Measures of distortion for machine learning

Leena Chennuru Vankadara
University of Tübingen
Max Planck Institute for Intelligent Systems, Tübingen
`leena.chennuru@tuebingen.mpg.de`

Ulrike von Luxburg
University of Tübingen
Max Planck Institute for Intelligent Systems, Tübingen
`luxburg@informatik.uni-tuebingen.de`

Abstract

Given data from a general metric space, one of the standard machine learning pipelines is to first embed the data into a Euclidean space and subsequently apply machine learning algorithms to analyze the data. The quality of such an embedding is typically described in terms of a distortion measure. In this paper, we show that many of the existing distortion measures behave in an undesired way, when considered from a machine learning point of view. We investigate desirable properties of distortion measures and formally prove that most of the existing measures fail to satisfy these properties. These theoretical findings are supported by simulations, which for example demonstrate that existing distortion measures are not robust to noise or outliers and cannot serve as good indicators for classification accuracy. As an alternative, we suggest a new measure of distortion, called σ-distortion. We can show both in theory and in experiments that it satisfies all desirable properties and is a better candidate to evaluate distortion in the context of machine learning.

1 Introduction

Given data from a general metric space, one of the standard machine learning pipelines is to first embed the data into a Euclidean space (for example using an unsupervised algorithm such as Isomap, locally linear embedding, maximum variance unfolding, etc) and subsequently apply out of the box machine learning algorithms to analyze the data. Typically, the quality of such an embedding is described in terms of a distortion measure that summarizes how the distances between the embedded points deviate from the original distances. Many distortion measures have been used in the past, the most prominent ones being worst case distortion, l_q-distortion (Abraham, Bartal, and Neiman, 2011), average distortion(Abraham, Bartal, and Neiman, 2011), ϵ-distortion (Abraham, Bartal, Kleinberg, et al., 2005), k-local distortion (Abraham, Bartal, and Neiman, 2007) and scaling distortion (Abraham, Bartal, Kleinberg, et al., 2005). Such distortion measures are sometimes evaluated in hindsight to evaluate the quality of an embedding, and sometimes used directly as objective functions in embedding algorithms, for example the stress functions that are commonly used in different variants of Multidimensional scaling (Cox and Cox, 2000). There also exist embedding algorithms with completely different objectives. For instance, t-SNE (Maaten and Hinton, 2008) employs an objective function that aims to enhance the cluster structure present in the data. In this paper, however, we restrict our analysis to distortion measures that evaluate the quality of distance preserving embeddings.

From a theoretical computer science point of view, many aspects of distance preservation of emeddings are well understood. For example, Bourgain's theorem (Bourgain, 1985) and the Johnson–Lindenstrauss Lemma (Johnson and Lindenstrauss, 1984) state that any finite metric space

of n points can be embedded into a Euclidean space of dimension $\mathcal{O}(\log n)$ with worst case distortion $\mathcal{O}(\log n)$. Many related results exist (Gupta, Krauthgamer, and Lee, 2003; Abraham, Bartal, and Neiman, 2008; Abraham, Bartal, and Neiman, 2011; Abraham, Bartal, Kleinberg, et al., 2005; Abraham, Bartal, and Neiman, 2007; Abraham, Bartal, and Neiman, 2011; Semmes, 1996).

However, from a machine learning point of view, these results are not entirely satisfactory. The typical distortion guarantees from theoretical computer science focus on a finite metric space. However, in machine learning, we are ultimately interested in consistency statements: given a sample of n points from some underlying space, we would like to measure the distortion of an embedding algorithm as $n \to \infty$. In particular, the dimension of the embedding space should be constant and not grow with n, because we want to relate the geometry of the original underlying space to the geometry of the embedding space. Hence, many of the guarantees that are nice from a theoretical computer science point of view (for example, because they provide approximation guarantees for NP hard problems) miss the point when applied to machine learning (either in theory or in practice, see below).

Ideally, in machine learning we would like to use the distortion measure as an indication of the quality of an embedding. We would hope that when we compare several embeddings, choosing the one with smaller distortion would lead to better machine learning results (at least in tendency). However, when we empirically investigated the behavior of existing distortion measures, we were surprised to see that they behave quite erratically and often do not serve this purpose at all (see Section 4).

In pursuit of a more meaningful measure of distortion in the context of machine learning, we take a systematic approach in this paper. We identify a set of properties that are essential for any distortion measure. In light of these properties, we propose a new measure of distortion that is designed towards machine learning applications: the σ-distortion. We prove in theory and through simulations that our new measure of distortion satisfies many of the properties that are important for machine learning, while all the other measures of distortion have serious drawbacks and fail to satisfy all of the properties. These results can be summarized in the following table (where each column corresponds to one measure of distortion and each row to one desirable property, see Section 2 for notation and definitions):

Property/Distortion measure	σ(sigma)	wc	avg(l_q)	navg	k-local	ϵ(epsilon)
Translation invariance	✓	✓	✓	✓	✓	✓
Monotonicity	✓	✓	✗	✓	✓	✓
Scale invariance	✓	✓	✗	✓	✓	✓
Robustness to outliers	✓	✗	✓	✗	✗	✓
Robustness to noise	✓	✗	✗	✗	✗	✓
Incorporation of probability	✓	✗	✗	✗	✗	✗
Constant distortion embeddings	✓	✗	✓	?	✓	✓

2 Existing measures of distortion

Let (X, d_X) and (Y, d_Y) be arbitrary finite metric spaces. Let $\binom{X}{2} := \{\{u,v\} \mid u, v \in X, \ u \neq v\}$ and for any $n \in \mathbb{N}$, let $[n]$ denote the set $\{1, 2, ..., n\}$. An embedding of X into Y is an injective mapping $f : (X, d_X) \to (Y, d_Y)$. Let P be a probability distribution on X, and $\Pi := P \times P$ the product distribution on $X \times X$. Distortion measures aim to quantify the deviation of an embedding from isometry. Intuitively, the distortion of such an embedding is supposed to measure how far the new distances $d_Y(f(u), f(v))$ between the embedded points deviate from the original distances $d_X(u, v)$. Virtually all the existing distortion measures are summary statistics of the pairwise ratios

$$\rho_f(u, v) = d_Y(f(u), f(v))/d_X(u, v)$$

with $u, v \in X$. The intention is to capture the property that if the ratios $d_Y(f(u), f(v))/d_X(u, v)$ are close to 1 for many pairs of points u, v, then the distortion is small. The most popular measures

of distortion are the following ones:

Worst case distortion: $\Phi_{wc}(f) := \left(\max\limits_{(u,v)\in\binom{X}{2}} \rho_f(u,v) \right) \cdot \left(\max\limits_{(u,v)\in\binom{X}{2}} \frac{1}{\rho_f(u,v)} \right).$

Average case distortion: $\Phi_{avg}(f) := \frac{2}{n(n-1)} \sum\limits_{(u,v)\in\binom{X}{2}} \rho_f(u,v).$

Normalized avg distortion: $\Phi_{navg}(f) := \frac{2}{n(n-1)} \sum\limits_{u\neq v\in X} \frac{\rho_f(u,v)}{\alpha}$ with $\alpha = \min\limits_{u\neq v\in X} \rho_f(u,v).$

l_q**-distortion** (with $1 \leq q < \infty$): $\Phi_{l_q}(f) := \mathbb{E}_\Pi(\rho_f(u,v)^q)^{\frac{1}{q}}.$

ϵ**-distortion** ($\forall\, 0 < \epsilon < 1$): $\Phi_\epsilon(f) := \min\limits_{S\subset\binom{X}{2},|S|\geq(1-\epsilon)\binom{n}{2}} \Phi_{wc}(f_S),$ where f_S denotes the

restriction of f to S.

k-local distortion: $\Phi_{klocal}(f) := \left(\max\limits_{u\in X, v\in \text{kNN}(u), u\neq v} \rho_f(u,v) \right) \cdot \left(\max\limits_{u\in X, v\in \text{kNN}(u) u\neq v} \frac{1}{\rho_f(u,v)} \right),$

where $\text{kNN}(u)$ denotes the set of k nearest neighbours of u.

The different measures of distortion put their focus on different aspects: the worst case among all pairs of points (Φ_{wc}), the worst case excluding pathological outliers (Φ_ϵ), the average case ($\Phi_{avg}, \Phi_{navg}, \Phi_{l_q}$) or distortions that are just evaluated between neighboring points (Φ_{klocal}).

From a conceptual level, all these measures of distortion make sense, and it is not obvious why one should prefer one over the other. However, when we studied them in practice, we found different sources of undesired behavior for many of them. For example, many of them behave in a quite unstable or even erratic manner, due to high sensitivity to outliers or because they are not invariant with respect to rescaling. To study these issues more systematically, we will now identify a set of properties that any measure of distortion should satisfy in the context of machine learning applications. In Section 3.2 we then prove which of the existing measures satisfies which properties and find that each of them has particular deficiencies. In Section 3.3 we then introduce a new measure of distortion that does not suffer from these issues, and demonstrate its practical behavior in Section 4.

