Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Metric Embeddings for Machine Learning

Master thesis
in Zusammenarbeit

Arbeitsbereich Theory of Machine Learning
Prof. Dr. U. v. Luxburg

Fachbereich Informatik (Wilhelm-Schickard-Institut)
Mathematisch-Naturwissenschaftliche Fakultät
Universität Tübingen

und

Arbeitsbereich Knowledge Technology, WTM
Dr. Victor Uc-Cetina

Department Informatik
MIN-Fakultät
Universität Hamburg

vorgelegt an der Universität Hamburg von
Leena Chennuru Vankadara
am
4.1.2018

Gutachter: Prof. Dr. U. v. Luxburg
Dr. Victor Uc-Cetina

Leena Chennuru Vankadara
Matrikelnummer: 6641141
Voechtingstrasse 17
72076 Tuebingen

Abstract

In this thesis, we initiate and perform an extensive study of the theory of metric embeddings in the context of Machine Learning. We begin by asking three questions that are fundamental to any systematic study of the theory of metric embeddings. 1) *What is the objective of an embedding in the context of Machine Learning*? 2) *What is a good evaluation metric for an embedding*? 3) *Given any metric space, what guarantees can be provided on the best possible dimension that can be achieved for embeddings of high quality into a well structured space*? In addition to providing preliminary answers to these questions, we also provide an assessment of the desirable properties required of a measure of the quality of an embedding (distortion). We show that the existing measures of distortion are ineffective in the context of Machine Learning and propose a novel measure of distortion, which we refer to as σ-*distortion* in order to overcome the limitations of the existing measures while retaining most of the desirable properties.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

It has been well established in *Statistical Machine Learning* that data from high dimensional spaces suffers from the curse of dimensionality [Hughes (1968), Marimont and Shapiro (1979)]. Machine Learning's success in dealing with this problem has partly been fuelled by the idea of dimensionality reduction by means of an embedding[Indyk and Motwani (1998), Hotelling (1933), Tenenbaum et al. (2000), Borg and Groenen (2005)]. In constructing such embeddings, it is completely natural to ask questions such as *What are some of the properties that need to be preserved in such embeddings*?, *How do we evaluate the goodness of an embedding*? and *Does such an embedding always exist between arbitrary metric spaces*? Similar questions have been thoroughly addressed in several other areas of research, notably in metric geometry, analysis as well as in theoretical computer science. Yet, to the best of our knowledge, any research in the theory of metric embeddings in the context of Machine Learning is practically non-existent.

In this thesis, we initiate the study of Metric embeddings for Machine Learning. We look at the existing literature on Metric Embeddings in tangential areas of research in an attempt to address the aforementioned questions. In Section 3.1, we answer the question *what properties do we wish to preserve in an embedding in the context of Machine Learning* for a very general setting and establish that approximate distance preservation through a continuous transformation is desirable in an embedding since it preserves the underlying geometry of the space as well as the underlying measure. In addressing the question, *How do we evaluate the goodness of an embedding*? we discuss the various measures of distortion (Section 3.2) that exist in literature notably, worstcase distortion, average distortion and l_q distortion. Worstcase distortion is the most well studied measure of distortion and exhibits nice theoretical properties. For instance, embeddings with finite worstcase distortion distortion exhibit the property of measure preservation. Hence we phrase the question (Section 3.3), What is the best possible dimension and worstcase distortion that can be achieved for embedding metric spaces with no restrictions on the underlying metric into l_p spaces for any $1 \leq p \leq \infty$? This question has been well studied in literature of Metric Geometry. It has been established that for an arbitrary metric, the best possible distortion and dimension that can be achieved is $O(\log n)$ and $O(\log n)$ for embedding into l_p spaces for all

$1 \leq p \leq \infty$. [Bourgain (1985), Johnson and Lindenstrauss (1984) Abraham et al. (2006)]. These guarantees are however, trivial in the context of Machine Learning since the dimension as well as distortion grow with the cardinality of the metric space. Embeddings for which the dimension or distortion grow with the cardinality of the space are rather uninteresting for Machine Learning since generalization to out of sample data is fundamental to the problem of learning. To obtain more relevant estimates, we consider two different approaches: 1) Restrict the class of metrics 2) Relax the measure of distortion. Following the first approach, we restrict the class of all metrics to doubling metrics (Section 3.4) in order to impose growth restrictions on the underlying metric space. Doubling metrics can be viewed as metrics with low intrinsic dimension. Since most high dimensional datasets such as images, speech signals lie on a low dimensional subspace, this restriction can be considered as meaningful in the context of ML. The question can now be rephrased as *Do doubling metrics embed into finite dimensional l_p spaces with finite distortion*? This question has been answered in the negative for $p > 2$ and $p = 1$ [Lee et al. (2005),Lafforgue and Naor (2014),Bartal et al. (2015)].

Following the second approach (Section 3.5) , we consider the other relaxed measures of distortion namely average distortion and l_q distortion and ask the question, do doubling metrics embed into a $O(1)$ dimensional normed space with $O(1)$ average or l_q distortion ? Abraham et al. (2006) showed that there exists such an embedding into a constant dimensional l_p spaces for all p. Although this result has a positive outlook, it is not clear if either average or l_q distortion are meaningful distortion measures in the context of Machine Learning.

In Chapter 4, we argue that for the most generic scenario in Machine Learning, the quality of an embedding can be characterized in terms of the sharpness of concentration of the distribution of the ratio of distances (embedding distance/original distance) upto a scale. Since any effective measure of distortion should resonate with the characterization of the quality of an embedding, we argue that any distortion measure should essentially satisfy the properties of scale and translation invariance, robustness to noise and robustness to outliers in addition to allowing for a probabilistic treatment of distortion. In light of these properties, we show that the existing measures of distortion in literature, worstcase distortion, average distortion and more generally l_q distortion are ineffective in estimating the quality of an embedding in the context of Machine Learning. In order to overcome the limitations posed by the existing distortion measures in literature, we propose a new measure of distortion (σ-*distortion*) and show using formal methods that it satisfies the properties desirable for an effective measure of distortion. We empirically demonstrate the same in Chapter 5.

Finally, we present discussion (Chapter 6) and future work (Chapter 7) in order to emphasize on further research directions of the work done in this thesis.

Chapter 2

Preliminaries and settings

2.1 Preliminaries

In this section, we list some basic notations and definitions that are necessary for the reader to navigate through the rest of the thesis.

Definition 1 (**Metric space**). *A metric space is an ordered pair* (X, d_X) *where* X *is a non-empty set and* d_X *is a metric. A metric* d_X *is a function* $d_X : X \times X \to \mathbb{R}$ *satisfying the following properties.*

(a) ***(Non-negativity)*** $\forall\, (u, v) \in X,\ d_X(u, v) \geq 0$ *and* $d_X(u, u) = 0$

(b) ***(Discrimination)*** $d_X(u, v) = 0 \implies u = v$

(c) ***(Symmetry)*** $d_X(u, v) = d_X(v, u)$

(d) ***(Triangle Inequality)*** *For any* $(u, v, w) \in X$ $d(u, v) \leq d(v, w) + d(u, w)$.

If the cardinality of the set X *is finite, then* (X, d_X) *is referred to as a finite metric space.*

Definition 2 (**p-Norm**). *Let* $\mathbb{R}^n$ *denote the standard n-dimensional vector space. Then the p-norm,* $l_p : \mathbb{R}^n \to \mathbb{R}$*, for any* $1 \leq p < \infty$ *is defined as*

$$\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}} \ \ \forall x \in \mathbb{R}^n$$

for $p = \infty$*,* l_p *norm is defined as*

$$\|x\|_\infty = \max_i(|x_i|)\ \forall x \in \mathbb{R}^n$$

where x_i *denotes the* i^{th} *coordinate of* x*.*

Definition 3 (**l_p-Space**). *Let* $\mathbb{R}^n$ *denote the standard n-dimensional vector space. Then the* l_p *Space is defined as the normed vector space* $(\mathbb{R}^n, \|.\|_p)$*. The distance metric induced by a* l_p *norm is referred to as the* l_p *metric.*

Definition 4 (**Homogeneity and translation invariance**). *Let X be a vector space. Then a metric d_X on X is said to be homogeneous if $\forall (u, v) \in X$ and for any $\alpha \in \mathbb{R}$.*

$$d_X(\alpha u, \alpha v) = |\alpha| \cdot d_X(u, v)$$

and translation invariant if for any $a \in X$,

$$d_X(u + a, v + a) = d_X(u, v)$$

Definition 5 (**Homeomorphism**). *A mapping $f : (X, d_X) \to (Y, d_Y)$ between two metric spaces is called a homeomorphism if it has the following properties:*

1. ***(Bijective)*** *f is a bijection,*

2. ***(Continuity)*** *f is continuous,*

3. ***(Continuity of inverse)*** *f^{-1} is continuous*

Note that we defined a homeomorphism only for mappings between metric spaces. The same definition is applicable to topological spaces.

Definition 6 (**Identity mapping**). *Let X be any set. Then the mapping $I : X \to X$ such that $I(x) = x$ for all $x \in X$ is referred to as the identity mapping.*

2.2 Settings

(a) In this thesis, we restrict our analysis to finite metric spaces and most of the analysis can be extended to continuous metric spaces as well.

(b) We also restrict our analysis to bijective mappings between metric spaces to avoid technical complications(An injective mapping can be converted into a bijective one by considering the mapping onto the image of the domain).

(c) It is straightforward to verify that a homogeneous and translation invariant metric induces a norm on the vector space and vice versa. In this thesis, we consider embeddings from arbitrary metric spaces into normed vector spaces and thereby the target metric is implicitly considered as homogeneous and translation invariant.

(d) We would like to differentiate the problem of metric embedding from that of metric learning where the underlying distance metric of the data is learned. We assume that the true underlying distance metric is known and focus on the problem of embedding data into a low dimensional normed vector space.

2.3 Notation

(a) Let (X, d_X) and (Y, d_Y) be arbitrary metric spaces. Then for a mapping $f : (X, d_X) \to (Y, d_Y)$, (X, d_X) is referred to as the ***original space*** and (Y, d_Y) is referred to as the ***target space***.

In this document, we use the terms measure space, topological space, borel σ algebra, borel spaces, borel functions and the borel σ algebra generated by a metric space. The definitions of these terms can be found in any standard textbook on topology or Measure theory. We refrain from defining these concepts since they are only tangential to the central argument of the thesis. For reference see Billingsley (2008) and Munkres (2000).

Chapter 3

Overview of the Literature

In this chapter we address the following three questions as mentioned in Chapter 1 in the context of Machine Learning.

(a) What structure/properties of the original metric space do we wish to preserve in an embedding ? (Section 3.1)

(b) How do we evaluate the quality of such an embedding ? (Section 3.2)

(c) What is the best possible dimension one can achieve for an embedding of high quality ? (Sections 3.3, 3.4 and 3.5)

3.1 Types of embeddings

A traditional Machine Learning algorithm can be broken down into three stages: Modelling(choosing a class of hypothesis), Learning(for e.g, parameter estimation) and Inference(e.g, prediction on new data). Learning is often posed as an optimization problem and optimization problems often require that the parameter space and as a consequence also the data/feature space is structured. Hence embeddings into inner product spaces(where the inner product is defined) or more generally into normed vector spaces are desired. Therefore, we only consider embeddings into normed vector spaces, specifically into l_p spaces.

Most Machine Learning algorithms rely either on a similarity function(the underlying distance metric) of the data(e.g, k-nearest neighbour, Support vector machine [Cortes and Vapnik (1995)]) or on the underlying probability distribution of the data(e.g, Bayes Classifier, Naive Bayes classifier [Hand and Yu (2001)], Generative classifiers) to generate learning rules. Hence, it is natural to seek embeddings that preserve exact distances or the underlying probability measure or ideally both. Mappings between metric spaces which preserve distances are referred to as ***isometries*** and mappings that preserve measure are referred to as ***(metric) isomorphisms***. Isometries and isomorphisms are formally defined in definitions 7 and 8 respectively. Under suitable conditions, distance preservation between two metric spaces implies preservation of the probability measure for an appropriate transformation of measure. This is formally expressed below.

Definition 7 (**Isometry**). *A bijective mapping between two metric spaces* $f : (X, d_X) \to (Y, d_Y)$ *is said to be an isometry iff* $\forall\ (u, v) \in X$, $d_Y(f(u), f(v)) = d_X(u, v)$.

Lemma 1. *Let* $f : (X, d_X) \to (Y, d_Y)$ *be an isometry. Then* f *is continuous.*

Proof (sketch). For any $x_0 \in X$ and for any $\epsilon > 0$, set $\delta = \epsilon$. Then $d_X(x_0, x) < \delta$ for some $x \in X$, implies $d_Y(f(x_0), f(x)) < \delta = \epsilon$ □

Definition 8 (**Metric Isomorphism or Measure preserving mapping**). *A bijective measurable mapping between two measure spaces* $f : (X, \mathcal{F}_X, \mu_X) \to (Y, \mathcal{F}_Y, \mu_Y)$ *is said to be a metric isomorphism iff* $\mu_X(f^{-1}(A)) = \mu_Y(A)$ *for every* $A \in \mathcal{F}_Y$.

Definition 9 (**Transformation of measure**). *Let* $(X, \mathcal{B}_X, \mu_X)$ *be a measure space and let* $(Y, \mathcal{B}_Y)$ *be a measurable space. For any measurable mapping* $f : (X, \mathcal{B}_X) \to (Y, \mathcal{B}_Y)$ *define a set function* μ_Y *on* $(Y, \mathcal{B}_Y)$ *as*

$$\mu_Y(A) = \mu_X(f^{-1}(A))\ \forall\ A \in \mathcal{B}_Y$$

Proposition 1 (**Distance preservation implies measure preservation**). *Let* $f : (X, d_X) \to (Y, d_Y)$ *be an isometry and let* μ_X *be a probability measure defined on the borel* σ *algebra* $\mathcal{B}_X$ *generated by the underlying metric* d_X *on* X. *Denote the corresponding probability space by* $(X, \mathcal{B}_X, \mu_X)$. *Let* $\mathcal{B}_Y$ *denote the borel* σ *algebra generated by* d_Y *on* Y. *Let* μ_Y *be the transformed measure associated with* f *as defined in Definition 9. Then* $f : (X, \mathcal{B}_X, \mu_X) \to (Y, \mathcal{B}_Y, \mu_Y)$ *is a* ***measure-preserving mapping***.

Proof (sketch). From Lemma 1, f is continuous. Since f is a continuous function between borel spaces $(X, \mathcal{B}_X)$ and $(Y, \mathcal{B}_Y)$, it is measurable. □

Due to proposition 1, we restrict our attention in this thesis to mappings that preserve the underlying metric and analysis on measure preservation would be left for future work.
Desiring an isometry is very ambitious since exact preservation of distances is extremely hard to achieve for a mapping (for e.g, see Lemma 2). Moreover in the context of Machine Learning, exact distance preservation is not essential. For illustrative purposes, consider the following example.

Example 3.1.1. *Consider the 4-cycle* C_4 *associated with the shortest path metric* d_G *and label its vertices in a cyclic order* v_1, v_2, v_3, v_4. *Let* $Y_i \in \{\pm 1\}$ *for* $i = 1$ *to* 4 *denote arbitrary labels assigned to* v_i. *From Lemma 2, we have that* C_4 *can not be isometrically embedded into Euclidean space. However, for any arbitrary embedding* $f : (C_4, d_G) \to (\mathbb{R}^3, l_2)$, *there exists a hyperplane* $< w, x > + b$ *for some* $w \in \mathbb{R}^3$, $b \in \mathbb{R}$, $\forall\ x \in \mathbb{R}^3$ *such that* $sgn(Y_i \langle w, f(v_i) \rangle + b) = 1\ \forall\ v_i$.

Proof (sketch). VC dimension of hyperplanes in R^3 is 4. □

Lemma 2. *C_4 can not be isometrically embedded into an Euclidean space no matter how high the dimension*

Proof (sketch). Consider an arbitrary mapping $f : (C_4, d_G) \to (\mathbb{R}^n, l_2)$ for an arbitrary n, where l_2 denotes the Euclidean metric. Consider the following entity: $\|f(v_1) - f(v_2) + f(v_3) - f(v_4)\|^2$. By the virtue of the norm, it is ≥ 0. This inequality can be rewritten to get the following form.

$$l_2(f(v_1), f(v_3))^2 + l_2(f(v_2), f(v_4))^2 \leq l_2(f(v_1), f(v_2))^2 + l_2(f(v_2), f(v_3))^2 + l_2(f(v_3), f(v_4))^2 + l_2(f(v_1), f(v_4))^2$$

However, d_G on C_4 does not satisfy the same inequality (Note the independence of the dimension of the R^n in the inequality). □

Lemma 2 shows that embedding even a metric space of cardinality 4 which preserves exact distances does not exist into Euclidean space of any dimension. This is a strong result since it demonstrates the difficulty of achieving an isometry. In the context of Machine Learning, exact distance preservation is also not essential due to the fact that more often than not we deal with noisy data. This persuades us to seek embeddings which preserve approximate distances rather than exact distances. In addition, we would also want the corresponding mapping to be continuous in order to retain the properties of measure preservation as described for isometries.

In order to make the notion of approximate embeddings more rigorous, a measure of quantification of *approximateness* is required. This measure of approximateness would be useful in the assessment of the quality of an embedding, thereby addressing our second question.

3.2 Quality of an embedding

In this section, we discuss some of the measures of *approximateness* widely used in literature. These measures are often referred to as distortion measures. It should be noted that these measures of distortion have been well studied in various contexts but their applicability and effectiveness in Machine Learning has not been explored. We present this analysis in Chapters 4 and 5.

The most well studied as well as widely used measure of distortion is ***worstcase distortion***. Worstcase distortion is formally described in definition 10.

Definition 10 (Worstcase distortion). *Let (X, d_X) and (Y, d_Y) be arbitrary finite metric spaces. Let $f : (X, d_X) \to (Y, d_Y)$ be an bijective mapping. Then the worstcase distortion, $\Phi_{wc}(f)$ is defined as the minimum of all such $D \geq 1$ for which there exists a $\gamma > 0$ such that $\forall (u, v) \in X$,*

$$\gamma d_X(u, v) \leq d_Y(f(u), f(v)) \leq D\gamma d_X(u, v)$$

If such a γ does not exist for any $D \geq 1$ then $\Phi_{wc}(f) = \infty$

Note that γ represents the scaling factor in Definition 10 thereby rendering the value of worstcase distortion D scale invariant (See Definition 18 for a formal expression of scale invariance). The definition can be better understood in terms of the ratio of distances d_Y/d_X. If an embedding has finite distortion D, then the ratio of distances for any pair of points in X satisfies $\gamma \leq (d_Y/d_X) \leq \gamma \cdot D$. Since worstcase distortion is given by D, it is independent of γ which represents the scaling factor. Also note that since X and Y are finite metric spaces, such a γ and D always exist. To see this simply set $\gamma = \frac{1}{max\left\{\frac{d_X(u,v)}{d_Y(f(u),f(v))}\right\}}$ and $D = max\left\{\frac{d_Y(f(u),f(v))}{d_X(u,v)}\right\} \cdot max\left\{\frac{d_X(u,v)}{d_Y(f(u),f(v))}\right\}, \forall (u,v) \in X \times X$. In essence, for $\Phi_{wc}(.)$ to be finite, the ratio of distances in the embedding space to the original space for any pair of points $(u,v) \in X \times X$, is bounded above and below by a finite value. Recall the definition of a bi-lipschitz map.

Definition 11 (**Bi-lipschitz map**). *A mapping $f : (X, d_X) \to (Y, d_Y)$ is said to be bi-lipschitz iff $\exists k_1, k_2 > 0$ s.t $\forall (u,v) \in X$*

$$k_1 d_X(u,v) \leq d_Y(f(u), f(v)) \leq k_2 d_X(u,v)$$

Hence, embeddings that have finite distortion are referred to as ***bi-lipschitz embeddings*** in the literature. Bi-lipschitz maps are homeomorphisms and are hence continuous maps. By virtue of their continuity, bi-lipschitz maps possess the property of measure preservation in the sense that was described earlier. In addition, bi-lipschitz embeddings have nice theoretical properties such as preservation of intrinsic dimension of manifolds and approximate volume preservation [Eftekhari and Wakin (2017)]. Nice theoretical properties added with the ease of evaluation(see below) of worstcase distortion makes it a very attractive measure of distortion and hence it has been the most popular definition of distortion in the literature of metric embeddings [Bourgain (1985), Assouad (1983), Rabinovich and Raz (1998), Matoušek (1996)].

