d-1$. Let Q^d denote the parallelepiped defined by the tangent vectors of Q^{d-1} and the vector $h\langle z_1, \ldots, z_{d-1}, \phi(z_1, \ldots, z_{d-1})\rangle$. Then, as the vector $\langle z_1, \ldots, z_{d-1}, \phi(z_1, \ldots, z_{d-1})\rangle$ is orthogonal to the tangent vectors $\Delta z_i \left(f_i - \frac{z_i}{\phi(z_1,\ldots,z_{d-1})} f_d \right)$ for $1 \le i \le d-1$, the volume of Q^d is that of the volume of Q^{d-1} scaled by its height h. That is,

$$\mathrm{Vol}_d\left(Q^d\right) = h \cdot \mathrm{Vol}_{d-1}\left(Q^{d-1}\right). \tag{3.2}$$

In addition, we note that Q^d is the image of P^d under the map $\mathrm{Jac}\,\Phi\big|_{D^{d-1}\times\{0\}}$. As the measure of the image is the measure of the preimage scaled by the determinant, the volume of Q^d can be computed in terms of P^d:

$$\mathrm{Vol}_d\left(Q^d\right) = \left| \det\left(\mathrm{Jac}\Phi\right)\Big|_{(z_1,\ldots,z_{d-1},0)} \right| \cdot \mathrm{Vol}_d\left(P^d\right). \tag{3.3}$$

The determinant of $\mathrm{Jac}(\Phi)$ at the point $(z_1, \ldots, z_{d-1}, 0)$ can be found by forming an upper triangular matrix by adding multiples of rows to the last row, resulting in a matrix with the product of its diagonal elements, $\frac{1}{\phi(z_1,\ldots,z_{d-1})}$. Hence, the determinant of $\mathrm{Jac}(\Phi)$ is $\frac{1}{\phi(z_1,\ldots,z_{d-1})}$

and is positive. Substituting this result in Expression (3.3), we obtain

$$\mathrm{Vol}_d\left(Q^d\right) = \frac{1}{\phi(z_1, \ldots, z_{d-1})} \cdot \mathrm{Vol}_d\left(P^d\right). \tag{3.4}$$

Replacing $\mathrm{Vol}_d\left(P^d\right)$ and $\mathrm{Vol}_d\left(Q^d\right)$ with their equivalents given in Expression (3.1) and Expression (3.2) and dividing by h, Equation (3.4) becomes

$$\mathrm{Vol}_{d-1}\left(Q^{d-1}\right) = \frac{1}{\phi(z_1, \ldots, z_{d-1})} \cdot \mathrm{Vol}_{d-1}\left(P^{d-1}\right).$$

Note that $\mathrm{Vol}\left(Q^{d-1}\right)$ and $\mathrm{Vol}\left(P^{d-1}\right)$ are products of their corresponding side lengths. As those side lengths approach zero, they become differential forms, and hence we obtain $d\Omega_{d-1}$ and $dz_1 \cdots dz_{d-1}$. Finally, observing that coordinates $(z_1, \ldots z_{d-1})$ of a disk D^{d-1} correspond to the first $d-1$ entries of coordinates $(y_1, \ldots, y_{d-1}, y_d)$ of a unit sphere, that is, $y_i = z_i$ for $1 \leq i \leq d-1$, we have

$$dz_1 \cdots dz_{d-1} = dy_1 \cdots dy_{d-1} \quad \text{and} \quad \phi(z_1, \ldots, z_{d-1}) = y_d.$$

Consequently, we have our desired result. □

As we are interested in reducing a d-dimensional vector to a k-dimensional vector, we find the surface area measure of the $(d-1)$-dimensional unit sphere in terms of a $(k-1)$-dimensional unit sphere and a $(d-k-1)$-dimensional unit sphere.

Theorem 18. Let $s \in [0, 1]$ and $f(s) = s^{\frac{k-2}{2}}(1-s)^{\frac{d-k-2}{2}}$. Let $\Psi : [0, 1] \times S^{k-1} \times S^{d-k-1} \longrightarrow S^{d-1}$ be defined by

$$s \times (x_1, \ldots, x_k) \times (y_1, \ldots, y_{d-k}) \mapsto (\sqrt{s}x_1, \ldots, \sqrt{s}x_k, \sqrt{1-s}y_1, \ldots, \sqrt{1-s}y_{d-k}).$$

The map Ψ is surjective and the surface area measure of S^{d-1}, in terms of the surface area measures of S^{k-1} and S^{d-k-1}, is

$$d\Omega_{d-1} = \frac{1}{2} f(s) ds \cdot d\Omega_{k-1} d\Omega_{d-k-1}.$$

Proof. Let $(w_1, \ldots, w_d)$, $(x_1, \ldots, x_k)$, and $(y_1, \ldots, y_{d-k})$ be coordinates of points on the $(d-1)$-dimensional unit sphere, $(k-1)$-dimensional unit sphere, and $(d-k-1)$-dimensional unit sphere, respectively.

To show Ψ is onto, consider an arbitrary $w \in S^{d-1}$. Let $s = \sum_{i=1}^{k} w_i^2$. Then, $1 - s = \sum_{i=k+1}^{d} w_i^2$. As a result, w can be decomposed into two unit vectors scaled by $\sqrt{s}$ and $\sqrt{1-s}$, respectively. That is, when $s \in (0,1)$,

$$w = \sqrt{s} \left(\frac{w_1}{\sqrt{s}}, \ldots, \frac{w_k}{\sqrt{s}} \right) \times \sqrt{1-s} \left(\frac{w_{k+1}}{\sqrt{1-s}}, \ldots, \frac{w_d}{\sqrt{1-s}} \right).$$

If $s = 0$, $w = 0 \cdot (1, 0, \ldots, 0) \times (w_{k+1}, \ldots, w_d)$. If $s = 1$, then $w = (w_1, \ldots, w_k) \times 0 \cdot (1, 0, \ldots, 0)$. Let $x = \left(\frac{w_1}{\sqrt{s}}, \ldots, \frac{w_k}{\sqrt{s}} \right)$ and $y = \left(\frac{w_{k+1}}{\sqrt{1-s}}, \ldots, \frac{w_d}{\sqrt{1-s}} \right)$ when $s \in (0,1)$, $x = e_1$ and $y = (w_{k+1}, \ldots, w_d)$ when $s = 0$, and $x = (w_1, \ldots, w_k)$ and $y = e_1$ when $s = 1$, where e_i is the i^{th} unit vector. We observe that $x \in S^{k-1}$ and $y \in S^{d-k-1}$. Then,

$$(s, x, y) \overset{\Psi}{\mapsto} w.$$

Let $(\hat{w}_1, \ldots, \hat{w}_{d-1})$, $(\hat{x}_1, \ldots, \hat{x}_{k-1})$, and $(\hat{y}_1, \ldots, \hat{y}_{d-k-1})$ be the coordinates of the points on the disks D^{d-1}, D^{k-1} and D^{d-k-1} respectively. Let

$$\varphi : [0,1] \times D^{k-1} \times D^{d-k-1} \longrightarrow D^{d-1}$$

be defined by

$$s \times (\hat{x}_1,\dots,\hat{x}_{k-1})\times(\hat{y}_1,\dots,\hat{y}_{d-k-1}) \mapsto \left(\sqrt{s}\hat{x}_1,\dots,\sqrt{s}\hat{x}_{k-1},\sqrt{s}\sqrt{1-\sum_{i=1}^{k-1}\hat{x}_i^2},\sqrt{1-s}\hat{y}_1,\dots,\sqrt{1-s}\hat{y}_{d-k-1}\right).$$

Observe that φ maps the disks D^{k-1} and D^{d-k-1} onto the disk D^{d-1} restricted to the set of points $\hat{w}$ such that the k^{th} coordinate $\hat{w}_k$ is positive. As the measure of the image is the measure of the preimage scaled by the determinant, the surface area measure of the disk D^{d-1} is

$$d\hat{w}_1\cdots d\hat{w}_{d-1} = |\det \operatorname{Jac}(\varphi)| \cdot ds \cdot d\hat{x}_1\cdots d\hat{x}_{k-1}\cdot d\hat{y}_1\cdots d\hat{y}_{d-k-1}. \tag{3.5}$$

The Jacobian of map φ is

$$\operatorname{Jac}\varphi = \left[\begin{array}{c|ccc|ccc}
\frac{\hat{x}_1}{2\sqrt{s}} & \sqrt{s} & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\frac{\hat{x}_{k-1}}{2\sqrt{s}} & 0 & \cdots & \sqrt{s} & 0 & \cdots & 0 \\
\hline
\frac{\sqrt{1-\sum_{i=1}^{k-1}\hat{x}_i^2}}{2\sqrt{s}} & \frac{-\sqrt{s}\cdot\hat{x}_1}{\sqrt{1-\sum_{i=1}^{k-1}\hat{x}_i^2}} & \cdots & \frac{-\sqrt{s}\cdot\hat{x}_{k-1}}{\sqrt{1-\sum_{i=1}^{k-1}\hat{x}_i^2}} & 0 & \cdots & 0 \\
\hline
\frac{-\hat{y}_1}{2\sqrt{1-s}} & 0 & \cdots & 0 & \sqrt{1-s} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\frac{-\hat{y}_{d-k-1}}{2\sqrt{1-s}} & 0 & \cdots & 0 & 0 & \cdots & \sqrt{1-s}
\end{array}\right].$$

Eliminating all but the k^{th} entry of the first column by adding multiples of columns to the first column, we obtain

$$\det \operatorname{Jac}(\varphi) = \frac{1}{2\hat{x}_k} s^{\frac{k-2}{2}}(1-s)^{\frac{d-k-1}{2}}.$$

Substituting this value into Expression (3.5), we have the surface area measure of D^{d-1} in terms of the disks D^{k-1} and D^{d-k-1}. That is,

$$d\hat{w}_1 \cdots d\hat{w}_{d-1} = \frac{1}{2\hat{x}_k} s^{(k-2)/2}(1-s)^{(d-k-1)/2} \cdot ds \cdot d\hat{x}_1 \cdots d\hat{x}_{k-1} \cdot d\hat{y}_1 \cdots d\hat{y}_{d-k-1}. \tag{3.6}$$

We observe that the coordinates of a disk D^{t-1} correspond to the first $t-1$ entries of coordinates of a unit sphere for $t = d$, k and $d-k$. Employing the results of Lemma 4, we define the surface measure of a unit sphere in terms of the surface measure of the corresponding disk:

$$\begin{aligned} d\hat{x}_1 \cdots d\hat{x}_{k-1} &= dx_1 \cdots dx_{k-1} = x_k d\Omega_{k-1} \\ d\hat{y}_1 \cdots d\hat{y}_{d-k-1} &= dy_1 \cdots dy_{d-k-1} = y_{d-k} d\Omega_{d-k-1} \\ d\hat{w}_1 \cdots d\hat{w}_{d-1} &= dw_1 \cdots dw_{d-1} = w_d d\Omega_{d-1}. \end{aligned}$$

Substituting these into Equation 3.6, we obtain

$$\begin{aligned} d\Omega_{d-1} = \frac{1}{w_d} dw_1 \cdots dw_{d-1} &= \frac{y_{d-k}}{w_d} \cdot \frac{1}{2} s^{(k-2)/2}(1-s)^{(d-k-1)/2} \cdot ds \cdot d\Omega_{k-1} d\Omega_{d-k-1} \\ &= \frac{1}{2} s^{(k-2)/2}(1-s)^{(d-k-2)/2} \cdot ds \cdot d\Omega_{k-1} d\Omega_{d-k-1}, \end{aligned}$$

where the last equality follows from the fact that $w_d = \sqrt{1-s} y_{d-k}$ by map ψ. Therefore, we have

$$d\Omega_d = \frac{1}{2} f(s) ds \cdot d\Omega_{k-1} d\Omega_{d-k-1},$$

where $f(s) = s^{\frac{k-2}{2}}(1-s)^{\frac{d-k-2}{2}}$, $0 \leq s \leq 1$. □

Parameters

We define the following parameters, which will be used throughout the remainder of the paper:

- Let ϵ and $\delta \in \left(0, \frac{1}{2}\right)$.
- Let $d \in \mathbb{Z}^+$ be even.