3 Properties of distortion measures

In this section we identify properties that a distortion measure is expected to satisfy in the context of machine learning. In addition to basic properties such as invariance to rescaling and translation, the most important properties should resonate with an appropriate characterization of the quality of an embedding. In the following, let (X, d_X) be an arbitrary metric space, let Y be an arbitrary vector space and let d_Y be a homogeneous and translation invariant metric on Y (See the supplement for the formal definitions). Let $f, g : (X, d_X) \rightarrow (Y, d_Y)$ be two embeddings and let Φ be any function that is supposed to measure the distortion of any injective mapping from X to Y.

3.1 Definitions

We start with a set of **basic properties** that should be satisfied by any function that is supposed to provide a measure of distortion, irrespective of the context in which it is applied.

Scale Invariance is an essential property for a measure of distortion since embeddings that are merely different in units of measurement (say, kilometers vs centimeters) should not be assigned different values of distortion. Formally, let $f : (X, d_X) \rightarrow (Y, d_Y)$ and $g : (X, d_X) \rightarrow (Y, d_Y)$ be two injective mappings. A distortion measure Φ is said to be scale invariant if for any $\alpha \in \mathbb{R}$,

$$\forall u \in X, f(u) = \alpha g(u) \implies \Phi(f) = \Phi(g). \tag{1}$$

Translation Invariance: A measure of distortion should clearly be invariant to translations: Let $f : (X, d_X) \rightarrow (Y, d_Y)$ and $g : (X, d_X) \rightarrow (Y, d_Y)$ be two injective mappings. A measure of

distortion Φ is said to be translation invariant if for any $y \in Y$,

$$\forall u \in X, f(u) = g(u) + y; \implies \Phi(f) = \Phi(g). \qquad (2)$$

Monotonicity captures the property that if distances are preserved more strictly, then the distortion of the corresponding embedding should be smaller. The formal definition is a bit tricky, because one has to be careful about scaling issues. We take care of it by standardizing the embeddings such that the average of the $\rho(u, v)$ is 1. Let $f : (X, d_X) \to (Y, d_Y)$ and $g : (X, d_X) \to (Y, d_Y)$ be embeddings. Define the scaling constants $\alpha(f) = (\frac{2}{n(n-1)}) \sum_{u \neq v \in X} \rho_f(u, v)$ and $\alpha(g) = (\frac{2}{n(n-1)}) \sum_{u \neq v \in X} \rho_g(u, v)$. Then a measure of distortion Φ is said to be monotonic if

$$\forall u, v \in X, \ \left(\left(\frac{\rho_f(u,v)}{\alpha(f)} \le \frac{\rho_g(u,v)}{\alpha(g)} \le 1 \right) \text{ or } \left(\frac{\rho_f(u,v)}{\alpha(f)} \ge \frac{\rho_g(u,v)}{\alpha(g)} \ge 1 \right) \right) \implies \Phi(f) \ge \Phi(g). \qquad (3)$$

After having introduced the basic properties that need to be satisfied such that a function Φ deserves the term "distortion", we now turn to some **advanced properties** that specifically identify the necessary characteristics of distortion measures in the context of machine learning applications.

Robustness to outliers: Outliers are inherent to data processed by machine learning algorithms, and hence a measure of distortion that is too volatile against outliers is not desirable. What we would like to achieve is rather that the influence of a single data point or a single distance value on the measure of distortion is very small. In the spirit of this interpretation, we create two test cases as necessary conditions to deem a measure of distortion robust to outliers.

Outliers in data: To verify that the effect of a single data point on the measure of distortion is small, we stipulate that the influence of this point should converge to 0 as the number n of points goes to infinity. To formulate this property, we compare an isometric embedding to an embedding that is "isometric except for one point". Formally, let $I : (X, d_X) \to (X, d_X)$ be an isometry. Fix arbitrary $x_0, x^* \in X$ and $\beta > 0$. For any $n \in \mathbb{N}$, let $X_n = \{x_1, x_2, ..., x_n\} \subset X \setminus B(x_0, \beta)$. Let $f_n : X_n \cup \{x_0\} \to X$ such that

$$f_n(x) = \begin{cases} x^*, & \text{if } x = x_0. \\ x, & \text{otherwise.} \end{cases} \qquad (4)$$

We say that 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 $X_n \cup \{x_0\}$. In the formal definition, one needs to make sure that the distortions do not grow arbitrarily fast, which can happen either if points in the original space are too close or if points in the image space are too far from each other. The ball of positive radius β prevents the first case, and the fact that we choose x^* as a fixed point prevents the second case.

Outliers in distances: To evaluate whether a measure of distortion is robust to outliers in distances, we consider mappings for which at most a constant (K) number of distances are distorted and compare the resulting distortion measure to the one of an isometry. Formally, let $I : (X, d_X) \to (X, d_X)$ be an isometry. Let $X_D = \{x_1, x_2, ..., \} \subset X$. Let $f : X_D \to X$ be an injective mapping such that there exists a constant $K \in \mathbb{N}$ for which the set $G = \{(u, v) \in X_D \times X_D : d_X(f(u), f(v)) \neq d_X(u, v)\}$ satisfies $|G| \le K$. For any $n \in \mathbb{N}$, let f_n and I_n denote the restriction of the mappings f and I, respectively, to $X_n = \{x_1, x_2, ..., x_n\} \subset X_D$. We say that a measure of distortion Φ is not robust to outliers if $\lim_{n \to \infty} \Phi(f_n) \neq \lim_{n \to \infty} \Phi(I_n)$.

Incorporation of the probability distribution: In machine learning, a standard assumption is that the data has been sampled according to some probability distribution from an underlying space. A measure of distortion should be able to take this probability distribution into account, in the sense that distortions of points in high density regions should be more costly than distortions of points in low density regions. We formalize this idea by stipulating that given two different embeddings which are "isometric except for one point", where the two embeddings distort two different points such that the ratios of distorted distances are the same for both the embeddings, then the embedding that distorts the point that occurs with higher probability needs to have a higher value of distortion.

Let (X, d_X) be an arbitrary metric space. Let $X_n = \{x_1, x_2, ..., x_n\}$ be a finite subset of X. Let P denote a probability distribution on X_n. Fix arbitrary $x^*, y^* \in X_n$ such that $P(x^*) > P(y^*)$.

Let $x^{'}, y^{'} \in X$ such that $\forall i \in [n], d_X(x_i, x^{'}) = \alpha d_X(x_i, x^*)$ and $d_X(x_i, y^{'}) = \alpha d_X(x_i, y^*)$. Let $f, g : X_n \to X$ be two embeddings such that:

$$f(x) = \begin{cases} x^{'}, & \text{if } x = x^*. \\ x, & \text{otherwise.} \end{cases} \quad , \quad g(x) = \begin{cases} y^{'}, & \text{if } x = y^*. \\ x, & \text{otherwise.} \end{cases}$$

Then a measure of distortion Φ is said to incorporate the probability distribution P if $\Phi(f) > \Phi(g)$.

Robustness to noise: Noisy observations, just as outliers, are common in machine learning. In machine learning applications we would expect that the measure of distortion is smaller if there is less noise on the data. For this property, we do not provide a formal definition. Rather, we conduct experiments to empirically verify whether this is the case in simple settings.

We believe that in order to be useful for machine learning, a measure of distortion should satisfy all the basic as well as the advanced properties. We would like to conclude this list with a last property that is perhaps not absolutely crucial, but **nice to have**: the ability to provide **constant-dimensional embeddings.**

In learning theory we often assume that we are given a set of data points that has been sampled according to some underlying probability distribution, and then we are interested in consistency statements: given a sample of n points, we study the behavior of algorithms as the sample size n goes to infinity. In particular, when we consider embeddings we would hope that as the sample size grows, the geometry of the embedded points "converges" to something that is close to the "true geometry" of the underlying space. In particular, if the underlying space is "simple", we would like to embed into a Euclidean space of constant dimension — the dimension is not supposed to grow with the sample size because we would then have to deal with an infinite-dimensional space in the limit case, which would allow for too complex geometries. For embeddings in a constant-dimensional space, we would then like to bound the distortion, ideally by a quantity that is bounded by a constant that is independent of n and just depends on the geometry of the true underlying space. In general, it is impossible to guarantee the existence of an embedding into Euclidean space with constant dimension and constant distortion (for all the standard measures of distortion, cf. (Semmes, 1999; Semmes, 1996; Abraham, Bartal, and Neiman, 2011)). However, guarantees can be given if we make assumptions on the underlying metric space. For example if (X, d_X) is doubling, it is possible to achieve an embedding into constant dimensional Euclidean space with $\mathcal{O}(1)$ average distortion (but $\Omega(\log n)$ worst case distortion) (Abraham, Bartal, and Neiman, 2011). Hence we stipulate that a measure of distortion that is nice for machine learning should satisfy that if the underlying geometry of the metric space is "nice" (according to an appropriate definition), then we can guarantee that there exists an embedding in **constant dimension with constant distortion.**

3.2 Theoretical results: existing distortion measures fail to satisfy all properties

In the following theorem we investigate which measure of distortion satisfies which of the properties mentioned above. All the proofs can be found in the appendix.

