

# Dimensionality reduction: beyond the Johnson-Lindenstrauss bound\*

Yair Bartal<sup>†</sup>

Ben Recht<sup>‡</sup>

Leonard J. Schulman<sup>§</sup>

## Abstract

Dimension reduction of metric data has become a useful technique with numerous applications. The celebrated Johnson-Lindenstrauss lemma states that any  $n$ -point subset of Euclidean space can be embedded in  $O(\epsilon^{-2} \log n)$ -dimension with  $(1 + \epsilon)$ -distortion. This bound is known to be nearly tight.

In many applications the demand that all distances should be nearly preserved is too strong. In this paper we show that indeed under natural relaxations of the goal of the embedding, an improved dimension reduction is possible where the target dimension is independent of  $n$ . Our main result can be viewed as a *local dimension reduction*. There are a variety of empirical situations in which small distances are meaningful and reliable, but larger ones are not. Such situations arise in source coding, image processing, computational biology, and other applications, and are the motivation for widely-used heuristics such as Isomap and Locally Linear Embedding.

Pursuing a line of work begun by Whitney, Nash showed that every  $C^1$  manifold of dimension  $d$  can be embedded in  $\mathbb{R}^{2d+2}$  in such a manner that the local structure at each point is preserved isometrically. Our work is an analog of Nash's for discrete subsets of Euclidean space. For perfect preservation of infinitesimal neighborhoods we substitute near-isometric embedding of neighborhoods of bounded cardinality.

We show that any finite subset of Euclidean space can be embedded in  $O(\epsilon^{-2} \log k)$ -dimension while preserving with  $(1 + \epsilon)$ -distortion the distances within a “core neighborhood” of each point. (The core neighborhood is a metric ball around the point, whose radius is a substantial fraction of the radius of the ball of cardinality  $k$ , the  $k$ -neighborhood.) When the metric space satisfies a weak growth rate property, the guarantee applies to the entire  $k$ -neighborhood (with some dependency of the embedding dimension on the growth rate). We also show how to obtain a global embedding that also keeps distant points well-separated (at the cost of dependency on the doubling dimension of the space).

As an application of our methods we obtain an (Assouad-style) dimension reduction for finite subsets of Euclidean space where the metric is raised to some fractional

power (the resulting metrics are known as snowflakes). We show that any such metric  $X$  can be embedded in dimension  $\tilde{O}(\epsilon^{-3} \dim(X))$  with  $1 + \epsilon$  distortion, where  $\dim(X)$  is the doubling dimension, a measure of the intrinsic dimension of the set. This result improves recent work by Gottlieb and Krauthgamer [20] to a nearly tight bound.

The new dimension reduction results are useful for applications such as clustering and distance labeling.

## 1 Introduction

Dimension reduction for high dimensional metric data has been an extremely important paradigm in many application areas. In particular, the celebrated Johnson-Lindenstrauss (JL) Lemma [25] has played a central role in a plethora of applications. The lemma states that every  $n$ -point subset of Euclidean space can be embedded in dimension  $O(\epsilon^{-2} \log n)$  with  $1 + \epsilon$  distortion. This bound is known to be nearly tight [5]. However, in practical instances it is often the case that the high-dimensional data is, locally, inherently low dimensional, and it is therefore desirable to reduce its dimension close to its inherent dimensionality, which is independent of the size of the data set. In this paper we offer a theoretical study of such “local” dimension reduction methods.

In many large-scale data processing applications, local distances convey more useful information than large distances and are sufficient for uncovering low-dimensional structure. Such situations would arise if the large distances are inaccurate or do not reflect the intrinsic geometry of the application. Moreover, there are a variety of situations that rely only on local distances, including nearest-neighbor search, the computation of vector quantization rate-distortion curves [18], and popular data-segmentation and clustering algorithms [40]. In all of these cases, it is often desirable to reduce the dimension of the data set for reductions of storage requirements or algorithm running times. If the long distances are unimportant, we may be able to reduce the dimensionality only preserving the local information, and such reduction can be into a far lower dimension than what is possible when attempting to preserve distances between all pairs of points.

Our main result is a *local dimension reduction* lemma which replaces the dependency in the global size of the data  $n$  in the JL bound with a local parameter.

\*A previous version of this paper was posted under the title: “A Nash-type Dimensionality Reduction for Discrete Subsets of  $L_2$ ” [10]. The work of the first and second authors was performed in part while at the Center for the Mathematics of Information, Caltech.

<sup>†</sup>yair@cs.huji.ac.il. School of Engineering and Computer Science, Hebrew University, Israel Supported in part by a grant from the Israeli Science Foundation (195/02) and in part by a grant from the National Science Foundation (NSF CCF-065253).

<sup>‡</sup>brecht@cs.wisc.edu. Computer Science Department, U Wisconsin, Madison.

<sup>§</sup>schulman@caltech.edu. Caltech, Pasadena, CA 91125. Supported in part by NSA H98230-06-1-0074, NSF CCF-0515342 and NSF CCF-0829909.

We then apply our lemma to provide dimension reduction for data with low “intrinsic dimension,” often measured by the doubling dimension [6, 21] of the data set. We show that the snowflake version of the data, where distances are raised to some fixed fractional power, can be embedded in dimension close to the doubling dimension. This provides a nearly tight bound for this problem, a variant of Assouad’s problem [6], recently raised and studied by Gottlieb and Krauthgamer [20].

For improvements to the JL lemma in another direction, see the bounds on the embedding dimension established in [19, 26] (and see [41] for an introduction to this approach); in contrast to ours, these embeddings are linear, but the embedding dimension may be larger as it depends not on the local behaviour of the data but upon a global parameter (known as  $\gamma_2$ ).

**1.1 Local Dimension Reduction** Two influential papers posited that if a high-dimensional data set lies on the embedding of a low-dimensional Riemannian manifold, the intrinsic dimensionality could then be found by examining only the nearest neighbor distances of the graph. The first algorithm, known as Isomap [42], uses Dijkstra’s algorithm on the nearest neighbors graph to compute the global distances and then applies multi-dimensional scaling to the computed distances to find a low dimensional embedding of the data. The second, Local Linear Embedding [37], computes the best linear approximation of each set of neighbors, and then stitches the neighborhoods together by solving an eigenvalue problem constraining the mappings of overlapping neighborhoods. Based on these initial results and their accompanying empirical examples, these two papers gave rise to an active field, commonly referred to as *manifold learning*, and the ensuing years have seen a multitude of applications of these algorithms in areas as diverse as protein folding [14], motion planning in robotics [24], data-mining microarray assays [33], and face recognition [22]. All of these applications use the  $L_2$  distance, even if it is not perfectly justified, because of its tractability and empirical power. Moreover, there have been a variety of alternative algorithms proposed to reduce dimensionality in nearest neighbor distances problems, employing kernel methods [11], generative probabilistic models [13], semidefinite programming [44] or neural networks [23].

Despite their wide appeal, all of these algorithms assume some sort of manifold model underlies the data, and make implicit assumptions about intrinsic curvature, Riemannian metrics, or volume. More importantly, not one of these manifold learning algorithms come with any provable guarantees for discrete data

sets, and many authors have pointed out that the geometric assumptions of these algorithms are not reasonable in practice. For example, the algorithms are quite sensitive to the determination of neighborhood structure [7], have problems recovering non-convex domains or manifolds with nontrivial homology [16], and cannot recover manifold structures that require more than one coordinate chart [34].

From a more theoretical perspective, the concept of a “local embedding” was first introduced in the context of metric space embedding in [2]. Local embeddings share the same objective as manifold learning: to find a mapping of a metric space into a low-dimensional metric space where distances of close neighbors are preserved more faithfully than those of distant neighbors. The field of metric embedding has been an active field of research both in mathematics and computer science and has emerged as a powerful tool in many algorithmic application areas. Two cornerstone theorems in this field are the theorem of Bourgain [12] stating that any  $n$ -point metric space embeds in  $L_2$  with  $O(\log n)$  distortion, and the JL [25] dimension reduction lemma. Both these theorems have many algorithmic consequences.

Abraham, Bartal and Neiman [2] show that many of the known classic embedding results can be extended to the context of local embeddings. In particular, generalizing Bourgain’s theorem (and [1]) they provide local embeddings requiring only  $O(\log k)$  dimensions to achieve distortion  $O(\log k)$  on the neighborhoods with at most  $k$ -points, assuming the metric obeys a certain *weak growth rate* condition, and [4] remove this assumption at the cost of increasing the dimension to  $O(\log^2 k)$ . This number  $k$  could have no relation to  $n$ , and in practice could be arbitrarily smaller than  $n$ . It should be emphasized that this type of embedding is an *immersion*, that is it preserves well the short distances but may arbitrarily distort the long ones. This is reasonable, for instance, if we desire a compact *distance oracle* [43] for close neighbors.

In this paper, we provide a local version of the JL lemma. Such a construction is challenging to achieve because all of the previously discussed algorithms based on this lemma require a globally consistent choice of random variables. For this reason, results extending the JL lemma to the projection of smooth manifolds end up depending on the dimension where the manifold is embedded, and both the volume and curvature of the manifold [8]. Here, we present an embedding of dimension that has no dependence on the volume. We show that for any  $\epsilon > 0$ , only  $O(\epsilon^{-2} \log k)$  dimensions are required for embedding with distortion  $1 + \epsilon$  within the neighborhoods with at most  $k$ -points, where the near-isometry is preserved inside a core neighborhood

of diameter at least  $\Omega(\epsilon^{1.5}/\log k)$  factor of the diameter of the  $k$ -neighborhood. As a consequence we get an alternative view of the result whereby if the metric obeys a weak growth rate condition (similar to the one defined by Abraham *et al.* [2]) then the near-isometry holds for the entire  $k$ -neighborhood (with some dependency of the dimension on the growth rate). Some assumption of this form is necessary, as follows from a lower bound by Schechtman and Shraibman [38] showing that there are worst case examples where no near-isometric local dimension reduction method can beat the JL bound. Our result overcomes this difficulty by showing that the near-isometry property can still be maintained within a *core* neighborhood. If the space has the weak growth rate property then the core neighborhood includes the  $k$ -neighborhood, and so it is fully preserved. In the general case the core neighborhood has smaller diameter compared to the  $k$ -neighborhood and so there is no guarantee that there will be many (or any) neighbors in it. Recall however that the main motivation for our result is preserving the *small* distances in the space, which is what our result guarantees. Our work provides the first result of this type. Prior to our work the only case where such a result was known is when the input set is isometric to an ultrametric [4].

For general metrics, this embedding is an immersion, but under the assumption that the metric has low intrinsic dimensionality (i.e., small doubling dimension) we can transform our immersion into a *global embedding* such that distances between far points can be bounded below so they don't intrude on the local structure. Unlike the results in manifold learning, we make no assumptions that our data lie on some compact manifold, and further assume nothing about the volume or cardinality of our data set.