Alternatively, for any bijective map $f : (X, d_X) \to (Y, d_Y)$, $\Phi_{wc}(f)$ can be defined as $\|f\|_{Lip} . \|f^{-1}\|_{Lip}$ where $\|f\|_{Lip}$ and $\|f^{-1}\|_{Lip}$ denote the Lipschitz norms of f and f^{-1} respectively. The definition of Lipschitz norm is given in definition 12:

Definition 12 (**Lipschitz norm**). *Let (X, d_X) and (Y, d_Y) be arbitrary finite metric spaces. If $f : (X, d_X) \to (Y, d_Y)$ is a bijective mapping, then the **Lipschitz norm** of f is defined as*

$$\|f\|_{Lip} := sup\left\{\frac{d_Y(f(u), f(v))}{d_X(u,v)} : u, v \in X, u \neq v\right\}$$

The following proposition 2 shows that both definitions of worstcase distortion are equivalent:

Proposition 2 (**Equivalence of distortions**). *Let (X, d_X) and (Y, d_Y) be arbitrary finite metric spaces. Let $f : (X, d_X) \to (Y, d_Y)$ be a bijective mapping. Let* $\Phi =$

$$min\{D \geq 1, \exists \gamma > 0 : \gamma d_X(u,v) \leq d_Y(f(u), f(v)) \leq \gamma.D.d_X(u,v), \forall (u,v) \in X \times X\}$$

and let

$$\Phi' = max\left\{\frac{d_Y(f(u), f(v))}{d_X(u,v)}\right\}.max\left\{\frac{d_X(u,v)}{d_Y(f(u), f(v))}\right\}, \forall (u,v) \in X \times X$$

then $\Phi = \Phi'$

Proof (sketch). To prove $\Phi = \Phi'$ we prove that $\Phi \geq \Phi'$ and $\Phi \leq \Phi'$. To prove $\Phi \geq \Phi'$, fix the γ for which $\gamma d_X(u,v) \leq d_Y(f(u), f(v)) \leq \gamma \cdot \Phi \cdot d_X(u,v), \forall (u,v) \in X \times X$. If such a γ does not exist then $\Phi = \infty$ by definition and hence $\Phi \geq \Phi'$. If such a γ and Φ do exist then it follows that $\max\left\{\frac{d_Y(f(u),f(v))}{d_X(u,v)}\right\} \leq \gamma \cdot \Phi$ and $\max\left\{\frac{d_X(u,v)}{d_Y(f(u),f(v))}\right\} \leq \frac{1}{\gamma}$. Hence $\Phi \geq \Phi'$. To prove $\Phi \leq \Phi'$, set $\gamma = \frac{\max\left\{\frac{d_Y(f(u),f(v))}{d_X(u,v)}\right\}}{\Phi'}$. Note that $\Phi' > 0$ since f is an injective mapping. This implies $\exists \gamma > 0 : \gamma d_X(u,v) \leq d_Y(f(u), f(v)) \leq \gamma \cdot \Phi' \cdot d_X(u,v) \forall (u,v) \in X \times X$. Since Φ is the minimum of all such D for which this condition holds true, $\Phi \leq \Phi'$. □

We now look at some of the other measures of distortion that exist in literature. More specifically we consider average distortion and l_q distortion. Average distortion is formally defined in definition 13.

Definition 13 (**Average distortion**). *Let (X, d_X) and (Y, d_Y) be arbitrary finite metric spaces such that $|X| = N$. Where d_Y is a homogeneous, translation invariant metric. Let $f : (X, d_X) \to (Y, d_Y)$ be an bijective mapping. Average distortion of f denoted by, $\Phi_{avg}(f)$ is defined as,*

$$\frac{1}{\binom{N}{2}} \sum_{u,v \in X, u \neq v} \frac{d_Y(f(u), f(v))}{d_X(u,v)}$$

$\Phi_{avg}(.)$ measures distortion by evaluating the mean of the distribution of the ratio of distances in the embedding space to the original space for any pair of points $(u, v) \in X \times X$.

A more general framework of a class of distortion measures referred to as l_q distortion has been proposed in [Abraham et al. (2006)]. l_q distortion encompasses $\Phi_{wc}(.)$ and $\Phi_{avg}(.)$ and is defined as follows:

Definition 14 (**l_q distortion**). *Let (X, d_X) and (Y, d_Y) be arbitrary finite metric spaces. Let $f : (X, d_X) \to (Y, d_Y)$ be a non contractive, bijective mapping. Given a distribution $\mathcal{P}$ over X, Let $\Pi = \mathcal{P} \times \mathcal{P}$ denote the distribution on the product space $X \times X$ (Note the inherent assumption of independence). Then $\forall 1 \leq q < \infty$ the l_q distortion of f with respect to Π is given by*

$$\Phi_q^{\Pi}(f) = [E_{\Pi}(\frac{d_Y(f(u), f(v))}{d_X(u,v)})^q]^{1/q}$$

Note that when $\mathcal{P}$ is the Uniform distribution over X, for $q = 1$, l_q distortion is equivalent to average distortion as defined in Definition 13. The worstcase distortion can be viewed as a special case of l_q distortion for $q = \infty$ defined as

$$\Phi^{\Pi}_{\infty}(f) \ = \max \left(\frac{d_Y(f(u), f(v))}{d_X(u, v)} \right)$$

The general framework of l_q distortion, improves upon the previous measures of distortion by allowing for the incorporation of a probability measure of the data space into the evaluation of distortion. However, note that the measure of l_∞ distortion is independent of the underlying probability measure.

$\Phi^{\Pi}_q(.) \ \ \forall 1 \leq q < \infty$ are merely normalized l_q norms of the ratio of distances. For any given embedding f and for any probability distribution Π over $X \times X$, the measures follow the following order of relations:

$$\forall p, q \geq 1, p \leq q \implies \Phi^{\Pi}_p(f) \ \leq \Phi^{\Pi}_q(f)$$

l_q distortions for $q \neq \infty$ can be viewed as relaxations of worstcase distortion since they capture various moments of the distribution of the ratio of distances instead of the range. An alternate relaxation of worstcase distortion referred to as ϵ-slack distortion has been proposed by Kleinberg et al. (2004) where worstcase distortion is computed on a subset of all pairs of points.

Definition 15 (**ϵ-slack distortion**). *Let $f : (X, d_X) \to (Y, d_Y)$ be an embedding between two arbitrary finite metric spaces, then given a $0 \leq \epsilon < 1$, f is said to have to have distortion Φ with ϵ slack, if there exists a set $E \subseteq \binom{X}{2}$ with $|E| \geq (1-\epsilon)|\binom{X}{2}|$ such that*

$$\Phi \leq \max_{(u,v) \in E} \left\{ \frac{d_Y(f(u), f(v))}{d_X(u, v)} \right\} . \max_{(u,v) \in E} \left\{ \frac{d_X(u, v)}{d_Y(f(u), f(v))} \right\} \tag{3.1}$$

Since ϵ-slack distortion depends on the parameter ϵ, a natural extension of this measure would be evaluating the distortion of the embedding as a function of ϵ. The following measure of distortion, referred to as scaling distortion [Abraham et al. (2005)] is defined in Definition 16.

Definition 16 (**Scaling distortion**). *Let $f : (X, d_X) \to (Y, d_Y)$ be an embedding between arbitrary finite metric spaces. Given a function $\alpha : (0, 1) \to \mathbb{R}^+$, f is said to incur scaling distortion α if for every $\epsilon \in [0, 1)$, f incurs distortion $\alpha(\epsilon)$ with ϵ slack.*

In other words, if an embedding has scaling distortion α, then for each $\epsilon \in [0, 1)$, there exists a subset containing at least $(1 - \epsilon)$ fraction of points whose worstcase distortion is at most $\alpha(\epsilon)$.

As mentioned earlier, worstcase distortion has been the most popular of these measures of distortion. Hence in addressing the third question, *what is the best possible dimension that can be achieved for an embedding of high quality*, we begin by asking for any l_p space, what is the best possible dimension and worstcase distortion ($\Phi_{\infty}(.)$) one can achieve for an embedding of an arbitrary metric space.

3.3 Guarantees on dimension and distortion

In this section, we present the current state of the art literature addressing the following question. Given any finite metric space with no restrictions on the underlying geometry of the metric. W*hat is the best possible dimension and distortion that can be achieved for an embedding into l_p space for all $1 \leq p \leq \infty$?*

3.3.1 Embeddings into Euclidean (l_2) space

A cornerstone result in the course of answering this question is a famous result known as the Johnson-Lindenstrauss lemma [Johnson and Lindenstrauss (1984)].

Theorem 1 (**JL-Lemma [Johnson and Lindenstrauss (1984)]**). *Let (X, l_2) be a subset of the $(\mathbb{R}^k, l_2)$ of cardinality n for some integer $k \geq 2$. Then for any $\epsilon \in (0, 1)$, there exists an embedding of X into $(\mathbb{R}^d, l_2)$ with distortion $1 + \epsilon$ where $d = O(\frac{logn}{\epsilon^2})$.*

JL lemma provides an upper bound on the dimension that can be achieved with $O(1)$ distortion for an embedding of an n point subset of Euclidean space of arbitrary dimension. In addition, the original paper also provides a lower bound $\Omega(log\, n)$ on the dimension required for embedding an n point subset of Euclidean space into l_2 with distortion $1 + \epsilon$ for $\epsilon \leq 1/2$. More recently Larsen and Nelson (2016) showed that the dimension provided by JL-lemma is optimal for nearly all ϵ.

Theorem 2 (**Optimality of JL-Lemma [Larsen and Nelson (2016)]**). *For any integers $n, k \geq 2$ and for $1/(\min\{n, k\})^{0.4999} < \epsilon \leq 1$, there exists an n point subset of $\mathbb{R}^k$ such that any embedding in $(\mathbb{R}^d, l_2)$ that has distortion $1 + \epsilon$ requires that $d = \Omega(\frac{log(\epsilon^2 n)}{\epsilon^2})$.*

For all practical purposes, JL-lemma is essentially tight. However, it is only applicable to finite subsets of Euclidean space and the techniques used in the original paper cannot be naturally extended to other l_p spaces. Another fundamental result in the theory of metric embeddings is provided in Bourgain (1985). Bourgain's Lemma, using a Fréchet style embedding, gives the best possible distortion for embedding an arbitrary metric space into Euclidean space. It however gives no estimate on the dimension required for achieving this distortion. This result can be used in conjunction with JL-lemma to provide lower and upper bounds on the dimension and distortion required for embedding arbitrary finite metric spaces into Euclidean space.

Theorem 3 (**Bourgain [Bourgain (1985)]**). *Any finite metric space of cardinality n can be embedded into Euclidean space with worstcase distortion $O(log\, n)$.*

Bourgain's result has been shown to be tight and the optimality of the distortion required as given by Bourgain's result has been provided in Linial et al. (1995).

Theorem 4 (**Optimality of Bourgain [Linial et al. (1995)]**). *Any embedding of an n vertex constant degree expander into Euclidean space requires that* $\Phi_{\infty}^{\cdot}(.) = \Omega(log\, n)$.

It follows from Theorems 1 and 3 that any finite metric space can be embedded into Euclidean space with dimension $O(log\, n)$ and distortion $(log\, n)$. It follows from theorems 2 and 4 that this result is essentially tight.

3.3.2 Embeddings into l_p spaces

Dvoretzky's theorem can be applied to show that every finite dimensional Euclidean space can be embedded with finite distortion [Matoušek (2002)] into any l_p space for $1 \leq p \leq \infty$.

Theorem 5 (**Dvoretzky's theorem [Dvoretzky (1964)]**). *For every* $k \in N$ *and every* $\epsilon > 0$ *there exists* $N(k, \epsilon) \in N$ *such that if* $(\mathbb{R}^N, l_p)$ *is an* l_p *space, then there exists a linear embedding* $f : (\mathbb{R}^k, l_2) \to (\mathbb{R}^N, l_p)$ *such that* $\Phi_{\infty}^{\mathcal{U}}(f) = (1 + \epsilon)$.

In fact, it can be shown that any finite dimensional Euclidean space can be isometrically embedded into any l_p space for all $1 \leq p \leq \infty$ [Matoušek (2013)].

Corollary 1. *Any arbitrary finite metric space of cardinality n can be embedded into an* l_p *space with distortion* $O(log\, n)$.

Corollary 1 follows from Dvorestzky's theorem (Theorem 5) used in conjunction with Bourgain's theorem (Theorem 3). Matoušek (1997) modified Bourgain's embedding to derive a slightly different bound (by a constant factor as $O(\frac{log\, n}{p})$). Matoušek (1997) also showed that this result is optimal by evaluating the minimum distortion required for embedding expanders into l_p and provided a matching lower bound.

Corollary 1 can be viewed as an analogue of Bourgain's embedding for l_p spaces. However, since JL-Lemma only addresses dimensionality reduction for l_2 a similar upper bound on the dimension required for such an embedding does not naturally follow for l_p spaces. An upper bound on the dimension required for embedding an arbitrary metric space into any l_p space $1 \leq p \leq \infty$ is given by Linial et al. (1995) as $O(log^2\, n)$ and is later improved in Abraham et al. (2006) to $O(log\, n)$ giving the following result.

Theorem 6 (**Upper bound on distortion and dimension[Abraham et al. (2006)]**). *Any arbitrary finite metric space can be embedded into an* l_p *space for* $1 \leq p \leq \infty$ *with dimension* $O(log\, n)$ *and distortion* $O(log\, n)$.

Abraham et al. (2006) also show that this result is optimal and for a constant degree expander equipped with the shortest path metric, the best possible dimension and distortion that can be similtaneously achieved is $O(log\, n)$ and $O(log\, n)$.

These results are encouraging in the sense of dimensionality reduction since an arbitrary n point subset of an infinite dimensional space can be embedded into a

space of $O(logn)$ dimensions with $O(logn)$ distortion into any l_p space. However, in the context of Machine Learning, we argue that the existing bounds can be deemed to be trivial. **Out of sample extension**s are fundamental to dimensionality reduction in Machine Learning and this implies that any upper bound on dimension or distortion that grows with n is a trivial upper bound.

Since **generalization** lies at the core of Machine Learning algorithms, embeddings into a fixed dimension are not only desired but also essential. In particular, both the dimension and distortion should not grow with the cardinality(n) of the original space. Since the existing bounds on $\Phi_\infty(.)$ grow with n and are tight, we look at two possible directions to attain more relevant estimates of the best possible dimension and distortion for a given metric space.

The first direction would be to analyze the embeddability of a subclass of metric spaces with some restriction on the growth rate of some underlying geometric property(e.g, volume). The parallel direction would be to consider an alternate distortion measure which is a relaxed version of $\Phi_\infty(.)$. Some of these distortion measures were already presented in section 3.2. Note that these approaches are not mutually exclusive and can be used in conjunction with each other.

3.4 Doubling spaces

Several high dimensional datasets processed by Machine Learning algorithms such as images, speech, etc are of relatively very low intrinsic dimensionality. Hence a common assumption while dealing with high dimensional data is the "Manifold assumption"[Tenenbaum et al. (2000)] which dictates that the data lies on a low dimensional manifold in a high dimensional space. Hence analyzing embeddability on a subclass of metric spaces with low intrinsic dimension could be considered as a reasonable restriction. There are several notions of intrinsic dimensionality that prevail in literature, notably, topological dimension, Assouad's dimension (a related notion to doubling dimension) and box counting dimension.

A particularly attractive notion of instrinsic dimension, in the context of Machine Learning, is that of doubling dimension. In addition to satisfying certain natural properties [Assouad (1983)], it has been demonstrated that algorithms such as nearest neighbour search can be made more efficient when restricted to doubling spaces [Karger and Ruhl (2002),Krauthgamer and Lee (2004)]. Doubling dimension is defined in definition 17.

Definition 17 (**Doubling dimension**). *A metric space (X, d_X) is said to have doubling dimension at most λ if every ball of radius 2r can be covered by the union of at most 2^λ balls of radius r.*

$$\forall x \in X, \forall r \geq 0, B(x, 2r) \subseteq \bigcup_{z \in \mathcal{D}} B(z, r), \text{ where } \mathcal{D} \subseteq X, \ |\mathcal{D}| \leq 2^\lambda$$

A metric space with finite doubling dimension is referred to as a doubling space. Another reason to focus our attention on doubling spaces can be viewed

as an adversary to the volume argument since it restricts the growth rate of the metric space. Volume argument is a technique used to obtain lower bounds on the distortion attained by embedding arbitrary finite metric spaces. Consider the following example for an illustration of the volume argument.

Example 3.4.1. *Let $(\mathcal{E}_n, d_\mathcal{E})$ denote the n point equilateral space where $d_\mathcal{E} : \mathcal{E}_n \times \mathcal{E}_n \to \mathbb{R}$ is defined as $d_\mathcal{E}(u, v) = 1 \, \forall u \neq v$ and $= 0$ otherwise. Then the distortion required to embed $\mathcal{E}_n$ into any finite dimensional l_p space $\forall 1 \leq p \leq \infty$ would be $\Omega(n^{\frac{1}{d}})$ where d is the dimension of the target Euclidean space.*

Proof (sketch). Assume w.l.o.g that the embedding $f : (\mathcal{E}_n, d_\mathcal{E}) \to (\mathbb{R}^d, l_2)$ is non-contractive. Let $\Phi_\infty(f) = \mathrm{D}$; Fix any point $x_0 \in \mathcal{E}_n$ and consider the ball of radius $(D + 1/2)$ around $f(x_0)$. Also consider balls of radius $1/2$ around the image in $\mathbb{R}$ of each point in $\mathcal{E}_n$. Its easy to see that the union of the balls of radius $1/2$ is contained in $B(f(x_0), D + 1/2)$. It follows that for any $x \in \mathcal{E}_n$, $n \cdot vol(B(f(x), 1/2)) \leq vol(B(f(x_0), D + 1/2))$. Therefore $D = \Omega(n^{\frac{1}{d}})$. □

It is easy to see from the example how the doubling dimension acts as an adversary to the volume argument. It is straightforward to verify that the equilateral space is not doubling. If the original space is restricted to be doubling, then the dependence on n in the above example would be replaced by the doubling dimension λ. Some relevant properties of doubling dimension and doubling spaces are listed below. The proofs are straightforward and are omitted here.

(a) Finite dimensional $l_p$spaces $\forall 1 \leq p \leq \infty$ are doubling.

(b) Doubling dimension of $(\mathbb{R}^n, l_p)$ $\forall 1 \leq p \leq \infty$ is $\Theta(n)$ [Assouad (1983)].

(c) Let (X, d_X) be a metric space of doubling dimension λ. Let $E \subseteq X$ and let d_E denote the metric d_X restricted to E. Then the doubling dimension of (E, d_E) is at most λ.

(d) If (X, d_X) is a finite metric space then doubling dimension of $X \leq log\, n$

(e) If (X, d_X) is a doubling metric space and if a mapping f admits a bi-lipschitz embedding of (X, d_X) into (Y, d_Y), then the image of X under f in Y is doubling.

With sufficient motivation, we seek an answer to the following question. *What is the best possible dimension and worstcase distortion that can be achieved in embedding finite doubling spaces into finite dimensional normed spaces? In particular, can we achieve embeddings with dimension and distortion that does not grow the cardinality of the original space. In other words, given a doubling space, is there a guarantee of existence of a bi-lipschitz embedding into some finite dimensional normed space?*

Observe that from properties (a), (c) and (e), it follows that for an arbitrary metric space to admit a bi-lipschitz embedding into any finite dimensional l_p space,

a necessary condition is that the original space is doubling. The question posed above asks if it is a sufficient condition.

Luckily, this is a very well addressed question in literature of metric embeddings both in the theoretical computer science community as well as in metric geometry and some of the important results are summarized here.

3.4.1 Embedding Doubling spaces into Euclidean space

An important and seminal result in answering this question is Assouad's embedding theorem[Assouad (1983)].