Theorem 1 (Properties of existing distortion measures). *For all choices of the parameters $1 \leq q < \infty$, $0 < \epsilon < 1$, $1 \leq k \leq n$, the following statements are true:*

(a) Φ_{wc}, Φ_{avg}, Φ_{navg}, Φ_{l_q}, Φ_ϵ *and* Φ_{klocal} *satisfy the property of* **translation invariance.**

(b) Φ_{wc}, Φ_{navg}, Φ_ϵ, Φ_{klocal} *satisfy the properties of* **scale invariance** *and* **monotonicity.** Φ_{avg} *and* Φ_{l_q} *fail to satisfy these properties.*

(c) Φ_ϵ, Φ_{avg}, Φ_{l_q} *satisfy the property of* **robustness to outliers.** *The measures* Φ_{wc}, Φ_{navg}, Φ_{klocal} *fail to do so.*

(d) *The distortion measures* Φ_{wc}, Φ_{avg}, Φ_{navg}, Φ_{l_q}, Φ_ϵ, Φ_{klocal} *fail to* **incorporate a probability distribution** *defined on the data space.*

At the current point in time, we do not yet have a formal guarantee towards **robustness to noise**. However, in our experiments we show that $\epsilon-$distortion is considerably robust to noise for certain values of ϵ and the rest of the distortion measures do not demonstrate robustness to noise.

Regarding the **constant-dimensional embeddings**, there exists a large literature. In the case of average distortion, ϵ-distortion, and k-local distortion for fixed values of k, ϵ, it has been shown

that any finite subset of a doubling metric space (see the supplement for a formal definition) can be embedded into a constant dimensional Euclidean space with $\mathcal{O}(1)$ distortion (Abraham, Bartal, and Neiman, 2011; Abraham, Bartal, and Neiman, 2009). Hence these measures of distortion also satisfy the "nice to have" property. Such a result for Normalized average distortion doesn't exist in the literature to the best of our knowledge. Worstcase distortion, however, fails to satisfy this property (Semmes, 1999; Semmes, 1996).

3.3 A new measure of distortion that satisfies all properties

We have seen that all the existing measures of distortion fail to satisfy at least one of the properties that we identified above. In the light of these results, we introduce a **new measure of distortion, the** $\sigma-$**distortion** (Φ_σ). The intuition for our definition is as follows. For a given data set X, consider a histogram of the ratios $\rho_f(u, v)$. An embedding of high quality should preserve *most* distances as well as possible, that is we would like to see that most of these ratios are close to 1. We characterize this by measuring the "concentration" of the distribution of the ratio of distances ($\rho_f(u, v)$) in terms of the variance. The fact that we consider the variance of this distribution makes our definition robust against outliers (one of the properties which most of the other distortion measures fail to satisfy). By a rescaling step we will achieve invariance with respect to scale. Furthermore we will see that also all the other properties are satisfied by our definition. Let X_n be a finite subset of X. Given a distribution P over X_n, let $\Pi = P \times P$ denote the distribution on the product space $X_n \times X_n$. For any embedding f, let $\widetilde{\rho_f}(u, v)$ denote the normalized ratio of distances given by

$$\widetilde{\rho_f}(u, v) := \frac{(n)(n-1)\rho_f(u, v)}{2\sum_{(u,v) \in \binom{X}{2}} \rho_f(u, v)}.$$

The σ-distortion is then defined as

$$\mathbb{E}_\Pi(\widetilde{\rho_f}(u, v) - 1)^2. \tag{5}$$

If P is a uniform probability distribution over X_n, then σ-distortion measures the variance of the distribution of the normalized ratio of distances, $\widetilde{\rho_f}(u, v)$.

Theorem 2 (Properties of σ-distortion). *The σ- distortion (a) is **invariant to scale and translations**, (b) satisfies the property of **monotonicity**, (c) is **robust to outliers in data** and **outliers in distances**, and (d) **incorporates a probability distribution** into its evaluation.*

In addition to satisfying all of the aforementioned properties, the proofs of Abraham, Bartal, and Neiman, 2011 can be extended to show that one can embed any finite subset of a doubling metric space into constant dimensional Euclidean space (or any l_p space) with $\mathcal{O}(1)$ distortion. So the σ-distortion also satisfies the nice-to-have property regarding constant dimensional embeddings with bounded distortion. The formal guarantees are given in the following two theorems:

Theorem 3 (General metric spaces: embeddable with constant σ-distortion in $\log n$ dimensions). *Given any finite sample $X_n = \{x_1, x_2, ..., x_n\}$ from an arbitrary metric space (X, d_X) and a probability distribution P on X_n, for any $1 \le p < \infty$ there exists an embedding $f : X_n \to l_p^D$, where $D = \mathcal{O}(\log n)$ with σ-distortion $= \mathcal{O}(1)$.*

Theorem 4 (Doubling metric spaces: embeddable with constant σ-distortion in constant dimensions). *Given any finite sample $X_n = \{x_1, x_2, ..., x_n\}$ from a doubling metric space (X, d_X) and a probability distribution P on X_n, for any $1 \le p < \infty$ there exists an embedding $f : X_n \to l_p^D$, where $D = \mathcal{O}(1)$ with σ-distortion $= \mathcal{O}(1)$.*

4 Experiments

We evaluate the behavior of various distortion measures by conducting experiments in two different settings: 1) Dimensionality reduction 2) A pipeline consisting of dimensionality reduction followed by classification. We start with simulated data for which we know all ground truth parameters. In order to generate datasets of dimension D, we sample each coordinate independently from a specified 1-dimensional distribution. Several different distributions such as Gaussian distribution, Gamma distribution, Beta distribution, Gaussian mixture distribution, Laakso Space (Bartal, Gottlieb, and Neiman, 2015) with many different parameter settings have been used to conduct the experiments. Embeddings are then generated by various dimensionality reduction algorithms. In particular, we

used Isomap (Tenenbaum, De Silva, and Langford, 2000), Maximum Variance Unfolding(Weinberger and Saul, 2006), Multidimensional Scaling, PCA (Hotelling, 1933), Probabilistic PCA (Tipping and Bishop, 1999), and Structure Preserving Embedding (Shaw and Jebara, 2009). All experiments have been repeated 10 times, the error bars in the plots depict the standard deviations over the different repetitions.

Embedding dimension vs distortion: For a dataset sampled from a Euclidean space with fixed dimension, it is natural to expect that any meaningful measure of distortion decreases with increasing embedding dimension. Intuitively, higher-dimensional spaces have more degrees of freedom to place points, and theoretical results confirm the general tendency (Chan, Gupta, and Talwar, 2010; Abraham, Bartal, and Neiman, 2008). In Figure 1, we show that this behavior can also be verified experimentally for most measures of distortion including $\sigma-$distortion, except for average distortion. The failure of average distortion to conform to this trend is clearly because it does not demonstrate invariance to scaling. To clarify, average distortion simply computes the sum of the ratios of distances $\rho_f(u,v)$. An embedding can deviate from isometry by either contracting the distances between pairs of points or expanding them relative to the original distances. An immediate consequence of scale invariance of a distortion is that it treats expansions and contractions symmetrically in the following sense: if an embedding expands all distances by α, a scale invariant distortion would assign the embedding the same value of distortion as an embedding that contracts them by α. Average distortion does not possess this property and places undue emphasis on expansions and underscores contractions. The balance of the weight of contractions and expansions influence the trend followed by average distortion, which is thereby unpredictable.

Distortion vs original dimension: For datasets generated from Euclidean spaces of increasing dimension, it is also natural to expect that for a fixed embedding dimension, the quality of an embedding decreases with increasing original dimension. The intuition here is that the data gets more complex, but the embedding space does not have the capacity to reflect this. To our surprise, this behavior cannot be observed for most of the traditional measures of distortion, as can be seen from Figures 2 and 3 (this observation was one of the starting points of this whole line of research). When looking closer into the data, we come to the conclusion that the reason for this failure is that these measures (except for average distortion, which shows erratic behaviour due to its dependence on the scale of the embedding) suffer from outliers, which disproportionately affect the distortion measures. There are only two measures of distortion that show the desired behavior: σ-distortion and ϵ-distortion. We attribute this to the fact that these two measures are robust to outliers (as also been shown in our theoretical results). We can also see from the error bars in Figures 2, 3 that the variability of the rest of the distortion measures is significantly larger compared to that of ϵ-distortion as well as σ-distortion. Again we attribute this behavior to the brittleness of the other distortions in the presence of outliers.
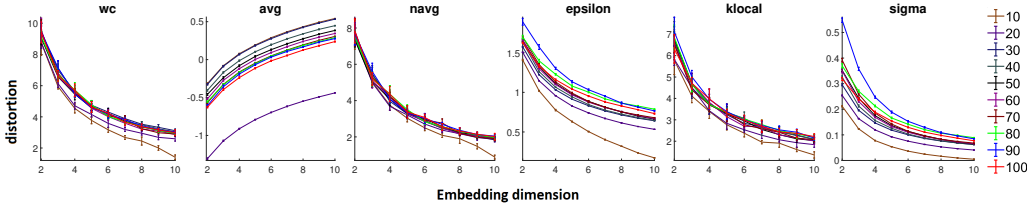

Figure 1: Embedding dimension vs various measures of distortion. From left to right: Φ_{wc}, Φ_{avg}, Φ_{navg}, Φ_ϵ for $\epsilon = 0.1$, Φ_{klocal} for k = 5, Φ_σ. The color of each curve indicates the dimension of the original space, the x-axis the dimension of the embedding space. We can clearly see that for all but the average distortion, distortion decreases with embedding dimension. Data was generated according to a standard normal distribution of dimension as indicated by the color, embeddings have been generated using Isomap. Results for other distributions and algorithms look similar.