As an example application that our embedding is suited to, the principal computational problem in vector quantization [18] is formally one of clustering (with  $\ell_2^2$  costs), but the parameters are different than in the clustering literature: primarily, one studies here the limit that the number of clusters,  $s$ , tends to  $\infty$ , while the distortion (the average distance to a codeword) tends to 0. This means that only the small distances between data points are germane to the problem. Known algorithms for construction of near-optimal clusterings are exponential in either  $s$  or the dimension of the space. Our embedding is well-suited to taking advantage of dimensionality reduction for vector quantization, since our target dimension depends only on the size of the small regions in which the  $L_2$  distance needs to be preserved. Using our embedding, the vector quantization algorithm can be run in a low-dimensional

space, and the clustering ("codebook") can then be lifted back to the original space.

Our approach for local dimension reduction combines several metric embedding techniques. We first employ probabilistic partitioning [9] of our metric space (Section 2). These partitions, developed in [1, 2, 4], decompose the metric space into clusters of bounded diameter and allow the coordinates of the embedding to smoothly transition between neighborhoods. As opposed to the standard decompositions where cluster diameters are similar, the partitions of [4] allow varying diameters to capture neighborhoods of similar cardinality. The idea is to apply for each of the clusters of the partition separately a dimension reduction method on the points within the cluster and then assemble these embedded neighborhoods into a global immersion.

While this idea sounds simple it in fact fails if we attempt to directly apply the JL embedding method in each of the clusters. The reason is that the values the embedding takes may be as large as the diameter of the cluster and that may temper the Lipschitz condition between points in separate clusters (that is the ratio of the embedded distance to the original distance may be unbounded). To avoid that we need to combine the dimension reduction method with a truncation mechanism. While there are several ways in which this may be done we introduce a natural and elegant mechanism for this aim which we call the randomized Nash device. To ensure the Lipschitz condition we finally apply a smoothing operator.

Our methods owe a substantial debt to seminal papers in several areas of mathematics. Pursuing a line of work begun by Whitney [45, 46], Nash showed that every Riemannian manifold of dimension  $D$  could be embedded in  $\mathbb{R}^{2D+2}$  by a  $C^1$  mapping such that the metric at each point is preserved isometrically [32]. Nash achieves this embedding using a device which locally perturbs a non-distance preserving embedding provided by Whitney. The randomized trigonometric embedding of Section 3.1 is adapted from Nash's deterministic embedding procedure, and we give a probabilistic analysis showing that with high probability this yields an embedding of the local distances in each neighborhood. As observed in [35] in the context of fast algorithms for pattern recognition, our random trigonometric functions form an embedding into a Euclidean space where the inner product approximates a positive definite shift-invariant kernel function. In our case, we sample frequencies from a Gaussian distribution and use the smoothness properties of the gaussian kernel  $k(x, y) = \exp(-\gamma\|x - y\|^2)$  to ensure the quality of our randomized Nash device. Our Nash device can also be viewed as a discretized version of the continuous trun-

cation technique of Schoenberg [39] which has appeared in the embedding literature (e.g. [30, 29, 20]). (These methods, combined with the JL dimension reduction, could have replaced the Nash device, but the latter is itself elegant, computationally efficient, simple to use, and has technical advantage in our proofs and so it may be of independent interest).

The existence of our embedding is guaranteed using the Lovász Local Lemma [17], and we rely on algorithmic implementation of the LLL by Moser and Tardos [31] to provide a randomized algorithm to generate our embeddings. The application of the LLL together with probabilistic partitions was first applied by Krauthgamer and Lee [27] and was later used in several papers in the context of low dimensional embedding (e.g. [3, 2]). Our work is closely related to that of [27] as they too give an embedding that preserves local structure under a (regular) growth rate assumption. However, there is a significant difference: while their work preserves distances only in a single scale we obtain a near-isometry in the entire local neighborhood.

Our main contribution is in the combination of these various ingredients to allow local dimension reduction. Following our work, this methodology has been applied in [20] in additional cases of dimensionality reduction. We mainly focus on applying these tools to obtain a *near optimal* local dimension reduction. Most notably, obtaining the near optimal bound requires a delicate probabilistic argument. The embedding must compose the coordinates associated with the probabilistic partitions and those associated with the Nash-type dimension reduction in an interlacing manner. The analysis follows with carefully balancing the contributions of the different components through the dependencies of the relevant probabilistic events.

In some applications it may be important that the dimension reduction procedure will keep the embedded distant pairs away from the local neighborhoods. In general, this is impossible if no further assumptions are made. However, under the additional assumption that the metric space has low doubling dimension [6, 21] we ensure that our mapping has this property.

**1.2 Dimension Reduction for Snowflakes** Let  $X$  be a subset of Euclidean space. The doubling constant of  $X$  is the minimum  $\lambda$  such that every ball can be covered by  $\lambda$  balls of half the radius. The *doubling dimension* of  $X$  is defined as  $\dim(X) = \log_2 \lambda$ . The question of whether the dimension bound in the JL lemma can be reduced to  $O(\epsilon^{-2} \dim(X))$  has been posed by several researchers [28, 21, 3]. While this question remains open, it has been recently asked by Gottlieb and Krauthgamer [20] if a result along

this line is possible for the “snowflake” version of the metric, i.e, if the distance function  $d(x, y) = \|x - y\|$  is replaced with  $d^\alpha(x, y) = \|x - y\|^\alpha$  for some  $0 < \alpha < 1$ . Such an embedding may suffice for certain applications. This problem is motivated by Assouad’s theorem [6] which states that the snowflake version of any metric space can be embedded in Euclidean space with dimension and distortion depending solely on the doubling dimension. Gottlieb and Krauthgamer [20] use a similar approach to ours to prove that such a dimension reduction is possible where the target dimension is  $\tilde{O}(\frac{1}{\alpha(1-\alpha)} \epsilon^{-4} (\dim(X))^2)$ . We observe that the main ingredient needed in the solution for this problem is a local dimension reduction theorem. Using a variant of our main local dimension reduction theorem (Theorem 5.2) we improve their result to a nearly tight bound:  $\tilde{O}(\frac{1}{\alpha(1-\alpha)} \epsilon^{-3} \dim(X))$ .

This theorem may be applicable to distance labeling schemes and to optimization problems where the objective function is composed of powers of distances, e.g., clustering problems.

**1.3 Structure of the Paper** In Section 2 we provide background on the probabilistic partitions that we use. Theorem 3.1 is proved in Section 3. The local Nash-device is described in Section 3.1. We first give the main component of the embedding in Section 3.2 which provides the guarantee for “close” pairs. Then in Section 3.3 we provide the complete definition of the embedding which now deals with farther pairs that are still within the range of application of our main theorem (Theorem 3.1). In Section 4 we show how to extend the embedding to deal with all pairs and maintain separation of local and distant pairs (Theorem 4.1). Finally, in Section 5 we prove the dimension reduction for snowflakes (Theorem 5.1).

## 2 Preliminaries

We start with some basic definitions: Let  $k \in \mathbb{N}$ . For a point  $x \in X$  and  $r \geq 0$ , the ball at radius  $r$  around  $x$  is defined as  $B(x, r) = \{z \in X \mid \|x - z\| \leq r\}$ . For a point  $x \in X$  let  $\Delta_k(x)$  be the smallest radius  $r$  such that  $|B(x, r)| \geq k$ . For a pair  $x, y \in X$ , define:  $\Delta_k(x, y) = \max\{\Delta_k(x), \Delta_k(y)\}$ .

For any point  $x \in X$  and a subset  $S \subseteq X$  let  $d(x, S) = \min_{s \in S} d(x, s)$ . The *diameter* of  $X$  is denoted  $\text{diam}(X) = \max_{x, y \in X} d(x, y)$ .

We require the definition of the Gaussian transform:  $G_r(z) = r(1 - \exp(-z^2/r^2))^{1/2}$ .

One of the tools we use are local probabilistic partitions. In particular, the following constructions are generalizations of the local probabilistic partitions of [2], and their analysis appears in [4]:

**DEFINITION 2.1. (PROBABILISTIC PARTITION)** A partition  $P$  of  $X$  is a collection of disjoint set of clusters  $\mathcal{C}(P) = \{C_1, C_2, \dots, C_t\}$  such that  $X = \cup_j C_j$ . A partition is called  $\Delta$ -bounded where  $\Delta : P \rightarrow \mathbb{R}^+$  if for all  $j$ ,  $\text{diam}(C_j) \leq \Delta(C_j)$ . For  $x \in X$  we denote by  $P(x)$  the cluster containing  $x$ . A probabilistic partition  $\hat{P}$  of a finite metric space  $(X, d)$  is a distribution over a set  $\mathcal{P}$  of partitions of  $X$ . Such a partition is  $\Delta$ -bounded if it is  $\Delta$ -bounded for every  $P \in \hat{P}$ .

**DEFINITION 2.2. (LOCALLY PADDED PP)** Let  $\hat{P}$  be a  $\Delta$ -bounded probabilistic partition of  $(X, d)$ . Let  $\mathcal{L}(x)$  denote the event that  $B(x, \eta \cdot \Delta(P(x))) \subseteq P(x)$ . For  $\delta \in (0, 1]$ ,  $\hat{P}$  is called  $(\eta, \delta)$ -locally padded if for any  $x \in X$  and  $Z \subseteq X \setminus B(x, 16\Delta(P(x)))$ :  $\Pr[\mathcal{L}(x) \wedge_{z \in Z} \mathcal{L}(z)] \geq \delta$ .

**LEMMA 2.3. (CARDINALITY-BASED LPPP)** Let  $(X, d)$  be a finite metric space. Let  $k \in \mathbb{N}$ . There exists a  $\Delta$ -bounded probabilistic partition  $\hat{P}$  of  $(X, d)$  with the following properties:

- For any  $P \in \mathcal{P}$  and any  $x \in X$ :  $|P(x)| \leq k$ .
- For any  $P \in \mathcal{P}$  and any  $x \in X$ :  $2^{-6} \leq \Delta(P(x))/\Delta_k(x) \leq 2^{-4}$ .
- $\hat{P}$  is  $(\eta^{(\delta)}, \delta)$ -locally padded for  $\eta^{(\delta)} = 2^{-11} \cdot \ln(1/\delta)/\ln k$ , for each  $\delta \in (1/k, 1]$ .

**Lemma 2.3** is a reformulation of Lemma 5 from [4]. A simple application of the Lovász Local Lemma implies:

**LEMMA 2.4.** Let  $(X, d)$  be a finite metric space. Let  $k \in \mathbb{N}$  and  $\xi > 0$ . Let  $\{\hat{P}^{(t)}\}_{t \in T}$  be a collection of size  $|T| \geq 8 \log k / \xi$  of independent  $\Delta$ -bounded probabilistic partitions of  $(X, d)$  as in Lemma 2.3. Let  $\delta = 1 - \xi$  and  $\mathcal{L}_t^{(\delta)}(x)$  denote the event that  $B(x, \eta^{(\delta)} \cdot \Delta(P^{(t)}(x))) \subseteq P^{(t)}(x)$ , where  $\eta^{(\delta)} = 2^{-11} \cdot \ln(1/\delta)/\ln k$ . Then with positive probability for every  $x \in X$  there exists a set  $T^{(\delta)}(x) \subseteq T$  of size  $|T^{(\delta)}(x)| \geq (1 - 2\xi)|T|$  such that  $\mathcal{L}_t^{(\delta)}(x)$  occurs for all  $t \in T$ .