Theorem 7 (**Assouads Embedding Theorem [Assouad (1983)]**). *Let (X, d_X) be a doubling metric space with doubling dimension λ. Then for every $\epsilon \in (0,1)$, there exists $d = d(\lambda, \epsilon) \in N$, $D = D(\lambda, \epsilon) \in (1, \infty)$ and a mapping $f : (X, d_X^\epsilon) \to (\mathbb{R}^d, l_2)$ such that $\Phi_\infty(f) \leq D$.*

If (X, d_X) is an arbitrary metric space, then the space (X, d_X^ϵ) for some $\epsilon \in (0,1)$ is referred to as a snowflake version of the metric (X, d_X). Assaouad's embedding theorem indicates a positive answer to our question. Moreover, Assoaud conjectured that the theorem could be extended for the case of $\epsilon = 1$. This conjecture was, however, disproved in [Semmes (1996)] giving a clear negative answer to our question posed earlier.

However, there are improvements in the bounds for the distortion and dimension when the class of arbitrary metric spaces is restricted to doubling spaces as we will see in the following results. For any n point doubling space, an upper bound on the best possible distortion that can be achieved is given by a theorem of [Gupta et al. (2003)].

Theorem 8 (**Upper bound [Gupta et al. (2003)]**). *For any n point doubling metric space, the distortion required to embed X into Euclidean space is $O(\sqrt{log\, n})$.*

Compare this upper bound by that given by Bourgain (see theorem 3) where the best possible distortion for embedding arbitrary metric into Euclidean space is $O(log\, n)$. By restricting the class of all metric spaces to doubling spaces, the best possible distortion that can be achieved for an embedding into Euclidean space is improved from $O(log\, n)$ to $O(\sqrt{log\, n})$ The original paper also shows that this result is optimal by providing a matching lower bound.

Theorem 9 (**Lower bound on $\Phi_\infty(.)$ [Gupta et al. (2003)]**). *There exists a family of metrics (L_k, d_G) which are uniformly doubling such that the minimum distortion required for an embedding f of (L_k, d_G) into any Euclidean space requires that $\Phi_\infty(f) = \Omega(\sqrt{log\, |L_k|})$, where $|A|$ denotes the cardinality of a set A.*

The lower bound on distortion here is provided by considering the distortion required to embed the Laakso graph into Euclidean space. These results in theorems 8 and 9 provide the best possible distortion required for embedding doubling

spaces into Euclidean space. However, they do not provide any guarantees on the best possible dimension of the target space.

This led Neiman (2016) to ask the following question: *Do finite dimensional doubling spaces admit embeddings into Euclidean space with worstcase distortion* $O(\sqrt{log\, n})$ *and dimension* $O(1)$? Although this is still an ***open question*** in literature, it is noteworthy that the Laakso graph (which is the doubling space that is used to obtain the lower bound on distortion) can be embedded in only 3 dimensions [Neiman (2016)]. This result can be viewed as an indication of a positive answer to this question.

Abraham et al. (2008) provided the following result in theorem 10, giving the first simultaneous upper bounds on the dimension of the embedding space and the worstcase distortion.

Theorem 10 (**Upper bound on dimension and** $\Phi_\infty^{\cdot}(.)$ [**Abraham et al. (2008)**])**.** *There exists a universal constant* C *such that for any* n *point doubling space* (X, d_X) *with doubling dimension* λ *and any* $\frac{C}{log\, log\, n} < \theta \leq 1$*, there exists an embedding* $f : X \to l_p^d$ *with* $\Phi_\infty^{\cdot}(f) = O(log^{1+\theta}\, n)$*, where* $d = O(\frac{\lambda}{\theta})$*.*

Note that this result holds more generally for all l_p spaces. Moreover, the upper bound on the number of dimensions required for embedding doubling spaces is independent of n but for say $\theta = 1$ the embedding costs a distortion of $O(log^2\, n)$. Also note that this result establishes a trade off between worstcase distortion and embedding dimension. Another result which provides a simultaneous upper bound on the distortion and dimension required for an embedding of an arbitrary doubling space into Euclidean spaces is given by Chan et al. (2010). Results in Chan et al. (2010) also exhibit a tradeoff between embedding dimension and worstcase distortion similar to that showed in Abraham et al. (2008). As dimension increased from $O(log\, log\, n)$ to $O(log\, n)$, the distortion decreases from $O(log\, n)$ to $O(\sqrt{log\, n})$.

Theorem 11 (**Upper bound on dimension and** $\Phi_\infty^{\cdot}(.)$ [**Chan et al. (2010)**])**.** *For any finite metric space* (X, d_X) *with doubling dimension* λ*, there exists a* $d = O(\lambda\ log\ log\ n)$ *and an embedding* $f : (X, d_X) \to (\mathbb{R}^d, l_2)$ *such that* $\Phi_\infty^{\cdot}(f) = O(\frac{log\, n}{\sqrt{log\, log\, n}})$

Non-embeddability of doubling spaces into Euclidean spaces has been demonstrated in Laakso (2000), Semmes (1999) and Pauls (2001). At this point it is worth mentioning that, to the best of our knowledge, only two known examples of doubling metric spaces exist which can not be embedded in a bi-lipschitz way into any Euclidean space namely the Laakso graph with the shortest path metric and the Heisenberg group in three dimensions with the Carnot-Caratheodry metric.

Lang and Plaut [Lang and Plaut (2001)] made an observation that both the Laakso graph with the shortest path metric and the Heisenberg group with the Carnot-Caratheodry metric do not admit bi-lipschitz embeddings into any other Hilbert space. This observation was the basis for the Lang and Plaut Conjecture [Lang and Plaut (2001)].

Lang and Plaut conjecture [Lang and Plaut (2001)]. If (X, d_X) is a doubling subset of a Hilbert space, then does there exist a bi-lipschitz embedding into some Euclidean space ? □

This is still an ***open question*** in literature. See Naor and Neiman (2010) for a detailed discussion of this conjecture. An analogous question (it can be viewed as a special case) to Lang and Plaut's conjecture is the following: *Do doubling subsets of Euclidean space embed in a bi-lipschitz way into a constant dimensional Euclidean space*? This question has been raised in [Lang and Plaut (2001), Gupta et al. (2003)] and is still an ***open question*** in literature.

3.4.2 Embedding doubling spaces into l_p space

Gupta et al. (2003) presents both an upper bound as well as a nearly matching lower bound on the distortion required for embedding doubling spaces into l_p spaces. The results are given in Theorems 12 and 13.

Theorem 12 (**Upper bound on distortion[Gupta et al. (2003)]**). *Let (X, d_X) be a finite doubling metric with doubling dimension λ, then $\forall p \in [1, \infty)$, there exists an embedding of (X, d_X) into l_p with distortion $\Phi_\infty(.) = O(\lambda log\, n^{min(\frac{1}{2}, \frac{1}{p})})$.*

Observe that the upper bound given by Theorem 12 for l_p spaces is dependent on the doubling dimension of the original space. For $p = 2$, the upper bound$(O(\lambda\sqrt{log\, n}))$ differs from the one given in Theorem 8 $(O(\sqrt{log\, n}))$.

Theorem 13 (**Lower bound on distortion[Gupta et al. (2003)]**). *There exists a family of metrics (L_k, d_G) which are uniformly doubling such that the minimum distortion required for an embedding f of (L_k, d_G) into an l_p space $\forall 2 \leq p \leq \infty$ requires that $\Phi_\infty(f) = \Omega((log\, |L_k|)^{\frac{1}{p}})$, where $|A|$ denotes the cardinality of a set A.*

The lower bound is derived based on the construction of the Laakso graph and is derived in a manner which is similar to that of Theorem 9. Note that this result only holds for $p \geq 2$. Gupta et al. (2003) provides guarantees on the distortion required to embed doubling spaces into l_p. However, no simultaneous guarantees on the dimension are provided. Krauthgamer and Lee (2004) prove the existence of an embedding of doubling spaces into l_p spaces with the best possible dependence of distortion on doubling dimension $(O((log\lambda)^{1-\frac{1}{p}}(log n)^{\frac{1}{p}}))$ but requiring $O(log^2 n)$ dimensions. Gupta et al. (2003) provide much stronger upper bounds on dimension and distortion for embedding arbitrary doubling spaces into l_∞ in Theorem 14.

Theorem 14 (**Upper bound for l_∞ [Gupta et al. (2003)**). *For any fixed $\epsilon > 0$, every doubling metric embeds into $(\mathbb{R}^d, l_\infty)$ with $\Phi_\infty(.) = (1 + \epsilon)$, where $d = O(log\, n)$.*

It follows that every finite doubling metric of cardinality n isometrically embeds into $(\mathbb{R}^{O(log\, n)}, l_\infty)$.(See Corollary 1.4.3 in Matoušek (2013)).

The best possible dependence of the embedding dimension on the doubling dimension of the original space is given by **Theorem 10** where the best possible embedding dimension that can be achieved is given by $O(\frac{\lambda}{\theta})$ with a distortion of $O(log^{1+\theta}n)$.

Simultaneous lower bounds on dimension and distortion for embedding doubling spaces in l_p spaces do not exist for all p. For l_1 spaces Lee et al. (2005) provide the simultaneous lower bounds on dimension and distortion derived based on the Laakso graph. This result is provided in Theorem 15.

Gupta et al. (2003) showed that there exists a doubling metric(Laakso graph) for which embedding into any l_p space $p \geq 2$ requires a distortion of $O(\lambda log^{\frac{1}{p}})$. Theorem 10 showed the existence of an embedding for an arbitrary doubling metric into l_p in $O(\frac{\lambda}{\theta})$ dimensions with $O(log^{1+\theta}n)$ distortion. Following the lead of Neiman (2016), we ask the following question. *Do doubling spaces embed into l_p ($p \geq 2$) spaces with distortion $O(log^{\frac{1}{p}} n)$ in $O(1)$ dimensions?*

Theorem 15 (**Lower bound for l_1 [Lee et al. (2005)]**). *There are arbitrarily large n-point subsets $X \subseteq L1$ with doubling dimension 6 such that any embedding of X into l_1 that has distortion D requires that the embedding dimension $d > n^{\Omega(\frac{1}{\alpha^2})}$.*

Observe that Theorem 15 in fact provides a negative answer to a much stronger question about dimensionality reduction in l_1: *Do doubling subsets of l_1 embed into constant dimensional l_1 in a bi-lipschitz way?* Recall that a similar question was raised in the case of l_2 spaces and more generally for all l_p spaces in Naor and Neiman (2010): *Do doubling subsets of l_p embed into constant dimensional l_p with constant worstcase distortion?*

While for $p = 1$, Theorem 15 answers the question in the negative, this is still an ***open problem*** in literature for $p = (1, 2]$. For $p > 2$, this question has been answered in the negative by the following impossibility result (Theorem 16) by Bartal et al. (2015).

Theorem 16 (**Lower bound for doubling subsets of l_p [Bartal et al. (2015)]**). *For any $p > 2$, there is a constant $c = c(p)$ such that for any positive integer n, there is a subset $A \subseteq l_p$ of cardinality n with doubling dimension $O(1)$, such that any embedding of A into $(\mathbb{R}^d, l_p)$ for any fixed $d \in \mathbb{N}$ with distortion at most D requires that $D \geq \Omega((\frac{clog\, n}{d})^{\frac{1}{2}-\frac{1}{p}})$.*

This result is derived using geometric methods based on a construction of the Laakso graph as a subset of l_p. Lafforgue and Naor (2014) simultaneously gave a result answering the aforementioned question in the negative using analytic methods based on the construction of the Heisenberg group as a subset of l_p.
A natural question that follows from this is: *Do doubling subsets of l_p embed into constant dimensional l_q ($p \neq q$) with constant distortion*? Bartal et al. (2015) also showed that the same construction could be used to derive the following result.

Theorem 17 (**Lower bound for l_p to l_q [Bartal et al. (2015)]**). *For any $p > 2$, there is a constant $c = c(p)$ such that for any positive integer n, there is a subset*

$A \subseteq l_p$ of cardinality n with doubling dimension $O(1)$, such that any embedding of A into $(\mathbb{R}^d, l_q)$ $(q \geq 1)$ $(p \neq q)$ for any fixed $d \in \mathbb{N}$ with distortion at most D requires that

$$D \geq \Omega(c(p)\frac{log^{\frac{1}{2}-\frac{1}{p}}n}{d^{\frac{max(q-2,2-q)}{2q}}})$$

Although these results seems to have a pessimistic outlook for dimensionality reduction even into l_p spaces, as we will demonstrate in Sections 4 and 5, l_∞ distortion which evaluates distortion in the worstcase scenario is not an appropriate measure of distortion in the context of Machine Learning. Hence we look at some of the other measures of distortions we discussed earlier and address the following question: *Can doubling spaces be embedded into constant dimensional l_p spaces with high quality, where the quality of an embedding is evaluated by relaxing the worstcase distortion measure and considering other measures of l_q distortion?*

3.5 Relaxed distortion measures

In this section, we consider the alternate approach mentioned in section 3.3.2 in order to obtain estimates on dimension and distortion which are more relevant for Machine Learning. Recall that so far, we analyzed the quality of an embedded by evaluating its worstcase distortion. In this section, we consider the other relaxed measures of distortion discussed in Section 3.2.

3.5.1 Average distortion of embeddings

In this section, we discuss some of the results which provide guarantees on the best possible embedding dimension and average distortion that exist in literature. Abraham et al. (2006) initiated a systematic study of the average distortion of embedding arbitrary finite metric spaces into l_p spaces. The first result concerns with embedding arbitrary finite metric spaces into Euclidean space.

Theorem 18 (**Average distortion into l_2 [Abraham et al. (2006)]**). *For any arbitrary finite metric space there exists an embedding into $(\mathbb{R}^d, l_2)$ with average distortion $O(1)$, where $d = O(log\, n)$. The worstcase distortion of this embedding is $O(log\, n)$.*

Note that the embedding provided by Theorem 18 gives the same upper bound on the worstcase distortion and the embedding dimension given by Bourgain's embedding [Bourgain (1985)] in conjunction with the JL-Lemma [Johnson and Lindenstrauss (1984)](which was shown to be tight) as mentioned in section 3.3.1. However, the average distortion of the embedding is $O(1)$. This clearly indicates that, informally speaking, only a constant fraction of distances have very high distortion. This result is extended to embedding arbitrary finite metrics into all l_p spaces by Abraham et al. (2006) which leads us to the following result.

Theorem 19 (**Average distortion into l_p [Abraham et al. (2006)]**). *Let (X, d_X) be an arbitrary finite metric space of cardinality n. For any $1 \leq p \leq \infty$, there exists an embedding $f : (X, d_X) \rightarrow (\mathbb{R}^d, l_p)$ such that the average distortion of f is $O(1)$ and worstcase distortion is $O(\lceil \frac{log\, n}{p} \rceil)$, where $d = O(e^{O(p)} log\, n)$.*

The embedding in Theorem 19 has the best possible worstcase distortion as given by Matoušek (1997) with constant average distortion. Moreover, the dimension of the embedding is only differs by constant factor from Theorem 6.

3.5.2 l_q distortion of Embeddings

Abraham et al. (2006) extend the results given in section 3.5.1 and provide general bounds for l_q distortion of embedding arbitrary finite metric spaces into l_2 and more generally to l_p spaces. Theorems 20 and 21 are generalized versions of Theorems 18 and 19.

Theorem 20 (**l_q distortion into l_2 [Abraham et al. (2006)]**). *For any arbitrary finite metric space there exists an embedding into $(\mathbb{R}^d, l_2)$ with l_q distortion $O(min(q, log n))$, where $d = O(log n)$. In particular, $\Phi^{\mathcal{U}}_{\infty}(f) = O(log n)$ and $\Phi^{\mathcal{U}}_{1}(f) = O(1)$.*

Theorem 21 (**l_q distortion into l_p [Abraham et al. (2006)]**). *Let (X, d_X) be an arbitrary finite metric space of cardinality n. For any $1 \leq p \leq \infty$, there exists an embedding $f : (X, d_X) \rightarrow (\mathbb{R}^d, l_p)$, where $d = O(e^{O(p)} log\, n)$, such that $\Phi^{\mathcal{U}}_{q}(f) = O(\lceil \frac{min(q, log\, n)}{p} \rceil)$. In particular $\Phi^{\mathcal{U}}_{\infty}(f) = O(\lceil \frac{log\, n}{p} \rceil)$ and $\Phi^{\mathcal{U}}_{1}(f) = O(1)$.*

For arbitrary finite metric spaces, relaxing the distortion measure from l_∞ to l_1 improved the upper bound on best possible distortion that can be achieved for an embedding into a space of $O(log\, n)$ dimensions from $O(log\, n)$ to $O(1)$.

3.5.3 ϵ-slack distortion of Embeddings

The following Theorem 22 provides a means by which(in some sense) simultaneous upper bounds on worstcase distortion and embedding dimension that can be achieved for an embedding between two metric spaces can be used to derive simultaneous upper bounds on ϵ-slack distortion and the embedding dimension.

Theorem 22 (**Upper bounds on ϵ-slack distortion [Abraham et al. (2005)]**). *Let (X, d_X) be a finite metric space such that $|X| = n$. If there exists an embedding $f : (X, d_X) \rightarrow (\mathbb{R}^{\alpha(n)}, l_p)$ for some $1 \leq p \leq \infty$ with $\Phi_{\infty}(f) = \beta(n)$, then there exists a constant $C > 0$ such that for any $\epsilon > 0$, there exists an embedding $g : (X, d_X) \rightarrow (\mathbb{R}^{\alpha(\frac{C}{\epsilon} log \frac{1}{\epsilon} + C log \frac{1}{\epsilon})}, l_p)$ such that g incurs ϵ-slack distortion at most $\beta(\frac{C}{\epsilon} log \frac{1}{\epsilon})$.*

Theorem 22 in conjunction with Theorem 6 immediately gives the following corollary.

Corollary 2 (ϵ-slack distortion for embeddings into l_p)**.** *Given any finite metric space (X, d_X), there exists an embedding $f : (X, d_X) \rightarrow (\mathbb{R}^d, l_p)$ for some $1 \leq p \leq \infty$ such that ϵ-slack distortion of f is $O(log \frac{1}{\epsilon})$, where $d = O(log \frac{1}{\epsilon})$.*

Corollary 2 is interesting since the simultaneous upper bounds on both ϵ-slack distortion and dimension are independent of the value of n. However, the bounds do depend on the value of the parameter ϵ. We provide a more detailed discussion of ϵ-slack distortion of embeddings in section . There are no known improvements of these bounds that are achieved by restricting the analysis to doubling spaces.

3.5.4 Scaling distortion of embeddings

The following theorem presents an upper bound on the best possible scaling distortion and embedding dimension that can be achieved in embedding arbitrary finite metric spaces into Euclidean space.

Theorem 23 (**Upper bound on scaling distortion into l_2)**. *Abraham et al. (2005) Let (X, d_X) be an arbitrary finite metric space. Then there exists an embedding $f : (X, d_X) \rightarrow (\mathbb{R}^d, l_2)$ such that scaling distortion of f is $O(log \frac{2}{\epsilon})$, where $d = O(log\, n)$.*

There are no known guarantees on the scaling distortion and dimension that can be simultaneously achieved for embedding arbitrary metric spaces or doubling spaces into l_p for $p \neq 2$ spaces. Also no known improvements exist for embedding doubling spaces into Euclidean space.

3.6 Relaxed distortions for embedding doubling metrics

Abraham et al. (2006) provide a similar result for doubling spaces where the existence of an embedding in a constant dimensional l_p space which incurs constant distortion.

Theorem 24 (**Upper bound on l_q distortion for doubling spaces**)**.** *For any arbitrary finite doubling space (X, d_X) with cardinality n and doubling dimension λ, there exists an embedding $f : (X, d_X) \rightarrow (\mathbb{R}^d, l_p)$, where $d = O(\lambda log\, \lambda)$ such that $\Phi_q^{\mathcal{U}}(f) = O(q^C)$ for some universal constant C. In particular $\Phi_1^{\mathcal{U}}(f) = O(1)$.*

This result is encouraging in the context of Machine Learning since it shows the existence of an embedding from any doubling space into a constant dimensional l_p space with finite l_q distortion $\forall 1 \leq q < \infty$. However, it is not clear if these relaxations of distortions are meaningful in the sense that they capture an appropriate interpretation of the quality of the embedding. In fact, it is not clear if worstcase distortion captures the essence of the quality if an embedding in the context of Machine Learning. We answer this question by presenting a detailed analysis of the properties of distortion measures in Sections 4 and 5.