Effect of noise: In order to test the effect of noise on various measures of distortion, we generated mixture of Gaussian data in $\mathbb{R}^2$ similar to that of the previous experiment and added normally distributed noise in $\mathbb{R}^{20}$ of increasing variance to the data to generate different datasets. Embeddings were then performed using various algorithms into $\mathbb{R}^2$. In a first evaluation, we investigated whether the distortion increases with increasing noise. In Figure 4 (left) we can see that while σ-distortion clearly shows the desired trend, all other measures of distortion fail to show the correct behavior.
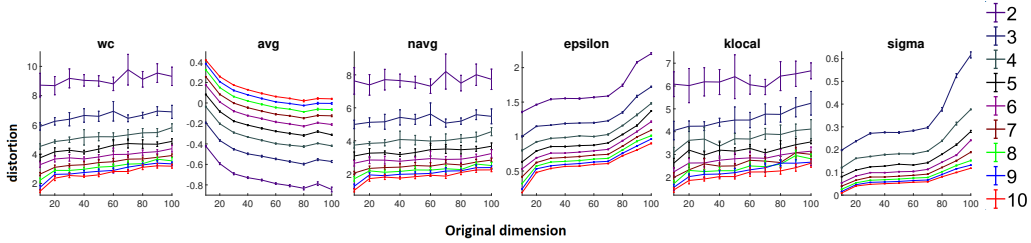
7

Figure 2: Original dimension vs measures of distortion. From left to right: $\Phi_{wc}, \Phi_{avg}, \Phi_{navg}, \Phi_{\epsilon}$ for $\epsilon = 0.1$, Φ_{klocal} for k = 5, Φ_{σ}. The x-axis shows the dimension of the original space, the color of the curve corresponds to the dimension of the embedding space. Each curve corresponds to Isomap embeddings of data generated according to gamma distribution (a = 1.5, b = 4) from Euclidean space of dimensions $(10 : 10 : 100)$. Results for other distributions and algorithms look similar.
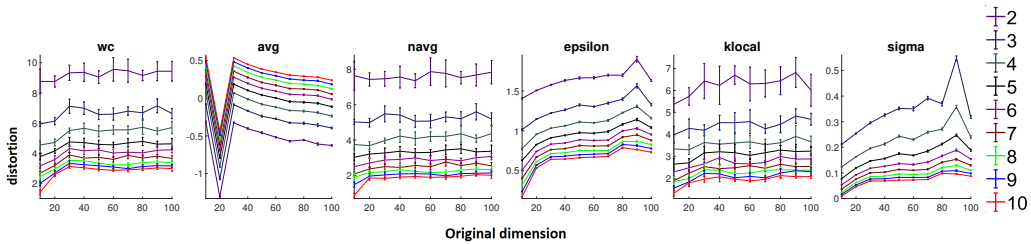

Figure 3: Same setting as in Figure 2, but data generated according to beta distribution (a = 0.75, b = 0.75).

In a second evaluation, we then performed classification on the embedded data. The corresponding SVM and kNN loss are plotted against the variance of the additive noise. Figure 4 clearly shows that the SVM and kNN classification loss increase with increasing variance of noise. This reiterates that the quality of the embedding indeed worsens with increasing additive noise. We performed this experiment using different embedding algorithms (Isomap, PCA, MVU) and the plots in all the experiments paint the same picture.
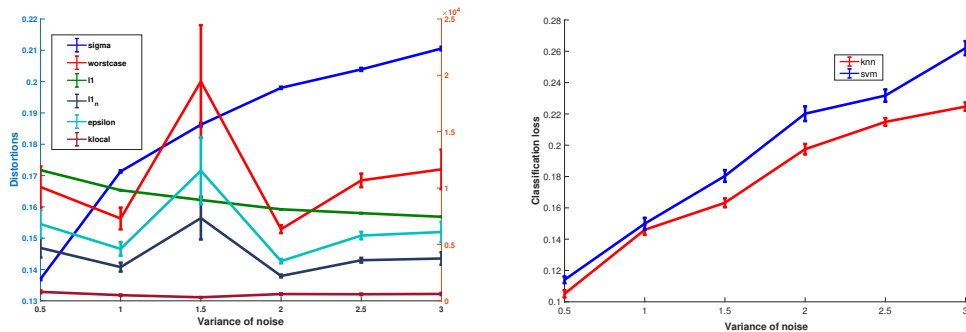

Figure 4: Left: Variance of noise vs distortion measures. The x-axis shows the variance of noise. As the measures of distortion are not all in the same range, we added two y-axes: the left (blue) one for σ-distortion, and the right (red) one for the values of the rest of the distortion measures. Right: Variance of noise vs. classification error. The x-axis shows the variance of noise, the y-axis the classification error. All embeddings here are created using Isomap. The behavior corresponding to the other embedding algorithms is similar.

Distortion vs classification accuracy: In this set of experiments, we want to investigate whether a measure of distortion is a good indicator for classification accuracy. To this end, we sampled data from various mixture of Gaussian distributions in $\mathbb{R}^2$ with different sets of parameters. Gaussian noise in $\mathbb{R}^{20}$ was then added to the data to generate various datasets. The datasets were then embedded into $\mathbb{R}^2$ using various embedding algorithms: PCA (Hotelling, 1933), GPLVM (Lawrence, 2004),

Isomap (Tenenbaum, De Silva, and Langford, 2000), MVU (Weinberger and Saul, 2006), SPE (Shaw and Jebara, 2009). Classification is performed on the resulting embeddings using kernel SVM (with RBF kernel) and kNN classification algorithms. In Figure 5, we plot the distortions incurred by the embeddings against the classification loss incurred by the classifier (where we sorted the outcome of all the simulations according to their resulting classification accuracy). Note that in this experiment, we compare the quality of embeddings across different embedding algorithms. The ideal behavior would be that distortion increases with increasing classification loss (in such a case, a measure of distortion could be used to select the best embedding, for example). This setting encapsulates the idea of using distortion measure as a means of evaluating the quality of an embedding in machine learning tasks. The plots clearly show that σ-distortion and ϵ-distortion consistently show the expected increasing trend with the classification loss, whereas the other measures of distortion fail to do so.
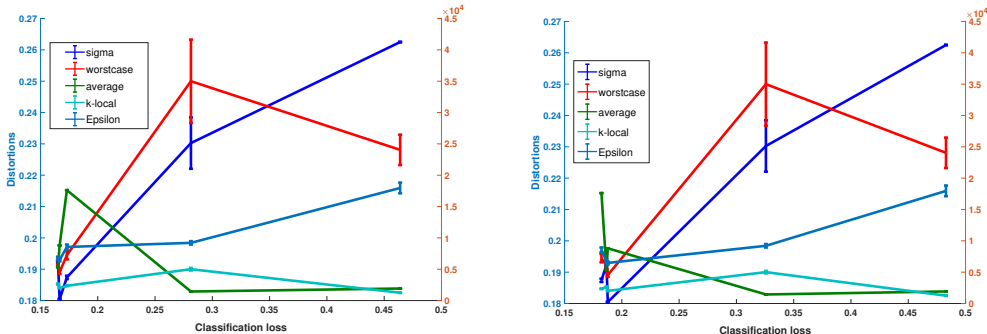

Figure 5: Classification error vs. distortion, for kNN (left) and SVM (right). x-axis: classification error and y-axis: distortion. The measures of distortion are not all in the same range, we added two y-axis: the left (blue) one for σ-distortion, and the right (red) one for the values of the rest of the distortion measures. Each curve corresponds to a distortion measure as indicated in the legend. The distortions are scaled appropriately for visualization.