- Let $k \in \mathbb{Z}^+$ be even such that $k - 4 \geq \frac{1}{\epsilon^2}$ and $d > 2.5k$.
- Let $A : \mathbb{R}^d \to \mathbb{R}^k$. Observe that $w \in S^{d-1}$ can be decomposed into the unit vectors μ_{k-1} and μ_{d-k-1} with length $\sqrt{s}$ and $\sqrt{1-s}$, respectively, such that $\mu_{d-k-1} \in \ker(A)$, where $\ker(A)$ denotes the kernel of map A. For the remainder of the paper, we let $s \in [0, 1]$, as described.

Let $\Gamma(z)$ denote the gamma function. That is,

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

when $z \in \mathbb{R}$. Recall that,

$$\Gamma(z) = (z-1)!$$

when $z \in \mathbb{Z}^+$. As a result of Theorem 18, we have the following corollary:

Corollary 1. Let $C > 0$ and $0 < \epsilon < 1/2$. Then,

$$\text{Prob}\left[|s \cdot C - 1| > \epsilon\right] = B \cdot \int_{s \notin \left(\frac{1-\epsilon}{C}, \frac{1+\epsilon}{C}\right)} f(s) ds,$$

where $B = \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{d-k}{2})}$.

Proof. Let $w \in S^{d-1}$. We observe that the probability may be defined in terms of the volume as follows:

$$\text{Prob}\left[|s \cdot C - 1| > \epsilon\right] = \frac{\text{Vol}(|s \cdot C - 1| > \epsilon)}{\text{Vol}(S^{d-1})}. \tag{3.7}$$

As $\text{Vol}(|s \cdot C - 1| > \epsilon) = \int_{|s \cdot C - 1| > \epsilon} dw_1 \cdots dw_d$ and

$$dw_1 \cdots dw_d = \frac{1}{2} f(s) ds d\Omega_{k-1} d\Omega_{d-k-1}$$

from Theorem 18, the volume $\text{Vol}(|s \cdot C - 1| > \epsilon)$ can be defined in terms of the variable s

and the unit spheres S^{k-1} and S^{d-k-1}:

$$\begin{aligned}\mathrm{Vol}(|sC-1|>\epsilon) &= \int_{|C\sum_{i=1}^{k} w_i^2-1|>\epsilon} dw_1\cdots dw_d \\ &= \int_{S^{d-k-1}}\int_{S^{k-1}}\int_{s\notin\left(\frac{1-\epsilon}{C},\frac{1+\epsilon}{C}\right)} \frac{1}{2}f(s)dsd\Omega_{k-1}d\Omega_{d-k-1}. \qquad (3.8)\end{aligned}$$

Recall that the volume of an $(n-1)$-dimensional unit sphere is $\frac{2\pi^{n/2}}{\Gamma\left(\frac{n}{2}\right)}$. Substituting $\frac{2\pi^{n/2}}{\Gamma\left(\frac{n}{2}\right)}$ for $\int_{S^{n-1}} d\Omega_{n-1}$ for $n=k,\ d-k$ into Equation 3.8, we obtain

$$\mathrm{Vol}(|s\cdot C-1|>\epsilon) = \frac{2\pi^{(d-k)/2}}{\Gamma\left(\frac{d-k}{2}\right)}\cdot\frac{2\pi^{k/2}}{\Gamma\left(\frac{k}{2}\right)}\int_{s\notin\left(\frac{1-\epsilon}{C},\frac{1+\epsilon}{C}\right)}\frac{1}{2}f(s)ds. \qquad (3.9)$$

Substituting in Expression (3.9) into Equation (3.7), we obtain our desired result

$$\mathrm{Prob}\left[|s\cdot C-1|>\epsilon\right] = B\cdot\int_{s\notin\left(\frac{1-\epsilon}{C},\frac{1+\epsilon}{C}\right)} f(s)ds,$$

where $B := \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{d-k}{2})}$. □

3.1.2 Bounds on Surface Area Measure

In this section, we bound $\mathrm{Prob}\left[s>C(1+\epsilon)\right]$ and $\mathrm{Prob}\left[s<C(1-\epsilon)\right]$ when $C=\frac{1}{s_0}=\frac{d}{k}$. From the proof of Corollary 1, we have

$$\mathrm{Prob}\left[s>s_0(1+\epsilon)\right] = B\int_{s>s_0(1+\epsilon)} f(s)ds \quad \text{and}$$

$$\mathrm{Prob}\left[s<(1-\epsilon)s_0\right] = \int_{s<s_0(1-\epsilon)} Bf(s)ds.$$

To bound the above expressions, we first bound B, as defined in Corollary 1.

3.1.2.1 Bounds for B

Lemma 5. Suppose k and d are even positive integers, $k < d$. Then

$$\frac{e^{-2}}{2\sqrt{\pi}} \cdot \frac{(d-2)^{(d-1)/2}}{(k-2)^{(k-1)/2}(d-k-2)^{(d-k-1)/2}} \leq B \leq \frac{e^{-43/42}}{2\sqrt{\pi}} \cdot \frac{(d-2)^{(d-1)/2}}{(k-2)^{(k-1)/2}(d-k-2)^{(d-k-1)/2}}.$$

Proof. Let k and d be even positive integers, $k < d$. Then, by the definition of the Γ function,

$$B = \frac{(\frac{d-2}{2})!}{(\frac{k-2}{2})!(\frac{d-k-2}{2})!}.$$

To find both a lower and an upper bound for B, we use a form of Stirling's approximation of $n!$ due to Robbins [42]:

$$\sqrt{2\pi} n^{n+1/2} e^{-n} e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi} n^{n+1/2} e^{-n} e^{\frac{1}{12n}}.$$

Using this bound, we obtain

$$C_0 \frac{(d-2)^{(d-1)/2}}{(k-2)^{(k-1)/2}(d-k-2)^{(d-k-1)/2}} \leq B \leq C_1 \frac{(d-2)^{(d-1)/2}}{(k-2)^{(k-1)/2}(d-k-2)^{(d-k-1)/2}},$$

where

$$C_0 = \frac{1}{2} \cdot \frac{1}{\sqrt{\pi}} \cdot e^{-1} e^{\frac{1}{6(d-2)+1}} e^{\frac{-1}{6(k-2)}} e^{\frac{-1}{6(d-k-2)}} \geq \frac{e^{-2}}{2\sqrt{\pi}}, \quad \text{and}$$

$$C_1 = \frac{1}{2} \cdot \frac{1}{\sqrt{\pi}} \cdot e^{-1} e^{\frac{1}{6(d-2)}} e^{\frac{-1}{6(k-2)+1}} e^{\frac{-1}{6(d-k-2)+1}} \leq \frac{e^{-43/42}}{2\sqrt{\pi}}.$$

□

Consequently, we have the next lemma:

Lemma 6. Let $f(s)$ be as in Theorem 18, $s_0 = \frac{k}{d}$, and $2k < d$. Then,

$$\frac{e^{-2}}{2\sqrt{\pi}} \cdot \sqrt{k} \leq B s_0 f(s_0) \leq \frac{9e^{-43/42}}{\sqrt{2\pi}} \cdot \sqrt{k}.$$

Proof. By evaluating f at s_0 and replacing B by its lower bound found in Lemma 5, we obtain the lower bound

$$
\begin{aligned}
Bs_0 f(s_0) &\geq \frac{e^{-2}}{2\sqrt{\pi}} \left(\frac{\sqrt{kd}}{\sqrt{d-k}} \right) \left(\frac{k}{k-2} \right)^{\frac{k-1}{2}} \left(\frac{d-k}{d-k-2} \right)^{\frac{d-k-1}{2}} \left(\frac{d-2}{d} \right)^{\frac{d-1}{2}} \\
&= \frac{e^{-2}}{2\sqrt{\pi}} \sqrt{k} \left(\frac{d-2}{d-k} \right)^{\frac{1}{2}} \left(\frac{k(d-2)}{d(k-2)} \right)^{\frac{k-1}{2}} \left(\frac{(d-k)(d-2)}{d(d-k-2)} \right)^{\frac{d-k-1}{2}} \\
&\geq \frac{e^{-2}}{2\sqrt{\pi}} \cdot \sqrt{k}.
\end{aligned}
$$

Similarly, by evaluating f at s_0 and replacing B by its upper bound found in Lemma 5, we obtain the upper bound

$$
\begin{aligned}
Bs_0 f(s_0) &\leq \frac{e^{-43/42}}{2\sqrt{\pi}} \cdot \left(\frac{d-2}{d} \right)^{\frac{d-1}{2}} \cdot \left(\frac{k}{k-2} \right)^{\frac{k-1}{2}} \cdot \left(\frac{d-k}{d-k-2} \right)^{\frac{d-k-1}{2}} \frac{\sqrt{kd}}{\sqrt{d-k}} \\
&\leq \frac{9e^{-43/42}}{\sqrt{2\pi}} \cdot \sqrt{k},
\end{aligned}
$$

where the last inequality follows from $\frac{d}{d-k} \leq 2$, for $d > 2k$, and $e \leq \left(\frac{x}{x-2} \right)^{\frac{x-1}{2}} \leq 3$, for $x \geq 3$.

□

3.1.2.2 Bounds on Prob $[s > s_0(1+\epsilon)]$

From Taylor's expansion on $\log(1+x)$ and $\log(1-x)$, we derive the following inequalities:

$$\log(1+x) \geq x - \frac{x^2}{2} \quad \text{for } 0 < x < 1, \tag{3.10}$$

$$\log(1-x) \geq -x - \frac{x^2}{2} - x^3 \quad \text{for } 0 < x < 0.815 \quad \text{and} \tag{3.11}$$

$$\log(1-x) \geq -x - x^2 \quad \text{for } 0 < x < 0.68. \tag{3.12}$$

With the bounds on B and $Bs_0 f(s_0)$ from Lemmas 5 and 6, respectively, and Inequalities (3.10) and (3.11), we find upper and lower bounds on Prob $[s > s_0(1+\epsilon)]$.

Lemma 7. Let $0 < \epsilon < \frac{1}{2}$ and $2.5k < d$. Then,

$$\text{Prob}\left[s > s_0(1+\epsilon)\right] \geq \frac{e^{-2}}{4\sqrt{\pi}} \cdot e^{-\frac{1}{4}(\sqrt{k}\epsilon+1)^2 \frac{1+s_0}{1-s_0}}.$$

Proof. From the proof of Corollary 1, it can be seen that Prob $\left[s > \frac{1+\epsilon}{C}\right] = B \int_{s>\frac{1+\epsilon}{C}} f(s)ds$. Substituting s_0 in for C^{-1}, we have

$$\text{Prob}\left[s > s_0(1+\epsilon)\right] = B \int_{s>s_0(1+\epsilon)} f(s)ds. \tag{3.13}$$

By the change of variables $s = s_0(1+x)$, $x > 0$, Equation (3.13) becomes

$$B \cdot \int_{s>s_0(1+\epsilon)} f(s)ds = Bs_0 \cdot \int_{\epsilon}^{\frac{1}{s_0}-1} f(s_0(1+x))dx. \tag{3.14}$$

Let $g(s) = s^{k/2}(1-s)^{(d-k)/2}$, then, $f(s_0(1+x))$ can be expressed in terms of $g(s)$, namely,

$$f(s_0(1+x)) = \frac{g(s_0(1+x))}{s_0(1+x)\left(1-s_0(1+x)\right)}. \tag{3.15}$$

To find the lower bound of Prob $[s > s_0(1+\epsilon)]$, we first find a lower bound for $g(s_0(1+x))$. Taking the log of $g(s_0(1+x))$, we find

$$\log\left(g(s_0(1+x))\right) = \log g(s_0) + \frac{d}{2}\left(s_0 \log(1+x) + (1-s_0)\log\left(1 - \frac{s_0}{1-s_0}x\right)\right). \tag{3.16}$$

Assuming $0 < \frac{s_0}{1-s_0}x < 0.68$, we bound the inner expression by utilizing Inequalities (3.10) and (3.12):

$$\begin{aligned} s_0\log(1+x) + &(1-s_0)\log\left(1 - \frac{s_0}{1-s_0}x\right) \\ &\geq s_0\left(x - \frac{x^2}{2}\right) + (1-s_0)\left(\frac{-s_0}{1-s_0}x - \left(\frac{s_0}{1-s_0}\right)^2 x^2\right) \\ &= -\left(\frac{s_0(1+s_0)}{2(1-s_0)}\right)x^2. \end{aligned} \tag{3.17}$$