Chapter 4

Properties of distortion measures

4.1 Characterization of a high quality embedding

In order to identify what properties a good measure of distortion should possess, it is important to answer the question, *what is a good embedding in the context of Machine Learning*? In Chapter 3, we have established that *approximate* distance preservation is the goal of an embedding in this context. However, since outliers and noisy observations are a commonplace in Machine Learning, roughly speaking, it is reasonable to state that a good embedding preserves **most** distances as well as possible while better preserving the distances between pairs of points which could be critical for a given task. We argue that the following characterization of a good quality embedding precisely captures the essence of the previous statement.

An embedding of high quality can essentially be described as a mapping between two metric spaces such that the distribution of ratio of distances is sharply concentrated around 1. In other words, the probability of the event that the ratio of distances is outside a certain ϵ interval around 1 (upto a scale) drops quickly with increasing ϵ. In order to see this, consider the following illustration.

Let f be an isometry. The distribution of the ratio of distances of f is a Dirac delta distribution around 1 (or around a constant C if the embedding is scaled by C). Consider an increasing sequence, ϵ_i, where $0 \leq \epsilon_i \leq 0.5$. We artificially create various distributions of ratio of distances, ρ_i (see Figure 4.1) from f by arbitrarily distorting $2{\cdot}\epsilon_i$ fraction of randomly chosen distances with equal number of arbitrary expansions and contractions. We arbitrarily distort the remaining points such that the ratio of distances for $(1-2\cdot\epsilon)$ fraction of points lies in the interval $[1-\gamma, 1+\gamma]$ for some $0 < \gamma < 1$.

If ϵ is small, then the fraction of arbitrarily distorted distances could be described as outliers in data. The outliers manifest themselves as the tails of the distribution. However, as ϵ increases, the tails become heavier and this implies that the probability of the event where the ratio of distances lies outside an ϵ interval around 1 decreases slowly with increasing ϵ irrespective of the value of γ. If ϵ is large, the distorted distances cannot be termed as outliers and it indicates

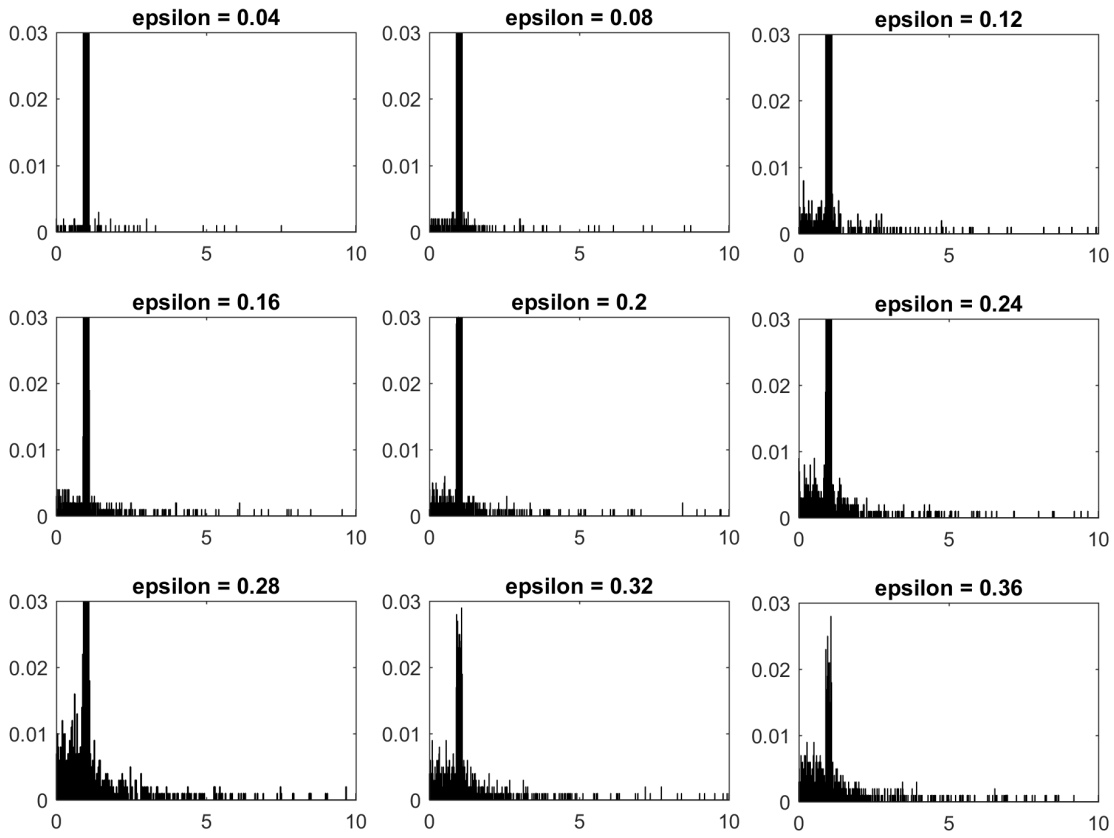


Figure 4.1: Each subfigure shows the distribution of ratio of distances corresponding to an embedding for which an ϵ proportion of distances are arbitrary distorted from an isometry. The corresponding values of ϵ are specified on top of each subplot. **Note:** Figure axes are truncated for better viewing.

that the quality of the embedding is systematically poor. The limiting distribution ($\epsilon = 0.5$), would resemble a plateaued distribution.

If γ is small, then it indicates that most of the distances are preserved as well as possible and the distribution sharply concentrated around one. However, if γ is large, it indicates that the distances are not well preserved and the distribution resembles a plateaued distribution. Figure 4.2 depicts the distributions of ratio of distances for increasing values of γ. It can be observed that the distributions transition from sharply concentrated distributions to a rather flat distributions.

Hence it can be stated that any embedding that well preserves most distances possesses the property that the distribution of the ratio of distances concentrates sharply around 1 upto a scale. In addition, since the characterization is probabilistic in nature, distances that are critical for preservation could be specified via the probability distribution.

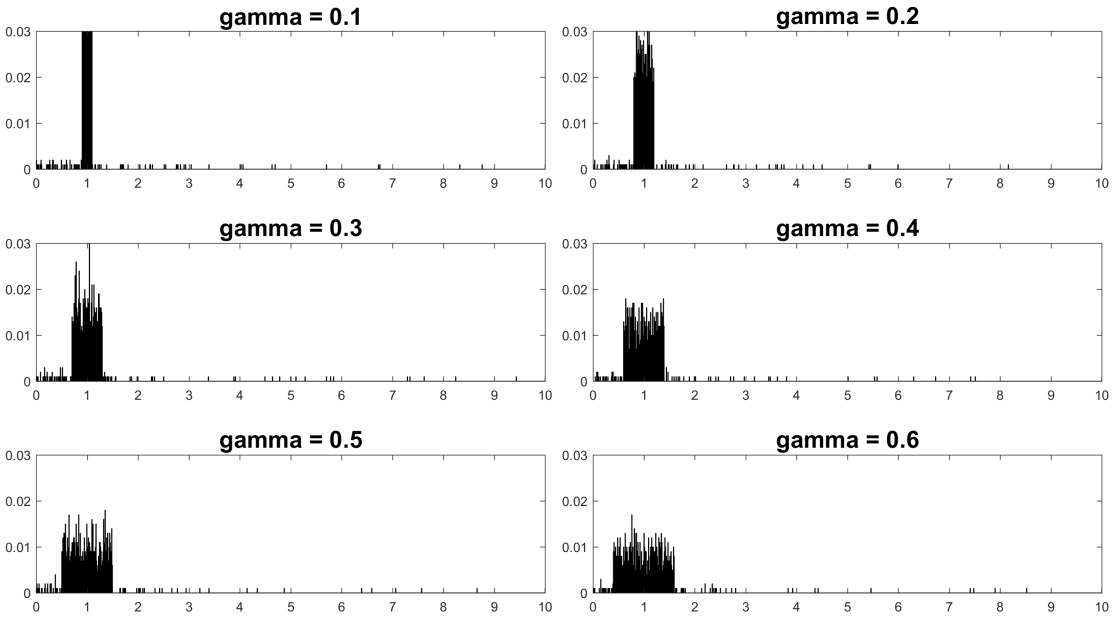


Figure 4.2: Each subfigure shows the distribution of ratio of distances corresponding to an embedding for which an ϵ=0.04 proportion of distances are arbitrary distorted from an isometry. (1-ϵ) are distortion within a range of [1-gamma, 1+gamma] around 1. The corresponding values of *gamma* are specified on top of each subplot. **Note:** Figure axes are truncated for better viewing.

4.2 What to expect from a measure of distortion?

In this section we give a preliminary assessment of what properties are expected of a distortion measure in the context of Machine Learning. An effective measure of distortion in addition to satisfying certain essential properties such as invariance to scale and translations, should resonate with the characterization of the quality of an embedding. In this spirit, we assert that any effective measure of distortion (or an estimate of the quality of an embedding) should have the aforementioned properties.

(a) **Scale Invariance:** Scale invariance is an essential property for a measure of distortion since mappings/embeddings which are merely different in units of measurement(e.g, Kilometers vs Centimeters) should not be assigned different values of distortion. Scale invariance of a measure of distortion is formally expressed in Definition 18.

Definition 18 (**Scale invariance**). *Let (X, d_X) and (Y, d_Y) be arbitrary metric spaces. Let $f : (X, d_X) \to (Y, d_Y)$ and $g : (X, d_X) \to (Y, d_Y)$ be two injective mappings. Let Φ be a measure of distortion, then Φ is said to be scale invariant if*

$$\exists \alpha \in \mathbb{R}, \forall u \in X, f(u) = \alpha g(u); \implies \Phi(f) = \Phi(g)$$

(b) **Translation Invariance:** A measure of distortion should clearly be invariant to translations since it should be independent of the point of origin. Translation invariance of a measure of distortion is formally expressed in Definition 19

Definition 19 (**Translation invariance**). *Let (X, d_X) and (Y, d_Y) be arbitrary metric spaces. Let $f : (X, d_X) \to (Y, d_Y)$ and $g : (X, d_X) \to (Y, d_Y)$ be two injective mappings. Let Φ be a measure of distortion, then Φ is said to be translation invariant if*

$$\exists \alpha \in \mathbb{R}, \forall u \in X, f(u) = g(u) + \alpha; \implies \Phi(f) = \Phi(g)$$

(c) **Robustness to outliers:** Outliers are inherent to data processed by Machine Learning algorithms and hence a measure of distortion which is volatile against outliers is not desirable. It is a challenging task to define outliers in general, but we define a framework for robustness to outliers by isolating its two possible sources: outliers in data space and outliers in the distance space. We argue that for a measure of distortion Φ to be deemed robust to outliers, Φ should be robust to outliers in the distance space(say if most but a constant order of distances are distorted) as well as outliers in data space(say outliers in data due to a measurement error) to a certain degree. We create test cases as necessary conditions to deem a measure of distortion robust to outliers.

Outlier distances: The following test case can be used to evaluate if a measure of distortion is not robust to outliers in the distance space. We consider mappings for which constant order of distances are distorted and compare its distortion measure with that of an isometry.

Scenario 1. *Let $I : (X, d_X) \to (Y, d_Y)$ be an isometry between two infinte metric spaces (X, d_X) and (Y, d_Y). Let $\{\mathcal{A}_n\} \subset X$ be an increasing sequence of sets such that $|\mathcal{A}_n| = n$. Let $f : (\bigcup_{n=1}^{\infty} \mathcal{A}_n, d_X) \to (X, d_X)$ be an injective mapping for which there exists a $K \in \mathbb{N}$ such that $|G| < K$, where $G = \left\{(u, v) \in \bigcup_{n=1}^{\infty} \mathcal{A}_n : d_Y(f(u), f(v)) \neq d_X(u, v)\right\}$. Then we say a measure of distortion Φ is not robust to outliers if $\lim_{n\to\infty} \Phi(f_n) \neq \lim_{n\to\infty} \Phi(I_n)$,where I_n denotes the restriction of the mapping I to $\mathcal{A}_n$.*

Outliers in data: The following test case can be used to evaluate if a measure of distortion is not robust to outliers in the data space. We consider mappings for which a single data point is remapped in deviation from an isometry (with minimal control on the distorted distances) and compare its distortion measure with that of the isometry.

Scenario 2. *Let $I : (X, d_X) \to (Y, d_Y)$ be an isometry between two infinite metric spaces (X, d_X) and (Y, d_Y). Fix an arbitrary $x_0 \in X$ and $\beta > 0$; Construct any increasing sequence of sets $\{\mathcal{A}_i\} \subset X$ by the following iterative procedure.*

$$\mathcal{A}_0 = \{x_0\}$$

$$\mathcal{A}_i = \mathcal{A}_{i-1} \cup \{x_i\}\,,\ x_i \in (X \setminus \mathcal{A}_{i-1} \cup B(x_0, \beta))$$

Let $f : (\bigcup_{n=1}^{\infty} \mathcal{A}_n, d_X) \to (Y, d_Y)$ *be an injective mapping such that* $\forall x_i \in \bigcup_{n=1}^{\infty} \mathcal{A}_n$, $f(x_i) = I(x_i)$, *if* $x_i \neq x_0$ *and* $f(x_0) = y^{'}$, *for some arbitrary* $y' \in Y$. *Let* f_n *denote the mapping* f *restricted to* $\mathcal{A}_n$. *A measure of distortion* Φ *is not robust to outliers if* $\lim_{n\to\infty} \Phi(f_n) \neq \lim_{n\to\infty} \Phi(I_n)$, *where* I_n *denotes the restriction of the mapping* I *to* $\mathcal{A}_n$.

(d) **Robustness to noise:** Noisy observations, just as outliers, are a common place in Machine Learning. Hence, informally speaking, we would like the measure of distortion to vary smoothly with noise. In this thesis, we provide qualitative arguments to analyze if a measure of distortion is robust to noise. In addition, we provide empirical evidence in Section 5 to further support our arguments. A systematic study of robustness to noise is deferred to future work.

(e) **Underlying probability measure:** A measure of distortion should be able to incorporate a probability measure on the data space into its evaluation. For instance, if known, the class conditional densities on the data space could provide valuable information concerning preservation of distances between which pairs of points is important for a classification task. This information could be leveraged in order to generate meaningful estimates of a measure of distortion.

The list of properties we mention here is neither exhaustive nor extensive. Nevertheless, we would argue that these properties are essential for any measure of distortion in the context of Machine Learning. In this Chapter, we formally analyze if the existing measures of distortion satisfy the aforementioned properties. This analysis is supported by empirical evidence presented in Section 5. In light of our analysis, we introduce a new measure of distortion, *σ-distortion* , and show that it overcomes most of the limitations faced by the existing measures of distortions and is indeed a meaningful measure of distortion in the paradigm of Machine Learning.

4.3 Properties of l_q distortion

In this section, we investigate if the different instances of l_q distortion possess the essential properties required for a measure of distortion as discussed in section 4.2.

Proposition 3 (**l_q distortions are translation invariant**). *All measures of l_q distortion are invariant to translations if the target metric is translation invariant.*

Proof. Follows from the definition of l_q distortion. □

4.3.1 Properties of l_∞ or worstcase distortion

Proposition 4 (l_∞ **is scale invariant**). *Let (X, d_X) and (Y, d_Y) be arbitrary finite metric spaces such that d_Y is homogeneous. Let $f : (X, d_X) \to (Y, d_Y)$ and $g : (X, d_X) \to (Y, d_Y)$ be two injective mappings such that $f(u) = \alpha \cdot g(u), \forall u \in X$ and for some $\alpha \in \mathbb{R}$. Then $\Phi_{wc}(f) = \Phi_{wc}(g)$. Hence $\Phi_{wc}(.)$ is scale invariant.*

Proof (sketch). $\Phi_{wc}(f) = \max\left\{\frac{d_Y(f(u),f(v))}{d_X(u,v)}\right\} \cdot \max\left\{\frac{d_X(u,v)}{d_Y(f(u),f(v))}\right\}, \forall (u, v) \in X \times X$, then $\Phi_{wc}(g) = \max\left\{\frac{\alpha \cdot d_Y(f(u),f(v))}{d_X(u,v)}\right\} \cdot \max\left\{\frac{d_X(u,v)}{\alpha \cdot d_Y(f(u),f(v))}\right\}, \forall (u, v) \in X \times X$ (From homogeneity of d_Y). Hence $\Phi_{wc}(f) = \Phi_{wc}(g)$. □

Proposition 5 ($\Phi_{wc}(.)$ **is not robust to outliers**). $\Phi_{wc}(.)$ *is not robust to outliers in the sense that it does not satisfy the necessary conditions specified in scenarios 1 and 2.*

Proof (sketch). We prove this proposition by showing the existence of a sequence of sets $\{\mathcal{A}_n\}$ and a corresponding mapping f conforming to the conditions specified in scenarios 1 and 2 for which the $\Phi_{wc}(f)$ diverges from the worstcase distortion of an isometry in the limit of $n \to \infty$.

Let $\{e_i\}_{\{i=1,..,d\}}$ denote the standard orthonormal basis for $(\mathbb{R}^d, l_2)$. Fix $x_0 = \frac{1-\alpha}{1+\alpha} e_d$ for some $0 < \alpha \ll 1$ and $x_1 = -e_d$. Set $\beta = \frac{1-\alpha}{1+\alpha}$. Construct a sequence of sets by the following iterative procedure.

$$\mathcal{A}_1 = \{x_0, x_1\}$$
$$\mathcal{A}_n = \mathcal{A}_{n-1} \cup \{x_i\} \ , \ x_i \in span\{e_1, ..., e_{d-1}\}$$

It is straightforward to verify that $\{\mathcal{A}_n\}$ satisfies all the conditions specified in scenarios 1 and 2. Let $I : (\mathbb{R}^d, l_2) \to (\mathbb{R}^d, l_2)$ denote the identity mapping. Let $f : (\mathbb{R}^d, l_2) \to (\mathbb{R}^d, l_2)$ be the mapping defined as $f(x) = x$, $\forall x \in \mathbb{R}^d$, $x \neq x_0$ and $f(x_0) = -x_0$. Let f_n denote the mapping f restricted to $\mathcal{A}_n$. It is easy to verify that the ratio of distances $\rho_{f_n}(x_0, x_1) = \alpha$ and $\rho_{f_n}(x, x') = 1$ for any $(x, x') \in A_n$ such that $(x, x') \neq (x_0, x_1)$. The sequence of worstcase distortions evaluated on mappings $f_n : (A_n, d_X) \to (X, d_X)$ is $\left\{(\frac{1}{\alpha})_n\right\}$ and thus $\lim_{n\to\infty} \Phi(f_n) = \frac{1}{\alpha} \gg \lim_{n\to\infty} \Phi(I_n) = 1$ where I_n denotes the restriction of the mapping I to $\mathcal{A}_n$.

Note that this construction shows that $\Phi_{wc}(.)$ is not robust to outliers in the distance space as well as outliers in the data space. □

Claim 1. ($\Phi_{wc}(.)$ **is not robust to noise**) . We claim that $\Phi_{wc}(.)$ is not robust to noisy observations since noisy observations potentially lead to outliers in the distribution of the ratio of distances and it follows from Proposition 5 that $\Phi_{wc}(.)$ is not robust to noisy observations. We provide empirical evidence in Section 5 to further support this claim.

Incorporation of the underlying probability measure: It is clear from Definition 10 that $\Phi_\infty(.)$ does not incorporate the underlying probability measure of the data space in its evaluation.

We showed that while $\Phi_\infty(.)$ is scale and translation invariant, it is not robust to outliers and does not incorporate the underlying probability measure into its evaluation. In addition, we provided a qualitative argument to support our claim that it is not robust to noisy observations.