### 3 Local Dimension Reduction

Given a discrete set of points  $X$  of cardinality  $n$  in  $U$ -dimensional Euclidean space we construct a low dimension local embedding, one that preserves distances to close neighbors with a  $1 + \epsilon$  multiplicative error. The main result of this paper is summarized by the following theorem.

Let  $k \in \mathbb{N}$ . Recall that for a point  $x \in X$ ,  $\Delta_k(x)$  denotes the smallest radius  $r$  such that  $|B(x, r)| \geq k$ , and for a pair  $x, y \in X$ :  $\Delta_k(x, y) = \max\{\Delta_k(x), \Delta_k(y)\}$ . Let  $\Delta_k^*(x) = c_1 \epsilon \Delta_k(x) / \log k$ , where  $c_1 < 1$  is a universal constant, and  $\Delta_k^*(x, y) = \max\{\Delta_k^*(x), \Delta_k^*(y)\}$ .

**THEOREM 3.1.** Let  $k \in \mathbb{N}$ . Given  $X$  a discrete subset of  $\mathbb{R}^U$ , then for any  $0 < \epsilon < 1/2$  there exists an embedding  $\hat{\Phi} : X \rightarrow \mathbb{R}^D$ , where  $D = O(\log k / \epsilon^2)$  with the following properties. For all  $x, y \in X$ :

- $\|\hat{\Phi}(x) - \hat{\Phi}(y)\| \leq (1 + \epsilon)\|x - y\|$
- if  $\|x - y\| \leq \Delta_k^*(x, y)$ :
 
$$\|\hat{\Phi}(x) - \hat{\Phi}(y)\| \geq \begin{cases} \frac{\|x-y\|}{1+\epsilon} & \text{if } \|x-y\| \leq \sqrt{\frac{\epsilon}{6}} \Delta_k^*(x, y) \\ \frac{\|x-y\|}{1+\epsilon'} & \text{if } \|x-y\| = \sqrt{\frac{\epsilon'}{6}} \Delta_k^*(x, y), \epsilon' > \epsilon \end{cases}$$
- if  $\Delta_k^*(x, y) < \|x - y\| \leq \frac{1}{2} \Delta_k(x, y)$ :
 
$$\|\hat{\Phi}(x) - \hat{\Phi}(y)\| \geq \frac{1}{8} \Delta_k^*(x, y).$$

We note that although Theorem 3.1 maintains  $(1 + \epsilon)$ -distortion only in a core neighborhood within the  $k$ -neighborhood of a point, this implies  $(1 + \epsilon)$ -distortion for all pairs within the entire  $k$ -neighborhood if we demand that  $X$  satisfies a condition (introduced in [2]) called the *weak growth rate condition*<sup>1</sup> with parameter  $\gamma$  ( $0 < \gamma < 1$ ):  $X$  satisfies  $\text{WGR}(\gamma)$  if for every  $x \in X$  and  $r_1, r_2 > 0$ ,  $|B(x, r_2)| \leq |B(x, r_1)|^{(r_2/r_1)^\gamma}$ .

**COROLLARY 3.1.** Let  $k \in \mathbb{N}$ . Given  $X$  a discrete subset of  $\mathbb{R}^U$ , satisfying  $\text{WGR}(\gamma)$ , then for any  $\epsilon > 0$  there exists an embedding  $\hat{\Phi} : X \rightarrow \mathbb{R}^D$ , where  $D = O(\epsilon^{-2} \epsilon^{-3/2} \frac{1}{1-\gamma} (\log k)^{\frac{1}{1-\gamma}})$  such that for all  $x, y \in X$ :

- $\|\hat{\Phi}(x) - \hat{\Phi}(y)\| \leq (1 + \epsilon)\|x - y\|$ .
- if  $\|x - y\| \leq \Delta_k(x, y)$ :
 
$$\|\hat{\Phi}(x) - \hat{\Phi}(y)\| \geq (1 + \epsilon)^{-1} \|x - y\|.$$

In particular for  $\gamma = O(1/(\log(\epsilon^{-3/2} \log k)))$  we have  $D = O(\log k / \epsilon^2)$ .

In the rest of this section we describe the embedding and analysis to prove Theorem 3.1<sup>2</sup>. The main ingredients are a set of probabilistic partitions described in Section 2, and a compact embedding, based on a randomization of a device of Nash, provided in Section 3.1. The core of the construction is presented in Section 3.2 where we prove the existence of an embedding  $\Phi$  satisfying all of the properties in Theorem 3.1 for all  $x, y \in X$  which are “close neighbors” in the sense that  $\|x - y\| \leq \Delta_k^*(x, y)$ , as well as the upper bound for all pairs. For farther neighbors, we use a simple additional construction in Section 3.3.

<sup>1</sup>The reason this condition is called weak is that it does not exclude rapidly expanding metrics.

<sup>2</sup>We note that the constants may differ but a rescaling of the parameter  $\epsilon$  would yield this formulation of the theorem.

**3.1 The Randomized Nash Device** In this section we introduce a new construct we call the randomized Nash device.

For any  $\omega \in \mathbb{R}^U$  and  $\sigma > 0$ , we define the function  $\varphi : \mathbb{R}^U \rightarrow \mathbb{R}^2$  as

$$(3.1) \quad \varphi(x; \sigma, \omega) = \frac{1}{\sigma} \begin{bmatrix} \cos(\sigma \omega' x) \\ \sin(\sigma \omega' x) \end{bmatrix}$$

where  $\omega' x$  denotes the inner product between  $\omega$  and  $x$ .  $\varphi(x; \sigma, \omega)$  maps onto a circle with radius  $\sigma^{-1}$  in  $\mathbb{R}^2$ . These functions were used by Nash in his construction of  $C^1$ -isometric embeddings of Riemannian manifolds [32], with the parameters chosen to correct errors in the metric. Note that as the parameter  $\sigma$  grows, the frequencies of the embedding function grow, but the amplitude becomes increasingly small.

In this section we present a sequence of *random* parameter settings for these functions  $\varphi$ , first studied in [35], that with high probability approximate small distances in discrete metrics and bound large distances away from zero. Fix  $\sigma > 0$  and let  $\omega$  be a sample from a  $U$ -dimensional Gaussian  $\mathcal{N}(0, I_U)$ . For this choice of parameters, one may interpret Equation (3.1) as a random projection wrapped onto the circle. Using the intuition provided by the JL lemma, one would expect nearby points  $x$  and  $y$  to be mapped to nearby points on the circle since the sine and cosine are Lipschitz. This intuition can be further reinforced by considering the expected distance between two points. Recall the definition of the Gaussian transform:  $G_r(z) = r(1 - \exp(-z^2/r^2))^{1/2}$ .

**CLAIM 3.1.** *For any  $x$  and  $y$  in  $\mathbb{R}^U$ ,  $|\varphi(x; \sigma, \omega) - \varphi(y; \sigma, \omega)|^2 = 2\sigma^{-2}(1 - \cos(\sigma \omega'(x - y)))$  and  $\mathbb{E}[|\varphi(x; \sigma, \omega) - \varphi(y; \sigma, \omega)|^2] = G_r(\|x - y\|)^2$ , where  $r = \sqrt{2}/\sigma$ .*

The main result of this section is to note that these random variables are very well concentrated about their expected value and hence inherit their distance preserving property from this Gaussian kernel function. Hence, a concatenation of several  $\varphi$  corresponding to different samples of  $\omega$  will provide a low-dimensional embedding.

Let  $\sigma_1, \dots, \sigma_D > 0$  be given real numbers bounded above by  $\sigma_{\mathbf{m}}$  and below by  $\sigma_{\mathbf{s}}$ , and let  $\omega_1, \dots, \omega_D$  be  $D$  samples from a  $U$ -dimensional Gaussian  $\mathcal{N}(0, I_U)$ . Let  $\varphi^{(t)}(x) := \varphi(x; \sigma_t, \omega_t)$  and, for  $x$  and  $y \in \mathbb{R}^U$ , let  $\Theta : X \rightarrow \mathbb{R}^{2D}$  denote the mapping  $\Theta = \frac{1}{\sqrt{D}} \bigoplus_{1 \leq t \leq D} \varphi^{(t)}$ . The main result of this section is the following lemma:

**LEMMA 3.2.** *Let  $0 < \epsilon < \frac{1}{2}$  and  $x$  and  $y \in \mathbb{R}^U$ .*

- a.  $\|\Theta(x) - \Theta(y)\|^2 \leq (1 + \epsilon)\|x - y\|^2$  with probability exceeding  $1 - \exp(-\frac{D}{2}(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}))$ .

- b.  $\|\Theta(x) - \Theta(y)\|^2 \geq (1 - \epsilon)G_r(\|x - y\|)^2$ , where  $r = \sqrt{2}/\sigma_{\mathbf{m}}$ , with probability exceeding  $1 - \exp(-\frac{D\epsilon^2}{6})$ .
- c.  $\|\Theta(x) - \Theta(y)\|^2 \leq (1 + \epsilon)G_r(\|x - y\|)^2$ , where  $r = \sqrt{2}/\sigma_{\mathbf{s}}$ , with probability exceeding  $1 - \exp(-\frac{D\epsilon^2}{24})$ .

The randomized embedding  $\Theta$  maps onto a product of circles of varying radii, a subset of the  $2D$ -sphere. The different values of  $\sigma$  will be necessary in the following sections to stitch together regions of the metric space with differing densities, but the important point is all of the concentration results are only a function of the largest value of the  $\sigma_t$ . Intuitively, one can interpret this as saying the high frequency information is the dominant source of error in the approximation. The analysis of Lemma 3.2 appears in Appendix A.

**3.2 Embedding Close Neighbors** We now turn to a recipe for combining multiple instances of these trigonometric embeddings into a global map that preserves local distances using the probabilistic partitions discussed in Section 2. Specifically, we concern ourselves with the “close neighbors,” pairs  $x$  and  $y$  satisfying  $\|x - y\| \leq \Delta_k^*(x, y)$  (for the lower bound, while the upper bound is proved for all pairs). Let  $D = C' \lceil \log k / \epsilon^2 \rceil$ , where  $C'$  is some universal constant to be determined later. We construct a locally padded cardinality-based probabilistic partition  $\tilde{\mathcal{P}}^{(t)}$  as in Lemma 2.4, where  $T = [D]$  and  $\xi = \epsilon$ . Now fix a partition  $P^{(t)} \in \mathcal{P}^{(t)}$ . We define a trigonometric embedding for every cluster  $C \in P^{(t)}$ .