We note that $0 < \frac{s_0}{1-s_0}x < 0.68$ since $d > 2.5k$ and we restrict x to $\left(\epsilon, \epsilon + k^{-1/2}\right)$. Hence, by substituting Inequality (3.17) into Equation (3.16), we obtain the following lower bound for $g(s_0(1+x))$:

$$g(s_0(1+x)) \geq g(s_0)e^{-\frac{k}{4}x^2 \cdot \frac{1+s_0}{1-s_0}}. \tag{3.18}$$

We next note that the remaining term of Equation (3.15) is bounded by the following:

$$\frac{1}{s_0(1+x)(1-s_0(1+x))} \geq \frac{1}{2}\frac{1}{s_0(1-s_0)}, \tag{3.19}$$

for $0 \leq x < 1$. We may assume $0 < x < 1$, since we will only be using x in the range from ϵ to $\epsilon + k^{-1/2}$. Substituting both Inequality (3.18) and Inequality (3.19) into Equation (3.15), we have

$$f(s_0(1+x)) \geq \frac{1}{2} \cdot \frac{g(s_0)}{s_0(1-s_0)}e^{-\frac{k}{4}x^2\frac{1+s_0}{1-s_0}} = \frac{1}{2}f(s_0)e^{-\frac{k}{4}x^2\frac{1+s_0}{1-s_0}},$$

for $0 < x < 1$. As a result, we now have a lower bound for the integrand and obtain

$$\begin{aligned} Bs_0 \cdot \int_{\epsilon}^{\frac{1}{s_0}-1} f(s_0(1+x))dx &\geq Bs_0 \cdot \int_{\epsilon}^{\epsilon+k^{-1/2}} f(s_0(1+x))dx \\ &\geq \frac{1}{2}Bs_0f(s_0) \cdot \int_{\epsilon}^{\epsilon+k^{-1/2}} e^{-\frac{k}{4}x^2\frac{1+s_0}{1-s_0}}dx. \end{aligned}$$

Replacing $Bs_0f(s_0)$ with its lower bound given in Lemma 6, the inequality above becomes

$$\frac{1}{2}Bs_0f(s_0) \cdot \int_{\epsilon}^{\epsilon+k^{-1/2}} e^{-\frac{k}{4}x^2\frac{1+s_0}{1-s_0}}dx \geq \frac{e^{-2}}{4\sqrt{\pi}}e^{-\frac{1}{4}(\sqrt{k}\epsilon+1)^2\frac{1+s_0}{1-s_0}},$$

$$\begin{aligned} \frac{1}{2}Bs_0f(s_0) \cdot \int_{\epsilon}^{\epsilon+k^{-1/2}} e^{-\frac{k}{4}x^2\frac{1+s_0}{1-s_0}}dx &\geq \frac{e^{-2}}{4\sqrt{\pi}} \cdot \sqrt{k}\int_{\epsilon}^{\epsilon+k^{-1/2}} e^{-\frac{k}{4}x^2\frac{1+s_0}{1-s_0}}dx \\ &\geq \frac{e^{-2}}{4\sqrt{\pi}}e^{-\frac{1}{4}(\sqrt{k}\epsilon+1)^2\frac{1+s_0}{1-s_0}}, \end{aligned}$$

which completes the proof.

□

Lemma 8. Let $0 < \epsilon < \frac{1}{2}$ and assume $k - 4 \geq \frac{1}{\epsilon^2}$ and $2k < d$. Then,

$$\text{Prob}\left[s > s_0(1+\epsilon)\right] \leq \frac{27e^{-43/42}}{\sqrt{2\pi}} \cdot e^{-\frac{k-2}{4}\epsilon^2(1-\frac{2}{3}\epsilon)}.$$

Proof. To find the upper bound of Prob $[s > (1+\epsilon)s_0]$, we first bound consider $f(s_0(1+x))$, which appears in Equation (3.14). Using Taylor's expansion of e^x, we bound $f(s_0(1+x))$ as follows:

$$\begin{aligned} f(s_0(1+x)) &= f(s_0)(1+x)^{\frac{k-2}{2}}\left(1 - \frac{s_0}{1-s_0}x\right)^{\frac{d-k-2}{2}} \\ &\leq f(s_0)(1+x)^{\frac{k-2}{2}}\left(\exp\left(\frac{-s_0}{1-s_0}x\right)\right)^{\frac{d-k-2}{2}} \\ &\leq f(s_0)(1+x)^{\frac{k-2}{2}}e^{-\frac{k-2}{2}x}, \end{aligned} \tag{3.20}$$

where (3.20) follows from $d > 2k$. Taking the integral of Expression (3.20) and employing integration by parts, we have for any constant ℓ, $1 \leq \ell \leq m$,

$$\int_{x=\epsilon}^{\infty}(1+x)^{\ell}e^{-mx}dx \leq \frac{1}{m}(1+\epsilon)^{\ell}e^{-m\epsilon} + \int_{\epsilon}^{\infty}(1+x)^{\ell-1}e^{-mx}dx. \tag{3.21}$$

Starting with $\ell = \frac{k-2}{2}$ and $m = \frac{k-2}{2}$ and repeatedly employing Inequality (3.21), we have

$$\begin{aligned} \int_{\epsilon}^{\infty}(1+x)^{\frac{k-2}{2}}e^{-\frac{k-2}{2}x}dx &\leq \frac{2e^{-\frac{k-2}{2}\epsilon}}{k-2}\left((1+\epsilon)^{\frac{k-2}{2}} + (1+\epsilon)^{\frac{k-4}{2}} + \cdots + (1+\epsilon)^{0}\right) \\ &\leq \frac{2(1+\epsilon)}{(k-2)\epsilon}(1+\epsilon)^{\frac{k-2}{2}}e^{-\frac{k-2}{2}\epsilon}. \end{aligned}$$

Noting that $(1+\epsilon)^{\frac{k-2}{2}} = e^{\frac{k-2}{2}\log(1+\epsilon)}$ and using Taylor's expansion of $\log x$, we obtain

$$\frac{2(1+\epsilon)}{(k-2)\epsilon}(1+\epsilon)^{\frac{k-2}{2}}e^{-\frac{k-2}{2}\epsilon} \leq \frac{2(1+\epsilon)}{(k-2)\epsilon}e^{\frac{k-2}{2}(\log(1+\epsilon)-\epsilon)} \leq \frac{2(1+\epsilon)}{(k-2)\epsilon}e^{-\frac{k-2}{4}\epsilon^2(1-\frac{2}{3}\epsilon)}.$$

Since $k-4 \geq \frac{1}{\epsilon^2}$, the expression $\frac{(k-2)^2}{k}$ is bounded below by $\frac{1}{\epsilon^2}$, namely,

$$\frac{(k-2)^2}{k} = k-4+\frac{4}{k} \geq k-4 \geq \frac{1}{\epsilon^2},$$

implying $(k-2)\epsilon \geq \sqrt{k}$. As a result,

$$\int_{\epsilon}^{\infty}(1+x)^{\frac{k-2}{2}}e^{-\frac{k-2}{2}x}dx \leq \frac{2(1+\epsilon)}{(k-2)\epsilon}e^{-\frac{k-2}{4}\epsilon^2(1-\frac{2}{3}\epsilon)} \leq \frac{2(1+\epsilon)}{\sqrt{k}}e^{-\frac{k-2}{4}\epsilon^2(1-\frac{2}{3}\epsilon)}. \tag{3.22}$$

Taking the upper bound to infinity of Expression (3.14) and substituting Inequality (3.20) and (3.22) into Expression (3.14), we obtain

$$\begin{aligned} Bs_0 \cdot \int_{\epsilon}^{\frac{1}{s_0}-1} f(s_0(1+x))dx &\leq Bs_0 f(s_0)\int_{\epsilon}^{\frac{1}{s_0}-1}(1+x)^{\frac{k-2}{2}}e^{-\frac{k-2}{2}x}dx \\ &\leq Bs_0 f(s_0)\frac{2(1+\epsilon)}{\sqrt{k}}e^{-\frac{k-2}{4}\epsilon^2(1-\frac{2}{3}\epsilon)}. \end{aligned}$$

Employing Lemma 6 we have our desired result. □

We collect Lemmas 7 and 8 into the following theorem:

Theorem 19. Let $0 < \epsilon < \frac{1}{2}$ and assume $k-4 \geq \frac{1}{\epsilon^2}$ and $d > 2.5k$. Then,

$$\frac{e^{-2}}{4\sqrt{\pi}} \cdot e^{-\frac{1}{4}(\sqrt{k}\epsilon+1)^2\frac{1+s_0}{1-s_0}} \leq \text{Prob}\left[s > s_0(1+\epsilon)\right] \leq \frac{27e^{-43/42}}{\sqrt{2\pi}} \cdot e^{-\frac{k-2}{4}\epsilon^2(1-\frac{2}{3}\epsilon)}.$$

3.1.2.3 Bounds on Prob $[s < s_0(1-\epsilon)]$

Lemma 9. Let $0 < \epsilon < \frac{1}{2}$. Then,

$$\text{Prob}\left[s < s_0(1-\epsilon)\right] \geq \frac{e^{-2}}{2\sqrt{\pi}} \cdot e^{-\frac{1}{4}\left(\frac{(\sqrt{k}\epsilon+1)^2}{1-s_0}+2(\sqrt[3]{k}\epsilon+k^{-1/6})^3\right)}.$$

Proof. Recalling the result of Corollary 1 and substituting s_0 in for C, we obtain

$$\text{Prob}\left[s < (1-\epsilon)s_0\right] = \int_{s<s_0(1-\epsilon)} Bf(s)ds. \tag{3.23}$$

By the change of variables $s = s_0(1-x)$, $x > 0$, Expression (3.23) becomes

$$\text{Prob}\left[s < (1-\epsilon)s_0\right] = Bs_0 \cdot \int_{\epsilon}^{1} f(s_0(1-x))dx. \tag{3.24}$$

Let $g(s) = s^{k/2}(1-s)^{(d-k)/2}$. Then $f(s_0(1-x))$ can be defined in terms of $g(s)$, namely,

$$f\left(s_0(1-x)\right) = \frac{g\left(s_0(1-x)\right)}{s_0(1-x)\left(1-s_0(1-x)\right)}. \tag{3.25}$$

To obtain a lower bound of $\text{Prob}\left[s < s_0(1-\epsilon)\right]$, we first bound $g\left(s_0(1-x)\right)$. Taking the log of $g(s_0(1-x))$, we obtain the expression

$$\log g(s_0(1-x)) = \log g(s_0) + \frac{d}{2}\left[s_0 \log(1-x) + (1-s_0)\log\left(1 + \frac{s_0}{1-s_0}x\right)\right]. \tag{3.26}$$

Employing Bounds (3.10) and (3.11), we find a lower bound for the inner expression:

$$\begin{aligned} s_0 \log(1-x) + &(1-s_0)\log\left(1+\frac{s_0}{1-s_0}x\right) \\ &\geq s_0(-x - \frac{x^2}{2} - x^3) + (1-s_0)\left(\frac{s_0}{1-s_0}x - \left(\frac{s_0}{1-s_0}\right)^2 \frac{x^2}{2}\right) \\ &= \frac{-s_0}{2(1-s_0)}x^2 - s_0 x^3. \end{aligned} \tag{3.27}$$

Substituting Inequality (3.27) into Expression (3.26), we get

$$\log g(s_0(1-x)) \geq \log g(s_0) - \frac{k}{4}\left[\frac{x^2}{1-s_0} + 2x^3\right],$$

implying

$$g(s_0(1-x)) \geq g(s_0)e^{-\frac{k}{4}\left[\frac{x^2}{1-s_0}+2x^3\right]}. \tag{3.28}$$

Since, for $0 < c < 1$ and $\epsilon < x < 1$, $(1-x)(1+c\cdot x) < 1$, the remaining term of Equation

(3.25) can be bounded by the following:

$$\frac{1}{s_0(1-x)\left(1-s_0(1-x)\right)} = \frac{1}{s_0(1-s_0)} \cdot \frac{1}{(1-x)\left(1+\frac{s_0 x}{1-s_0}\right)} \geq \frac{1}{s_0(1-s_0)}. \tag{3.29}$$