4.3.2 Properties of l_1 distortion

Proposition 6 ($\Phi_1(.)$ **is not scale invariant**). $\Phi_1(.)$ *is not invariant to scaling according to Definition 18.*

Proof (sketch). If f and g are two mappings such that $f(x) = \alpha \cdot g(x)$ for all $x \in X$, then homogeneity of d_Y implies that $\forall (u, v) \in X$, $\frac{d_Y(f(u),f(v))}{d_X(u,v)} = \frac{\alpha(d_Y(g(u),g(v)))}{d_X(u,v)}$. Thus $\Phi_1(f) = \alpha \cdot \Phi_1(g)$. □

Proposition 7 ($\Phi_1(.)$ **is robust to outliers in distances**). $\Phi_1(.)$ *satisfies the necessary condition as specified in Scenario 1 to be deemed robust to outlier distances.*

Proof (sketch). Let $I : (X, d_X) \to (Y, d_Y)$ be an isometry. Let $\{A_n\} \subset X$ be any sequence of sets and let $f : (\bigcup_{n=1}^{\infty}, d_X) \to (Y, d_Y)$ be a mapping such that the number of distances distorted by f is bounded from above by K. Let f_n denote the restriction of f to A_n. Then it is easy to see that(recall that the mappings are injective) $\exists$ some $0 < C_1 < C_2 < \infty$ such that $\frac{(\binom{n}{2}-K)+C_1}{\binom{n}{2}} \leq \Phi_1(f_n) \leq \frac{(\binom{n}{2}-K)+C_2}{\binom{n}{2}}$. Hence $\lim_{n\to\infty} \Phi_1(f_n) = 1 = \lim_{n\to\infty} \Phi_1(I_n)$. □

Proposition 8 ($\Phi_1(.)$ **is robust to outliers in data**). $\Phi_1(.)$ *satisfies the necessary condition as specified in Scenario 2 to be deemed robust to outliers in data.*

Proof (sketch). Let $I : (X, d_X) \to (Y, d_Y)$ be an isometry. Let $\{\mathcal{A}_n\} \subset X$ be an increasing sequence of sets and let f be the corresponding mapping constructed in conformation to the conditions specified in Scenario 2. For any $n \in \mathbb{N}$, assign an arbitrary ordering of the elements of $\mathcal{A}_n$ as $\{x_0, x_1, ..., x_{n-1}\}$ such that $x_0 \in \bigcap_{n=1}^{\infty} \mathcal{A}_n$. Then average distortion of f_n (f restricted to $\mathcal{A}_n$) is evaluated as

$$\Phi_1(.) = \frac{\binom{n}{2} - (n-1) + \sum_{i=1}^{n-1} \alpha_i}{\binom{n}{2}} \text{ , where } \alpha_i = \frac{d_Y(f(x_i), f(x_0))}{d_X(x_i, x_0)} = \frac{d_Y(f(x_i), f(x_0))}{d_Y(f(x_i), I(x_0))}$$

From the subadditivity of d_Y, it follows that

$$|\frac{d_Y(I(x_0), f(x_0))}{d_Y(I(x_0), f(x_i))} - 1| < \alpha_i < \frac{d_Y(I(x_0), f(x_0))}{d_Y(I(x_0), f(x_i))} + 1$$

By construction, we have that $d_Y(I(x_0), f(x_i)) = d_Y(I(x_0), I(x_i)) = d_X(x_0, x_i) > \beta$ for some $\beta > 0$ and it follows that:

$$\implies 0 < |\frac{d_Y(I(x_0), f(x_0))}{d_Y(I(x_0), f(x_i))} - 1| < \alpha_i < \frac{d_Y(I(x_0), f(x_0))}{\beta} + 1$$
$$\implies \frac{\binom{n}{2} - (n-1)}{\binom{n}{2}} < \Phi_1(f_n) < \frac{\binom{n}{2} - (n-1) + \frac{(n-1)d_Y(I(x_0), f(x_0))}{\beta}}{\binom{n}{2}}$$
$$\implies \lim_{n\to\infty} \Phi_1(f_n) = 1 = \lim_{n\to\infty} \Phi_1(I_n)$$

□

Claim 2. ($\Phi_1(.)$ is robust to noise) . We claim that $\Phi_1(.)$ is robust to noisy observations to a certain degree since noisy observations potentially lead to outliers in the distribution of the ratio of distances and it follows from Propositions 7 and 8 that $\Phi_1(.)$ is(arguably) robust to noisy observations.

Average distortion exhibits a trade off between scale invariance(under certain conditions) and robustness to outliers. Average distortion as defined in definition 13 under the current set of assumptions is not invariant to scaling as shown in Proposition 6. If however, f is restricted to either non-contractive(or non-expansive embeddings) such that

$$min_{u,v\in X, u\neq v} \frac{d_Y(f(u), f(v))}{d_X(u, v)} = 1 (max_{u,v\in X, u\neq v} \frac{d_Y(f(u), f(v))}{d_X(u, v)} = 1) \tag{4.1}$$

then the measure would be invariant to scaling (see analysis below). Note that for any embedding that does not satisfy these properties, it can be normalized such that Equation 4.1 holds. We refer to l_1 distortion of this normalized embedding as the normalized l_1 distortion. However, this restriction would mean that the measure would be volatile against outliers when evaluating the quality of the embedding(see Proposition 10). Non contractive embeddings would be hugely affected by the minima and vice versa. We observe the same in our experiments. In fact the measures of the worst case distortion and average distortion follow the same trend(see figure 4.3) except for a handful of exceptions. An alternative way of making a distribution invariant to scaling is by dividing each value by the mean of the distribution. If the distribution of the ratio of distances ($\rho_f(u, v)$) is normalized by dividing each value by its mean, average distortion would always be 1 irrespective of the embedding.

4.3.3 Properties of normalized l_1 distortion

Proposition 9 (Normalized $\Phi_1(.)$ is scale invariant). *Normalized $\Phi_1(.)$ is invariant to scaling.*

Proof (sketch). If f and g are two mappings such that $f(x) = \alpha \cdot g(x)$ for some $\alpha \in \mathbb{R}$ for all $x \in X$, from homogeneity of d_Y,

$$\implies \forall (u, v) \in X, \frac{d_Y(f(u), f(v))}{d_X(u, v)} = \frac{|\alpha|(d_Y(g(u), g(v)))}{d_X(u, v)}.$$

$$\implies \min_{u,v\in X,u\neq v} \frac{d_Y(f(u),f(v))}{d_X(u,v)} = \min_{u,v\in X,u\neq v} \alpha\frac{(d_Y(g(u),g(v)))}{d_X(u,v)}$$

$$\implies \Phi_1^{\cdot}(f) \ = \Phi_1^{\cdot}(g)$$

□

Proposition 10 (**Normalized $\Phi_1^{\cdot}(.)$ is not robust to outliers**). *Noramlized $\Phi_1^{\cdot}(.)$ is not robust to outlier distances or to outliers in data.*

Proof (sketch). Consider the example constructed in scenario 1. It is easy to see that the sequence of average distortions evaluated on mappings $\{f_n\}$ is given by

$$\Phi_1^{\cdot}(f_n) \ = \frac{(\binom{n}{2}-1)\frac{1}{\alpha}+1}{\binom{n}{2}}$$
$$\lim_{n\to\infty} \Phi_1^{\cdot}(f_n) \ = \frac{1}{\alpha} \ll \Phi_1^{\cdot}(I_n) \ = 1$$

□

Claim 3. Robustness to noise We claim that normalized $\Phi_1^{\cdot}(.)$ under the assumption of non-contraction is not robust to noisy observations and this follows from similar arguments made in the analysis of l_∞ distortion.

Incorporation of the underlying probability measure: l_1 distortion/average distortion(both the general definition as well as the definition with the assumption of non-contraction) under the framework of l_q distortion, has a natural way of incorporating the underlying probability distribution into its evaluation.(see definition 14).

4.3.4 Properties of l_q distortion $\forall 1 < q < \infty$

Similar to the case of average distortion the assumption of non-contraction along with the condition given in Equation 4.2 needs to be imposed on the embedding in order to make the distortion measure scale invariant(See analysis below). We refer to the l_q distortion of this normalized embedding as the normalized l_q distortion.

$$min_{u,v\in X,u\neq v} \frac{d_Y(f(u),f(v))}{d_X(u,v)} = 1 \tag{4.2}$$

l_q distortions $\forall 1 \leq q < \infty$ are merely normalized l_q norms of the ratio of distances. For any given embedding f and for any probability distribution Π over $\binom{X}{2}$, the measures follow the following order of relations.

$$\forall p,q \geq 1, p \leq q \implies dist_p^{\Pi}(f) \leq dist_q^{\Pi}(f) \tag{4.3}$$

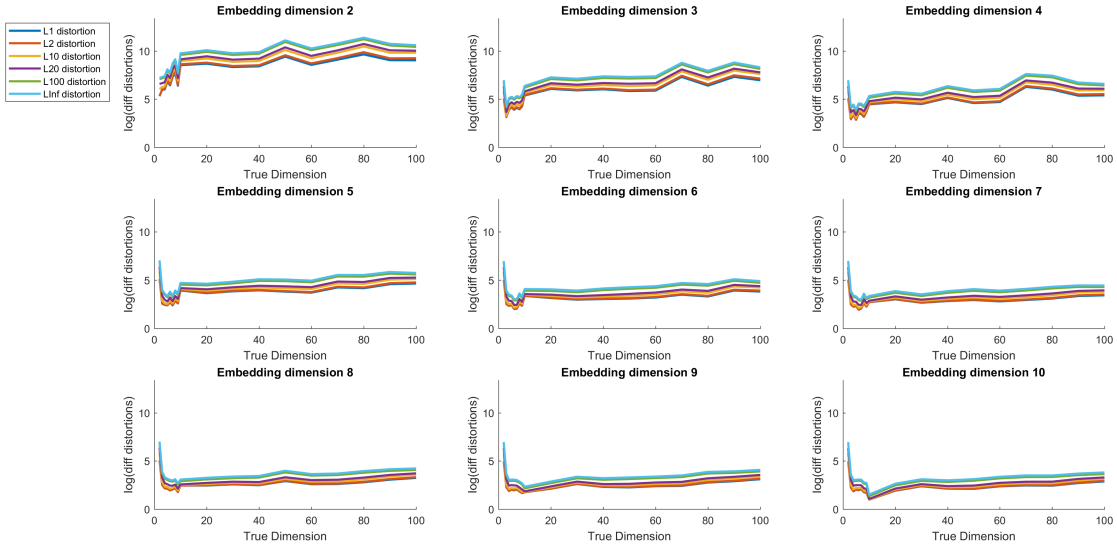


Figure 4.3: The plots correspond to data sampled according to a multivariate standard normal distribution from different dimensions as indicated by the variable True dimension on *x-axes*. Each subplot in the figure corresponds to an embedding to a fixed embedding dimension (created by Isomap) as indicated by the title of each subplot. Different l_q distortions for $q = \{1, 2, 10, 20, 100, \infty\}$ are plotted in each subplot, as indicated in the legend. The plots indicate that the trends of l_q distortion are very similar $\forall 1 \leq q < \infty$

.

It is easy to verify that l_q distortion $\forall 1 < q < \infty$ exhibits the same trade off between scale invariance and robustness to noise and outliers. l_q distortions $\forall 1 \leq q < \infty$ which are not normalized are not scale invariant (Follows from the analysis of l_1 distortion in section 4.3.2).

Since normalized l_1 distortion (or average distortion) and l_∞ distortion are affected by noisy observations and outliers by extension similar arguments can be made to all other definitions of normalized l_q distortion. A sample plot demonstrating that the trends of the normalized l_q distortions for all $1 \leq q < \infty$ are very similar is shown in Figure 4.3. **Note:** In the rest of the thesis, we refer to normalized l_q distortion as simply l_q distortion.

4.4 ϵ-slack distortion and Scaling Distortion

Recall that for any given $\epsilon > 0$, ϵ-slack distortion is defined as worstcase distortion over an $(1-\epsilon)$ fraction of all pairs of points. Observe that as a consequence of this formulation, ϵ-slack distortion is robust to outliers in the sense that was described in Scenarios 1 and 2 for some appropriate choice of ϵ. However, it is not clear what an appropriate choice of ϵ is for a given task. ϵ-slack distortion inherits the properties of scale and translation invariance from worstcase distortion.

In addition, it fails to incorporate the underlying probability measure of the

data space. We assert that this renders ϵ-slack distortion as an ineffective measure of distortion in the context of Machine Learning. To see this clearly, recall that this measure of distortion ignores ϵ fraction of distances and hence provides no guarantees on the distortion incurred on these distances. This is an issue since the ignored distances could be critical for a task. For illustrative purposes, consider the problem of binary classification with highly imbalanced classes (for e.g, classification of cancer cells vs non cancer cells). Treating all pairs of distances uniformly in evaluating ϵ slack distortion could be detrimental if the distances that are ignored are highly important for the task at hand.

Scaling distortion overcomes the limitation due to the paramaterization of distortion by ϵ by defining distortion as a function of ϵ. However, there is no straightforward way of comparing the quality of two embeddings using scaling distortion since it is defined as a function of ϵ.

4.5 σ - distortion

We have so far shown that the existing measures of distortion in literature ineffective in estimating the quality of an embedding in the context of Machine Learning. Driven by our characterization of the quality of an embedding, we propose an alternate measure of distortion which we refer to as *σ-distortion* (see Definition 20) which measures the width of concentration of the ratio of distances, upto a scale. We will show in chapters 4 and 5 that *σ-distortion* overcomes the limitations of the existing measures of distortions while retaining many of the desirable properties.

Definition 20. *Let (X, d_X) and (Y, d_Y) be arbitrary finite metric spaces. Let $f : (X, d_X) \to (Y, d_Y)$ be an injective mapping. Let $\rho_f(u, v)$ denote the ratio $\frac{d_Y(f(u),f(v))}{d_X(u,v)}$ for any $(u, v) \in X \times X$. Let $\widetilde{\rho}_f(u, v)$ denote the normalized ratio of distances given by $\rho_f(u, v)/\sum_{\substack{u,v \in X \times X \\ u \neq v}} \rho_f(u, v)$. Given a distribution $\mathcal{P}$ over X, Let $\Pi = \mathcal{P} \times \mathcal{P}$ denote the distribution on the product space $X \times X$ (Note the inherent assumption of independence), σ-distortion is defined as*

$$\mathbb{E}_{\Pi}(\widetilde{\rho}_f(u, v) - 1)^2$$

If $\mathcal{P}$ is a uniform probability measure over X, then *σ-distortion* measures the variance of the distribution of the normalized ratio of distances, $\widetilde{\rho}_f(u, v)$.

4.5.1 Properties of σ distortion

Proposition 11 (σ **distortion is scale invariant**). *σ-distortion is invariant to scaling.*

Proof. For any two finite metric spaces (X, d_X) and (Y, d_Y), let $f : (X, d_X) \to (Y, d_Y)$ and $g : (X, d_X) \to (Y, d_Y)$ be two mappings such that $\forall u \in X$, $f(u) = \alpha g(u)$, for some $\alpha \in \mathbb{R}$. From homogeneity of d_Y

$$\rho_f(u,v) = \frac{d_Y(f(u),f(v))}{d_X(u,v)} = \frac{d_Y(\alpha g(u),\alpha g(v))}{d_X(u,v)} = |\alpha|\rho_g(u,v)$$
$$\widetilde{\rho}_f(u,v) = \frac{\rho_f(u,v)}{\sum_{i=1}^{n}\rho_f(u,v)} = \frac{|\alpha|\rho_g(u,v)}{\sum_{i=1}^{n}|\alpha|\rho_g(u,v)} = \widetilde{\rho}_g(u,v)$$
$$\implies \mathbb{E}(\widetilde{\rho}_f(u,v)-1)^2 = \mathbb{E}(\widetilde{\rho}_g(u,v)-1)^2$$

□

Proposition 12 (*σ*-***distortion*** **is translation invariant**). *σ-distortion is invariant to translations if the target metric is translation invariant.*

Proof. This follows from the definition of *σ-distortion* . □

Proposition 13 (*σ*-***distortion*** **is robust to outliers in distances**). *σ-distortion is robust to outlier distances in the sense that it satisfies the necessary condition required to deem a measure of distortion robust to outlier distances.*

Proof. For any $f_n : (\mathcal{A}_n, d_X) \to (X, d_X)$, *σ-distortion* (f_n) can be expressed as

$$\sigma\text{-}distortion\ (f_n) = \frac{\sum_{(u,v)\in\binom{\mathcal{A}_n}{2}}[\rho_{f_n}(u,v) - \sum_{(u,v)\in\binom{\mathcal{A}_n}{2}}\rho_{f_n}(u,v)]^2}{\binom{n}{2}[\sum_{(u,v)\in\binom{\mathcal{A}_n}{2}}\rho_{f_n}(u,v)]^2} \tag{4.4}$$

By definition, for any

$$\forall n \in \mathbb{N}, \exists K \in \mathbb{N}, s.t\ |\{(u,v)\in\mathcal{A}_i :\ d_X(f_n(u),f_n(v)) \neq d_X(u,v)\}| \leq K$$
$$\implies \exists n_0 \in \mathbb{N}\ \text{s.t}\ \forall n \geq n_0 |\{(u,v)\in\mathcal{A}_i :\ d_X(f_n(u),f_n(v)) \neq d_X(u,v)\}| = K$$

$$\implies \forall n \geq n_0,\ \rho_{f_n} = [\underbrace{1,1,...,1}_{\binom{n}{2}-K\ \text{times}}, \underbrace{\alpha_1,\alpha_2,...,\alpha_K}_{K}] \tag{4.5}$$

From 4.4 and 4.5, after substitution and simplification, $\forall n > n_0, \sigma\text{-}distortion\ (f_n)$,

$$= \frac{(K-\sum_{i=1}^{K}\alpha_i)^2(\binom{n}{2}-K) + \sum_{i=1}^{K}(\binom{n}{2}(\alpha_i-1)+K-\sum_{i=1}^{K}\alpha_i)^2}{\binom{n}{2}(\binom{n}{2}-K+\sum_{i=1}^{K}\alpha_i)^2}$$

$$\implies \lim_{n\to\infty}\sigma\text{-}distortion\ (f_n) = 0 = \lim_{n\to\infty}\sigma\text{-}distortion\ (I_n)$$ □

Proposition 14 (*σ*-***distortion*** **is robust to outliers in data**). *σ-distortion is robust to outliers in data in the sense that it satisfies the necessary condition required to deem a measure of distortion robust to outliers in data.*

Proof (sketch). For a sequence of f_n constructed as specified in scenario 2, let $\mu(\rho)$ denote $\frac{\sum\limits_{(u,v)\in\binom{\mathcal{A}_n}{2}} \rho_{f_n}(u,v)}{\binom{n}{2}}$, then σ-*distortion* (f_n) is evaluated as:

$$\begin{aligned}
&= \frac{\sum\limits_{(u,v)\in\binom{\mathcal{A}_n}{2}} [\rho_{f_n}(u,v) - \mu(\rho)]^2}{\binom{n}{2}\mu(\rho)^2} \\
&= \frac{\sum\limits_{i=1}^{\binom{n}{2}-(n-1)} [1-\mu(\rho)]^2 + \sum\limits_{i=1}^{n-1}[\alpha_i - \mu(\rho)]^2}{\binom{n}{2}[\mu(\rho)]^2} \text{ , where } \alpha_i = \frac{d_Y(f(x_i), f(x_0))}{d_X(x_i, x_0)}
\end{aligned}$$

By substituting $\mu(\rho) = \dfrac{\binom{n}{2} - (n-1) + \sum\limits_{i=1}^{n-1}\alpha_i}{\binom{n}{2}}$ we have,

$$= \frac{\binom{n}{2}^2(\sum\limits_{i=1}^{n-1}\alpha_i^2) - (2\binom{n}{2}(\binom{n}{2} - (n-1)))\sum\limits_{i=1}^{n-1}\alpha_i - \binom{n}{2}(\sum\limits_{i=1}^{n-1}\alpha_i)^2}{\binom{n}{2}[(\binom{n}{2} - (n-1))^2 + (\sum\limits_{i=1}^{n-1}\alpha_i)^2 + 2\sum\limits_{i=1}^{n-1}\alpha_i(\binom{n}{2} - (n-1))]^2}$$

Since I is an isometry and by definition $f(x_i) = I(x_i)$,

$$\frac{d_Y(f(x_i), f(x_0))}{d_X(x_i, x_0)} = \frac{d_Y(f(x_i), f(x_0))}{d_Y(I(x_i), I(x_0))} = \frac{d_Y(f(x_i), f(x_0))}{d_Y(f(x_i), I(x_0))}$$

$$|\frac{d_Y(I(x_0), f(x_0))}{d_Y(I(x_0), f(x_i))} - 1| < \alpha_i < \frac{d_Y(I(x_0), f(x_0))}{d_Y(I(x_0), f(x_i))} + 1$$

By construction, we have that $d_Y(I(x_0), f(x_i)) = d_Y(I(x_0), I(x_i)) = d_X(x_0, x_i) > \beta$ for some $\beta > 0$ and it follows that:

$$\begin{aligned}
&\implies 0 < |\frac{d_Y(I(x_0), f(x_0))}{d_Y(I(x_0), f(x_i))} - 1| < \alpha_i < \frac{d_Y(I(x_0), f(x_0))}{\beta} + 1 \\
&\implies \exists C \in \mathbb{R}, \text{ s.t } \forall i \in \mathbb{N}\ 0 < \alpha_i < C \\
&\implies 0 < \sum_{i=1}^{n-1}\alpha_i < (n-1)C \\
&\implies 0 < \sum_{i=1}^{n-1}\alpha_i^2 < (n-1)C^2 \\
&\implies 0 < (\sum_{i=1}^{n-1}\alpha_i)^2 < ((n-1)C)^2
\end{aligned}$$

By substitution and simplification, it follows that $\forall n > 2$,

$$= \frac{-(2\binom{n}{2}(\binom{n}{2} - (n-1)))((n-1)C) - \binom{n}{2}((n-1)C)^2}{\binom{n}{2}[(\binom{n}{2} - (n-1))^2 + ((n-1)C)^2 + 2(n-1)C(\binom{n}{2} - (n-1))]^2}$$

$$< \sigma\text{-}distortion\ (f_n) < \frac{\binom{n}{2}^2 C^2(n-1)}{\binom{n}{2}(\binom{n}{2} - (n-1))^2}$$

$$\lim_{n\to\infty} \frac{-(2\binom{n}{2}(\binom{n}{2} - (n-1)))((n-1)C) - \binom{n}{2}((n-1)C)^2}{\binom{n}{2}[(\binom{n}{2} - (n-1))^2 + ((n-1)C)^2 + 2(n-1)C(\binom{n}{2} - (n-1))]^2} = 0$$

$$\lim_{n\to\infty} \frac{\binom{n}{2}^2 C^2(n-1)}{\binom{n}{2}(\binom{n}{2} - (n-1))^2} = 0$$

Hence it follows that $\lim_{n\to\infty} \sigma\text{-}distortion\ (f_n) = 0 = \lim_{n\to\infty} \sigma\text{-}distortion\ (I_n)$ □

Claim 4. *σ-distortion* is robust to noise By virtue of similar arguments made in the analysis of the existing measures of distortion, we claim that *σ-distortion* is(arguably) robust to noise.