5   Discussion

We investigate the properties of various measures of distortion for machine learning. Both in theory and experiments we can demonstrate that many of the existing measures of distortion behave in an undesired way: in simulations they show the wrong tradeoff with respect to the dimension of the original space, and they are not robust to noise or outliers, and cannot serve as a good indicator for classification accuracy. As an alternative, we define a new measure of distortion, called σ-distortion. In a nutshell, it measures the variance of the pairwise distortion ratios (rather than a norm of the vector of these ratios). We can show in theory and in experiments that it satisfies all our desirable properties. There is only one existing measure of distortion that comes close to our new σ-distortion, namely the ϵ-distortion. It explicitly excludes an ϵ fraction of outlier points from the distortion computation. For most properties it behaves nice as well, but it fails to take the probability measure into account. This is important because ϵ-distortion provides no guarantees on ϵ fraction of the pairwise distances, which could be critical for a given machine learning task. One of its drawbacks is that it has an important parameter to tune, the value of ϵ (fraction of outliers), whereas σ-distortion does not have a parameter. Our work clearly shows the need to study measures of distortion from a more systematic point of view, both in theory and practice.

9

References

[1] I. Abraham, Y. Bartal, J. Kleinberg, T-H. Chan, O. Neiman, K. Dhamdhere, A. Slivkins, and A. Gupta, "Metric embeddings with relaxed guarantees", *IEEE Symposium on Foundations of Computer Science*, pp. 83–100, 2005.

[2] I. Abraham, Y. Bartal, and O. Neiman, "Advances in metric embedding theory", *Advances in Mathematics*, vol. 228, no. 6, pp. 3026–3126, 2011.

[3] I. Abraham, Y. Bartal, and O. Neiman, "Embedding metric spaces in their intrinsic dimension", *ACM-SIAM Symposium on Discrete Algorithms*, pp. 363–372, 2008.

[4] I. Abraham, Y. Bartal, and O. Neiman, "Local embeddings of metric spaces", *ACM Symposium on Theory of Computing*, pp. 631–640, 2007.

[5] I. Abraham, Y. Bartal, and O. Neiman, "On low dimensional local embeddings", *ACM-SIAM Symposium on Discrete Algorithms*, pp. 875–884, 2009.

[6] Y. Bartal, L. Gottlieb, and O. Neiman, "On the impossibility of dimension reduction for doubling subsets of lp", *SIAM Journal on Discrete Mathematics*, vol. 29, no. 3, pp. 1207–1222, 2015.

[7] J. Bourgain, "On Lipschitz embedding of finite metric spaces in Hilbert space", *Israel Journal of Mathematics*, vol. 52, no. 1, pp. 46–52, 1985.

[8] T. Chan, A. Gupta, and K. Talwar, "Ultra-low-dimensional embeddings for doubling metrics", *Journal of the ACM*, vol. 57, no. 4 2010.

[9] T Cox and M. Cox, *Multidimensional scaling*, Chapman and hall/CRC, 2000.

[10] A. Gupta, R. Krauthgamer, and J. Lee, "Bounded geometries, fractals, and low-distortion embeddings", *Foundations of Computer Science*, pp. 534–543, 2003.

[11] H. Hotelling, "Analysis of a complex of statistical variables into principal components.", *Journal of Educational Psychology*, vol. 24, no. 6, p. 417, 1933.

[12] W. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space", *Contemporary Mathematics*, vol. 26, no. 189-206 1984.

[13] N. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data", *Neural Information Processing Systems*, pp. 329–336, 2004.

[14] L. Maaten and G. Hinton, "Visualizing data using t-SNE", *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[15] S. Semmes, "Bilipschitz embeddings of metric spaces into Euclidean spaces", *Publicacions Matematicks*, pp. 571–653, 1999.

[16] S. Semmes, "On the nonexistence of bilipschitz parameterizations and geometric problems about A-infinity weights", *Revista Matematica Iberoamericana*, vol. 12, no. 2, pp. 337–410, 1996.

[17] B. Shaw and T. Jebara, "Structure preserving embedding", *International Conference on Machine Learning*, pp. 937–944, 2009.

[18] J. Tenenbaum, V. De Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction", *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[19] M. Tipping and C. Bishop, "Probabilistic principal component analysis", *Journal of the Royal Statistical Society: Series B*, vol. 61, no. 3, pp. 611–622, 1999.

[20] K. Weinberger and L. Saul, "An introduction to nonlinear dimensionality reduction by maximum variance unfolding", *Association for the Advancement of Artificial Intelligence (AAAI)*, vol. 6, pp. 1683–1686, 2006.

Supplement to "Measures of distortion for machine learning"

A   Proof of Theorems 1 and 2

Homogeneous metric: A metric, d_X, on a vector space, X, is said to satisfy the property of homogeneity if for all $u, v \in X$ and for any scalar $\alpha \in \mathbb{R}$, d_X satisfies the following condition: $d_X(\alpha \cdot u, \alpha \cdot v) = |\alpha| \cdot d_X(u, v)$. We refer to such a metric, d_X, as a homogeneous metric and the corresponding metric space, (X, d_X), as a homogeneous metric space.

Translation invariant metric: A metric d_X on a vector space X is said to be translation invariant if for any u, v and $w \in X$, d_X satisfies the following condition: $d_X(u + w, v + w) = d_X(u, v)$.

Recall that (X, d_X) is an arbitrary metric space and (Y, d_Y) is assumed to be a homogeneous and a translation invariant metric. Let $f, g : (X, d_X) \to (Y, d_Y)$ be two injective mappings. For any $S \subseteq \binom{X}{2}$, let f_S denote the restriction of the mapping f to the set S. Formally, $f_S := \{(f(u), f(v)) | (u, v) \in S\}$.

A.1   Scale and translation invariance

A distortion measure Φ is said to be **scale invariant** if for any $\alpha \in \mathbb{R}$

$$\forall u \in X, f(u) = \alpha g(u); \implies \Phi(f) = \Phi(g).$$

Φ is said to be **translation invariant** if for any $y \in Y$,

$$\forall u \in X, f(u) = g(u) + y; \implies \Phi(f) = \Phi(g).$$

(a) Φ_{wc}, Φ_{navg}, Φ_ϵ, Φ_{klocal} and Φ_σ are invariant to scaling:

Proof. The proofs naturally follow from the definitions of these distortion measures and rely on the assumption of homogeneity of the target metric.

worstcase distortion: By the virtue of homogeneity of d_Y,

$$\Phi_{wc}(g) = \max_{u \neq v \in X} \left\{ \frac{|\alpha| \cdot d_Y(f(u), f(v))}{d_X(u, v)} \right\} \cdot \max_{u \neq v \in X} \left\{ \frac{d_X(u, v)}{|\alpha| \cdot d_Y(f(u), f(v))} \right\}$$
$$= \Phi_{wc}(f).$$

normalized average distortion:

$$\Phi_{navg}(g) = \frac{2}{n(n-1)} \sum_{u \neq v \in X} \frac{\rho_g(u, v)}{\min\limits_{u \neq v \in X} \rho_g(u, v)}, \qquad \text{where } \rho_g(u, v) = \frac{d_Y(g(u), g(v))}{d_X(u, v)}.$$

$$\rho_g(u, v) = \frac{|\alpha| \cdot d_Y(f(u), f(v))}{d_X(u, v)} = |\alpha| \cdot \rho_f(u, v). \quad \text{(from homogeneity of } d_Y\text{)}.$$

$$\Phi_{navg}(g) = \frac{2}{n(n-1)} \sum_{u \neq v \in X} \frac{|\alpha| \cdot \rho_f(u, v)}{\min\limits_{u \neq v \in X} |\alpha| \cdot \rho_f(u, v)}$$
$$= \Phi_{navg}(f).$$

$\boldsymbol{\epsilon}$**-distortion:** The scale invariance of ϵ-distortion follows from the fact that if f, g are two embeddings that are scaled versions of each other, then for any subset S of $\binom{X}{2}$, the worstcase distortions of f_S, g_S are equal: For any $\epsilon \in (0, 1)$,

$$\Phi_\epsilon(g) = \min_{S \subset \binom{X}{2}, |S| \geq (1-\epsilon)\frac{n(n-1)}{2}} \Phi_{wc}(g_S)$$

and thus, for any $S \subset \binom{X}{2}$,

$$\Phi_{wc}(g_S) = \max_{\{\{u,v\} \in S\}} \left\{ \frac{d_Y(g(u), g(v))}{d_X(u, v)} \right\} \cdot \max_{\{\{u,v\} \in S\}} \left\{ \frac{d_X(u, v)}{d_Y(g(u), g(v))} \right\}$$

$$= \max_{\{\{u,v\} \in S\}} \left\{ \frac{|\alpha| \cdot d_Y(f(u), f(v))}{d_X(u, v)} \right\} \cdot \max_{\{\{u,v\} \in S\}} \left\{ \frac{d_X(u, v)}{|\alpha| \cdot d_Y(f(u), f(v))} \right\}$$

$$= \Phi_{wc}(f_S).$$

k-local distortion: For any $k \in I\!N$ and for any $u \in X$, let $kNN(u)$ denote the k-nearest neighbours of u in X according to d_X. The set S defined as $\{\{u,v\} \mid u, v \in X, v \in kNN(u)\}$ is a subset of $\binom{X}{2}$. As shown in the case of ϵ-distortion, for any two embeddings f, g which are scaled versions of each other, we have $\Phi_{wc}(g_S) = \Phi_{wc}(f_S)$ and hence $\Phi_{klocal}(g) = \Phi_{klocal}(f)$.