Substituting both Inequality (3.28) and Inequality (3.29) into Expression (3.15), we obtain a lower bound for $f\left(s_0(1-x)\right)$:

$$f(s_0(1-x)) \geq f(s_0) e^{-\frac{k}{4}\left(\frac{x^2}{1-s_0}+2x^3\right)}, \tag{3.30}$$

for $\epsilon < x < 1$. Consequently, substituting Inequality (3.30) into Equation (3.24) and lowering the upper bound to $\epsilon + k^{-1/2}$ result in

$$\begin{aligned} Bs_0 \cdot \int_\epsilon^1 f(s_0(1-x))dx &\geq Bs_0 \cdot \int_\epsilon^{\epsilon+k^{-1/2}} f(s_0(1-x))dx \\ &\geq Bs_0 f(s_0) \cdot \int_\epsilon^{\epsilon+k^{-1/2}} e^{-\frac{k}{4}\left(\frac{x^2}{1-s_0}+2x^3\right)} dx. \end{aligned}$$

Replacing $Bs_0f(s_0)$ with its lower bound given in Lemma 6 and evaluating the integral, we obtain

$$\begin{aligned} Bs_0 f(s_0) \cdot \int_\epsilon^{\epsilon+k^{-1/2}} e^{-\frac{k}{4}\left(\frac{x^2}{1-s_0}+2x^3\right)} dx &\geq \frac{e^{-2}}{2\sqrt{\pi}} \cdot \sqrt{k} \int_\epsilon^{\epsilon+k^{-1/2}} e^{-\frac{k}{4}\left(\frac{x^2}{1-s_0}+2x^3\right)} dx \\ &\geq \frac{e^{-2}}{2\sqrt{\pi}} \cdot e^{-\frac{1}{4}\left(\frac{(\sqrt{k}\epsilon+1)^2}{1-s_0}+2(\sqrt[3]{k}\epsilon+k^{-1/6})^3\right)}, \end{aligned}$$

which completes the proof. □

Lemma 10. Let $0 < \epsilon < \frac{1}{2}$ and assume $k \geq \frac{1}{\epsilon^2}$. Then,

$$\text{Prob}\left[s < s_0(1-\epsilon)\right] \leq \frac{18\sqrt{2}e^{10/21}}{\sqrt{\pi}} \cdot e^{-\left(\frac{k}{4}\right)\epsilon^2}.$$

Proof. To find the upper bound of $\text{Prob}\left[s < s_0(1-\epsilon)\right]$, we first consider $f(s_0(1-x))$ from

Expression (3.24). Observe that

$$f(s_0(1-x)) = f(s_0)(1-x)^{\frac{k-2}{2}}\left(1+\frac{s_0}{1-s_0}x\right)^{\frac{d-k-2}{2}}.$$

Taking the log of the expression $(1-x)^{\frac{k-2}{2}}\left(1+\frac{s_0}{1-s_0}x\right)^{\frac{d-k-2}{2}}$ and using Taylor's expansion of $\log x$, we obtain

$$\begin{aligned}\log\left((1-x)^{\frac{k-2}{2}}\left(1+\frac{s_0}{1-s_0}x\right)^{\frac{d-k-2}{2}}\right) &\leq \frac{k-2}{2}\left[-x-\frac{x^2}{2}\right]+\frac{d-k-2}{2}\left[\frac{k}{d-k}x\right]\\ &\leq x-\left(\frac{k-2}{4}\right)x^2 \leq \frac{3}{2}-\left(\frac{k}{4}\right)x^2,\end{aligned}$$

for $\epsilon \leq x \leq 1$. Hence, for $\epsilon \leq x \leq 1$, we have

$$f(s_0(1-x)) \leq f(s_0)\cdot e^{3/2}\cdot e^{-\left(\frac{k}{4}\right)x^2} \leq f(s_0)\cdot e^{3/2}\cdot e^{-\left(\frac{k}{4}\right)\epsilon x}. \tag{3.31}$$

Substituting Inequality (3.31) into the integral $\int_\epsilon^1 f(s_0(1-x))dx$ and expanding the bounds of integration, we get

$$\begin{aligned}\int_\epsilon^1 f(s_0(1-x))dx &\leq f(s_0)e^{3/2}\int_\epsilon^1 e^{-\left(\frac{k}{4}\right)\epsilon x}dx\\ &\leq f(s_0)e^{3/2}\int_\epsilon^\infty e^{-\left(\frac{k}{4}\right)\epsilon x}dx\\ &\leq f(s_0)\frac{4e^{3/2}}{\sqrt{k}}\cdot e^{-\frac{k}{4}\epsilon^2},\end{aligned}$$

where the last inequality holds since $k \geq \frac{1}{\epsilon^2}$. Substituting the above result and the upper bound of $Bs_0f(s_0)$ from Lemma 6 into Expression (3.24), we obtain the following upper bound:

$$Bs_0\cdot\int_\epsilon^1 f(s_0(1-x))dx \leq Bs_0f(s_0)\frac{4e^{3/2}}{\sqrt{k}}\cdot e^{-\frac{k}{4}\epsilon^2} \leq \frac{18\sqrt{2}e^{10/21}}{\sqrt{\pi}}\cdot e^{-\frac{k}{4}\epsilon^2},$$

which completes the proof. □

Hence, we have a lower and upper bound for Prob $[s < s_0(1-\epsilon)]$. We collect Lemmas 9 and 10 into the following:

Theorem 20. Let $0 < \epsilon < \frac{1}{2}$ and assume $k \geq \frac{1}{\epsilon^2}$. Then,

$$\frac{e^{-2}}{2\sqrt{\pi}} \cdot e^{-\frac{1}{4}\left(\frac{(\sqrt{k}\epsilon+1)^2}{1-s_0}+2(\sqrt[3]{k}\epsilon+k^{-1/6})^3\right)} \leq \text{Prob}\,[s < s_0(1-\epsilon)] \leq \frac{18\sqrt{2}e^{10/21}}{\sqrt{\pi}} \cdot e^{-\left(\frac{k}{4}\right)\epsilon^2}.$$

3.2 An Explicit Threshold

In this section, we prove our main theorem, Theorem 2. First, we provide an explicit threshold $k_1 = k_1(\epsilon, \delta)$ such that for any $w \in S^{d-1}$, there exists a JL transformation $A : \mathbb{R}^d \to \mathbb{R}^k$ for $k > k_1$. That is, A preserves the Euclidean norm of $w \in S^{d-1}$ for a given error factor ϵ and probability of failure δ. We then provide an explicit threshold $k_2 = k_2(\epsilon, \delta)$ such that for any $w \in S^{d-1}$, there does not exist a JL transformation $A : \mathbb{R}^d \to \mathbb{R}^k$ for $k < k_2$.

Let $s_0 := \frac{k}{d}$ and $s \in [0,1]$, as defined in Section 3.1.1. To obtain the threshold k_0, we find bounds on the probabilitites

$$\text{Prob}\,[s < s_0(1-\epsilon)] \quad \text{and} \quad \text{Prob}\,[s > s_0(1+\epsilon)]$$

employing the results of Corollary 1.

3.2.1 Lower Bound Guaranteeing a Distribution

To obtain the threshold such that there is a JL transformation $A : \mathbb{R}^d \to \mathbb{R}^k$ for all k above that point, we use the construction provided by Gupta and Dasgupta in [13]. Employing the bounds given in Theorems 19 and 20, we find the possible dimensions k for which it is a JL distribution, and hence, providing an upper bound on the smallest possible projected dimension k_0 for any arbitrary JL distribution.

Theorem 21. If $k > 4\epsilon^{-2} \log\left(\frac{1}{\delta}\right)[1 + o(1)]$, then there exists a distribution D on $k \times d$ real matrices such that for any $w \in S^{d-1}$

$$\text{Prob}_{A \sim \mathcal{D}}\left[|\|Aw\|^2 - 1| < \epsilon\right] \geq 1 - \delta.$$

The term $o(1)$ is dependent on the error factor ϵ and probability of failure δ, and approaches zero as ϵ and δ approach zero.

Proof. Let $A : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ be a linear transformation such that $A = \sqrt{\frac{1}{s_0}}\Sigma V^\top$, where

$$\Sigma = \left(\begin{array}{c|c} I_k & 0 \end{array}\right)_{k \times d}$$

and V is orthonormal. Observe that orthonormal transformations are both measure and norm invariant. That is, for $w \in S^{d-1}$, uniform random, and an orthonormal matrix $B \in \mathbb{R}^{d \times d}$, the projection $Bw \in S^{d-1}$, uniform random, and $\|Bw\| = \|w\|$. Hence, for uniform random $w \in S^{d-1}$, we have $V^\top w \in S^{d-1}$, uniform random. Let $x := V^\top w \in S^{d-1}$, uniform random. Then,

$$\|Aw\|^2 = \|U\Sigma x\|^2 = \|\Sigma x\|^2 = \frac{1}{s_0}\sum_{i=1}^{k} x_i^2,$$

resulting in

$$\text{Prob}\left[|\|Aw\|^2 - 1| > \epsilon\right] = \text{Prob}\left[\left|\frac{1}{s_0}\sum_{i=1}^{k} x_i^2 - 1\right| > \epsilon\right].$$

Let $s = \sum_{i=1}^{k} x_i^2$. Then, $s \in [0, 1]$ and the probability above becomes

$$\text{Prob}\left[|\|Aw\|^2 - 1| > \epsilon\right] = \text{Prob}\left[s < s_0(1 - \epsilon) \cup s > s_0(1 + \epsilon)\right].$$

Replacing $\text{Prob}\left[s < s_0(1-\epsilon)\right]$ and $\text{Prob}\left[s > s_0(1+\epsilon)\right]$ with the upper bounds provided in Theorem 19 and Theorem 20, we obtain

$$\begin{aligned}\text{Prob}\left[\left|\|Aw\|^2 - 1\right| > \epsilon\right] &\leq \frac{18\sqrt{2}e^{10/21}}{\sqrt{\pi}} e^{-\frac{k}{4}\epsilon^2} + \frac{27e^{-43/42}}{\sqrt{2\pi}} \cdot e^{-\frac{k-2}{4}\epsilon^2\left(1-\frac{2}{3}\epsilon\right)} \\ &\leq \left[\frac{18\sqrt{2}e^{10/21}}{\sqrt{\pi}} + \frac{27e^{-43/42}}{\sqrt{2\pi}}\right] e^{-\frac{k-2}{4}\epsilon^2\left(1-\frac{2}{3}\epsilon\right)} \\ &\leq 27e^{-\frac{k-2}{4}\epsilon^2\left(1-\frac{2}{3}\epsilon\right)}\end{aligned}$$

Observe that $27e^{-\frac{k-2}{4}\epsilon^2\left(1-\frac{2}{3}\epsilon\right)} < \delta$ when

$$k > 4\epsilon^{-2}\log\left(\frac{1}{\delta}\right)\left[1 + \frac{2\epsilon}{3-2\epsilon} + \frac{\log(27)}{\log\left(\frac{1}{\delta}\right)} \cdot \frac{1}{1-2\epsilon/3} + \frac{2\epsilon^2}{4\log\left(\frac{1}{\delta}\right)}\right] = 4\epsilon^{-2}\log\left(\frac{1}{\delta}\right)\left[1 + o(1)\right],$$

where $o(1) \longrightarrow 0$ as $\epsilon, \delta \to 0$. Hence, for $k > 4\epsilon^{-2}\log\left(\frac{1}{\delta}\right)\left[1 + o(1)\right]$, there is a distribution of transformations such that for $0 < \epsilon < \frac{1}{2}$ and $\delta > 0$,

$$\text{Prob}\left[\left|\|Aw\|^2 - 1\right| < \epsilon\right] \geq 1 - \delta. \qquad \square$$

3.2.2 Lower Bounds for Arbitrary Projection

We now provide an explicit threshold for the projected dimension k, under which there does not exist a JL transformation. Let $\mathcal{D}$ be an arbitrary JL distribution. Then, for any $w \in S^{d-1}$,

$$\text{Prob}_{A\sim\mathcal{D}}\left[\left|\|Aw\|^2 - 1\right| > \epsilon\right] < \delta,$$

which implies that the following inequality must hold:

$$\text{Prob}_{A\sim\mathcal{D}}\left[\text{Prob}_{w\in S^{d-1}}\left[\left|\|Aw\|^2 - 1\right| > \epsilon\right]\right] < \delta.$$

So, if, for all $A \in \mathbb{R}^{k \times d}$, $\text{Prob}_{w \in S^{d-1}}\left[|\|Aw\|^2 - 1| > \epsilon\right] > \delta$, then we have the result that there does not exist a distribution $\mathcal{D}$ such that

$$\text{Prob}_{A \sim \mathcal{D}}\left[|\|Aw\|^2 - 1| > \epsilon\right] < \delta$$

for any $w \in S^{d-1}$. We show for $k < 4\epsilon^{-2}\log\left(\frac{1}{\delta}\right)[1 - o(1)]$,

$$\text{Prob}_{w \in S^{d-1}}\left[|\|Aw\|^2 - 1| > \epsilon\right] > \delta$$

for any $A \in \mathbb{R}^{k \times d}$.