Let  $\sigma_C = 2^{12} \ln k / \epsilon \cdot \Delta(C)^{-1}$ , and let  $\{\omega_C | C \in P^{(t)}, 1 \leq t \leq D\}$  be i.i.d. samples from a  $U$ -dimensional Gaussian  $\mathcal{N}(0, I_U)$ . For  $x \in C$  define  $\sigma^{(t)}(x) = \sigma_C$ ,  $\omega^{(t)}(x) = \omega_C$ , and  $A^{(t)}(x) = \min \{d(x, X \setminus C), \sigma^{(t)}(x)^{-1}\}$ , and let

$$\Phi^{(t)}(x) = A^{(t)}(x) \hat{\varphi}^{(t)}(x)$$

where,

$$\hat{\varphi}^{(t)}(x) = \sigma^{(t)}(x) \varphi^{(t)}(x) = \begin{bmatrix} \cos(\sigma^{(t)}(x) \omega^{(t)}(x)' x) \\ \sin(\sigma^{(t)}(x) \omega^{(t)}(x)' x) \end{bmatrix}.$$

The function  $A^{(t)}$  serves as the amplitude of the embedding. For padded  $x$ , this number is equal to the amplitude defined in Section 3.1, and the amplitude rolls off to zero near the boundary of each cluster. In each cluster, we have a different trigonometric embedding, and continuity is maintained because the amplitude is zero at the boundaries of the clusters.

We define our embedding  $\Phi : X \rightarrow l_2^{2D}$  by concatenating  $D$  instances of  $\Phi^{(t)}$ :  $\Phi = \frac{1}{\sqrt{D}} \bigoplus_{1 \leq t \leq D} \Phi^{(t)}$ .

**Analysis Overview:** Our goal is to show that the embeddings  $\Phi$  and the Nash-device based embeddings

of Section 3.1 have similar distortion guarantees. The purpose of the padded probabilistic partitions and the smoothing amplitude function is to allow a smooth transition between the different local embeddings in different clusters. For a close pair the padded probabilistic partition guarantees that in  $\approx 1 - \epsilon$  of the coordinates they fall in the same cluster and therefore their distortion is governed by the local Nash-device based embedding, which still maintains its distortion guarantees over the random set of successful coordinates. With probability  $\approx \epsilon$  that this fails we rely on the Lipschitz property (that the smoothing amplitude function provides) to make sure the distortion only deviates slightly and the overall distortion remains  $1 + O(\epsilon)$ . To enable this probabilistic argument our proof utilizes the Lovász Local Lemma, showing that the necessary constraints are satisfied everywhere with positive probability. The rest of this section is devoted to carrying out this proof strategy.

**Embedding Analysis.** We start with the following lemma which will be useful to bound the distance between embedded points:

LEMMA 3.3. *Let  $x, y \in X$ . Then,*

1. *If  $P^{(t)}(x) \neq P^{(t)}(y)$ ,  $\|\Phi^{(t)}(x) - \Phi^{(t)}(y)\| \leq 2\|x - y\|$ .*
2. *If  $P^{(t)}(x) \neq P^{(t)}(y)$ ,  $d(x, X \setminus P^{(t)}(x)) \geq 2\sigma^{(t)}(x)^{-1}$  and  $d(y, X \setminus P^{(t)}(y)) \geq 2\sigma^{(t)}(y)^{-1}$ , then  $\|\Phi^{(t)}(x) - \Phi^{(t)}(y)\| \leq \|x - y\|$ .*
3. *If  $P^{(t)}(x) = P^{(t)}(y)$ ,  $\|\Phi^{(t)}(x) - \Phi^{(t)}(y)\|^2 \leq \|x - y\|^2 + \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2$ .*
4. *If  $C := P^{(t)}(x) = P^{(t)}(y)$ ,  $\sigma_C^{-1} \leq d(x, X \setminus P^{(t)}(x))$  and  $\sigma_C^{-1} \leq d(y, X \setminus P^{(t)}(y))$ , then  $\|\Phi^{(t)}(x) - \Phi^{(t)}(y)\| = \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|$ .*

*Proof.* First, we observe that for all  $x$  and  $y$

$$\|\Phi^{(t)}(x) - \Phi^{(t)}(y)\| = \|A^{(t)}(x)\hat{\varphi}^{(t)}(x) - A^{(t)}(y)\hat{\varphi}^{(t)}(y)\|.$$

We now proceed case by case.

For claims (1) and (2), note that since  $\|\hat{\varphi}^{(t)}(u)\| = 1$ , we have

$$\begin{aligned} \|\Phi^{(t)}(x) - \Phi^{(t)}(y)\| &\leq A^{(t)}(x)\|\hat{\varphi}^{(t)}(x)\| \\ &\quad + A^{(t)}(y)\|\hat{\varphi}^{(t)}(y)\| \leq A^{(t)}(x) + A^{(t)}(y). \end{aligned}$$

For claim (1) we have that  $A^{(t)}(x) + A^{(t)}(y) \leq d(x, X \setminus P^{(t)}(x)) + d(y, X \setminus P^{(t)}(y))$ . Now if  $x$  and  $y$  fall in different clusters,  $\|x - y\| \geq d(y, X \setminus P^{(t)}(y))$  and  $\|x - y\| \geq d(x, X \setminus P^{(t)}(x))$ , and the assertion follows. Claim (2) follows as  $A^{(t)}(x) + A^{(t)}(y) \leq \sigma^{(t)}(x)^{-1} + \sigma^{(t)}(y)^{-1} \leq 2 \max\{\sigma^{(t)}(x)^{-1}, \sigma^{(t)}(y)^{-1}\} \leq \max\{d(x, X \setminus P^{(t)}(x)), d(y, X \setminus P^{(t)}(y))\} \leq \|x - y\|$ .

We now turn to claims (3) and (4). Assume  $C := P^{(t)}(x) = P^{(t)}(y)$ . Then

$$\begin{aligned} \|\Phi^{(t)}(x) - \Phi^{(t)}(y)\|^2 &= (A^{(t)}(x) - A^{(t)}(y))^2 \\ &\quad + A^{(t)}(x)A^{(t)}(y)\|\hat{\varphi}^{(t)}(x) - \hat{\varphi}^{(t)}(y)\|^2, \end{aligned}$$

using  $\|\hat{\varphi}^{(t)}(u)\| = 1$ . In this case we have that  $A^{(t)}(x)A^{(t)}(y) \leq \sigma_C^{-2}$ . For claim (3) also need to show that  $|A^{(t)}(x) - A^{(t)}(y)| \leq \|x - y\|$  for all  $x, y \in P^{(t)}(x)$ . We show that  $A^{(t)}(x) - A^{(t)}(y) \leq \|x - y\|$  and the claim holds by reversing the roles of  $x$  and  $y$ . There are two cases: if  $A^{(t)}(y) = \sigma_C^{-1}$  then  $A^{(t)}(x) \leq \sigma_C^{-1}$  and  $A^{(t)}(x) - A^{(t)}(y) \leq 0$ . Otherwise  $A^{(t)}(y) = d(y, X \setminus P^{(t)}(y))$  and  $A^{(t)}(x) \leq d(x, X \setminus P^{(t)}(x))$  implying  $A^{(t)}(x) - A^{(t)}(y) \leq d(x, X \setminus P^{(t)}(x)) - d(y, X \setminus P^{(t)}(y)) \leq \|x - y\|$  since  $P^{(t)}(x) = P^{(t)}(y)$ .

Finally, for claim (4), we only need use the fact that  $A^{(t)}(x) = A^{(t)}(y) = \sigma_C^{-1}$ .

We now proceed to proving Theorem 3.1. For  $x, y \in X$ , let us now classify the different coordinates  $t$  according to the cases of Lemma 3.3. Define the sets

$$\begin{aligned} T_{\neq}(x, y) &= \{t | P^{(t)}(x) \neq P^{(t)}(y)\} \\ T_{=} (x, y) &= \{t | P^{(t)}(x) = P^{(t)}(y)\} \\ T_{\circ}(x, y) &= \{t | d(x, X \setminus P^{(t)}(x)) \geq 2\sigma^{(t)}(x)^{-1} \\ &\quad \wedge d(y, X \setminus P^{(t)}(y)) \geq 2\sigma^{(t)}(y)^{-1}\} \end{aligned} \quad (3.2)$$

so that we have the upper and lower bounds for our embedded distances

$$\|\Phi(x) - \Phi(y)\|^2 \geq \frac{1}{D} \sum_{t \in T_{=} (x, y) \cap T_{\circ}(x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2, \quad (3.3)$$

and

$$\begin{aligned} \|\Phi(x) - \Phi(y)\|^2 &\leq \\ (3.4) \quad \frac{1}{D} \sum_{t \in T_{=} (x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2 \\ &\quad + \frac{1}{D} \left[ \sum_{t \in T_{\neq}(x, y)} \|x - y\|^2 + \sum_{t \in T \setminus T_{\circ}(x, y)} \|x - y\|^2 \right]. \end{aligned}$$

We now turn to show that the properties of the embedding hold with positive probability. For  $t \in T$ , let  $\sigma^{(t)}(x, y) = \min\{\sigma^{(t)}(x), \sigma^{(t)}(y)\}$ . Recall that we have applied Lemma 2.4 with  $\xi = \epsilon$ , so that  $\delta = 1 - \epsilon$ .

Consider  $t \in T^{(\delta)}(x)$  then  $B(x, \eta^{(\delta)}) \cdot \Delta(P^{(t)}(x)) \subseteq P^{(t)}(x)$ , where  $\eta^{(\delta)} = 2^{-11}\epsilon/\ln k$ . It follows that  $d(x, X \setminus P^{(t)}(x)) \geq \eta^{(\delta)} \cdot \Delta(P^{(t)}(x)) \geq 2\sigma^{(t)}(x)^{-1}$ , by definition. Similarly, if  $t \in T^{(\delta)}(y)$  then  $d(y, X \setminus P^{(t)}(y)) \geq 2\sigma^{(t)}(y)^{-1}$ . Hence,  $T^{(\delta)}(x) \cap T^{(\delta)}(y) \subseteq T_{\circ}(x, y)$ , implying that  $|T \setminus T_{\circ}(x, y)| \leq |T \setminus (T^{(\delta)}(x) \cap T^{(\delta)}(y))|$ .