σ-distortion: For any $\{u, v\} \in \binom{X}{2}$, let $\widetilde{\rho_f}(u, v)$ denote the normalized ratio of distances defined as $\rho_f(u, v)/\alpha_f$, where $\alpha_f = \sum_{\{u,v\} \in \binom{X}{2}} \rho_f(u, v)/\binom{n}{2}$. Observe that, by the virtue of homogeneity of d_Y, $\widetilde{\rho_f}(u, v) = \widetilde{\rho_g}(u, v)$. Therefore,

$$\Phi_\sigma(g) = \mathbb{E}_\Pi(\widetilde{\rho_g} - 1)^2 = \mathbb{E}_\Pi(\widetilde{\rho_f} - 1)^2 = \Phi_\sigma(f).$$

$\square$

(b) Φ_{avg} and Φ_{l_q} are not invariant to scaling.

Proof. The proof follows from the linearity of the expectation. We prove the statement for the l_q distortion, the case of the average distortion then is the special case of $q = 1$.

$$\Phi_{l_q}(g) = \mathbb{E}_\Pi(\rho_g(u, v)^q)^{\frac{1}{q}} = \mathbb{E}_\Pi(|\alpha| \cdot \rho_f(u, v)^q)^{\frac{1}{q}} = |\alpha| \cdot \Phi_{l_q}(f).$$

$\square$

(c) All the distortion measures, $\Phi_{wc}, \Phi_{l_q}, \Phi_{navg}, \Phi_\epsilon, \Phi_{klocal}$ and Φ_σ are invariant to translations.

Proof. It is straightforward to see that all the distortion measures derive this property from the translation invariance of d_Y. $\square$

A.2 Monotonicity

For any $f : (X, d_X) \to (Y, d_Y)$, let $\alpha_f = \binom{n}{2}/\sum_{u \neq v \in X} \rho_f(u, v)$. Observe that $\alpha_f > 0$. Let $f, g : (X, d_X) \to (Y, d_Y)$ be two embeddings such that for all $u, v \in X$,

$$\alpha_f \cdot \rho_f(u, v) \leq \alpha_g \cdot \rho_g(u, v) \leq 1 \quad \text{or} \quad \alpha_f \cdot \rho_f(u, v) \geq \alpha_g \cdot \rho_g(u, v) \geq 1.$$

(a) $\Phi_{wc}, \Phi_{navg}, \Phi_\epsilon, \Phi_{klocal}$ and Φ_σ satisfy the property of monotonicity.

Proof. The proofs follow directly from the definitions and utilize the scale invariance property of the corresponding distortion measure:

worstcase distortion:

$$\Phi_{wc}(f) = \Phi_{wc}(\alpha_f \cdot f) \qquad\qquad \text{(Scale invariance of } \Phi_{wc})$$

$$= \max_{u \neq v \in X} \{\alpha_f \cdot \rho_f(u, v)\} \cdot \frac{1}{\min_{u \neq v \in X} \{\alpha_f \cdot \rho_f(u, v)\}}$$

$$\geq \max_{u \neq v \in X} \{\alpha_g \cdot \rho_g(u, v)\} \cdot \frac{1}{\min_{u \neq v \in X} \{\alpha_g \cdot \rho_g(u, v)\}}$$

$$= \Phi_{wc}(g).$$

The inequality follows since $\max\limits_{u\neq v\in X}\{\alpha_f\cdot\rho_f(u,v)\}\geq 1$ and $\min\limits_{u\neq v\in X}\{\alpha_f\cdot\rho_f(u,v)\}\leq 1$.

normalized average distortion:

$$\Phi_{navg}(f)=\Phi_{navg}(\alpha_f\cdot f) \qquad\qquad\text{(Scale invariance of }\Phi_{navg})$$

$$=\frac{\sum\limits_{u\neq v\in X}\{\alpha_f\cdot\rho_f(u,v)\}}{\min\limits_{u\neq v\in X}\{\alpha_f\cdot\rho_f(u,v)\}}$$

$$\geq\frac{2}{(n)\cdot(n-1)\min\limits_{u\neq v\in X}\{\alpha_g\cdot\rho_g(u,v)\}}$$

$$=\Phi_{navg}(g).$$

The inequality follows since $\min\limits_{u\neq v\in X}\{\alpha_f\cdot\rho_f(u,v)\}\leq 1$.

ϵ**-distortion:** Let $\Psi=\arg\min\limits_{S\subset\binom{X}{2},|S|>(1-\epsilon)\frac{n(n-1)}{2}}\Phi_{wc}(f_S)$. Then,

$$\Phi_\epsilon(f)=\Phi_{wc}(f_\Psi)\geq\Phi_{wc}(g_\Psi)\geq\min\limits_{S\subset\binom{X}{2},|S|>(1-\epsilon)\frac{n(n-1)}{2}}\Phi_{wc}(g_S)=\Phi_\epsilon(g).$$

The first inequality follows due to the monotonicity of Φ_{wc}.

k-local distortion: For any $k\in I\!\!N$, the set $S=\{(u,v)|u,v\in X,v\in kNN(u)\}$ is a subset of $\binom{X}{2}$. Then,

$$\Phi_{klocal}(f)=\Phi_{wc}(f_S)\geq\Phi_{wc}(g_S)=\Phi_{klocal}(g).$$

The first equality follows by definition and the first inequality follows due to the monotonicity of Φ_{wc}.

σ**-distortion** For any $f:(X,d_X)\to(Y,d_Y)$ and for all $u\in X$, let $f'(u)=\frac{(n)(n-1)f(u)}{2\sum\rho_f(u,v)}$. Then

$$\Phi_\sigma(f)=\mathbb{E}_\Pi(\rho_{f'}(u,v)-1)^2.$$

Observe that, by the definition of monotonicity, we have that for all $u,v\in X$, $(\rho_{f'}(u,v)-1)^2\geq(\rho_{g'}(u,v)-1)^2$. Hence it follows that,

$$\Phi_\sigma(f)=\mathbb{E}_\Pi(\rho_{f'}(u,v)-1)^2)\geq\mathbb{E}_\Pi(\rho_{g'}(u,v)-1)^2)=\Phi_\sigma(g).$$

$\square$

(b) Φ_{avg},Φ_{l_q} fail to satisfy the property of monotonicity.

Proof. Proof by contradiction.

average distortion, l_q distortion:

Let $f,g:(X,d_X)\to(Y,d_y)$ be two embeddings such that there exists a constant $\beta>1$ and for all $u\in X$, $\beta f(u)=g(u)$. Let $\alpha_f=(n)(n-1)/2\sum\limits_{(u\neq v\in X)}\rho_f(u,v)$. Then for all $\{u,v\}\in\binom{X}{2}$, f,g satisfy the following condition:

$$\alpha_f\rho_f(u,v)\leq\alpha_g\rho_g(u,v)\leq 1\quad\text{or}\quad\alpha_f\rho_f(u,v)\geq\alpha_g\rho_g(u,v)\geq 1.$$

However $\Phi_{avg}(f)=\frac{\Phi_{avg}(f)}{\beta}<\Phi_{avg}(g)$. $\square$

A.3 Robustness to outliers

(a) Φ_{avg},Φ_ϵ and Φ_σ are robust to outliers in distances.

Proof. Let X,d_X be an arbitrary metric space. Let $X_D=\{x_1,x_2,....,\}$ be a subset of X. Let $f:X_D\to X$ be an injective mapping such that there exists a constant $K\in I\!\!N$ such that the set $G=\left\{\{u,v\}\in\binom{X_D}{2}\mid d_X(f(u),f(v))\neq d_X(u,v)\right\}$ satisfies $|G|\leq K$. Let f_n denote the

restriction of f to $X_n = \{x_1, x_2, ..., x_n\}$. Let $I : (X, d_X) \to (X, d_X)$ be an isometry and let I_n denote the restriction of I to X_n.

average distortion: For any $n \in \mathbb{N}$, let $G_n = \left\{ \{u, v\} \in \binom{X_n}{2} : d_X(f_n(u), f_n(v)) \neq d_X(u, v) \right\}$. By definition, it follows that for all $n \in \mathbb{N}$, there exists a $K \in \mathbb{N}$ such that, $|G_n| \leq |G| \leq K$. This implies that there exists a $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$, $G_n = G$ and $|G| = K$.

Therefore for any fixed $n \geq n_0$, ρ_{f_n} can be expressed as

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

where for all $i = \{1, 2, ..., K\}$, α_i's denote the ratio of distances $\rho_f(u, v)$ corresponding to each $\{u, v\} \in G$. Therefore for any fixed $n \geq n_0$, average distortion of f_n can be expressed as

$$\Phi_{avg}(f_n) = \frac{\left(\binom{n}{2} - K\right) + \sum\limits_{i=1}^{K} \alpha_i}{\binom{n}{2}}.$$

Hence $\lim\limits_{n \to \infty} \Phi_{avg}(f_n) = 1 = \lim\limits_{n \to \infty} \Phi_{avg}(I_n)$.