The layout of this section is as follows: we let $A : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ be a fixed arbitrary linear transformation and let w be random. We then find a lower bound on the probability $\text{Prob}\left[|\|Aw\|^2 - 1| > \epsilon\right]$ and find the values of k for which this lower bound is greater than δ. As a result, we obtain the values of k for which there does not exist a JL distribution.

Theorem 22. Let $A : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ be a linear transformation with $d > 2k$ and let $0 < \epsilon < \frac{1}{2}$. Then, for a random $w \in S^{d-1}$,

$$\text{Prob}\left[|\|Aw\|^2 - 1| > \epsilon\right] \geq \frac{e^{-2}}{2\sqrt{\pi}} \cdot e^{-\frac{1}{4}\left(\frac{(\sqrt{k}\epsilon+1)^2}{1-s_0} + 2(\sqrt[3]{k}\epsilon+1)^3\right)}.$$

Proof. Let $A : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ be a linear transformation. We may assume A is onto; otherwise, we may change the image accordingly. Let $S = U\Sigma V^\top$ be the singular value decomposition of S. Recall that U and V are orthonormal matrices and Σ is a diagonal matrix. Recall that orthonormal transformations are both measure and norm invariant. That is, for $w \in S^{d-1}$, uniform random, and an orthonormal matrix $B \in \mathbb{R}^{d \times d}$, we have $Bw \in S^{d-1}$, uniform random, and $\|Bw\| = \|w\|$.

Let $w \in S^{d-1}$. Then, since V is an orthonormal matrix,

$$V^\top w = \begin{pmatrix} v_1^\top w \\ \vdots \\ v_d^\top w \end{pmatrix} \in S^{d-1},$$

uniform random, where v_i is the i^{th} column of V. Let $x := V^\top w \in S^{d-1}$, uniform random. Then, as U is orthonormal,

$$\|Aw\|^2 = \|U\Sigma x\|^2 = \|\Sigma x\|^2 = \lambda_1^2 x_1^2 + \cdots + \lambda_k^2 x_k^2,$$

where λ_i's are the diagonal entries of Σ. Let $\langle H \rangle$ denote the set of elements generated by a set H. Let $W^\perp$ denote the orthogonal complement of subspace W. We observe that

$$\ker(A) = \langle v_{k+1}, \ldots, v_d \rangle \quad \text{and} \quad (\ker(A))^\perp = \langle v_1, \ldots, v_k \rangle,$$

where v_i is the i^{th} column of V. Let $K := \ker(A)$. Then, for $w \in S^{d-1}$, w can be decomposed into the components $w_k \in K^\perp$ and $w_{d-k} \in K$, where w_k and w_{d-k} can be represented by their length and a unit vector in their corresponding space. In particular, since $\|w\| = 1$,

$$w_k = \sqrt{s} \cdot \Omega_k \quad \text{and} \quad w_{d-k} = \sqrt{1-s} \cdot \Omega_{d-k},$$

where $s \in [0,1]$, $\Omega_k \in S^{k-1}$ and $\Omega_{d-k} \in S^{d-k-1}$. Since w_{d-k} lies in the kernel of A, it follows that

$$\|Aw\|^2 = s\|A\Omega_k\|^2 = \lambda_1^2 x_1^2 + \cdots + \lambda_k^2 x_k^2.$$

As a result, $\|A\Omega_k\|^2 = \frac{\|\Sigma x\|^2}{s}$ and

$$\text{Prob}\left[|\|Aw\|^2 - 1| > \epsilon\right] = \text{Prob}\left[|s \cdot \|A\Omega_k\|^2 - 1| > \epsilon\right].$$

To find a lower bound for Prob $\left[|s\|A\Omega_k\|^2 - 1| > \epsilon\right]$, we first consider the probability

$$\text{Prob}\left[|s \cdot C - 1| > \epsilon\right]$$

for some constant $C > 0$. From Corollary 1,

$$\text{Prob}\left[|s \cdot C - 1| > \epsilon\right] = B \cdot \int_{s \notin \left(\frac{1-\epsilon}{C}, \frac{1+\epsilon}{C}\right)} f(s)ds.$$

Let $\widehat{C} = \frac{1}{s_0} = \frac{d}{k} > 0$. Then, we have two cases: $\widehat{C} \leq C$ or $\widehat{C} \geq C$. If $\widehat{C} \leq C$,

$$\text{Prob}\left[s \notin \left(\frac{1-\epsilon}{C}, \frac{1+\epsilon}{C}\right)\right] \geq \text{Prob}\left[\frac{1+\epsilon}{C} < s \leq 1\right] \geq \text{Prob}\left[s_0(1+\epsilon) < s \leq 1\right].$$

Replacing Prob $[s \in (s_0(1+\epsilon), 1]]$ with its lower bound from Theorem 19, we have

$$\text{Prob}\left[s \notin \left(\frac{1-\epsilon}{C}, \frac{1+\epsilon}{C}\right)\right] \geq \frac{e^{-2}}{4\sqrt{\pi}} \cdot e^{-\frac{1}{4}(\sqrt{k}\epsilon+1)^2 \frac{1+s_0}{1-s_0}}.$$

Similarly, if $\widehat{C} \geq C$,

$$\text{Prob}\left[s \notin \left(\frac{1-\epsilon}{C}, \frac{1+\epsilon}{C}\right)\right] \geq \text{Prob}\left[0 \leq s < \frac{1-\epsilon}{C}\right] \geq \text{Prob}\left[0 \leq s < s_0(1-\epsilon)\right].$$

Replacing Prob $[s \in [0, s_0(1-\epsilon))]$ with its lower bound from Theorem 20, we have

$$\begin{aligned} \text{Prob}\left[s \notin \left(\frac{1-\epsilon}{C}, \frac{1+\epsilon}{C}\right)\right] &\geq \frac{e^{-2}}{2\sqrt{\pi}} \cdot e^{-\frac{1}{4}\left(\frac{(\sqrt{k}\epsilon+1)^2}{1-s_0} + 2(\sqrt[3]{k}\epsilon + k^{-1/6})^3\right)} \\ &\geq \frac{e^{-2}}{2\sqrt{\pi}} \cdot e^{-\frac{1}{4}\left(\frac{(\sqrt{k}\epsilon+1)^2}{1-s_0} + 2(\sqrt[3]{k}\epsilon + 1)^3\right)}. \end{aligned}$$

Taking the minimum of both cases, we have for any $C > 0$,

$$\text{Prob}\left[|s \cdot C - 1| > \epsilon\right] \geq \min\left\{\frac{e^{-2}}{2\sqrt{\pi}} \cdot e^{-\frac{1}{4}\left(\frac{(\sqrt{k}\epsilon+1)^2}{1-s_0} + 2(\sqrt[3]{k}\epsilon+1)^3\right)}, \frac{e^{-2}}{4\sqrt{\pi}} \cdot e^{-\frac{1}{4}(\sqrt{k}\epsilon+1)^2 \frac{1+s_0}{1-s_0}},\right\}.$$

Hence, as $s \in [0, 1]$, chosen independent random, and $\Omega_k \in S^{k-1}$, chosen independent random,

$$\text{Prob}\left[|\|Aw\|^2-1|>\epsilon\right]\geq\min\left\{\frac{e^{-2}}{2\sqrt{\pi}}\cdot e^{-\frac{1}{4}\left(\frac{(\sqrt{k}\epsilon+1)^2}{1-s_0}+2(\sqrt[3]{k}\epsilon+1)^3\right)},\frac{e^{-2}}{4\sqrt{\pi}}\cdot e^{-\frac{1}{4}(\sqrt{k}\epsilon+1)^2\frac{1+s_0}{1-s_0}},\right\}.$$

□

As a result of Theorem 22, we have a threshold for the projected dimension, under which there does not exist a JL distribution of linear transformations.

Corollary 2. Let $A:\mathbb{R}^d\to\mathbb{R}^k$ be a linear transformation, and $\delta>0$, and $0<\epsilon<\frac{1}{2}$. Then, for $k<4\epsilon^{-2}\log\left(\frac{1}{\delta}\right)[1-o(1)]$,

$$\text{Prob}_{w\in S^{d-1}}\left[|\|Aw\|^2-1|>\epsilon\right]>\delta.$$

Proof. Let $A:\mathbb{R}^d\to\mathbb{R}^k$ be a linear transformation. From Theorem 22, we have for random $w\in S^{d-1}$,

$$\text{Prob}\left[|\|Aw\|^2-1|>\epsilon\right]\geq\min\left\{\frac{e^{-2}}{2\sqrt{\pi}}\cdot e^{-\frac{1}{4}\left(\frac{(\sqrt{k}\epsilon+1)^2}{1-s_0}+2(\sqrt[3]{k}\epsilon+1)^3\right)},\frac{e^{-2}}{4\sqrt{\pi}}\cdot e^{-\frac{1}{4}(\sqrt{k}\epsilon+1)^2\frac{1+s_0}{1-s_0}},\right\}.$$

Suppose $k<\eta\epsilon^{-2}\log\frac{1}{\delta}$ where $\eta<4$. Then, in the case when the minimum is $\frac{e^{-2}}{2\sqrt{\pi}}\cdot e^{-\frac{1}{4}\left(\frac{(\sqrt{k}\epsilon+1)^2}{1-s_0}+2(\sqrt[3]{k}\epsilon+1)^3\right)}$,

$$\begin{aligned}\text{Prob}\left[|\|Aw\|^2-1|>\epsilon\right]&\geq\frac{e^{-2}}{2\sqrt{\pi}}\exp\left[-\frac{1}{4}\left(\frac{(\sqrt{\eta\log\frac{1}{\delta}}+1)^2}{1-s_0}+2\left(\sqrt[3]{\eta\log\frac{1}{\delta}}\epsilon^{1/3}+1\right)^3\right)\right]\\&\geq\frac{e^{-2}}{2\sqrt{\pi}}\exp\left[-\frac{1}{4}\eta\log\frac{1}{\delta}\cdot\left(\frac{\left(1+\frac{1}{\sqrt{\eta\log\frac{1}{\delta}}}\right)^2}{1-s_0}+2\left(\epsilon^{1/3}+\frac{1}{\sqrt[3]{\eta\log\frac{1}{\delta}}}\right)^3\right)\right]\\&=\frac{e^{-2}}{2\sqrt{\pi}}\delta^{\frac{\eta}{4}\cdot\gamma},\end{aligned}$$

where $\gamma = \left(\frac{\left(1+\frac{1}{\sqrt{\eta \log \frac{1}{\delta}}}\right)^2}{1-s_0} + 2\left(\epsilon^{1/3} + \frac{1}{\sqrt[3]{\eta \log \frac{1}{\delta}}}\right)^3 \right)$. We observe that γ approaches 1 from above as ϵ, δ, and s_0 approach zero. We note that when $d \geq \epsilon^{-3} \log \frac{1}{\delta}$, then γ approaches 1 from above as ϵ and δ approach zero.

In the case when the minimum is $\frac{e^{-2}}{4\sqrt{\pi}} \cdot e^{-\frac{1}{4}(\sqrt{k}\epsilon+1)^2 \frac{1+s_0}{1-s_0}}$,

$$\begin{aligned} \text{Prob}\left[|\|Aw\|^2 - 1| > \epsilon\right] &\geq \frac{e^{-2}}{4\sqrt{\pi}} \exp\left[-\frac{1}{4}\eta \log \frac{1}{\delta} \cdot \left(\frac{1}{\sqrt{\eta \log \frac{1}{\delta}}} + 1\right)^2 \left(\frac{1+s_0}{1-s_0}\right)\right] \\ &= \frac{e^{-2}}{4\sqrt{\pi}} \delta^{\frac{\eta}{4} \cdot \gamma}, \end{aligned}$$

where $\gamma = \left(\frac{1}{\sqrt{\eta \log \frac{1}{\delta}}} + 1\right)^2 \left(\frac{1+s_0}{1-s_0}\right)$. We observe that γ approaches 1 from above as δ and s_0 approach zero.