$|T^{(\delta)}(y)| \leq |T \setminus T^{(\delta)}(x)| + |T \setminus T^{(\delta)}(y)| \leq 4\epsilon D$ , by Lemma 2.4. Plugging this bound into (3.4) we conclude that:

$$(3.5) \quad \begin{aligned} & \|\Phi(x) - \Phi(y)\|^2 \leq \\ & \frac{1}{D} \cdot |T_{=}(x, y)| \cdot \frac{\sum_{t \in T_{=}(x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2}{|T_{=}(x, y)|} \\ & + \frac{1}{D} \cdot |T_{\neq}(x, y)| \cdot \|x - y\|^2 + 4\epsilon \|x - y\|^2. \end{aligned}$$

Now consider pairs  $x, y$  that are close neighbors, that is:  $\|x - y\| \leq \Delta_k^*(x, y)$  where  $\Delta_k^*(x, y) = c_1 \epsilon / \ln k \cdot \Delta_k(x, y)$ , and  $c_1 = 2^{-19}$ . Note that  $c_1$  is chosen so that  $\Delta_k^*(x, y) \leq \frac{1}{2} \sigma^{(t)}(x, y)^{-1}$  (this follows from Lemma 2.3). Assume w.l.o.g that  $\sigma^{(t)}(x, y) = \sigma^{(t)}(x)$  (otherwise switch the roles of  $x$  and  $y$ ). Consider  $t \in T^{(\delta)}(x)$  then we've seen that  $d(x, X \setminus P(x)) \geq 2\sigma^{(t)}(x)^{-1}$ . Now consider  $y \in X$  such that  $\|x - y\| \leq \Delta_k^*(x, y) \leq \frac{1}{2} \sigma^{(t)}(x)^{-1}$  then  $P^{(t)}(y) = P^{(t)}(x)$ , implying that  $T^{(\delta)}(x) \cap T^{(\delta)}(y) \subseteq T_{=}(x, y) \cap T_{\circ}(x, y)$  implying that  $|T_{=}(x, y) \cap T_{\circ}(x, y)| \geq |T^{(\delta)}(x) \cap T^{(\delta)}(y)| \geq (1 - 4\epsilon)D$ . Plugging this bound into (3.3) yields:

$$(3.6) \quad (1 - 4\epsilon) \cdot \frac{\sum_{t \in T_{=}(x, y) \cap T_{\circ}(x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2}{|T_{=}(x, y) \cap T_{\circ}(x, y)|} \geq \|\Phi(x) - \Phi(y)\|^2$$

We will next apply the Local Lemma again over events related to the Nash-type embeddings in Section 3.1 for the different clusters. Define:

$$\begin{aligned} L(x, y) &= \frac{\sum_{t \in T_{=}(x, y) \cap T_{\circ}(x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2}{|T_{=}(x, y) \cap T_{\circ}(x, y)|}, \\ U(x, y) &= \frac{\sum_{t \in T_{=}(x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2}{|T_{=}(x, y)|}. \end{aligned}$$

We define the following events for pairs. Let  $A_U(x, y)$  be the event that  $U(x, y) > (1 + \epsilon)\|x - y\|^2$ . For pairs  $x, y$  that are close neighbors, that is:  $\|x - y\| \leq \Delta_k^*(x, y)$ . Let  $\sigma_{\mathbf{m}} = \max_{t \in T_{=}(x, y) \cap T_{\circ}(x, y)} \sigma^{(t)}(x, y)$  and  $r = \sqrt{2}/\sigma_{\mathbf{m}}$  be as in Lemma 3.2. Define  $A_L(x, y)$  to be the event that  $L(x, y) > (1 - \epsilon)G_r(\|x - y\|)^2$ . Let  $A(x, y) = A_L(x, y) \vee A_U(x, y)$ . If  $x, y$  are not close neighbors then  $A(x, y) = A_U(x, y)$ . The rest of the argument utilizes the Lovász Local Lemma to prove that there is positive probability that none of the events  $A(x, y)$  occurs.

We create a dependency graph  $G_A$  whose vertices are the events  $A(x, y)$ . Let  $d_{G_A}$  denote its maximum degree. Note that the event  $A(x, y)$  depends only on the random variables associated with clusters  $C \in P^{(t)}$  where  $P^{(t)}(x) = P^{(t)}(y)$ . We place an edge between two events  $A(x, y)$  and  $A(x', y')$  if  $P^{(t)}(x) = P^{(t)}(x')$  for

some  $t \in T_{=}(x, y) \cap T_{=}(x', y')$ . Note that if there is no edge between the two events then they are independent. On the other hand assume if there is an edge then for some  $t$ ,  $P^{(t)}(x) = P^{(t)}(y) = P^{(t)}(x') = P^{(t)}(y')$ . Then  $\max\{\|x - x'\|, \|x - y'\|\} \leq \Delta(P^{(t)}(x)) \leq \Delta_k(x)/16$ , by Lemma 2.3, and hence  $x', y' \in B(x, \Delta_k(x))$ . This implies that the number of such pairs is bounded by  $d_{G_A} \leq \binom{k}{2}$ .

Now, by part (a) of Lemma 3.2 the probability that  $U(x, y) > (1 + \epsilon)\|x - y\|^2$  is at most  $e^{-D(\epsilon^2/4 - \epsilon^3/6)} \leq k^{-2}/4$ . For pairs  $x, y$  that are not close neighbors this implies that the probability that event  $A(x, y)$  occurs is at most  $1/(e(\binom{k}{2} + 1)) \leq 1/(e \cdot d_{G_A} + 1)$ .

Consider now pairs  $x, y$  that are close neighbors:  $\|x - y\| \leq \Delta^*$ . By part (b) of Lemma 3.2 that the probability that  $A_L(x, y)$  holds is at most  $e^{-D\epsilon^2/6} \leq k^{-2}/4$ . Hence the probability the event  $A(x, y)$  occurs is at most  $k^{-2}/2 < 1/(e \cdot d_{G_A} + 1)$ . This completes the proof that the events  $A(x, y)$  satisfy the conditions of the Local Lemma, implying that there is positive probability that none of these events occur. Therefore, by (3.5) we have for any pair  $x, y \in X$ :

$$\begin{aligned} \|\Phi(x) - \Phi(y)\|^2 &\leq \\ & \frac{|T_{=}(x, y)|}{D} \cdot U(x, y) + \frac{|T_{\neq}(x, y)|}{D} \cdot \|x - y\|^2 \\ & + 4\epsilon \|x - y\|^2 \leq (1 + 5\epsilon)\|x - y\|^2, \end{aligned}$$

and by (3.6), for all close neighbors  $x, y$  such that  $\|x - y\| \leq \Delta_k^*(x, y)$  we have:

$$\begin{aligned} \|\Phi(x) - \Phi(y)\|^2 &\geq (1 - 4\epsilon)L(x, y) \\ &\geq (1 - 4\epsilon)(1 - \epsilon)G_r(\|x - y\|)^2 \geq (1 - 5\epsilon)G_r(\|x - y\|)^2. \end{aligned}$$

Using the following property:

CLAIM 3.4.  $G_r(z)^2 \geq (1 - \frac{1}{2}(z/r)^2)z^2$ ,

and  $\sigma_{\mathbf{m}} \leq \max_{t \in T} \sigma^{(t)}(x, y) \leq \Delta_k^*(x, y)^{-1}/2$  or  $r^{-2} \leq \Delta_k^*(x, y)^{-2}/8$ , we obtain:

$$\begin{aligned} \|\Phi(x) - \Phi(y)\|^2 &\geq (1 - 5\epsilon)(1 - \frac{1}{16}\Delta_k^*(x, y)^{-2}\|x - y\|^2)\|x - y\|^2 \\ &\geq (1 - 6 \cdot \max\{\epsilon, \epsilon'(x, y)\})\|x - y\|^2, \end{aligned}$$

where  $\hat{\epsilon}(x, y) = \Delta_k^*(x, y)^{-2}\|x - y\|^2$ . The statement of the theorem is obtained through an appropriate re-scaling of  $\epsilon$ .

**3.3 Embedding Farther Neighbors** In this section, we extend the embedding to cover all pairs such that  $\|x - y\| \leq \frac{1}{2}\Delta_k(x, y)$ . To this end, we add another component to the embedding  $\Psi : X \rightarrow \mathbb{R}^D$ . The embedding  $\Psi$  is based on ideas similar to those of [36, 1]. For

each  $1 \leq t \leq D$ , define a function  $\Psi^{(t)} : X \rightarrow \mathbb{R}^2$  and let  $\{\nu^{(t)}(C) | C \in P^{(t)}, t \in T\}$  be i.i.d symmetric  $\{0, 1\}$ -valued Bernoulli random variables. The embedding is defined for each  $x \in X$  as  $\Psi(x) = \frac{1}{\sqrt{D}} \bigoplus_{1 \leq t \leq D} \Psi^{(t)}(x)$  with  $\Psi^{(t)} = \sqrt{\epsilon} \cdot \nu^{(t)}(P(x)) \cdot d(x, X \setminus P^{(t)}(x))$ . Our final embedding will be  $\hat{\Phi} = \Phi \oplus \Psi$ .

For the analysis of  $\hat{\Phi}$ , first observe that the upper bound on the distance in the embedding is maintained with only small loss. This follows since  $\|\Psi(x) - \Psi(y)\| \leq \sqrt{\epsilon} \|x - y\|$ , as follows by a standard argument (see, e.g., [1]), and we have

$$\begin{aligned} \|\hat{\Phi}(x) - \hat{\Phi}(y)\|^2 &= \|\Phi(x) - \Phi(y)\|^2 + \|\Psi(x) - \Psi(y)\|^2 \\ &\leq (1 + 5\epsilon) \|x - y\|^2 + \epsilon \|x - y\|^2 \\ &= (1 + 6\epsilon) \|x - y\|^2. \end{aligned}$$

We now turn to show that the embedding provides a lower bound on the distance between images of neighbors which are not ‘‘close’’. We can partition the pairs  $x, y$  such that  $\Delta_k^*(x, y) \leq \|x - y\| \leq \frac{1}{2} \Delta_k(x, y)$  into two sets as follows:  $W_- = \{\{x, y\} \mid |T_=(x, y)| \geq D/2\}$  and  $W_{\neq} = \{\{x, y\} \mid |T_{\neq}(x, y)| > D/2\}$ . For pairs in  $W_-$  we show that the  $\Phi$  component of the embedding gives a good lower bound on the distance, whereas for pairs in  $W_{\neq}$  such a contribution is obtained from the  $\Psi$  component of the embedding.

Consider first a pair in  $W_-$ . Recall that

$$\begin{aligned} (3.7) \quad \|\Phi(x) - \Phi(y)\|^2 q &\geq \frac{\sum_{t \in T_=(x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2}{D} \\ &\geq \frac{1}{2} \cdot \frac{\sum_{t \in T_=(x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2}{|T_=(x, y)|}. \end{aligned}$$

Let  $L_B(x, y) = \frac{\sum_{t \in T_=(x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(x)\|^2}{|T_=(x, y)|}$  and define the event  $B(x, y)$  that  $L_B(x, y) < \frac{1}{2} G_r (\|x - y\|)^2$ , where  $\sigma_{\mathbf{m}} = \max_{t \in T_=(x, y)} \sigma^{(t)}(x, y)$  and  $r = \sqrt{2}/\sigma_{\mathbf{m}}$ , as in Lemma 3.2.

As before we create a dependency graph  $G_B$  whose vertices are these events and place an edge between two events  $B(x, y)$  and  $B(x', y')$  if  $P^{(t)}(x) = P^{(t)}(x')$  for some  $t \in T_=(x, y) \cap T_=(x', y')$ . Note that if there is no edge between the two events then they are independent. By the same argument made before we can bound the degree of  $G_B$  as  $d_{G_B} \leq \binom{k}{2}$ .