Incorporation of the underlying probability measure: The underlying probability measure of the data space can be naturally incorporated into the measure of distortion to attain meaningful estimates of the quality of an embedding.

We showed that *σ-distortion* possesses all the properties required of a measure of distortion as discussed in section 4.2. Our experiments (Chapter 5) support our claims and suggest that *σ-distortion* is a much more meaningful and a stable estimate of the quality of an embedding compared to the other measures of distortion in the context of Machine Learning. However, theoretical properties of this measure of distortion have not yet been fully explored in the course of this thesis and would be part of future work (see Chapters 6 and 7).

Chapter 5

Experiments

We conducted several experiments of exploratory nature to investigate whether the existing measures of distortion are well suited to evaluate the quality of an embedding. We designed experiments to examine if the various measures of distortion exhibit patterns that deviate from expected behaviour. These experiments are categorized as follows:

(a) **Distortion vs Embedding Dimension:** For a dataset originally sampled from a fixed dimension, it is natural to assume that the quality of an embedding generated by an algorithm which aims to preserve distances improves with increasing embedding dimension. This tradeoff has been theoretically established in literature [Abraham et al. (2008), Chan et al. (2010)] in the case of worst case distortion. However, no known results exist which demonstrate the tradeoff between embedding dimension and distortion for any other measure of distortion (including l_q distortions and σ distortion). We conduct tests to verify if the various measures of distortions exhibit this tradeoff. In Section 5.2.1, we observe that all the measures of distortion, including σ-distortion, exhibit the tradeoff between distortion and embedding dimension. We also utilize this experiment as a means of validation for σ distortion as a meaningful measure of distortion.

(b) **Distortion vs Original Dimension:** Similarly, for a fixed embedding dimension, it is also natural to expect that, under suitable conditions, the quality of an embedding decreases with increasing original dimension. This tradeoff is supported in the case of worstcase distortion by a volume argument similar to the one in Example 3.4.1, which shows that the minimum dimension required for embedding a doubling metric into any l_p space with worstcase distortion Φ is $\Omega(\frac{\lambda}{log\,\Phi})$, where λ denotes the doubling dimension of the original space. This indicates that the distortion incurred for embedding into a fixed dimension is expected to increase with the intrinsic dimension of the original space. We conduct experiments in order to investigate this hypothesis and empirically show that l_q distortions do not always conform to the expected trend. In addition, we also provide empirical evidence to show that σ distortion consistently follows an increasing trend with the original dimension(conforming to

the expectation) even in the cases where l_q distortions fail to conform to the expected trend.

(c) **Correlation with the concentration of the ratio distribution:** A measure of distortion is expected to act as a means of quantification of the quality of an embedding. As mentioned earlier, we argue that in the context of Machine Learning, a good quality embedding can be characterized as a mapping between two metric spaces such that the ratio of distances(embedding distance by original distance or vice versa) concentrates very sharply around 1 up to a scaling factor. In the third set of experiments, we conduct tests to verify whether l_q distortions as well as σ distortion conform to this characterization of the quality of an embedding.

(d) **Robustness to Noise and Outliers:** In section 4.3, we demonstrated that l_q distortions are not robust to outliers and noisy observations and hence are not appropriate distortion measures in the Machine Learning context. In addition, we also claimed that σ distortion is robust to noise and outliers. In the last set of experiments, we garner empirical evidence to support our theory.

5.1 Experimental Setup

In this section, we describe the experimental setup including the various datasets and algorithms used for creating the embeddings. In order to conduct the experiments, we constructed several datasets generated from various probability distributions in different original dimensions as follows:

1. **Gaussian distribution:** We created normally distributed data in low dimensions 2 to 9 with sample size $N = 1000$ and in high dimensions $\{10, 20, ..., 100\}$ with sample size $N = 10000$ where a datapoint of dimension d is generated by independently sampling each coordinate from a standard normal distribution.

2. **Gamma distribution:** We created datasets sampled from gamma distributions (with parameters $(a, b) \in \{(100, 10), (10, 10), (1, 4), (1.5, 4)\}$) in low dimensions 2 to 9 with sample size $N = 1000$ and in high dimensions $\{10, 20, ..., 100\}$ with sample size $N = 10000$ where a data point of dimension d is generated by independently sampling each coordinate from gamma distributions of the corresponding parameters. Figure 5.1 shows probability density functions(PDF) of the various gamma distributions used to generate the data.

3. **Beta distribution:** We created datasets sampled from beta distributions (with parameters $(a, b) \in \{(1.5, 4), (0.75, 0.75), (0.75, 4)\}$) in low dimensions 2 to 9 with sample size $N = 1000$ and in high dimensions $\{10, 20, ..., 100\}$ with sample size $N = 10000$ where a data point of dimension d is generated by independently sampling each coordinate from beta distributions of

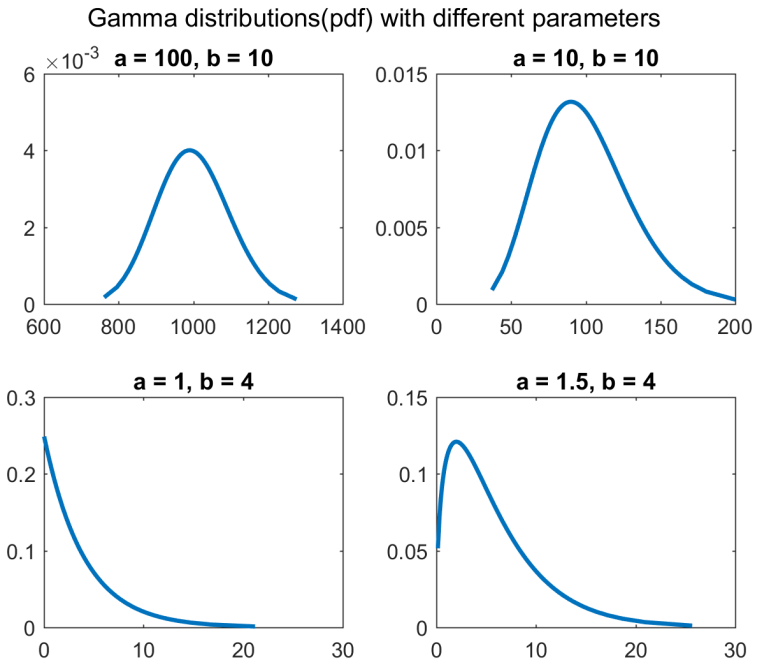


Figure 5.1: Each figure represents the PDF of a Gamma distribution with parameters as shown in the title of each subplot.

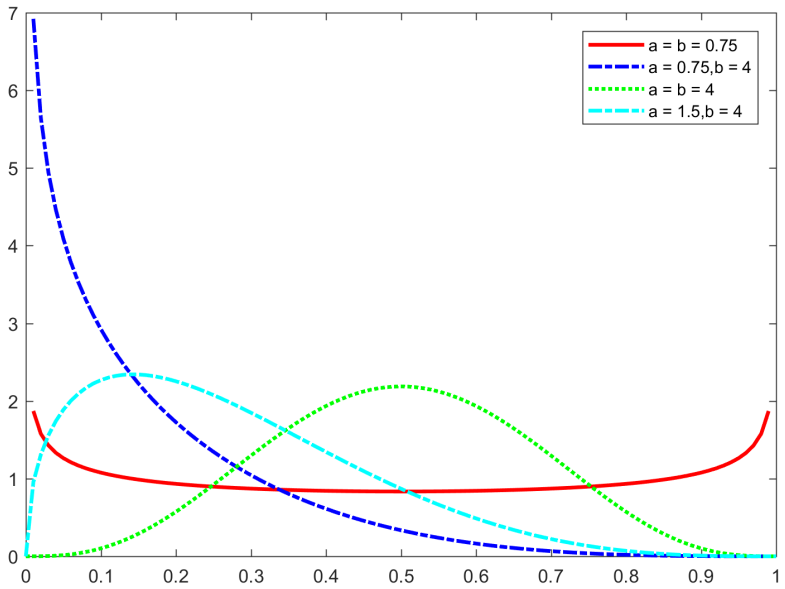


Figure 5.2: Each figure represents the PDF of a Beta distribution with parameters as shown in the title of each subplot.

the corresponding parameters. Figure 5.2 shows PDF of the various beta distributions used to generate the data.

4. **Gaussian mixture distribution:** We created datasets sampled from a

multivariate Gaussian mixture distribution with means 0, 4 and variance 1 with covariance between any two random variables 0 . The data is generated in low dimensions 2 to 9 with sample size $N = 1000$ and in high dimensions $\{10, 20, ..., 100\}$ with sample size $N = 10000$.

5. **Laakso space:** We created complete Laakso spaces in dimensions 2 to 10

In addition, for each dataset in dimensions 2 to 9, we created a noisy version of the dataset in $\mathbb{R}^{10}$ by adding 10 dimensional Gaussian noise with mean 0 and low variance (relative to the variances of the original dataset).

For each dataset, embeddings were created using three different algorithms Isomap [Tenenbaum et al. (2000)], Maximum variance unfolding(MVU) [Weinberger and Saul (2006)] and Laplacian Eigenmaps [Belkin and Niyogi (2003)] into dimensions 2 to min(10, dimension of the original space).

5.2 Results

In this section, we present the results from various experiments conducted as specified earlier.

5.2.1 Distortion vs Embedding dimension

A preliminary experiment was designed to verify if the various measures of distortion conform to the tradeoff between embedding dimension and distortion as discussed earlier. In addition to this experiment being a test of verification for the existing measures of distortion, it also acts as a means of validation for σ-*distortion*. For a dataset sampled from a fixed dimension, we would expect a measure of distortion to decrease (if the true dimension is greater than the embedding dimension) or stay approximately the same (if the true dimension is nearly the same or lesser than the embedding dimension) as the embedding dimension increases.

Our experiments (for e.g, see figure 5.3) indicate that all the measures of distortion including σ-*distortion* exhibit the tradeoff between embedding dimension and the measure of distortion (Note that in this example, the original dimension is greater or equal to the embedding dimension and hence we would expect that the curves corresponding to all original dimensions follow a decreasing trend).

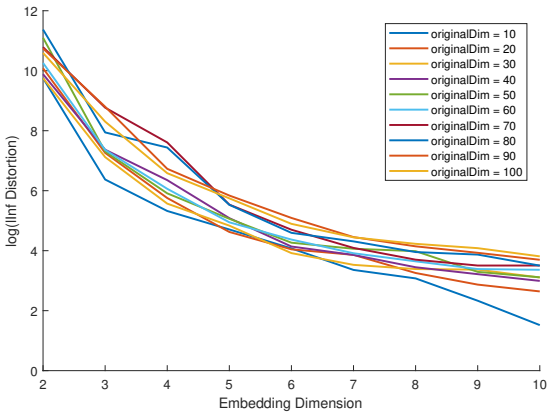
5.2.2 Distortion vs Original dimension

The next important experiment conducted was to investigate if, for any fixed embedding dimension, the various measures of distortion follow an increasing trend with the original dimension. In each experiment, we fix a univariate base probability distribution(for instance Gaussian distribution with mean 0 and unit variance). In order to generate a dataset of a particular original dimension(d), we create data points by sampling each of the d coordinates independently from the base distribution.

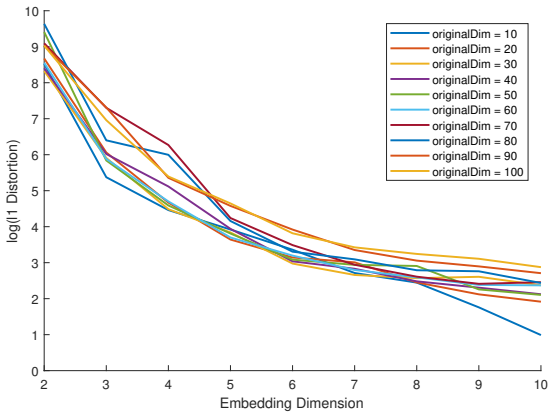
This experiment(see for e.g Figure 5.4 and Figure 5.5) revealed that the existing measures of distortion(specifically all l_q distortions) exhibit anomalous behaviour in the sense that they can change erratically as opposed to increasing with the original dimension(which is the expected behaviour). However, σ-*distortion* follows the expected increasing trend with original dimension. These patterns were predominant across the entire spectrum of our experiments. This anomalous behaviour can be explained by means of the analysis presented in Sections 5.2.3 and 5.2.4.

5.2.3 Correlation with concentration of the ratio distribution

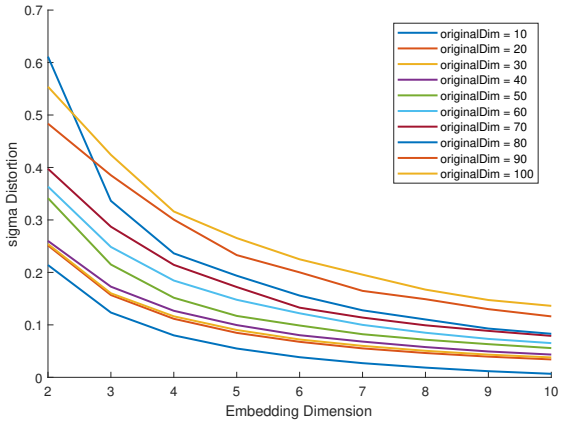
In this Section we analyze the set of experiments conducted to investigate whether the various measures of distortion assess the quality of an embedding appropriately. As mentioned earlier, we characterize the quality of an embedding by the sharpness



(a) l_∞ distortion vs embedding dimension



(b) l_1 distortion vs embedding dimension

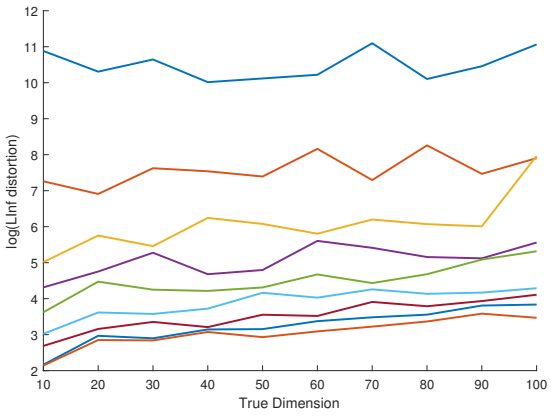


(c) σ distortions vs embedding dimension

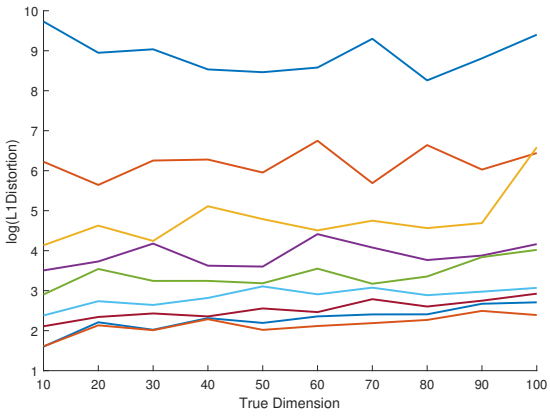
Figure 5.3: Figure **(a)** shows the decreasing trend of worstcase distortion with increasing embedding dimension for a fixed dataset($N = 10000$) generated by a standard normal distribution of a dimension as indicated by *originalDim*. Figure **(b)** corresponds to average distortion. Figure **(c)** corresponds to σ-*distortion*. **Note:** Embeddings were computed using Isomap and the logarithm of worstcase and average distortion measures are plotted for better viewing of the results.

of concentration of the ratio distribution. The first experiment in this setting was to compare the quality of embeddings across different embedding algorithms. The ratio distributions are normalized for each embedding to ensure that the sharpness of concentration of the distribution is not affected by scaling. Our experiments show that the existing measures of distortion often fail to reflect the quality of an embedding according to our characterization. In contrast, σ-*distortion* captures this characterization precisely.