ϵ**-distortion:** For any $\epsilon \in (0, 1)$,

$$\Phi_\epsilon(f_n) = \min_{S \subset \binom{X_n}{2}, |S| \geq (1-\epsilon)\frac{n(n-1)}{2}} \Phi_{wc}(f_{n_S}).$$

Observe that for all $n > 1 + \sqrt{1 + \frac{8K}{\epsilon}}$, there exists a set $S \subset \binom{X_n}{2}$ such that $|S| \geq (1 - \epsilon)\frac{n(n-1)}{2}$ and $\Phi_{wc}(f_{n_S}) = 1$. Hence $\lim\limits_{n \to \infty} \Phi_\epsilon(f_n) = 1 = \lim\limits_{n \to \infty} \Phi_\epsilon(I_n)$.

σ**-distortion:** The proof follows along the same lines as that of average distortion.

For any $f_n : (X_n, d_X) \to (X, d_X)$, $\Phi_\sigma(f_n)$ can be expressed as

$$\Phi_\sigma(f_n) = \frac{\sum\limits_{(u,v) \in \binom{X_n}{2}} [\binom{n}{2}\rho_{f_n}(u, v) - \sum\limits_{(u,v) \in \binom{X_n}{2}} \rho_{f_n}(u, v)]^2}{\binom{n}{2}[\sum\limits_{(u,v) \in \binom{X_n}{2}} \rho_{f_n}(u, v)]^2}. \tag{1}$$

For any $n \in \mathbb{N}$, let $G_n = \{(u, v) \in X_n : d_X(f_n(u), f_n(v)) \neq d_X(u, v)\}$. By definition, it follows that for all $n \in \mathbb{N}$, there exists a $K \in \mathbb{N}$ such that, $|G_n| \leq |G| \leq K$. This implies that there exists a $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$, $G_n = G$ and $|G| = K$. Therefore for any fixed $n \geq n_0$, ρ_{f_n} can be expressed as

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

where for all $i = \{1, 2, ..., K\}$, α_i's denote the ratio of distances $\rho_f(u, v)$ corresponding to each $\{u, v\} \in G$. Therefore for any fixed $n \geq n_0$, σ-distortion of f_n can be expressed as

$$\Phi_\sigma = \frac{\binom{n}{2}(K - \sum\limits_{i=1}^{K} \alpha_i)^2 (\binom{n}{2} - K) + \sum\limits_{i=1}^{K}(\binom{n}{2}(\alpha_i - 1) + K - \sum\limits_{i=1}^{K} \alpha_i)^2}{\binom{n}{2}(\binom{n}{2} - K + \sum\limits_{i=1}^{K} \alpha_i)^2}.$$

Hence, $\lim\limits_{n \to \infty} \Phi_\sigma(f_n) = 0 = \lim\limits_{n \to \infty} \Phi_\sigma(I_n)$. $\qquad \square$

(b) $\Phi_{avg}, \Phi_\epsilon$ and Φ_σ are robust to outliers in data

Proof. The proofs follow by definition and utilize the subadditivity of the target metric d_Y.

average distortion: Let $f_n : X_n \to X$ be an embedding as specified in the premise of the definition. Then the average distortion of f_n is evaluated as

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

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

$$0 < |\frac{d_Y(I(x_0), f(x_0))}{d_Y(I(x_0), f(x_i))} - 1| < \alpha_i < \frac{d_Y(I(x_0), f(x_0))}{\beta} + 1.$$

This implies,

$$\frac{\binom{n}{2} - (n-1)}{\binom{n}{2}} < \Phi_{avg}(f_n) < \frac{\binom{n}{2} - (n-1) + \frac{(n-1)d_Y(I(x_0), f(x_0))}{\beta}}{\binom{n}{2}}.$$

Note that $\beta > 0$ is a constant and for a fixed $x_0 \in X$, $d_Y(I(x_0), f(x_0))$ is also a constant. Therefore, $\lim\limits_{n\to\infty} \Phi_{avg}(f_n) = 1 = \lim\limits_{n\to\infty} \Phi_{avg}(I_n)$.

ϵ-distortion:

Let $f_n : X_n \to X$ be an embedding as specified in the premise of the definition. Its easy to verify that for any $\epsilon \in (0, 1)$ and for all $n > \frac{8}{\epsilon}$, there exists a set $S \subset \binom{X}{2}$ such that, $|S| > (1 - \epsilon)\frac{n(n-1)}{2}$ and the worstcase distortion of f_n restricted to S is $1 = \Phi_\epsilon(I_n)$.

σ-distortion:

Let $f_n : X_n \to X$ be an embedding as specified in the premise of the definition. Let $\mu(\rho)$ denote $\sum\limits_{u \neq v \in X} \rho_{f_n}(u, v)/\binom{n}{2}$ and for all $i \in [n]$, let $\alpha_i = d_Y(f(x_i), f(x_0))/d_X(x_i, x_0)$. Then $\Phi_\sigma(f_n)$ is evaluated as

$$\frac{\sum\limits_{(u,v) \in \binom{X_n}{2}} [\rho_{f_n}(u, v) - \mu(\rho)]^2}{\binom{n}{2}\mu(\rho)^2} = \frac{\sum\limits_{i=1}^{\binom{n}{2} - (n-1)} [1 - \mu(\rho)]^2 + \sum\limits_{i=1}^{n-1} [\alpha_i - \mu(\rho)]^2}{\binom{n}{2}[\mu(\rho)]^2}.$$

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

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

Since I is an isometry, and since we have $f(x_i) = I(x_i)$ by definition , we obtain that

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

By subadditivity of d_Y, we have that, $\forall i \in [n]$

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

5

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

$$0 < |\frac{d_Y(I(x_0), f(x_0))}{d_Y(I(x_0), f(x_i))} - 1| < \alpha_i < \frac{d_Y(I(x_0), f(x_0))}{\beta} + 1.$$

Hence there exists a constant $C \in \mathbb{R}$ such that for all $i \in \mathbb{N}$, $0 < \alpha_i < C$. It follows that

$$0 < \sum_{i=1}^{n-1} \alpha_i < (n-1)C, \ 0 < \sum_{i=1}^{n-1} \alpha_i^2 < (n-1)C^2 \text{ and } 0 < (\sum_{i=1}^{n-1} \alpha_i)^2 < ((n-1)C)^2.$$

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

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

$$\Phi_\sigma(f_n) > \frac{-(2\binom{n}{2}(\binom{n}{2} - (n-1)))((n-1)C) - \binom{n}{2}((n-1)C)^2}{\binom{n}{2}[(\binom{n}{2} - (n-1))^2 + ((n-1)C)^2 + 2(n-1)C(\binom{n}{2} - (n-1))]^2}.$$

Observe that

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

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

Hence it follows that $\lim_{n \to \infty} \Phi_\sigma(f_n) = 0 = \lim_{n \to \infty} \Phi_\sigma(I_n)$. $\qquad \square$

(c) Φ_{wc}, Φ_{navg} and Φ_{klocal} are not robust to outliers in data or distances.

Proof. Proof by a counterexample. We first construct a sequence of embeddings for which the number of distances that are distorted is bounded by a constant K and the distances distorted stem from mapping a single point away from an isometry. Then we show that in the limit, as the size of the metric space tends to infinity, Φ_{wc}, Φ_{navg} and Φ_{klocal} do not have the same distortion as that of an isometry.