Now, by part (b) of Lemma 3.2, the probability that  $L_B(x, y) < \frac{1}{2} G_r (\|x - y\|)^2$  is at most  $e^{-D(1/2)^2/6} < k^{-2}/2$ . Hence, the probability that event  $B(x, y)$  occurs is at most  $k^{-2}/2 < 1/(e(\binom{k}{2} + 1)) \leq 1/(e(d_{G_B} + 1))$ , which satisfies the conditions of the Local Lemma,

implying that there is positive probability that none of these event occur.

We make use of the property:

CLAIM 3.5.  $G_r(z) \geq \sqrt{1 - 1/e} \cdot \min\{|z|, r\}$ .

Recall that  $\Delta_k^*(x, y) \leq \frac{1}{2} \sigma^{(t)}(x, y)^{-1}$  for all  $t \in T$ , implying that  $\Delta_k^*(x, y) \leq \frac{1}{2} \sigma_{\mathbf{m}}^{-1}$  and therefore  $r = \sqrt{2} \sigma_{\mathbf{m}}^{-1} \geq 2\sqrt{2} \Delta_k^*(x, y)$ . Since  $\|x - y\| \geq \Delta_k^*(x, y)$  we get that  $G_r(\|x - y\|)^2 \geq (1 - 1/e) \Delta_k^*(x, y)^2$ . We conclude that for every pair  $x, y$  in  $W_-$ ,

$$\begin{aligned} \|\hat{\Phi}(x) - \hat{\Phi}(y)\|^2 &\geq \|\Phi(x) - \Phi(y)\|^2 \\ &\geq \frac{1}{2} L_B(x, y) \geq \frac{1}{4} G_r(\|x - y\|)^2 \geq \frac{1}{4} (1 - 1/e) \Delta_k^*(x, y)^2, \end{aligned}$$

so that  $\|\hat{\Phi}(x) - \hat{\Phi}(y)\| \geq \frac{1}{4} \Delta_k^*(x, y)$ .

Next we deal with pairs in  $W_{\neq}$ . Here we will make use of the  $\Psi$  component of the embedding. By applying Lemma 2.4 with  $\xi = 1/4$  we infer that with positive probability for every  $x \in X$  there exists a set  $T'(x) = T^{(7/8)}(x)$  such that  $|T'(x)| \geq (1 - \frac{2}{8})D = \frac{3}{4}D$  and for each  $t \in T'(x)$ ,  $B(x, \eta^{(3/4)} \Delta(P^{(t)}(x))) \subseteq P^{(t)}(x)$ , and therefore  $d(x, X \setminus P^{(t)}(x)) \geq \sigma^{(t)}(x)^{-1}/(4\epsilon)$ , by definition. We note that this event is positively correlated with the former application of the lemma and so this assertion holds in conjunction with our analysis of  $\Phi$ . Assume w.l.o.g that  $\sigma^{(t)}(x, y) = \sigma^{(t)}(x)$  (otherwise switch the roles of  $x$  and  $y$ ), then we have that:  $\epsilon \cdot d(x, X \setminus P^{(t)}(x)) \geq \Delta_k^*(x, y)$ .

For such a pair  $x, y$  define  $B'(x, y)$  to be the event that  $\|\Psi(x) - \Psi(y)\| < \frac{1}{8} \Delta_k^*(x, y)$ . Define a dependency graph  $G_{B'}$  whose vertices are these events. We place an edge between two events  $B'(x, y)$  and  $B'(x', y')$  if one of  $\{x, y\}$  is in the same cluster as  $\{x', y'\}$  for some  $t \in T$ . Note that if there is no edge between two events then they are independent. On the other hand assume there exists  $t \in T$  such that  $P^{(t)}(x) = P^{(t)}(x')$ . As before we have that  $\|x - x'\| \leq \Delta(P^{(t)}(x)) \leq \Delta_k(x)/16$ , by Lemma 2.3, and hence  $x' \in B(x, \Delta_k(x))$  and therefore there are at most  $k$  such points  $x'$ . Now consider all such pairs including  $x'$ . Denote the other points in these pairs  $y'_1, \dots, y'_s$ . Let  $z$  be the point which maximizes  $\Delta_k(z)$  over all  $y'_j$ s and  $x'$ . Since  $\|x' - y'_j\| < \frac{1}{2} \Delta_k(x', y'_j) = \frac{1}{2} \max\{\Delta_k(x'), \Delta_k(y'_j)\} \leq \frac{1}{2} \Delta_k(z)$ . We conclude that  $\|z - y'_j\| \leq \|z - x'\| + \|x' - y'_j\| < \Delta_k(z)$  and therefore all  $y'_j$ s are in a ball around  $z$  containing fewer than  $k$  points so that  $s < k$ . We conclude that there are at most  $k^2$  such pairs. The same calculation can be made for the case that  $P^{(t)}(y) = P^{(t)}(x')$ , giving a total bound of  $2k^2$  pairs, which provides an upper bound on the degree  $d_{G_{B'}}$  of the dependency graph  $G_{B'}$ .

Now, let  $T''(x) = T'(x) \cap W_{\neq}$  then  $|T''(x)| \geq D/4$ . Then for each  $t \in T''(x)$  with probability at

least  $1/4$ ,  $\nu(P^{(t)}(x)) = 1$  and  $\nu(P^{(t)}(y)) = 0$ , as  $P^{(t)}(x) \neq P^{(t)}(y)$ . Applying a Chernoff bound we have that the probability that there are less than  $1/8$  fraction of the coordinates  $t \in T''(x)$  such that  $|\Psi^{(t)}(x) - \Psi^{(t)}(y)| \geq \sqrt{\epsilon} \cdot d(x, X \setminus P^{(t)}(x)) \geq \Delta_k^*(x, y)$  is at most  $e^{-D/16}$ . But this means that with probability  $1 - e^{-D/16}$ ,  $\|\Psi(x) - \Psi(y)\| \geq \frac{1}{\sqrt{8.4}} \Delta_k^*(x, y) > \frac{1}{8} \Delta_k^*(x, y)$ . Therefore the probability that event  $B'(x, y)$  occurs is at most  $e^{-D/16} \leq k^{-2}/4 < 1/(e(k^2 + 1)) \leq 1/(e(d_{G_{B'}} + 1))$ , satisfying the condition for the Local Lemma. We can therefore conclude that with positive probability none of the events  $B'(x, y)$  occur. Therefore for every  $x, y \in W_{\neq}$  we have:  $\|\hat{\Phi}(x) - \hat{\Phi}(y)\| \geq \|\Psi(x) - \Psi(y)\| \geq \frac{1}{8} \Delta_k^*(x, y)$ , completing the proof of Theorem 3.1.

Finally, for property (c) of Theorem 3.1 note that it follows directly from the definition of  $\Phi$  and Lemma 2.3.

#### 4 Maintaining Separation of Distant Pairs

In many applications it is desirable that not only our distortion for neighbors is small but also that the distant pairs (non-neighbors) will not become too close in the embedding so that the local structure is preserved. If we assume nothing about the metric space  $X$  there is no such low dimensional embedding that will give good guarantees. However, in this section we show that under reasonable assumptions on the local growth structure of the space there exists an embedding that provides reasonable bounds and in particular guarantees that the local structure of the space would be preserved.

To obtain this type of property we can use any non-expansive embedding  $\Upsilon : X \rightarrow \ell_2^D$  that provides guarantees for the distortion of the distant pairs via a similar trick to the one in Section 3.3, i.e., add a component  $\sqrt{\epsilon}\Upsilon$  to the embedding  $\hat{\Phi}$ . Let  $\bar{\Phi} = \hat{\Phi} \oplus (\sqrt{\epsilon}\Upsilon)$  then:

$$\begin{aligned} \|\bar{\Phi}(x) - \bar{\Phi}(y)\|^2 &= \\ \|\hat{\Phi}(x) - \hat{\Phi}(y)\|^2 + \epsilon\|\Upsilon(x) - \Upsilon(y)\|^2 &\leq \\ (1 + \epsilon)\|x - y\|^2 + \epsilon\|x - y\|^2 &= (1 + 2\epsilon)\|x - y\|^2, \end{aligned}$$

whereas the lower bound for neighbors given by  $\hat{\Phi}$  still holds and the lower bound for far neighbors is given by  $\Upsilon$  with just an additional  $\sqrt{\epsilon}$  factor loss.

In recent work [3] it is shown that every metric space embeds in  $\ell_2^D$  where  $D = O(\dim(X)/\theta)$  with distortion  $O(\log^{1+\theta} n)$ , where  $\dim(X)$  is the doubling dimension of  $X$ . Hence a possible choice for the component  $\Upsilon$  could be this embedding, and combining it with  $\hat{\Phi}$  as described above, we obtain a global embedding in dimension  $O(\epsilon^{-2} \log k + \theta^{-1} \dim(X))$  that guarantees that the distance between distant pairs does not shrink below  $\Delta_k(x, y) \cdot \Omega(1/\log^{1+\theta} n)$ . However, as this bound depends on the global size of the set this still does

not promise full preservation of the local structure. To overcome this we give a refinement of this embedding using ideas from [2].<sup>3</sup>

In Appendix C we give a local scaling embedding for doubling metrics satisfying the weak growth rate condition<sup>4</sup>. By using this embedding for the component  $\Upsilon$  as explained above we obtain the following theorem:

**THEOREM 4.1.** *Let  $k \in \mathbb{N}$ , and  $X$  a discrete subset of  $\mathbb{R}^U$ . Suppose that  $X$  satisfies a weak growth rate condition then for any  $0 < \epsilon, \theta \leq 1$  there exists an embedding  $\bar{\Phi} : X \rightarrow \mathbb{R}^D$ , where  $D = O(\log k/\epsilon^2 + \dim(X)/\theta)$  such that Theorem 3.1 holds, and additionally if  $\|x - y\| \geq \frac{1}{2} \Delta_k(x, y)$  then:*

$$(4.8) \quad \|\bar{\Phi}(x) - \bar{\Phi}(y)\| \geq \Delta_k(x, y) \cdot c_2 \theta \sqrt{\epsilon} / \log^{1+\theta} k,$$

for some universal constant  $c_2$ .

#### 5 Dimension Reduction for Euclidean Snowflakes

In this section we provide a dimension reduction for snowflakes of finite subsets of Euclidean space.

**THEOREM 5.1.** *Given a subset  $X$  of Euclidean space, for every  $0 < \alpha < 1$  and  $\epsilon > 0$  there exists an embedding  $\Phi : X \rightarrow \mathbb{R}^D$ , where  $D = O(\frac{\log(1/\epsilon)}{\alpha(1-\alpha)} \epsilon^{-3} \dim(X) (\log(\dim(X)) + \log(1/\epsilon)))$  such that for all  $x, y \in X$ :*

$$(1 + \epsilon)^{-1} \|x - y\|^\alpha \leq \|\Phi(x) - \Phi(y)\| \leq (1 + \epsilon) \|x - y\|^\alpha$$

The proof follows the methodology of [20]. The first step in their work is to obtain a special single scale embedding (Theorem 3.1 of [20]). They then show, via a delicate Assouad-type argument [6], that such an embedding implies that there exists an embedding which preserves small distortion for snowflakes in all scales simultaneously. We will replace Theorem 3.1 of [20] by a lemma which obtains a single scale embedding with the same properties but with lower dimension.