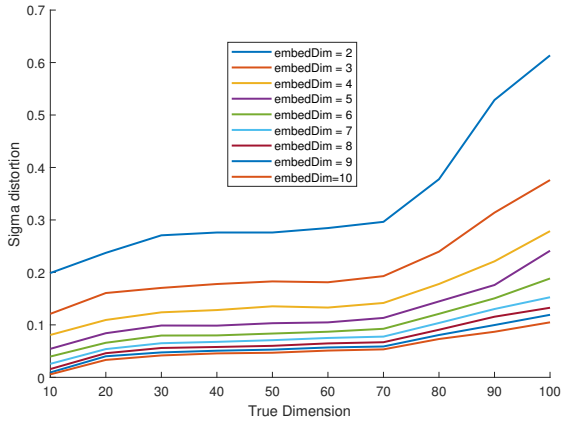
The following example illustrates this discrepancy. Consider data sampled from $\mathbb{R}^2$ according to a normal distribution with additive gaussian noise of low variance in $\mathbb{R}^{10}$. This data is embedded into $\mathbb{R}^6$ using Isomap and MVU. The distribution of the ratio of distances $(\widetilde{\rho}_f(u, v))$ and the corresponding empirical CDF's for the embeddings created by Isomap and MVU are shown in figure 5.6. It can be clearly observed from the figure that distribution due to Isomap concentrates much more



(a) Original Dimension vs l_∞ distortion



(b) Original Dimension vs l_1 distortion

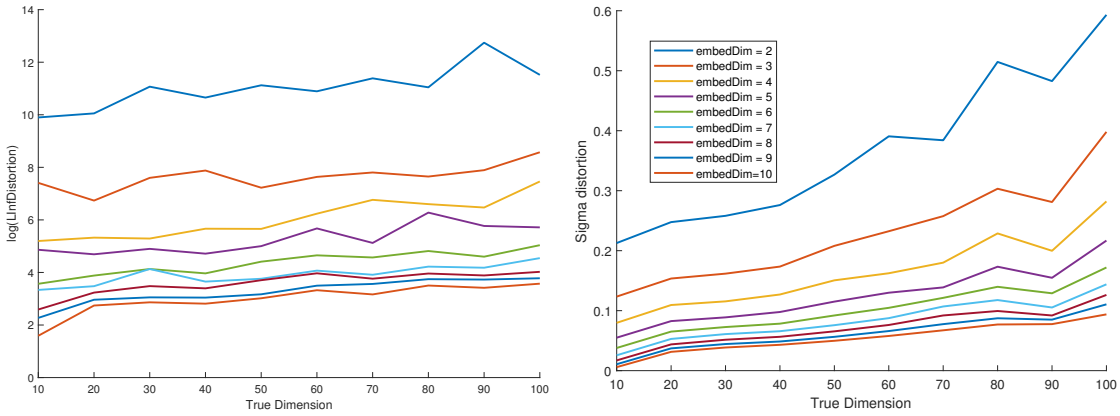


(c) Original Dimension vs σ distortion

Figure 5.4: These plots correspond to datasets($N = 10000$) sampled according to independent gamma distributions with original dimension shown on the x-axes and with parameters a = 1.5 and b = 4. Embeddings were created using Isomap. **(a)** and **(b)** corresponding to l_∞ and l_1 distortions demonstrate the erratic behaviour as described and **(c)** corresponding to σ distortion follows the expected behaviour.

sharply around 1 in comparison to the distribution due to MVU. However, as shown in Figure 5.7, the measures of worstcase distortion as well as average distortion(by extension l_q *distortion* $\forall 1 \leq q < \infty$) show the opposite trend and hence fail to capture this.

To understand why this happens, consider the minimum and the maximum values of the ratios for the two embeddings shown in Figure 5.6. Recall from Definition 10 that worst case distortion can be computed as $max(\widetilde{\rho}_f(u,v)) * \frac{1}{min(\widetilde{\rho}_f(u,v))}$ This implies that the value of the minimum in this example, disproportionately affects the value of worstcase distortion(Observe that the right tail of the distribution due to Isomap is shorter in comparison to the distribution due to MVU). Similarly, in the case of average distortion, in accordance to our speculation in Section 4.3.3, it is observed that the difference in the minima disproportionately affects the measure of average distortion. As with the other results we presented in



(a) Original Dimension vs l_∞ distortion **(b)** Original Dimension vs σ distortion

Figure 5.5: These plots correspond to datasets($N = 10000$) sampled according to independent beta distributions with original dimension shown on the x-axes and with parameters a = 1.5 and b = 4. Embeddings were created using Isomap. **(a)** corresponding to l_∞ distortion demonstrate the erratic behaviour as described and **(b)** corresponding to σ distortion follows the expected behaviour.

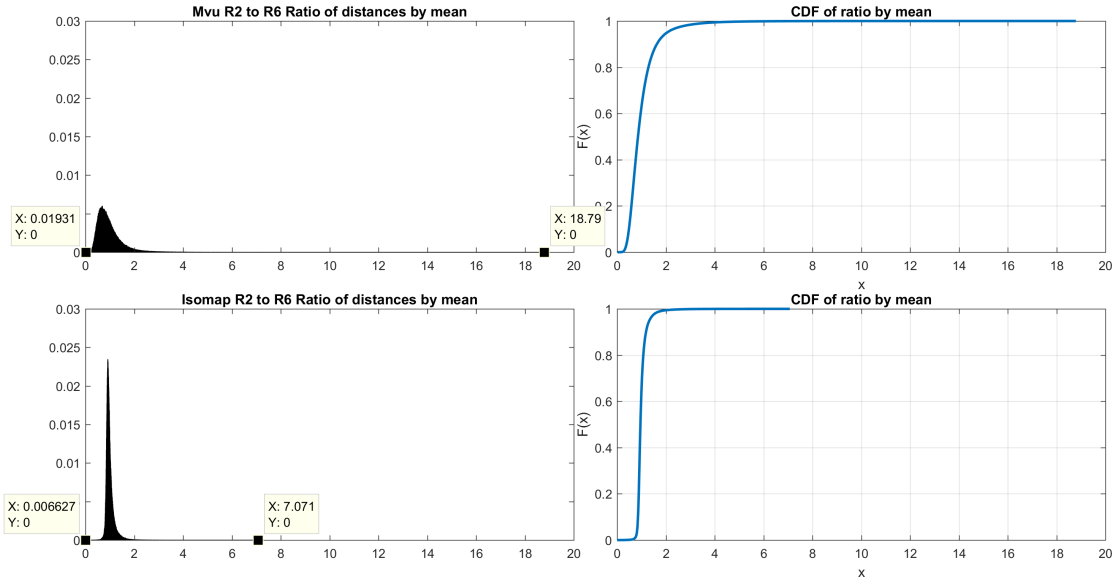


Figure 5.6: The plots correspond to embeddings of normally distributed data($N = 1000$ with mean 0 and variance 1) sampled from $\mathbb{R}^2$ embedded into $\mathbb{R}^6$ using Isomap(**top**) and MVU(**bottom**). **Left:** Histogram of normalized ratio of distances of the embeddings. **Right:** The empirical CDF of the ratio of distances. The datatips show the minimum and the maximum values of the support of each distribution.

this section, anomalous trends similar to that of this example have been prevalent through out the entire spectrum of our experiments. As shown in Figure 5.7, σ-*distortion* precisely captures the expected trend. More examples illustrating this behaviour can be seen in Section 5.2.4. These observations positively support our hypothesis that all measures of l_q distortion are not reflective of the quality of an

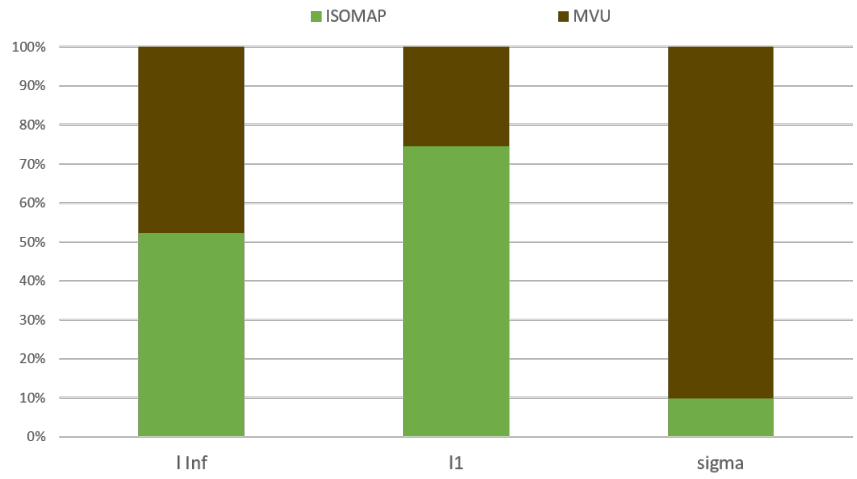


Figure 5.7: Values of l_∞, l_1 and σ distortions for the embeddings due to Isomap and MVU up to a scale. It corresponds to Figure 5.6. **Note:**Since each measure of distortion generates values at a different scale, the plot shows the proportion of the distortions contributed by each embedding and can be used to interpret the trend.

embedding.

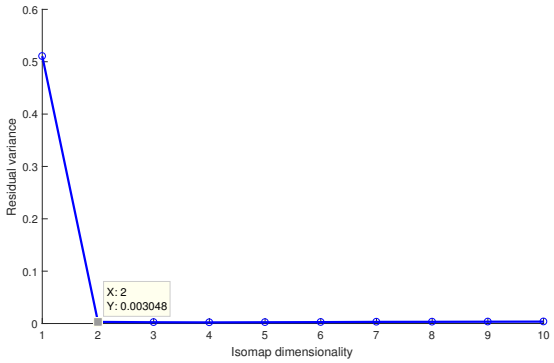
5.2.4 Effect of noise and outliers on distortions

The next set of experiments were designed specifically to test the effect of noise and outliers on the measures of distortion.

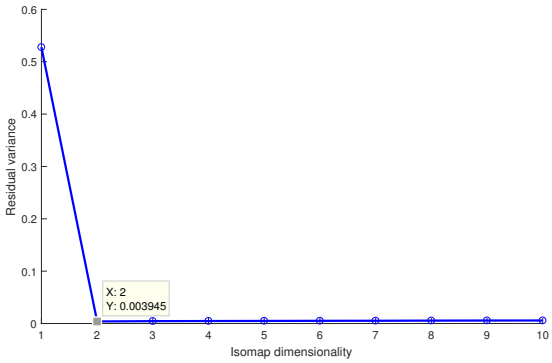
For illustrative purposes, consider data(X) sampled from a 2 dimensional subspace in $\mathbb{R}^{10}$ generated by independent standard normal distributions. This dataset is embedded into various embedding dimensions using Isomap. A dataset(X') is created from X by adding 10 dimensional Gaussian noise of relatively low variance to X. X' is similarly embedded using Isomap. The residual variance after embedding X and X' into Euclidean spaces of dimensions 2 to 10 is shown in figure 5.8. Consider the embeddings of X and X' into $\mathbb{R}^2$. The figure indicates that the embeddings of both X and X' into $\mathbb{R}^2$ are fairly similar with the embedding of X resulting in a slightly lower residual variance than the embedding of X'. In addition, it also shows that the manifold dimension of X' is close to 2, indicating that the effect of noise on the intrinsic dimension of X' is minimal if any.

Figure 5.9 shows the distributions of the ratios of distances($\widetilde{\rho}(u, v)$) and their corresponding Empirical CDFs for the embeddings $X \to \mathbb{R}^2$ and $X' \to \mathbb{R}^2$. It can be seen in the figure that the distribution of ratio of distances of X is more sharply concentrated around 1 compared to that of X'. However, the two distributions are fairly similar and this can be observed clearly by comparing the visualizations via the corresponding CDFs.

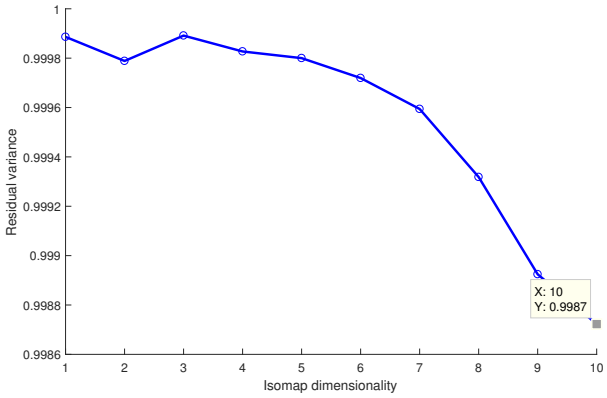
As shown in figure 5.10 all the distortion measures (including *σ-distortion*) follow the expected trend. Meaning that the embedding of $X \to \mathbb{R}^2$ has a lower value of distortion in comparison to the embedding $X' \to \mathbb{R}^2$. However, we argue



(a) Residual variance $\mathbb{R}^2 \rightarrow \mathbb{R}^2$ without noise



(b) Residual variance $\mathbb{R}^2 \rightarrow \mathbb{R}^2$ with noise



(c) Residual variance $\mathbb{R}^{100} \rightarrow \mathbb{R}^{10}$

Figure 5.8: This shows the residual variances of various embeddings (a) $\mathbb{R}^2 \rightarrow \mathbb{R}^2$ without noise, (b) $\mathbb{R}^2 \rightarrow \mathbb{R}^2$ with noise and (c) $\mathbb{R}^{100} \rightarrow \mathbb{R}^{10}$. The datatips show corresponding residual variance of the embedding.

that the increase in distortion from embedding of X to that of X' is disproportionate in the case of l_∞ and l_1 distortions. In order to see this, we compare the embeddings of X and X' with a third embedding. To clearly demonstrate this effect, we consider an embedding of very low quality both in terms of the residual variance as well as the concentration of the ratio of distances. In our example, we consider normally distributed data in $\mathbb{R}^{100}$ and its embedding generated by Isomap into $\mathbb{R}^{10}$.

Figure 5.9 shows the distribution of the ratio of distances corresponding to all three embeddings and the corresponding CDFs. It is abundantly clear that the embedding $\mathbb{R}^{100} \rightarrow \mathbb{R}^{10}$ is significantly worse in comparison to the other two embeddings. Figure 5.8 also shows the residual variance of the embedding $\mathbb{R}^{100} \rightarrow \mathbb{R}^{10}$, painting the same picture. However, as shown in figure 5.10, the trends of l_∞ and l_1 distortion tell a different story. The values of distortion of the embedding of $X' \rightarrow \mathbb{R}^2$, is disproportionately higher than the distortion of $X \rightarrow \mathbb{R}^2$ as well as the embedding $\mathbb{R}^{100} \rightarrow \mathbb{R}^{10}$. This behaviour can be easily explained by similar arguments made previously by taking a closer look at the values of the minima

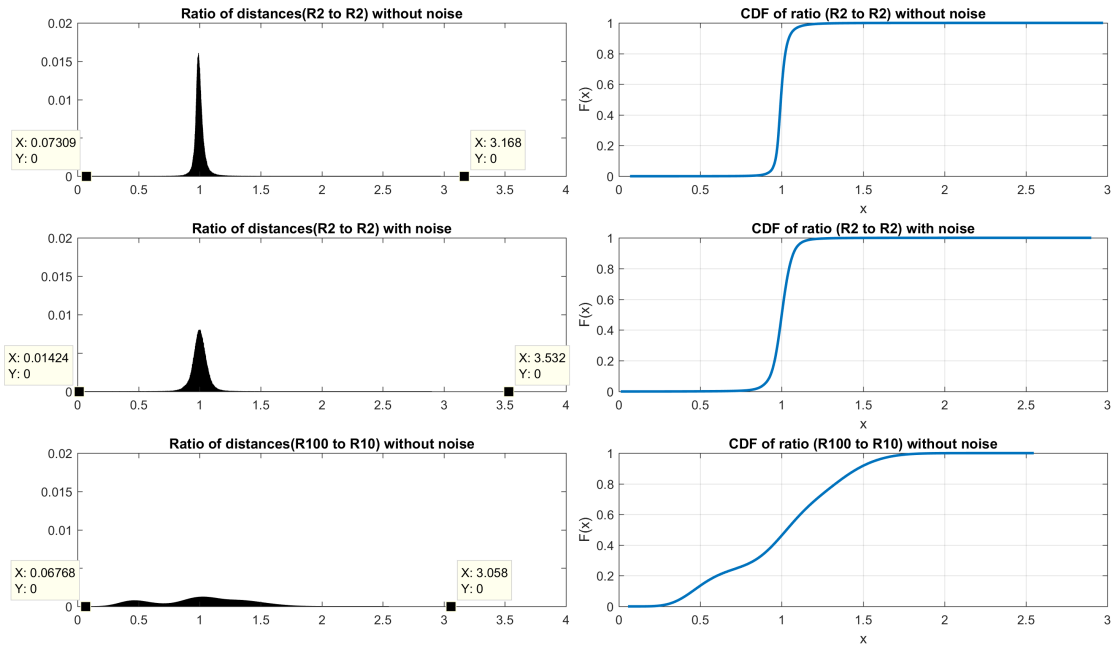


Figure 5.9: Left: Distribution of the ratio of embedding distance to original distance for different datasets. The datatips show the values of the minima and the maxima of the support of the distribution. **Right:** Corresponding Empirical CDF's for better visualization. **Top:** Data sampled from $\mathbb{R}^2$ embedded into $\mathbb{R}^2$. **Middle:** Data sampled from $\mathbb{R}^2$ additive gaussian noise of low variance in $\mathbb{R}^{10}$ embedded into $\mathbb{R}^2$. **Bottom:** Data sampled from $\mathbb{R}^{100}$ embedded into $\mathbb{R}^{10}$. The base distribution used in this Experiment is the standard normal distribution.

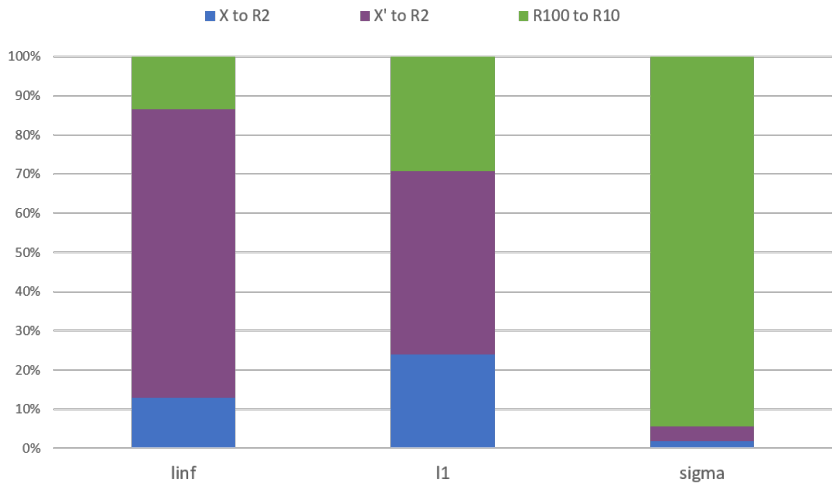


Figure 5.10: Values of l_∞, l_1 and σ distortions for the three embeddings up to a scale. Corresponds to figure 5.9. Since each distortion generates values at a different scale, the plot shows the proportion of the distortions contributed by each embedding and can be used to interpret the trend.

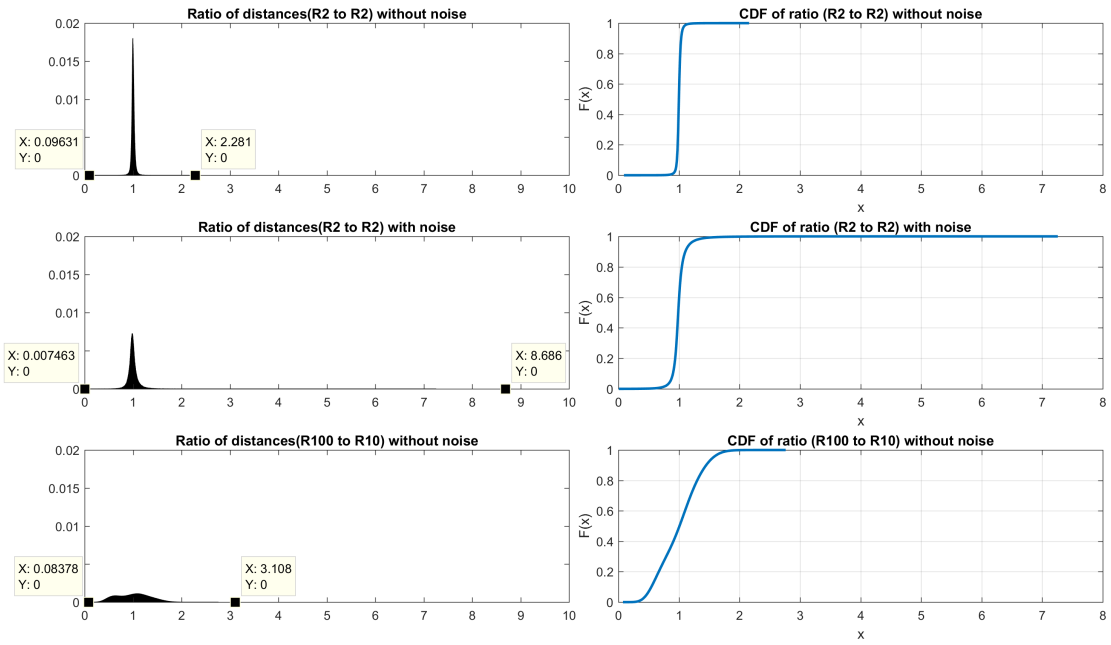


Figure 5.11: Left: Distribution of the ratio of embedding distance to original distance for different datasets. The datatips show the values of the minima and the maxima of the support of the distribution. **Right:** Corresponding Empirical CDF's for better visualization. **Top:** Data sampled from $\mathbb{R}^2$ embedded into $\mathbb{R}^2$. **Middle:** Data sampled from $\mathbb{R}^2$ with additive gaussian noise of low variance with mean 0 in $\mathbb{R}^{10}$ embedded into $\mathbb{R}^2$. **Bottom:** Data sampled from $\mathbb{R}^{100}$ embedded into $\mathbb{R}^{10}$. **Note:** The base distribution used in this experiment is the gamma distribution with parameters $a = 100$ and $b = 10$

.

and the maxima in figure 5.9. However, σ-*distortion* consistently remains robust against outliers and noisy observations and reflects the expected trend by assigning a very high value of distortion to the embedding $\mathbb{R}^{100} \to \mathbb{R}^{10}$, a moderately low value to $X' \to \mathbb{R}^2$ and a lower value to $X \to \mathbb{R}^2$.