Hence, in either case,

$$\frac{e^{-2}}{2\sqrt{\pi}} \delta^{\frac{\eta}{4} \cdot \gamma} > \delta \tag{3.32}$$

for sufficiently small ϵ, δ, and s_0. That is, Prob $\left[|\|Aw\|^2 - 1| > \epsilon\right] \geq \delta$ for sufficiently small δ, ϵ and s_0. We note as η approaches 4, ϵ and δ must approach zero for Inequality (3.32) to hold.

□

Consequently, we have a lower bound, $4\epsilon^{-2} \log\left(\frac{1}{\delta}\right)[1 - o(1)]$, for the smallest possible projected dimension. That is,

Corollary 3. If $k < 4\epsilon^{-2} \log\left(\frac{1}{\delta}\right)[1 - o(1)]$, then there is no distribution $\mathcal{D}$ on $k \times d$ real matrices such that

$$\text{Prob}_{A \sim \mathcal{D}}\left[|\|Aw\|^2 - 1| < \epsilon\right] \geq 1 - \delta.$$

Proof. Suppose there did exist such a distribution. Then, it follows that

$$\text{Prob}_{A \sim \mathcal{D}}\left[\text{Prob}_{w \in S^{d-1}}\left[|\|Aw\|^2 - 1| > \epsilon\right]\right] < \delta.$$

However, from Corollary 2, we have for $k < 4\epsilon^{-2} \log\left(\frac{1}{\delta}\right)[1 - o(1)]$,

$$\text{Prob}\left[|\|Aw\|^2 - 1| > \epsilon\right] > \delta. \qquad \square$$

Combining Corollary 3 and Theorem 21, we have the following theorem:

Theorem 17. Let $0 < \epsilon, \delta < \frac{1}{2}$.

a.) If $k > 4\epsilon^{-2} \log\left(\frac{1}{\delta}\right)[1 + o(1)]$, then there exists a JL distribution.

The term $o(1)$ is dependent on the error factor ϵ and probability of failure δ, and approaches zero as ϵ and δ approach zero. The exact term $o(1)$ may be found in the proof of Theorem 21.

b.) If $k < 4\epsilon^{-2} \log\left(\frac{1}{\delta}\right)[1 - o(1)]$, then there does not exist a JL distribution.

The term $o(1)$ is dependent on the error factor ϵ, probability of failure δ, and the ratio $\frac{k}{d}$, and approaches zero as ϵ, δ and the ratio $\frac{k}{d}$ approach zero.

Consequently, we have our main result as stated in Theorem 2.

Theorem 2. For ϵ and δ sufficiently small, $k_0 \approx 4\epsilon^{-2} \log \frac{1}{\delta}$. That is,

$$\frac{k_0}{4\epsilon^{-2} \log \frac{1}{\delta}} \to 1 \text{ as } \epsilon, \delta \to 0.$$

Chapter 4

Applications

4.1 Approximate Nearest Neighbors

The nearest neighbor (NN) problem can be summarized by the following: given a set P of points and a query point in a d-dimensional space, find a point in set P closest to the query point. To speed up the search process, the problem is relaxed to the approximate nearest neighbor (ANN) search, which finds the closest point up to some $\epsilon > 0$. More formally,

Definition 6. Given a metric space (X, d_X), a finite collection of points $P \subset X$, and a query point $x \in X$, an ϵ-ANN search finds a point $p \in P$ satisfying the inequality

$$d_X(x,p) \leq (1+\epsilon) d_X(x,p')$$

for all $p' \in P$.

Due to numerous applications of ANN, such as pattern recognition, computer vision, and coding theory, there have been many publications on the subject [6, 10, 24, 32, 38, 48]. In [32], Kushilevitz, Ostrovsky, and Rabani implemented low dimensional embedding in their ANN algorithm to obtain a point in $\mathcal{O}(d\epsilon^{-2} \log d \log n)$ time for any data set where n is the number of points and d is the original dimension. To improve the time to $\mathcal{O}(d \log d +$

$\epsilon^{-3}\log n)$ time while using $n^{\mathcal{O}(\epsilon^{-2})}$ storage, Ailon and Chazelle [2] integrated JL distributions, specifically FJLT, into their algorithm.

Their algorithm consists of two stages, the second of which utilizes the properties of FJLT [2]. As a brief overview, stage one runs an $\mathcal{O}(n)$-ANN query and returns a point q. Then, stage two returns the closest point to x by answering an $\mathcal{O}(\epsilon)$-ANN query restricted to a set $P_x \subset P$ dependent on the point q found in stage one. In particular, both stages can be defined as follows.

Stage 1:

With input $P \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$, the algorithm chooses a random vector $v \in \mathbb{R}^d$ and returns the nearest neighbor with respect to pseudometric

$$D_v(x,p) = |v^\top x - v^\top p|.$$

The returned point q is an answer to an $\mathcal{O}(n)$-ANN for query x. To increase the probability of success to $1-\delta$, the procedure must be repeated $\mathcal{O}(\log \delta^{-1})$ times.

Stage 2:

Let q be the point returned in stage one. Let $P_x \subset P$ be defined as

$$P_x = P \cap B_2\left(q, 2\|x-q\|_2\right),$$

where $B_2(s,t)$ is the set of points within ℓ_2-distance t of point s. Let $R(P_x)$ denote the largest possible distance between query x and a point in P_x. The algorithm first applies a JL transform to query x and the set P_x, and then, searches for a point $p \in P_x$ such that

$$d(x,p) \leq (1+\epsilon)\ell, \tag{4.1}$$

where $\ell = R(P_x)$. If the search returns $p \in P$ such that (4.1) holds, the constant ℓ is decreased and the process is repeated.

Consequently, we have the following theorem given by Ailon and Chazelle:

Theorem 23 ([2])**.** Given a set P of n points in ℓ_2^d, for any $\epsilon > 0$, there is a randomized data structure of size $n^{\mathcal{O}(\epsilon^{-2})}$ that can answer any ϵ-ANN query in time $\mathcal{O}(d \log d + \epsilon^{-3} \log n)$ with high probability.

4.2 Linear Algebra Applications

Embedding data into a lower dimension has applications in linear algebra techniques such as matrix multiplication, linear ℓ_2 regression, and rank k approximation.

4.2.1 Matrix Multiplication

A common method to increase the speed of matrix multiplication is to approximate the product by a sampling process [17]. Sampling approximates the product of two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ by constructing submatrices $\hat{A} \in \mathbb{R}^{m \times r}$ and $\hat{B} \in \mathbb{R}^{r \times p}$ from A and B, respectively, and computing the product of $\hat{A}\hat{B}$. Let

$$p_k = \frac{\|A^{(k)}\|_2 \|B_{(k)}\|_2}{\sum_{k'=1}^{n} \|A^{(k')}\|_2 \|B_{(k')}\|_2}$$

for $1 \leq k \leq n$ be the sampling probability for index k, where $A^{(i)}$ and $B_{(j)}$ are the i^{th} column of A and j^{th} row of B respectively. Let $1 \leq i_1 < \cdots < i_r \leq n$ be r indices with the largest sampling probabilities. Then, $\hat{A}$ consists of the columns $A^{(i_1)}, \ldots, A^{(i_r)}$ and $\hat{B}$ consists of the corresponding r rows $B_{(i_1)}, \ldots, B_{(i_r)}$.

Deviating from the common process of sampling, Sarlós in [44] implemented JL distributions in approximating matrix multiplication, further reducing the computational complexity. In Lemma 6 of [44], Sarlós showed for $A \in \mathbb{R}^{n \times m}$, $B \in \mathbb{R}^{n \times p}$, and a JL distribution $\mathcal{D}$,

$$\text{Prob}_{S \sim \mathcal{D}} \left[\|AB - AS^\top SB\|_F \leq \epsilon \|A\|_F \|B\|_F \right] \geq 1 - \delta \tag{4.2}$$

by observing for a given JL distribution $\mathcal{D}$ and any set $V \subset \mathbb{R}^d$ of n points,

$$\text{Prob}_{S\sim\mathcal{D}}\left[\langle u - v\rangle - \epsilon\|u\|_2\|v\|_2 \leq \langle Su, Sv\rangle \leq \langle u - v\rangle + \epsilon\|u\|_2\|v\|_2\right] \geq 1 - \delta$$

for all pairs $u, v \in V$. Hence, by combining (4.2) and Theorem 3, approximations

$$\hat{A} = AS^\top \quad \text{and} \quad \hat{B} = SB$$

can be found such that

$$\text{Prob}_{S\sim\mathcal{D}}\left[\|AB - \hat{A}\hat{B}\|_F \leq \epsilon\|A\|_F\|B\|_F\right] \geq 1 - \delta.$$

Approximations $\hat{A}$ and $\hat{B}$ can be found with a one-pass algorithm using $\mathcal{O}\left(\epsilon^{-2}(m+p)\log(m+p)\log\frac{1}{\delta}\right)$ space and $\mathcal{O}\left(\epsilon^{-2}M\log(m+p)\log\frac{1}{\delta}\right)$, where M is the combined number of nonzero entries in A and B.

4.2.2 Linear Regression

For $A \in \mathbb{R}^{n\times d}$, $n > d$, and $b \in \mathbb{R}^d$, linear ℓ_2 regression finds an optimal vector x_{opt} minimizing $\|Ax - b\|_2$. Often, x_{opt} is given by A^+b, where $A^+ = V\Sigma^{-1}U^\top$ is the Moore Penrose matrix and $U\Sigma V^\top$ is the SVD of A. Analogous to matrix multiplication, sampling based on sampling probabilities is used to approximate and speed up the computation of regression. Drineas et al. [16] showed if k rows of A and b are sampled according to the sampling probabilities, where $k = poly(\epsilon^{-1}, d)$, the lower-dimensional problem gives an ϵ-approximation to the original regression problem. In [44], Sarlós replaced sampling with random embeddings implementing FJLT (Section 2.3.1). In particular,

Theorem 24 (Theorem 12, [44])**.** Suppose $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$. Let $\mathcal{Z} = \min_{x\in\mathbb{R}^d}\|b - Ax\|_2 = \|b - Ax_{opt}\|_2$, where $x_{opt} = A^+b$. Let $0 < \epsilon < 1$, $S \in \mathbb{R}^{r\times n}$ be a transform from a JL distribution and $\tilde{\mathcal{Z}} = \min_{x\in\mathbb{R}^d}\|Sb - SAx\|_2 = \|Sb - SA\tilde{x}_{opt}\|_2$, where $\tilde{x}_{opt} = (SA)^+Sb$.

If $r = \Omega(\epsilon^{-2} d \cdot \log d)$, then

$$\text{Prob}\left[\|x_{opt} - \tilde{x}_{opt}\|_2 \leq \frac{\epsilon}{\sigma_{min}(A)} \mathcal{Z}\right] \geq \frac{1}{3},$$

where $\sigma_{min}(A)$ is the first diagonal entry of Σ such that $A = U\Sigma V^\top$.

Computing $\tilde{x}_{opt}$ via the FJLT takes $\mathcal{O}\left(nd \log n + d^2\epsilon^{-2}(d + \log^2 n) \log d\right)$ time. Observe that repeating the process $\log \frac{1}{\delta}$ times provides the probability of at least $1 - \delta$.

4.2.3 Rank k-Approximation

Rank k-approximation is a minimization problem, which seeks to find a rank k matrix A_k for a given matrix $A \in \mathbb{R}^{m \times n}$ such that $\|A - A_k\|$ is minimized over all rank k matrices, where $\|\cdot\|$ is either the Frobenius norm or spectral norm. Some applications include facial recognition, web search, latent semantic indexing, text analysis, and lossy data compression. Let $A = U\Sigma V^\top$ be the SVD of A. By the Eckar Young Theorem, we have that the best rank k-approximation of A is

$$A_k = U_k \Sigma_k V_k^\top,$$

where U_k and V_k consist of the first k rows of U and V, respectively, and Σ_k is a diagonal matrix consisting of the first k diagonal entries of Σ. Hence, to speed up rank k-approximation, obtaining the SVD of a matrix must be fast. To speed up the process, SVD is relaxed to relative error SVD, which was first introduced by Har-Peled [25] in 2006. Shortly thereafter, Deshpande and Vampala in [14] presented an algorithm, which takes $\Theta(k \log k)$ passes to obtain a rank k-approximation achieving relative error

$$(1 + \epsilon)\|A - A_k\|_F$$

with probability at least $\frac{3}{4}$ in time $\mathcal{O}\left(M(\frac{k}{\epsilon} + k^2 \log k) + (m + n)(\frac{k^2}{\epsilon^2} + \frac{k^3 \log k}{\epsilon} + k^4 \log^2 k)\right)$. Sarlós in [44] cut the number of passes down to two by implementing a JL distribution, with uniform random ±1 entries such as the construction given by Achlioptas (Section 2.2.5). In particular, he proved the following theorem.