Let $\{e_i\}_{\{i=1,...,d\}}$ denote the standard orthonormal basis for $(\mathbb{R}^d, l_2)$. Fix $x_0 = (\alpha - 1) \cdot e_d$ for some $1 < \alpha < 2$. Let $x_1 = e_d$. For any $n \in \mathbb{N}$, set $X_n = \{x_2, x_3, ..., x_n\}$, such that for all $i \in \{2, 3, ..., n\}$, x_i is sampled according to some distribution $\mathcal{P}$ on $span\{e_1, ..., e_{d-1}\}$. 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 $X_n \cup \{x_0, x_1\}$. It is easy to verify that the ratio of distances $\rho_{f_n}(x_0, x_1) = \alpha$ and $\rho_{f_n}(x_i, x_j) = 1$ for any $\{x_i, x_j\} \in \binom{X_n}{2} \setminus \{x_0, x_1\}$.

worstcase distortion:

The worstcase distortion evaluated on $f_n : (X_n \cup \{x_0, x_1\}, d_X) \to (X, d_X)$ for any $n \in \mathbb{N}$ is α and thus $\lim_{n \to \infty} \Phi_{wc}(f_n) = \alpha > \lim_{n \to \infty} \Phi_{wc}(I_n) = 1$, where I_n denotes the restriction of the mapping I to $X_n \cup \{x_0, x_1\}$.

normalized average distortion:

The sequence of average distortions evaluated on mappings $\{f_n\}$ is given by:

$$\Phi_{navg}(f_n) = \frac{(\binom{n}{2} - 1)\alpha + 1}{\binom{n}{2}}$$

Thus, $\lim_{n \to \infty} \Phi_{navg}(f_n) = \alpha > \Phi_{navg}(I_n) = 1$.

k-local distortion:

Since x_0 lies in the set of k-nearest neighbours of x_1, for any $n \geq 2$, the k-local distortion evaluated on the mapping $f_n : (X_n \cup \{x_0, x_1\}, d_X) \to (X, d_X)$ is α and thus $\lim_{n \to \infty} \Phi_{klocal}(f_n) = \alpha > \lim_{n \to \infty} \Phi_{klocal}(I_n) = 1$. $\qquad \square$

6

A.4 Incorporation of a probability distribution

Let (X, d_X) be an arbitrary metric space. Let $X_n = \{x_1, x_2, ..., x_n\}$ be a finite subset of X. Let P denote a probability distribution on X_n and let $\Pi = P \times P$ denote the product distribution on $X_n \times X_n$. Fix any arbitrary $x^*, y^* \in X_n$ such that $P(x^*) > P(y^*)$. Let $x', y' \in X$ such that for all $i \in [n]$, $d_X(x_i, x') = \alpha d_X(x_i, x^*)$ and $d_X(x_i, y') = \alpha d_X(x_i, y^*)$. Let $f, g : X_n \to X$ be two embeddings such that:

$$f(x) = \begin{cases} x', & \text{if } x = x^*. \\ x, & \text{otherwise.} \end{cases} \quad , \quad g(x) = \begin{cases} y', & \text{if } x = y^*. \\ x, & \text{otherwise.} \end{cases}$$

(a) $\Phi_{wc}, \Phi_{avg}, \Phi_{navg}, \Phi_\epsilon$, and Φ_{klocal} fail to incorporate a probability distribution into their evaluation.

Proof. The proofs follow directly from the definitions of the distortion measures.

worstcase, normalized avg, ϵ-distortion and k-local distortion: The above distortion measures, by definition are independent of the probability distribution over the data space.

average (l_q) distortion: The proof follows from explicit evaluation of the average distortion of any two embeddings (f, g) that satisfy the conditions as specified in the definition. Average distortion of f can be expressed as:

$$\Phi_{avg}(f) = \mathbb{E}_\Pi[\rho_f(u, v)] = [\sum_{i,j \neq 1,2} \Pi_{ij} + \sum_{i \neq 1} \alpha \Pi_{i1} + \sum_{i \neq 2} \Pi_{i2}].$$

$$\Phi_{avg}(f) - \Phi_{avg}(g) = [\sum_{i \neq 1,2} (\Pi_{i1} - \Pi_{i2})(\alpha - 1)]$$

$$= (\sum_{i \neq 1,2} (\Pi_i)(\Pi_1 - \Pi_2)(\alpha - 1))$$

$$< 0 \quad \text{if } \alpha < 1.$$

The case of l_q-distortion follows similarly. $\square$

(b) Φ_σ incorporates a probability distribution into its evaluation.

Proof. Observe that $\sum_{u \neq v \in X} \rho_f(u, v) = \sum_{u \neq v \in X} \rho_g(u, v) = \binom{n}{2} - (n - 1) + (n - 1)\alpha$. Let $\kappa = \binom{n}{2} / \sum_{u \neq v \in X} \rho_f(u, v) = \binom{n}{2} / \sum_{u \neq v \in X} \rho_g(u, v)$. Then,

$$\Phi_\sigma(f) - \Phi_\sigma(g) = \mathbb{E}_\Pi[(\kappa \rho_f(u, v) - 1)^2] - \mathbb{E}_\Pi[(\kappa \rho_g(u, v) - 1)^2]$$

$$= \mathbb{E}_\Pi[\kappa^2 \rho_f(u, v)^2 - 2\kappa \rho_f(u, v)] - \mathbb{E}_\Pi[\kappa^2 \rho_g(u, v)^2 - 2\kappa \rho_g(u, v)].$$

$$= \kappa^2(\alpha^2 - 1) \sum_{i,j \neq 1,2} (\Pi_{1i} - \Pi_{2i}) - 2\kappa(\alpha - 1) \sum_{i,j \neq 1,2} (\Pi_{1i} - \Pi_{2i})$$

$$= K^2(\Pi_1 - \Pi_2)(\alpha - 1)(\kappa(\alpha + 1) - 2) \sum_{i \neq 1,2} \Pi_i.$$

It is easy to verify that for any $\alpha \geq 0$ and for all $n > 4$, $(\kappa(\alpha + 1) - 2)(\alpha - 1) \geq 0$ and by definition $\Pi_1 > \Pi_2$. Therefore, $\Phi_\sigma(f) \geq \Phi_\sigma(g)$. $\square$

B Proofs of Theorems 3 and 4

Doubling space: A metric space (X, d_X) is referred to as a doubling space if there exists a doubling constant $\lambda > 0$ such that for any $u \in X$ and $r > 0$, the ball $B(u, r) = \{v \mid d_X(u, v) < r\}$ can be covered by at most λ balls of radius $r/2$.

B.1 Theorem 3:

Abraham, Bartal, and Neiman, 2011 showed the existence of an embedding from any arbitrary metric space into an Euclidean space with properties as stated in Theorem A. Combined with Lemma A, this provides the sought upper bound on σ-distortion.

Theorem A. *(Abraham, Bartal, and Neiman, 2011) Given any arbitrary finite metric space (X, d_X), there exists an embedding $f : X \to l_p^D$ where $D = O(\log n)$ and for any $\epsilon \in (0, 1)$, there exists a set G_ϵ and constants C_1 and C_2 (independent of ϵ,n) such that $|G_\epsilon| \geq (1 - \epsilon) \cdot \binom{|X|}{2}$ and for any $x, y \in G_\epsilon$: $C_1 \leq \frac{\|f(x) - f(y)\|_p}{d_X(x,y)} \leq C_2 \cdot \log(\frac{2}{\epsilon})$.*

Lemma A. *(Abraham, Bartal, and Neiman, 2011) Given any finite metric spaces (X, d_X) and (Y, d_Y) and an embedding $f : X \to Y$ satisfying properties described in Theorem A, for any distribution Π over $X \times X$, there exists a constant $K = K(\Pi)$ such that $\mathbb{E}_\Pi \left(\frac{d_Y(f(x), f(y))}{d_X(x,y)} \right)^2 < K$ and $\mathbb{E}_\Pi \left(\frac{d_Y(f(x), f(y))}{d_X(x,y)} \right) < K$.*

Proof of Theorem 3: From theorem A, by choosing any $\epsilon < \frac{2}{n(n-1)}$, we have that $\forall x, y \in X$, $C_1 \leq \frac{\|f(x) - f(y)\|_p}{d_X(x,y)}$ implies $C_1 \leq \mathbb{E}_\Pi \left(\frac{\|f(x) - f(y)\|_p}{d_X(x,y)} \right)$. Recall that σ-distortion is defined as

$$\frac{\mathbb{E}_\Pi[\rho_f(u,v) - \frac{2}{n(n-1)} \sum \rho_f(u,v)]^2}{(\frac{2}{n(n-1)} \sum \rho_f(u,v))^2}.$$

Combined with Lemma A, this completes the proof of Theorem 3. $\qquad\square$

B.2 Theorem 4:

(Abraham, Bartal, and Neiman, 2011) also showed the existence of an embedding from any arbitrary doubling metric space into an Euclidean space with properties as stated in Theorem B and Lemma B. These results provide an upper bound on the σ-distortion evaluated on this embedding.

Theorem B. *(Abraham, Bartal, and Neiman, 2011) Given any finite metric space (X, d_X) with doubling constant λ, there exists an embedding $f : X \to l_p^D$ and a constant $K = K(\lambda)$ such $D < K$ and for any $\epsilon \in (0, 1)$, there exists a set G_ϵ and constants C_1 and C_2 (independent of ϵ) such that $|G_\epsilon| \geq (1 - \epsilon) \cdot \binom{|X|}{2}$ and for any $x, y \in G_\epsilon$: $C_1 \leq \frac{\|f(x) - f(y)\|_p}{d_X(x,y)} \leq C_2 \cdot \log^{26}(\frac{1}{\epsilon})$.*

Lemma B. *(Abraham, Bartal, and Neiman, 2011) Given a finite metric space (X, d_X) with doubling constant λ, another metric space (Y, d_Y) and an embedding $f : X \to Y$ satisfying the properties described in Theorem B, then for any distribution Π over $X \times X$, there exists a constant $K = K(\Pi)$ such that $\mathbb{E}_\Pi \left(\frac{d_Y(f(x), f(y))}{d_X(x,y)} \right)^2 < K$.*

Proof of Theorem 4: Same as the proof of Theorem 4. $\qquad\square$