We present a few additional examples to further demonstrate the volatility of l_∞ and l_1 distortions(and by extension all other l_q distortions) as well as the stability of σ distortion to noise and outliers. Figure 5.11 shows the same phenomenon described in the previous illustration for a dataset sampled according to a gamma distribution(See figure caption for more details).

Figure 5.11 clearly shows that the sharpness of concentration of the distribution of the ratio of distances decreases from top to bottom. However, similar to the previous example, Figure 5.12 shows that l_1 and l_∞ distortions are disproportionately affected by the addition of noise demonstrating the volatility of l_q distortions. The figure also shows that σ distortion remains stable under noise.

To highlight the effect of outliers in isolation, we present the following illustration. Figure 5.13 clearly shows that the sharpness of concentration of the distributions are nearly the same for the embeddings $\mathbb{R}^2 \to \mathbb{R}^2$ without noise and $\mathbb{R}^2 \to \mathbb{R}^2$ with noise and it is very low for the embedding $\mathbb{R}^{100} \to \mathbb{R}^{10}$. However, as shown

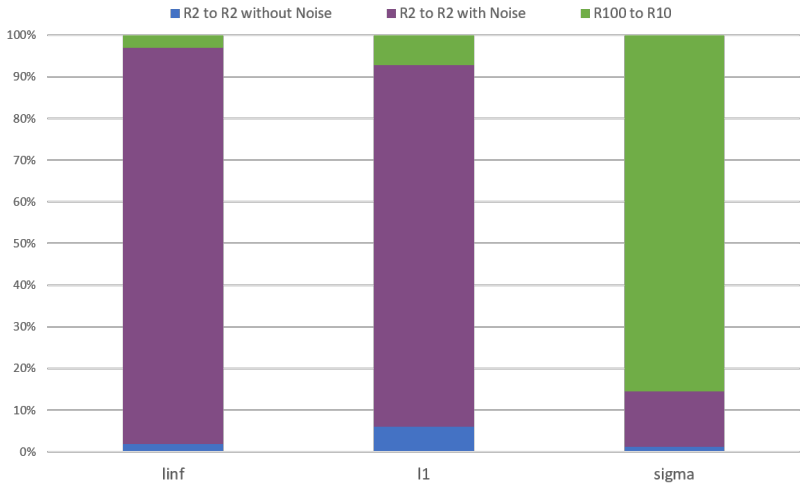


Figure 5.12: Values of l_∞, l_1 and σ distortions for the three embeddings up to a scale. Corresponds to Figure 5.11. Since each distortion generates values at a different scale, the plot shows the proportion of the distortions contributed by each embedding and can be used to interpret the trend.

.

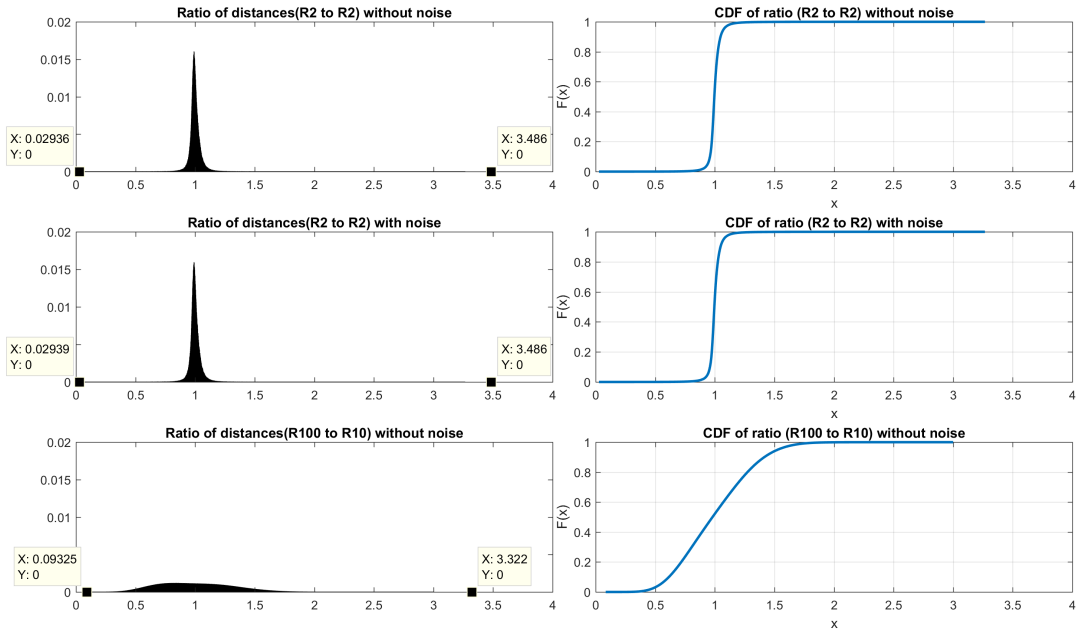


Figure 5.13: Left: Distribution of the ratio of embedding distance to original distance for different datasets. The datatips show the values of the minima and the maxima of the support of the distribution. **Right:** Corresponding Empirical CDF's for better visualization. **Top:** Data from $\mathbb{R}^2$ embedded into $\mathbb{R}^2$. **Middle:** Data sampled from $\mathbb{R}^2$ with additive noise of low variance in $\mathbb{R}^{10}$ embedded into $\mathbb{R}^2$. **Bottom:** Data sampled from $\mathbb{R}^{100}$ embedded into $\mathbb{R}^{10}$. The base distribution used in this experiment is a beta distribution with parameters $a = 1.5$ and $b = 4$

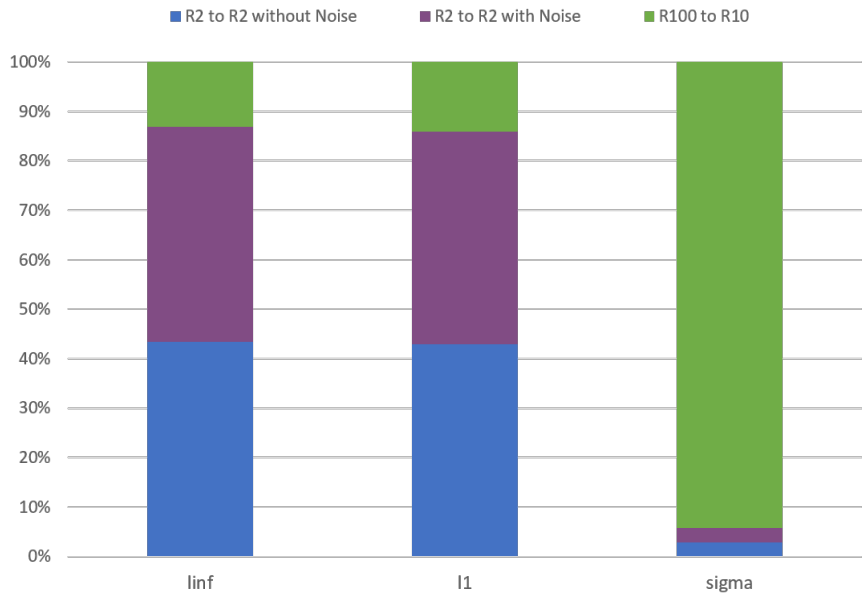


Figure 5.14: Values of l_∞, l_1 and σ distortions for the three embeddings up to a scale. Corresponds to figure 5.13. Since each distortion generates values at a different scale, the plot shows the proportion of the distortions contributed by each embedding and can be used to interpret the trend.

.

by Figure 5.14 both l_∞ and l_1 distortions of the embedding $\mathbb{R}^{100} \to \mathbb{R}^{10}$ are lower than the corresponding distortions of the other two embeddings. It is easy to see from minima and the maxima values in Figure 5.13 that the presence of outliers cause this volatility in the measures of l_q distortion $\forall 1 \leq q \leq \infty$. Once again it can be seen from Figure 5.14 that σ distortion consistently follows the expected trend and remains stable against outliers.

5.2.5 Summary of Results

We presented experimental analysis of the measures of σ distortion, l_∞ and l_1 distortions(and by extension all l_q distortions). The first experiment we conducted was a preliminary test to validate the different measures of distortion as viable measures of quantification of the quality of an embedding. For a fixed dataset, we plotted each measure of distortion vs the corresponding embedding dimension. We observed that all the measures of distortion exhibit a decreasing trend with the embedding dimension thereby conforming to the expected behaviour. Then we presented experiments to assess the functional properties of the various distortion measures. The first of these experiments was to verify if the various measures of distortion exhibit an increasing trend with the original dimension for a fixed embedding dimension. This experiment revealed that the measures of l_q distortion fail to conform to the expected trend while σ distortion consistently follows the expected trend. The next set of experiments test if the estimates generated by

various measures of distortion correlate with the sharpness of concentration of the ratio distribution (which arguably is a plausible interpretation of the quality of an embedding in the context of Machine Learning as discussed before). We showed that the measures of l_q distortion do not consistently correlate with this characterization. In the last set of experiments we verify if the various distortion measures are stable against noise and outliers. Our experiments revealed that while l_q distortions often are volatile against noise and outliers, σ distortion showed robustness to the same.

These results validate our hypothesis that the existing measures of distortion do not vary smoothly with noise and are volatile against outliers. They also support our claim that σ-*distortion* is a stable and a more suitable measure of distortion in the context of Machine Learning.

Chapter 6

Discussion

In this thesis, we initiated a systematic study of the theory of metric embeddings in the context of Machine Learning. We emphasized on the three fundamental questions that broadly encompass different research aspects in the theory of metric embeddings for Machine Learning. The first question asks what are some of the desirable properties that one wishes to preserve in an embedding? Of course, any answer to this question would be highly subjective to the underlying context/task. For instance, preserving the class conditional densities could already be sufficient if the task at hand is classification. In contrast, preserving the underlying geometry is essential if the corresponding task is nearest neighbour search. We would like to emphasize that this question has never been systematically studied in literature and can be treated as a separate research direction in itself. We address this question in the broadest possible sense and establish that approximate distances preservation(preferably via a continuous transformation in order to retain the properties of measure preservation) is the goal of an embedding in the domain of Machine Learning.

The next pressing question which naturally follows is, for any given embedding, how can a measure of *approximateness* (a measure of deviation from isometry) be formally realized. The first step in answering this question is to characterize what entails an embedding of high quality since any meaningful distortion measure needs to resonate with this characterization. We argued that the characterization defined in terms of the sharpness of concentration of the ratio of distances(embedding distance/original distance) encapsulates the quality of an embedding in the context of Machine Learning since it exhibits robustness to noisy observations and outliers while gracefully accommodating a probabilistic treatment of the quality of an embedding.

In the spirit of this characterization, we established that any effective measure of distortion should essentially exhibit the properties of robustness to outliers and noisy observations in addition to certain basic properties such as invariance to scaling and translations. Furthermore, it should allow for the incorporation of the underlying probability measure.

We discussed various measures of distortion that exist in literature, notably the measures of worstcase distortion, average distortion and more generally l_q

distortion. We showed that the measures of l_q distortion do not capture the essence of our characterization of a high quality embedding in the context of Machine Learning. To overcome the limitations of the existing measures of distortion, we proposed a novel measure of distortion, which we refer to as *σ-distortion* . We showed that *σ-distortion* precisely captures the essence of our characterization of the quality of an embedding by means of formal arguments as well as empirical evidence. Hence, we suggest that *σ-distortion* could be utilized as an effective evaluation metric for any embedding/dimensionality reduction task. However, it is noteworthy that *σ-distortion* is effective in scenarios where the goal of an embedding is to preserve most distances as well as possible. If distances between all pairs of points need to be well preserved then *σ-distortion* fails to capture the essence of this characterization of the quality of an embedding and worstcase distortion would be a better measure.

However, theoretical properties of *σ-distortion* have not been thoroughly explored in this thesis. Some of the important aspects that need to be addressed for *σ-distortion* are as follows. Can one characterize mappings between infinite metric spaces that incur finite *σ-distortion* ? Do mappings that incur finite *σ-distortion* also preserve the doubling property? In other words, is doubling property invariant under mappings of finite *σ-distortion* ? Do mappings of finite *σ-distortion* possess the property of continuity ? Also, the next key question which constitutes a fundamental research aspect encompassing the theory of metric embeddings is What is the best possible *σ-distortion* and dimension that can simultaneously be achieved for embeddings of arbitrary finite metric spaces into normed spaces? More importantly, do doubling metrics embed into finite dimensional normed spaces with finite *σ-distortion* ?

One approach in answering this question is to establish the relationship between *σ-distortion* and l_q distortion. Since this question has been reasonably answered for some measures of l_q distortion in the affirmative, the relationship between the distortion measures could potentially be used to derive guarantees on the best possible *σ-distortion* and dimension that can be achieved. Another approach that builds on the existing literature on l_q distortions would rely on the existing explicit constructions of embeddings that are used to derive guarantees on the best possible dimension and distortion.

Chapter 7

Future work

As discussed earlier, theoretical properties of mappings that incur finite σ-*distortion* would be fundamental to the analysis of σ-*distortion* . In addition, extensive analysis of what properties does one wish to preserve in Machine Learning for a given task such as Classification or clustering is necessary. In this thesis, we gave a preliminary assessment of the properties required of a good distortion measure in the context of Machine Learning. The next direction of work in this line is to formally present an exhaustive list of properties desirable for any effective measure of distortion.

One of the limitations of σ-*distortion* in its current form is that it does not treat expansions and contractions in a symmetric fashion. This could potentially be remedied by defining σ-*distortion* of an embedding f as the maximum of σ-*distortion* of f and σ-*distortion* of f^{-1}. However, it is not entirely clear to what extent, a symmetric treatment of contractions and expansions is essential for a measure of distortion. This analysis would be part of our future work. Another important aspect of our future work is to systematically analyze the effect of sample size, noise and outliers on σ-*distortion* more generally in a formal setting as well as through extensive experimentation.

Bibliography

Ittai Abraham, Yair Bartal, J Kleinberg, T-HH Chan, O Neiman, Kedar Dhamdhere, A Slivkins, and Anupam Gupta. Metric embeddings with relaxed guarantees. In *Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on*, pages 83–100. IEEE, 2005.

Ittai Abraham, Yair Bartal, and Ofer Neimany. Advances in metric embedding theory. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 271–286. ACM, 2006.

Ittai Abraham, Yair Bartal, and Ofer Neiman. Embedding metric spaces in their intrinsic dimension. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 363–372. Society for Industrial and Applied Mathematics, 2008.

Patrice Assouad. Plongements lipschitziens dans n. *Bull. Soc. Math. France*, 111 (4):429–448, 1983.

Yair Bartal, Lee-Ad Gottlieb, and Ofer Neiman. On the impossibility of dimension reduction for doubling subsets. *Ariel*, 1609:11, 2015.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

Jean Bourgain. On lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, 52(1):46–52, 1985.

T-H Hubert Chan, Anupam Gupta, and Kunal Talwar. Ultra-low-dimensional embeddings for doubling metrics. *Journal of the ACM (JACM)*, 57(4):21, 2010.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Aryeh Dvoretzky. Some results on convex bodies and banach spaces. *Matematika*, 8(1):73–102, 1964.

Armin Eftekhari and Michael B Wakin. What happens to a manifold under a bi-lipschitz map? *Discrete & Computational Geometry*, 57(3):641–673, 2017.

Anupam Gupta, Robert Krauthgamer, and James R Lee. Bounded geometries, fractals, and low-distortion embeddings. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 534–543. IEEE, 2003.

David J Hand and Keming Yu. Idiot's bayesnot so stupid after all? *International statistical review*, 69(3):385–398, 2001.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Gordon F. Hughes. On the Mean Accuracy of Statistical Pattern Recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, 1968.

Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 1998.

William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

David R Karger and Matthias Ruhl. Finding nearest neighbors in growth-restricted metrics. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 741–750. ACM, 2002.

Jon Kleinberg, Aleksandrs Slivkins, and Tom Wexler. Triangulation and embedding using small sets of beacons. In *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, pages 444–453. IEEE, 2004.

Robert Krauthgamer and James R Lee. Navigating nets: simple algorithms for proximity search. In *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 798–807. Society for Industrial and Applied Mathematics, 2004.

Tomi J Laakso. Ahlfors q-regular spaces with arbitrary q¿ 1 admitting weak poincaré inequality. *Geometric and functional Analysis*, 10(1):111–123, 2000.

Vincent Lafforgue and Assaf Naor. A doubling subset of l_p for p¿ 2 that is inherently infinite dimensional. *Geometriae Dedicata*, 172(1):387–398, 2014.

Urs Lang and Conrad Plaut. Bilipschitz embeddings of metric spaces into space forms. *Geometriae Dedicata*, 87(1):285–307, 2001.

Kasper Green Larsen and Jelani Nelson. Optimality of the johnson-lindenstrauss lemma. *arXiv preprint arXiv:1609.02094*, 2016.

James R Lee, Manor Mendel, and Assaf Naor. Metric structures in l1: Dimension, snowflakes, and average distortion. *European Journal of Combinatorics*, 26(8): 1180–1190, 2005.

Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

RB Marimont and MB Shapiro. Nearest neighbour searches and the curse of dimensionality. *IMA Journal of Applied Mathematics*, 24(1):59–70, 1979.

Jiří Matoušek. On the distortion required for embedding finite metric spaces into normed spaces. *Israel Journal of Mathematics*, 93(1):333–344, 1996.

Jiří Matoušek. On embedding expanders into p spaces. *Israel Journal of Mathematics*, 102(1):189–197, 1997.

Jiří Matoušek. *Lectures on discrete geometry*, volume 212. Springer Science & Business Media, 2002.

Jirı Matoušek. Lecture notes on metric embeddings. Technical report, 2013.

James R Munkres. *Topology*. Prentice Hall, 2000.

Assaf Naor and Ofer Neiman. Assouad's theorem with dimension independent of the snowflaking. *arXiv preprint arXiv:1012.2307*, 2010.

Ofer Neiman. Low dimensional embeddings of doubling metrics. *Theory of Computing Systems*, 58(1):133–152, 2016.

Scott D Pauls. The large scale geometry of nilpotent lie groups. In *Comm. Anal. Geom.* Citeseer, 2001.

Yuri Rabinovich and Ran Raz. Lower bounds on the distortion of embedding finite metric spaces in graphs. *Discrete & Computational Geometry*, 19(1):79–94, 1998.

Stephen Semmes. On the nonexistence of bilipschitz parameterizations and geometric problems about $a - infty$-weights. *Revista Matemática Iberoamericana*, 12(2):337–410, 1996.

Stephen Semmes. Metric spaces and mappings seen at many scales, appendix in metric structures for riemannian and non-riemannian spaces, m. gromov et al. *Progress in Mathematics*, 152, 1999.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Kilian Q Weinberger and Lawrence K Saul. Unsupervised learning of image manifolds by semidefinite programming. *International journal of computer vision*, 70 (1):77–90, 2006.

Erklärung der Urheberschaft

Ich versichere an Eides statt, dass ich die Master thesis im Studiengang Intelligent Adaptive Systems selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Ort, Datum Unterschrift

Erklärung zur Veröffentlichung

Ich erkläre mein Einverständnis mit der Einstellung dieser Master thesis in den Bestand der Bibliothek.

Ort, Datum Unterschrift