Theorem 25 (Theorem 14, [44])**.** Let $A \in R^{m \times n}$. Let $\Pi_V(A)$ denote the projection of A onto subspace $V \subset \mathbb{R}^m$ and $\Pi_{V,k}(A)$ denote the best rank-k approximation of $\Pi_V(A)$. Let $0 < \epsilon \leq 1$ and $r = \Theta(k/\epsilon + k \log k)$. If S is an r-by-n JL transform with i.i.d. zero-mean ± 1 entries, then

$$\text{Prob}\left[\|A - \Pi_{AS^\top,k}\|_F \leq (1+\epsilon)\|A - A_k\|_F\right] \geq \frac{1}{2}.$$

Computing the singular vectors spanning $\Pi_{AS^\top,k}(A)$ in two passes over the data requires $\mathcal{O}(Mr + (m+n)r^2)$ time and $\mathcal{O}((m+n)r^2)$ space, where M denotes the number of nonzero entries in A.

Observe the probability of success can be increased to $1 - \delta$ for a predetermined δ by repeating the process $\mathcal{O}(\log \frac{1}{\delta})$ times in parallel and choosing the instance with maximal Frobenius norm $\|\Pi_{AS^\top,k}\|_F^2$.

4.3 Machine Learning

4.3.1 Background

The goal of machine learning is to construct an algorithm that can independently learn from data and adapt accordingly to make accurate predictions on future data. Applications include email filtering, computer vision, optical character recognition, online marketing, and detection of network intruders. There are three main types of learning:

1. Supervised: In supervised learning, the learner is provided data labeled either positive or negative, from which, it must determine a way to accurately label future data. Typical tasks of supervised learning include classification, e.g., email filtering, and regression, e.g., prediction of future sales. For instance, given emails classified as spam or not spam, the email provider desires to find a way to properly classify future emails that is consistent with the labeled emails

2. Unsupervised: In unsupervised learning, the learner is not provided with any labeled data. The learner is expected to label the given data and determine a way to accurately label future examples. A typical task of unsupervised learning is that of clustering. In clustering, a data set is subdivided into groups without any prior knowledge of the groups.

3. Reinforcement: Reinforcement learning is predominantly used in gaming. In reinforcement, the goal is to make decisions that maximize reward in the allotted time given.

In learning, data is represented by vectors in $\mathbb{R}^d$ and labeled either positive, $+1$, or negative, -1. A concept is defined to be either a vector or a subset of vectors in $\mathbb{R}^d$, and is used by the learner to label the data. A concept class is a set of concepts. A common algorithm for classification is the Perceptron Algorithm and was presented by Rosenblatt [43] in 1962. If given a set of labeled vectors $x_i \in \mathbb{R}^d$ for $1 \leq i \leq m$, which can be separated by a hyperplane, the algorithm will return a vector $w \in \mathbb{R}^d$ such that

- $w \cdot x < 0$ for negative labeled vectors x, and
- $w \cdot x > 0$ for positive labeled vectors x.

The vector w is called a concept and is used in the labeling of future data according to the sign of the inner product: if $w \cdot x > 0$, x is labeled positive and otherwise, negative.

Algorithm 1. Perceptron Algorithm

```
Input: set $S$ of vectors $x_i \in \mathbb{R}^d$ with true label $y_i \in \{\pm 1\}$ for $1 \leq i \leq m$
Set $w = 0 \in \mathbb{R}^d$
For some fixed number of iterations or until all labels are correct:
    Pick a vector $x_i \in S$
        $\hat{y} = \text{sgn}(w \cdot x_i)$
        If $\hat{y} \neq y_i$:
            $w \leftarrow w + y_i x_i$
Output: concept $w$
```

A learning algorithm, such as the Perceptron Algorithm, that constructs a concept w from given labeled vectors is said to (ϵ, δ)-learn if w classifies at least $(1-\epsilon)$ fraction of the data distribution $\mathcal{E}$ with probability at least $1-\delta$. A lower bound to the number of labeled examples needed in order for a learning algorithm to (ϵ, δ)-learn was given by Kearns and Vazirani in [30].

Theorem 26 ([30])**.** Let C be any concept class in $\mathbb{R}^d$. Let w be a concept from C that is consistent with m labeled examples of some concept in C. Then, with probability at least $1-\delta$, w correctly classifies at least $(1-\epsilon)$ fraction of $\mathcal{E}$ provided

$$m > \frac{4}{\epsilon} \log C(2m, d) + \frac{4}{\epsilon} \log \frac{2}{\delta},$$

where $C(m, d)$ denotes the maximum number of distinct labelings of m points in $\mathbb{R}^d$ obtainable from a concept.

4.3.2 Implementation of JL Transformations in Machine Learning

To reduce the number of samples needed to determine a concept w, Arriaga and Vempala [5] implemented dimension reduction techniques. They restricted the case to that of robust concept classes.

Definition 7. For any real number $\ell > 0$, a concept class C in conjuction with a distribution $\mathcal{D}$ in $\mathbb{R}^d$ is said to be ℓ-robust if

$$\text{Prob}\left[x : \exists\, y : \text{label}(x) \neq \text{label}(y),\ \|x-y\|_2 \leq \ell\right] = 0$$

That is, the probability that there exists two points with different labels within distance ℓ of each other is zero. As a result, a few attributes of the samples can be altered without affecting a concept from a robust concept class.

Arriaga and Vempala [5] observed that robust target concepts are preserved when the data is projected to a lower dimension. Projecting the data to a lower dimension results in a smaller number of labeled vectors required and a smaller computational complexity. Projections implemented in [5] come from JL distributions given by Indyk and Motwani (Section 2.2.3) and Achlioptas (Section 2.2.5). Let $\mathcal{D}_1$ be the construction of Indyk and Motwani. That is, for $A \sim \mathcal{D}_1$, $a_{ij} \overset{iid}{\sim} N(0,1)$ for $1 \leq i \leq k$ and $1 \leq j \leq d$. Let $\mathcal{D}_2$ be the first construction of Achlioptas. That is, for $A \sim \mathcal{D}_2$, $A = \frac{1}{\sqrt{k}} B \in \mathbb{R}^{k \times d}$ where $b_{ij} \overset{iid}{\sim} \{\pm 1\}$.

Learning of a Half-Space:

We first focus on the well-studied problem, learning of a half-space, and restrict the vectors to the $(n-1)$-dimensional unit sphere, S^{n-1}. The algorithm presented in [5] to obtain concept w is as follows:

Algorithm 2. Half-Space Algorithm

Input: vectors $x_1, \ldots, x_m \in \mathbb{R}^d$ with true labels $y_1, \ldots, y_m \in \{\pm 1\}$

1. Choose $k \times d$ random matrix R from JL distribution $\mathcal{D}_1$ or $\mathcal{D}_2$
2. Project $x_1, \ldots, x_m \in \mathbb{R}^d$ to $\mathbb{R}^k$ by Rx_i for $1 \leq i \leq m$.
3. Run the Perceptron Algorithm in $\mathbb{R}^k$.

Output: JL transform R and concept w.

With output R and w, a future example x is labeled by projecting x to $\mathbb{R}^k$, Rx, and labeling x positive if $w \cdot Rx > 0$ and otherwise, negative.

Theorem 27 ([5])**.** For an ℓ-robust half-space, the algorithm will (ϵ, δ)-learn when given m samples in $n \cdot poly\left(\frac{1}{\ell}, \frac{1}{\epsilon}, \log \frac{1}{\delta}\right)$ time, where

$$k = \frac{100}{\ell^2} \ln \frac{100}{\epsilon \ell \delta} \quad \text{and} \quad m = \frac{8k}{\epsilon} \log \frac{48}{\epsilon} + \frac{4}{\epsilon} \log \frac{4}{\delta}.$$

Learning an Intersection of Half-Spaces:

In addition to the learning half-space problem, Arriaga and Vempala [5] applied JL transformations to learning an intersection of half-spaces. In this scenario, a concept is a set of vectors $Q = \{w_1, \ldots, w_t\}$ in $\mathbb{R}^d$. A vector is labeled positive if it lies in the intersection of the t half-spaces $h_1, \ldots, h_t$, where h_i is the half-space $\{x \in \mathbb{R}^d : w_i \cdot x > 0\}$ for $1 \leq i \leq t$, and negative if it lies outside the region. In [5], it is assumed the hyperplane defining the half-spaces contains the origin, i.e., the half-spaces are homogenous. The algorithm to obtain concept $Q = (w_1, \ldots, w_t)$ is similar to that of learning a half-space. It first projects the given set of labeled vectors to a smaller dimension using a JL transformation and then finds a concept in the lower dimension.

Algorithm 3. Intersection of t Half-Spaces Algorithm

Input: vectors $x_1, \ldots, x_m \in \mathbb{R}^d$ with true labels $y_1, \ldots, y_m \in \{\pm 1\}$

1. Choose $k \times d$ random matrix R from JL distribution $\mathcal{D}_1$ or $\mathcal{D}_2$
2. Project $x_1, \ldots, x_m \in \mathbb{R}^d$ to $\mathbb{R}^k$ by Rx_i for $1 \leq i \leq m$.
3. Find a concept $Q = \{w_1, \ldots, w_t\} \subset \mathbb{R}^k$ such that the intersection of the half-spaces $\{x \in \mathbb{R}^d : w_i \cdot x \geq 0\}$ for $1 \leq i \leq t$ is consistent with the given labels of the projected vectors.

Output: JL transform R and concept Q

With the output R and $Q = \{w_1, \ldots w_t\}$, a future vector $x \in \mathbb{R}^d$ is labeled by projecting x to $\mathbb{R}^k$, Rx, and labeling x positive if $w_i \cdot Rx > 0$ for all $i \in \{1, \ldots, t\}$ and otherwise, negative.

Theorem 28 ([5])**.** For an ℓ-robust intersection of t half-spaces, the algorithm will (ϵ, δ)-learn when given m samples in $\mathcal{O}(nmk) + \left(\frac{48t}{\epsilon} \log \frac{4t}{\epsilon\delta}\right)^{kt}$ time, where

$$k = \frac{100}{\ell^2} \ln \frac{100t}{\epsilon \ell \delta} \quad \text{and} \quad m = \frac{8kt}{\epsilon} \log \frac{48t}{\epsilon} + \frac{4}{\epsilon} \log \frac{4}{\delta}.$$

Those interested in the problem learning of balls are directed to [5]. We note the process is similar to that of learning a half-space and an intersection of t half-spaces.

4.4 Compressed Sensing

The goal of compressed sensing is to efficiently reconstruct a sparse signal from a limited number of measurements by exploiting the sparseness property. A recovery algorithm $\mathcal{R}$ satisfies a recovery guarantee called the ${}^{\ell_2}\!/\!_{\ell_1}$ guarantee if, given $\Phi \in \mathbb{C}^{k \times d}$,

$$\|R(\Phi x) - x\|_2 \leq \frac{c}{\sqrt{m}} \inf_{\substack{y \in \mathbb{C}^d \\ \|y\|_0 \leq m}} \|x - y\|_1$$

for all $x \in \mathbb{R}^d$ for some global constant $c > 0$ where $\|x\|_0 = m$. That is, the error between the recovered signal and the original signal is less than the scaled distance between the original signal and the best m-sparse approximation. A matrix Φ with the restricted isometry property (RIP) is sufficient for the ${}^{\ell_2}\!/\!_{\ell_1}$ guarantee.