We do this by first providing a simple variant of Theorem 3.1. We then show that this theorem can be used to obtain the required lemma with an improved dimension of  $\tilde{O}(\epsilon^{-2} \dim(X))$ .

Recall the definition of the Gaussian transform:  $G_r(z) = r(1 - \exp(-z^2/r^2))^{1/2}$ .

<sup>3</sup>Note that an alternate choice for  $\Upsilon$  could be our snowflake embedding of Section 5, which would provide lower bound on the contraction of distant pairs which is a function of their distance. However, we prefer a function of  $k$ .

<sup>4</sup> $X$  satisfies a weak growth rate (cf. [2]):  $\text{WGR}(\gamma)$  for some constant  $\gamma < 1$  if for every  $x \in X$  and  $r_1, r_2 > 0$ ,  $|B(x, r_2)| \leq |B(x, r_1)|^{(\gamma r_2/r_1)^\gamma}$ , and further assume  $\gamma < 0.2$ .

**THEOREM 5.2.** *Given  $X$  a discrete subset of  $\mathfrak{R}^U$ , let  $k \in \mathbb{N}$  and  $\Delta > 0$  such that for each  $x \in X$ :  $\Delta \leq \Delta_k(x)$ . Let  $\Delta^* = c_1 \epsilon \Delta / \log k$ , where  $c_1 < 1$  is a universal constant. Let  $0 < \tau < 1$  and let  $r = \tau \Delta^*$ . Then for any  $\epsilon > 0$  there exists an embedding  $\hat{\Phi} : X \rightarrow \mathfrak{R}^D$ , where  $D = O(\log k / \epsilon^2)$  with the following properties:*

- a. For all  $x, y \in X$ ,  $\|\hat{\Phi}(x) - \hat{\Phi}(y)\| \leq (1 + \epsilon)\|x - y\|$ .
- b. For all  $x, y \in X$  such that  $\|x - y\| \leq \Delta^*$ :

$$(1 + \epsilon)^{-1} \leq \frac{\|\bar{\Phi}(x) - \bar{\Phi}(y)\|}{G_r(\|x - y\|)} \leq (1 + \epsilon).$$

- c. For all  $x \in X$ ,  $\|\hat{\Phi}(x)\| \leq r$ .

The proof of [Theorem 5.2](#) is given in [Appendix D](#). The theorem provides the following lemma which improves [Theorem 3.1](#) of [\[20\]](#):

**LEMMA 5.1.** *Given a subset  $X$  of Euclidean space, for every  $r > 0$ ,  $\epsilon > 0$  and  $0 < \tau < 1$ , there exists an embedding  $\bar{\Phi} : X \rightarrow \mathfrak{R}^D$ , where  $D = O(\epsilon^{-2} \dim(X)(\log(\dim(X)) + \log((\epsilon\tau)^{-1})))$ , with the following properties:*

1.  $\|\bar{\Phi}(x) - \bar{\Phi}(y)\| \leq \|x - y\|$ .
2. For all  $x, y \in X$  such that  $\tau r \leq \|x - y\| \leq r/\tau$ :

$$(1 + \epsilon)^{-1} \leq \frac{\|\bar{\Phi}(x) - \bar{\Phi}(y)\|}{G_r(\|x - y\|)} \leq (1 + \epsilon).$$

3. For all  $x \in X$ ,  $\|\bar{\Phi}(x)\| \leq r$ .

*Proof.* Let  $\hat{X}$  be an  $\epsilon\tau r$ -net of  $X$ . We show the theorem holds for  $\hat{X}$ . As in [\[20\]](#) claim (1) of the theorem can be easily obtained by using Kirszbraun's extension theorem<sup>5</sup>, and observing that if  $x, y \in X$  are such that  $\tau r \leq \|x - y\| \leq r/\tau$  then there exist  $x', y' \in \hat{X}$  such that  $\tau(1 - 2\epsilon)r \leq \|x' - y'\| \leq r/\tau(1 + 2\epsilon)$  and a small adaptation of the parameters provides the statement in the theorem.

Let  $k = 2^{c' \dim(X)(\log(\dim(X)) + \log((\epsilon\tau)^{-1}))}$ , where  $c'$  is an appropriate constant to be determined, and let  $\Delta = \log k / (c_1 \epsilon) \cdot r/\tau$ . Let  $x$  be an arbitrary point  $x \in \hat{X}$  then  $|B_{\hat{X}}(x, \Delta)| \leq 2^{\dim(X) \log(\Delta / (\epsilon\tau r))} \leq 2^{\dim(X) \log(\log k / (c_1 \epsilon^3 \tau^2))} < k$  (for an appropriate choice of  $c'$ ) and therefore for all  $x \in \hat{X}$ ,  $\Delta_k(x) > \Delta$ . The lemma now follows from [Theorem 5.2](#).

<sup>5</sup>Kirszbraun's theorem is not really necessary as an extension property can be shown to hold directly for the embedding in [Theorem 3.1](#).

## References

- [1] I. Abraham, Y. Bartal, and O. Neiman. Advances in metric embedding theory. In *Proceedings of the thirty-eighth annual ACM Symposium on Theory of Computing*, pages 271–286, New York, NY, USA, 2006. ACM Press.
- [2] I. Abraham, Y. Bartal, and O. Neiman. Local embedding of metric spaces. In *Proceedings of the thirtieth annual ACM Symposium on Theory of Computing*, pages 631–640, 2007.
- [3] I. Abraham, Y. Bartal, and O. Neiman. Embedding metric spaces in their intrinsic dimension. In *Proceedings of the 19th ACM-SIAM Symp. on Discrete Algorithms*, pages 363–372, 2008.
- [4] I. Abraham, Y. Bartal, and O. Neiman. On low dimensional local embeddings. In *Proceedings of the 20th ACM-SIAM Symp. on Discrete Algorithms*, pages 875–884, 2009.
- [5] N. Alon. Perturbed identity matrices have high rank: proof and applications. *Combinatorics, Probability and Computing*, 18:3–15, 2009.
- [6] P. Assouad. Plongements lipschitziens dans  $\mathbb{R}^n$ . *Bull. Soc. Math. France*, 111(4):429–448, 1983.
- [7] M. Balasubramanian and E. L. Schwartz. The Isomap algorithm and topological stability. *Science*, 295(5552), 2002.
- [8] R. G. Baraniuk and M. B. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, February 2009.
- [9] Y. Bartal. Probabilistic approximation of metric spaces and its algorithmic applications. In *37th Annual Symposium on Foundations of Computer Science (Burlington, VT, 1996)*, pages 184–193. IEEE Comput. Soc. Press, Los Alamitos, CA, 1996.
- [10] Y. Bartal, B. Recht, and L. J. Schulman. A Nash-type dimensionality reduction for discrete subsets of  $L_2$ . Manuscript, 2007.
- [11] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
- [12] J. Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel J. Math.*, 52(1-2):46–52, 1985.
- [13] M. Brand. Charting a manifold. In *Neural Information Processing Systems (NIPS)*, 2002.
- [14] P. Das, M. Moll, H. Stamati, L. E. Kavradi, and C. Clementi. Low-dimensional, free-energy landscapes of protein folding reactions by nonlinear dimensionality reduction. *Proceedings of the National Academy of Science*, 103(26):9885–9890, 2006.
- [15] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2003.
- [16] D. L. Donoho and C. Grimes. Hessian eigenmaps: new locally linear embedding techniques for high dimensional data. Technical report, TR2003-08, Dept. of Statistics, Stanford University, 2003.

- [17] P. Erdős and L. Lovász. Problems and results on 3-chromatic hypergraphs and some related questions. In A. Hajnal et. al, editor, *Infinite and Finite Sets*, volume 11. Colloquia Mathematica Societas János Bolyai, North Holland, Amsterdam, 1975.
- [18] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*, volume 159 of *The Springer International Series in Engineering and Computer Science*. Springer, 1992.
- [19] Y. Gordon. On Milman’s inequality and random subspaces which escape through a mesh in  $\mathbb{R}^n$ . In *Geom. Aspects of Funct. Anal., Israel seminar, Lecture Notes in Mathematics 1317*, pages 84–106. Springer-Verlag, 1988.
- [20] L. Gottlieb and R. Krauthgamer. A nonlinear approach to dimension reduction. In *SODA ’11: Proceedings of the 24th annual ACM-SIAM Symposium on Discrete Algorithms*, 2011. arXiv:0907.5477v1.
- [21] A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *FOCS ’03: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, page 534, Washington, DC, USA, 2003. IEEE Computer Society.
- [22] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang. Face recognition using Laplacianfaces. *IEEE. Trans. Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.
- [23] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
- [24] O. C. Jenkins and M. J. Mataric. Deriving action and behavior primitives from human motion data. In *IEEE/RSK International Conference on Intelligent Robots and Systems*, 2002.
- [25] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984.
- [26] B. Klartag and S. Mendelson. Empirical processes and random projections. *J. Functional Analysis*, 225(1):229–245, 2005.
- [27] R. Krauthgamer and J. R. Lee. The intrinsic dimensionality of graphs. In *Annual ACM Symposium on Theory of Computing*, pages 438–447, 2003.
- [28] U. Lang and C. Plaut. Bilipschitz embeddings of metric spaces into space forms. *Geom. Dedicata*, 87(1-3):285–307, 2001.
- [29] J. R. Lee. On distance scales, embeddings, and efficient relaxations of the cut cone. In *SODA ’05: Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms*, pages 92–101, 2005.
- [30] M. Mendel and A. Naor. Euclidean quotients of finite metric spaces. *Advances in Mathematics*, 189(2):451–494, 2004.
- [31] R. A. Moser and G. Tardos. A constructive proof of the general lovász local lemma. *Journal of the ACM*, 57(2), 2010.
- [32] J. Nash.  $C^1$  isometric embeddings. *The Annals of Mathematics*, 60(3):383–396, 1954.
- [33] J. Nilsson, T. Fioretos, M. Hoglund, and M. Fontes. Approximate geodesic distances reveal biologically relevant structures in microarray data. *Bioinformatics*, 20(6):874–880, 2004.
- [34] A. Rahimi and B. Recht. Estimating observation functions in dynamical systems using unsupervised regression. In *Neural Information Processing Systems*, 2006.
- [35] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Neural Information Processing Systems*, 2007.
- [36] S. Rao. Small distortion and volume preserving embeddings for planar and Euclidean metrics. In *Proceedings of the Fifteenth Annual Symposium on Computational Geometry*, pages 300–306, New York, 1999. ACM.
- [37] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
- [38] G. Schechtman and A. Shraibman. Lower bounds for local versions of dimension reductions, 2007. Tech Report.
- [39] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.
- [40] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- [41] M. Talagrand. *The generic chaining*. Springer, 2005.
- [42] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- [43] M. Thorup and U. Zwick. Approximate distance oracles. In *33<sup>rd</sup> Annual ACM Symposium on Theory of Computing*, pages 183–192, Hersonissos, Crete, Greece, July 2001.
- [44] K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.
- [45] H. Whitney. Differentiable manifolds. *Annals of Mathematics*, 31:645–680, 1936.
- [46] H. Whitney. Self-intersection of a smooth  $n$ -manifold in  $2n$ -space. *Annals of Mathematics*, 45:220–246, 1944.