Definition 8. Matrix $\Phi \in \mathbb{C}^{k \times d}$ has the (ϵ, m)-RIP if for all $x \in \mathbb{C}^d$ such that $\|x\|_0 \leq m$,

$$(1 - \epsilon)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \epsilon)\|x\|_2^2.$$

In particular, the ${}^{\ell_2}\!/\!_{\ell_1}$ guarantee is satisfied for $\Phi \in \mathbb{C}^{k \times d}$ if Φ is an (ϵ, m)-RIP for $\epsilon < \sqrt{2} - 1$ [11]. As the definitions of matrices with RIP and those from a JL distribution are similar, there has been effort to connect these two. In [7], Baraniuk, et al., showed a transform from a JL distribution is a $k \times d$ matrix with RIP for sparsity up to $m \leq {}^{c_1 \delta^2 k}\!/\!_{\log \frac{d}{m}}$. In particular, they prove the following theorem:

Theorem 29 (Theorem 5.2, [7])**.** Suppose that k, d and $0 < \epsilon < 1$ are given. If the probability distribution generating the $k \times d$ matrices Φ satisfies the concentration inequality

$$\text{Prob}\left[(1 - \epsilon)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \epsilon)\|x\|_2^2\right] \geq 1 - 2\exp\left(-c_0 \epsilon^2 k\right) \tag{4.3}$$

where $\epsilon = \delta$ and c_0 is an absolute constant, then there exists absolute constants c_1, c_2 such that with probability at least $1 - 2\exp\left(-c_0\delta^2 k\right)$, the RIP

$$(1-\delta)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\delta)\|x\|_2^2$$

holds for Φx with the prescribed δ and any $m \leq c_1 k/\log(d/m)$.

Observe that JL distributions satisfy (4.3), and as a result, Φ chosen from a JL distribution is a matrix with RIP. A variation on the converse was provided by Krahmer and Ward in [31].

Theorem 30 (Theorem 3.1, [31])**.** Fix $\delta > 0$ and $0 < \epsilon < 1$. Let S be an n-point set in $\mathbb{R}^d$. Set $m \geq 40 \log \frac{4n}{\eta}$, and suppose that $\Phi \in \mathbb{R}^{k \times d}$ satisfies the (η, m)-RIP where $\eta \leq \frac{\epsilon}{4}$. Let D be a $d \times d$ diagonal matrix consisting of diagonal entries uniformly at random chosen from $\{\pm 1\}$. Then for all $x \in S$,

$$\mathrm{Prob}_D\left[(1-\epsilon)\|x\|_2^2 \leq \|\Phi D x\|_2^2 \leq (1+\epsilon)\|x\|_2^2\right] \geq 1 - \delta.$$

That is, given a matrix Φ with (ϵ, m)-RIP, a JL distribution can be constructed by multiplying each of the columns of Φ by a random sign change.

Example 8. An example of a matrix with RIP is that with Gaussian or subgaussian entries. Consider $\Phi \in \mathbb{R}^{k\times d}$ such that $\Phi_i \overset{iid}{\sim} N(0,1)$. Then, a JL distribution $\mathcal{D}$ can be constructed from Φ with RIP in the following way: for $A \sim \mathcal{D}$, $A = \Phi D$ where $D \in \mathbb{R}^{d\times d}$ is a diagonal matrix with diagonal entries uniformly at random chosen from $\{\pm 1\}$.

Observe that the RIP matrices from Example 8 are dense and have computation time $\mathcal{O}(kd)$. An example of a RIP matrix allowing for computation $\mathcal{O}(d \log d)$ is a submatrix of a Hadamard or discrete Fourier transform.

Those interested in the recovery of a signal are encouraged to read [9, 15, 19, 23, 40, 39, 46].

4.5 Differential Privacy

One application of JL distributions discovered by Blocki et al. [8] is that of differential privacy.

Definition 9. We say matrices A and B are neighboring if the matrix $(A - B)$ has rank 1 and $\|A - B\|_1 \leq 1$. A probabilistic distribution $\mathscr{A}$ of maps gives (α, β)-differential privacy, if for all neighboring matrices A and B and all subsets S in the range of $\mathscr{A}$,

$$\text{Prob}_{f \sim \mathscr{A}}\left[f(A) \in S\right] \leq \exp(\alpha)\text{Prob}_{f \sim \mathscr{A}}\left[f(B) \in S\right] + \beta.$$

That is, for neighboring matrices A and B, a map f preserves differential privacy if $f(A)$ and $f(B)$ are indistinguishable. Blocki et al. [8] showed a JL distribution with random Gaussian entries as in Section 2.2.3 preserves differential privacy; whereas, Upadhyay showed in [47], that the sparse distributions defined by Nelson et al. [41], Kane and Nelson [29], Dasgupta et al. [12], and Ailon and Liberty [3] as given in Section 2.3 do not preserve differential privacy.

To prove these constructions do not preserve differential privacy, Upadhyay in [47] gave a counterexample and provided two neighboring matrices for which they are not differentially private.

Answering an open question, Upadhyay [47] provided a sparse JL distribution that preserves differential privacy. As done for other sparse constructions, the vector is first preconditioned by a randomized Walsh-Hadamard matrix HD, where H is the Walsh-Hadamard matrix as defined in (2.4) and D is a $d \times d$ diagonal matrix such that the diagonal entries are uniform random $\{\pm 1\}$. The entries are then permuted by a permutation matrix Π before being projected by a sparse matrix P. The sparse matrix P is constructed as follows:

1. Choose d random subgaussian samples g_i for $1 \leq i \leq d$ with mean 0 and variance 1.

2. Set $g = (g_1, \ldots, g_d)$ and divide g into blocks of d/k elements:

$$\phi_i = i^{th} \text{ block } = (g_{(i-1)d/k+1}, \ldots, g_{ik}).$$

3. Let P be the block diagonal matrix:

$$P = \begin{pmatrix} \phi_1 & & & \\ & \phi_2 & & \\ & & \ddots & \\ & & & \phi_k \end{pmatrix}.$$

Then, let $\mathcal{D}$ be a distribution on $k \times d$ matrices such that, for $A \sim \mathcal{D}$, $A = \frac{1}{k} P \Pi H D$, where both matrices P and D are random. Each matrix $A \in \mathcal{D}$ can be generated with $3d$ random samples, and computing Ax takes $\mathcal{O}(d \log d)$ time. Distribution $\mathcal{D}$ is shown in [47] to be differentially private for a restricted set of matrices.

Theorem 31 ([47])**.** If the singular values of a matrix $B \in \mathbb{R}^{m \times d}$ are at least $\frac{\ln\left(\frac{4}{\beta}\right)\sqrt{16k \log \frac{2}{\beta}}}{\alpha}$, then for $A \sim \mathcal{D}$, $AB^\top$ is $\left(\alpha, \beta + 2^{-\Omega((1-\epsilon)^2 d^{2/3})}\right)$-differentially private.

Bibliography

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput.Syst. Sci*, 66(4):671–687, 2003.

[2] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the 38th ACM Symposium on Theory of Computing (STOC)*, pages 557–563, 2006.

[3] Nir Ailon and Edo Liberty. Fast dimension reduction using Rademacher series on dual bch codes. *Discrete and Computational Geometry*, 42(4):615–630, 2009.

[4] Noga Alon. Problems and results in extremal combinatorics. *Discrete Mathematics*, 273:31–53, 2003.

[5] Rosa I. Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Mach Learn*, 63:161–182, 2006.

[6] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *J. ACM*, 45(6):891–923, 1998.

[7] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28:253–263, 2008.

[8] J. Blocki, A. Blum, A. Datta, and O. Sheffet. The Johnson-Lindenstrauss transform itself preserves differential privacy. In IEEE Computer Society, editor, *FOCS*, pages 410–419, 2012.

[9] T. Blumensatch and M. E. Davies. Iterative hard thresholding for compressed sensing. *J. Fourier Anal. Appl.*, 14(629-654), 2008.

[10] J. Brandt. Transform coding for fast approximate nearest neighbor search in high dimensions. In *CVPR*, 2010.

[11] E. J. Candés. The restricted isometry property and its implications for compressed sensing. *C. R. Acad. Sci. Paris*, 346:589–592, 2008.

[12] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, pages 341–350, 2010.

[13] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.

[14] A. Deshpande and S. Vempala. Adaptive sampling and fast low-rank matrix approximation. In *Proceedings of the 10th RANDOM*, 2006.

[15] D. L. Donoho, Y. Tsaig, I. Drori, and J. Starck. Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. *IEEE Trans. Inf. Theory*, 58:1094–1121, 2012.

[16] P. Drineas, R. Kannan, and M. W. Mahoney. Fast Monte Carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM Journal on Computing*, 36:132–157, 2006.

[17] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the 17th SODA*, pages 1127–1136, 2006.

[18] V.G. Drinfeld and S.G. Vladut. The number of points of an algebraic curve. *Func. Anal.*, 17(53-54), 1983.

[19] S. Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM J. Numer. Anal.*, 49(6):2543–2563, 2011.

[20] Peter Frankl and Hiroshi Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *J. Comb. Theory Ser. B*, 44(3):355–362, 1988.

[21] S. Gao, F. Knoll, Y. Mao, and L. You. Practical Johnson-Lindenstrauss transformations via algebraic geometry codes. In Preparation.

[22] A. Garcia and H. Stichtenoth. On the asymptotic behavior of some towers of function fields over finite fields. *Journal of Number Thoery*, 61(0147):248–273, 1996.

[23] R. Garg and R. Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 337–344, 2009.

[24] K. Hajebi, Y. Abbasi-Yadkori, H. Shahbazi, , and H. Zhang. Fast approximate nearest neighbors with k-nearest neighbor graph. In *Proceedings of the 22nd Int. Joint Conf. Artif. Intell.*, pages 1312–1317, 2011.

[25] S. Har-Peled. Low rank matrix approximation in linear low rank matrix approximation in linear time. *Manuscript*, 2006.

[26] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th ACM Symposium on Theory of Computing (STOC)*, pages 604–613, 1998.

[27] T. S. Jayram and David P. Woodruff. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *ACM Trans. Algor.*, 9(26), 2013.

[28] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[29] Daniel M. Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. *J. ACM*, 61(1), 2014.

[30] M.R. Kearns and U. Vazirani. *Introduction to computational learning theory.* MIT Press, 1994.

[31] F. Krahmer and R. Ward. New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. *SIAM J. Math. Anal.*, 43(3):1269–1281, 2011.

[32] E. Kushilevitz, R. Ostrovsky, and Y. Rabini. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM J. Comput.*, 30:457–474, 2000.

[33] E. Liberty, N. Ailon, and A. Singer. Dense fast random projections and lean walsh transforms. In *APPROX-RANDOM*, pages 512–522, 2008.

[34] F.J. MacWilliams and N.J.A. Sloane. *The Theory of Error Correcting Codes.* North-Holland, 1983.

[35] J. Matousek. *Lectures on Discrete Geometry.* Springer-Verlag, New York, 2002.

[36] Jiri Matousek. On variants of the Johnson-Lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, 2008.

[37] R. Meka. Almost optimal explicit Johnson-Lindenstrauss transformations. *ECCC*, 183, 2010.

[38] M. Muja and D. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *ICCV*, 1009.

[39] D. Needell and R. Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of Computational Mathematics*, 9(3):317–334, 2009.

[40] D. Needell and R. Vershynin. Signal recovery from inaccurate and incomplete measurements via regularized orthogonal matching pursuit. *IEEE Journal of Selected Topics in Signal Processing*, 4:310–316, 2010.

[41] Jelani Nelson, Eric Price, and Mary Wootters. New constructions of rip matrices with fast multiplication and fewer rows. In SIAM, editor, *SODA*, pages 1515–1528, 2014.

[42] H. Robbins. A remark on Stirling formula. *Amer. Math. Monthly*, 62:26–29, 1955.

[43] F. Rosenblatt. *Principles of neurodynamics.* Spartan Books, 1962.

[44] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[45] Henning Stichtenoth. *Algebraic Function Fields and Codes.* Springer, 2009.

[46] J.A. Tropp and A.C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory*, 53(12):4655–4666, 2007.

[47] J. Upadhyay. Randomness efficient Fast-Johnson-Lindenstrauss transform with applications in differential privacy and compressed sensing. *arXiv preprint arXiv:1410.2470*, 2014.

[48] H. Xu, J. Wang, Z. Li, G. Zeng, S. Li, and N. Yu. Complementary hashing for approximate nearest neighbor search. In *Proceedings of IEEE Int. Conf. Comput. Vis.*, pages 1631–1638, 2011.