## A Randomized Nash Device Analysis

In this section we prove [Lemma 3.2](#).

**Part (a):** To prove part (a) of the lemma note that  $1 - \cos(\alpha) \leq \alpha^2/2$  for all  $\alpha$ . Let  $\ell = \|x - y\|$ .  $\tau_i := \omega_i'(x - y)$  is distributed as a one-dimensional Gaussian distribution  $\mathcal{N}(0, \ell^2)$  and  $\tau_1, \dots, \tau_D$  are independent

and we have

$$\begin{aligned}
\text{(A.1)} \quad & \|\Theta(x) - \Theta(y)\|^2 \\
&= \frac{1}{D} \sum_{t=1}^D \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2 \\
&= \frac{1}{D} \sum_{t=1}^D \frac{2}{\sigma_t^2} (1 - \cos(\sigma_t \tau_t)) \leq \frac{1}{D} \sum_{t=1}^D \tau_t^2.
\end{aligned}$$

It therefore follows that

$$\begin{aligned}
\text{(A.2)} \quad & \Pr [\|\Theta(x) - \Theta(y)\|^2 \geq (1 + \epsilon)\ell^2] \\
&\leq \Pr \left[ \frac{1}{D} \sum_{t=1}^D \tau_t^2 \geq (1 + \epsilon)\ell^2 \right] \\
&\leq e^{-\frac{D}{2} \left( \frac{\epsilon^2}{2} - \frac{\epsilon^3}{3} \right)}
\end{aligned}$$

where the second inequality is a well known concentration inequality a  $\chi$ -squared random variable (see, e.g., [15]).

Parts (b) and (c) require a more detailed verification, but follow from a Chernoff Bound type analysis.

**Part (b):** We explicitly bound the moment generating function of the everywhere non-positive process  $\cos(\sigma\omega'(x - y)) - 1$  by using the upper bound  $\exp(\alpha) \leq 1 + \alpha + \alpha^2/2$  for all  $\alpha \leq 0$ . Using this upper bound allows us to bound  $\mathbb{E}_\omega[s(\cos(\sigma\omega'(x - y)) - 1)]$  by employing Claim 3.1.

Using the identity  $\|\Theta(x) - \Theta(y)\|^2 = \frac{1}{D} \sum_{t=1}^D \frac{2}{\sigma_t^2} (1 - \cos(\sigma_t \tau_t))$  we have for any  $u > 0$

$$\begin{aligned}
\text{(A.3)} \quad & \mathbf{P} [\|\Theta(x) - \Theta(y)\|^2 \leq u] \\
&= \mathbf{P} \left[ \frac{1}{D} \sum_{t=1}^D \frac{2}{\sigma_t^2} (1 - \cos(\sigma_t \tau_t)) \leq u \right] \\
&= \mathbf{P} \left[ \sum_{t=1}^D \frac{2}{\sigma_t^2} (\cos(\sigma_t \tau_t) - 1) + uD \geq 0 \right] \\
&= \mathbf{P} \left[ \exp \left( s \sum_{t=1}^D \frac{2}{\sigma_t^2} (\cos(\sigma_t \tau_t) - 1) + uDs \right) \geq 1 \right] \quad (\forall s > 0) \\
&\leq \mathbb{E} \left[ \exp \left( s \sum_{t=1}^D \frac{2}{\sigma_t^2} (\cos(\sigma_t \tau_t) - 1) + uDs \right) \right] \quad (\text{Markov}) \\
&= \exp(uDs) \prod_{t=1}^D \mathbb{E}_{\tau_t} \left[ \exp \left( s \frac{2}{\sigma_t^2} (\cos(\sigma_t \tau_t) - 1) \right) \right].
\end{aligned}$$

We first bound the expectations with respect to  $\tau_t$ . Let  $\tau$  be a zero-mean Gaussian random variable with variance  $\ell^2$ . Since  $\exp(t) \leq 1 + t + t^2/2$  for all  $t \leq 0$ , we

have, for all  $s, \sigma > 0$ ,

$$\begin{aligned}
\text{(A.4)} \quad & \exp \left( s \frac{2}{\sigma^2} (\cos(\sigma\tau) - 1) \right) \\
&\leq 1 + \frac{2}{\sigma^2} (\cos(\sigma\tau) - 1)s + \frac{2}{\sigma^4} [1 - 2\cos(\sigma\tau) + \cos^2(\sigma\tau)]s^2 \\
&= 1 + \frac{2}{\sigma^2} (\cos(\sigma\tau) - 1)s + \frac{1}{\sigma^4} [3 - 4\cos(\sigma\tau) + \cos(2\sigma\tau)]s^2.
\end{aligned}$$

Using the fact that  $\mathbb{E}[\cos(z\tau)] = \exp(-\ell^2 z^2/2)$  for all  $z \in \mathfrak{R}$ , we can compute the expectation of (A.4)

$$\begin{aligned}
\text{(A.5)} \quad & \mathbb{E} [\exp(s \frac{2}{\sigma^2} [\cos(\sigma\tau) - 1])] \\
&\leq 1 + \frac{2}{\sigma^2} (\exp(-\frac{1}{2}\sigma^2\ell^2) - 1) s \\
&\quad + \frac{1}{\sigma^4} (3 - 4\exp(-\frac{1}{2}\sigma^2\ell^2) + \exp(-2\sigma^2\ell^2)) s^2.
\end{aligned}$$

The negative of the term linear in  $s$  is equal

$$\text{(A.6)} \quad b(\sigma) := \frac{2}{\sigma^2} (1 - \exp(-\frac{1}{2}\sigma^2\ell^2))$$

and the term quadratic in  $s$  is equal to

$$\text{(A.7)} \quad a(\sigma) := \frac{1}{4} b(\sigma)^2 ((1 + \exp(-\frac{1}{2}\sigma^2\ell^2))^2 + 2).$$

Both  $b(\sigma)$  and  $a(\sigma)$  are positive decreasing functions of  $\sigma > 0$ .

To complete the proof, suppose we can find an  $s_0 > 0$  such that

$$\text{(A.8)} \quad b(\sigma_t)s_0 - a(\sigma_t)s_0^2 - us_0 \geq \gamma \quad \text{for all } 1 \leq t \leq D.$$

for some constant  $\gamma > 0$ . Then, using the inequality  $1 + t \leq e^t$  and the preceding analysis, we would have the probability of  $\|\Theta(x) - \Theta(y)\|^2 \leq u$  being at most

$$\begin{aligned}
\text{(A.9)} \quad & \exp(uDs_0) \prod_{t=1}^D (1 - b(\sigma_t)s_0 + a(\sigma_t)s_0^2) \\
&\leq \exp \left( \sum_{t=1}^D (us_0 - b(\sigma_t)s_0 + a(\sigma_t)s_0^2) \right) \\
&\leq \exp(-\gamma D).
\end{aligned}$$

Part (b) would be proven if we find an  $s_0$  for which (A.8) holds with  $u = (1 - \epsilon)G_r(\ell)^2 = (1 - \epsilon)b(\sigma_{\mathbf{m}})$  and  $\gamma = \frac{1}{6}\epsilon^2$ .

We show that choosing  $s_0$  such that the equality is attained in (A.8) when  $\sigma = \sigma_{\mathbf{m}}$ . That is, we set

$$\text{(A.10)} \quad s_0 = \frac{b(\sigma_{\mathbf{m}}) - u - \sqrt{(b(\sigma_{\mathbf{m}}) - u)^2 - 4a(\sigma_{\mathbf{m}})\gamma}}{2a(\sigma_{\mathbf{m}})}.$$

If this choice of  $s_0$  is positive, then (A.8) is automatically satisfied. To see that note that  $a$  and  $b$  are both decreasing functions of  $\sigma$  so we have

$$\begin{aligned}
\text{(A.11)} \quad & (b(\sigma_t) - u)s_0 - a(\sigma_t)s_0^2 \\
&\geq (b(\sigma_{\mathbf{m}}) - u)s_0 - a(\sigma_t)s_0^2 \\
&= \gamma + a(\sigma_{\mathbf{m}})s_0^2 - a(\sigma_t)s_0^2 \geq \gamma
\end{aligned}$$

All that remains is to verify that  $s_0$  is positive for the values of  $u$  and  $\gamma$  defined above. Note that  $s_0$  is positive as long as it is well define, i.e. if  $b(\sigma_{\mathbf{m}}) - u \geq 2\sqrt{\gamma a(\sigma_{\mathbf{m}})}$ .

Recall that  $u = (1 - \epsilon)b(\sigma_{\mathbf{m}})$  and  $\gamma = \frac{1}{6}\epsilon^2$ . Rearranging terms, we must show  $\epsilon b(\sigma_{\mathbf{m}}) \geq 2\epsilon\sqrt{\frac{1}{6}a(\sigma_{\mathbf{m}})}$  or that  $b(\sigma)^2 \geq 4\frac{1}{6}a(\sigma_{\mathbf{m}})$ , which clearly holds using the bound  $(1 + \exp(-\frac{1}{2}\sigma_{\mathbf{m}}^2\ell^2))^2 + 2 \leq 6$ .

**Part (c):** We distinguish two cases. Assume first that  $\ell = \|x - y\| \leq \sqrt{\epsilon/2r}$ . In this case  $G_r(\ell)^2 \geq (1 - \epsilon/4)\ell^2$  (using Claim 3.4). By part (a) of Lemma 3.2 we have that with probability exceeding  $1 - e^{-(\epsilon/2)^2 D(1/4-1/24)} \geq 1 - e^{-\epsilon^2 D/24}$ ,

$$(A.12) \quad \|\Theta(x) - \Theta(y)\|^2 \leq (1 + \epsilon/2)\ell^2 \leq \frac{1 + \epsilon/2}{1 - \epsilon/4} G_r(\ell)^2 \leq (1 + \epsilon)G_r(\ell)^2.$$

Now, consider the case that  $\ell > \sqrt{\epsilon/2r}$ . In this case we follow an argument very similar to that for part (b) above. We have for any  $u > 0$

$$(A.13) \quad \mathbf{P} [\|\Theta(x) - \Theta(y)\|^2 \geq u] \leq \exp(-uD_s) \prod_{t=1}^D \mathbb{E}_{\tau_t} \left[ \exp \left( s \frac{2}{\sigma_t^2} (1 - \cos(\sigma_t \tau_t)) \right) \right].$$

We first bound the expectations with respect to  $\tau_t$ . Let  $\tau$  be a zero-mean Gaussian random variable with variance  $\ell^2$ . We will require that  $0 < s \leq \sigma_s^2/2$ , ensuring that for all  $t \leq D$ ,  $s \frac{2}{\sigma_t^2} (1 - \cos(\sigma_t \tau_t)) \leq 2$ . Since  $\exp(t) \leq 1 + t + 2t^2$  for all  $t \leq 2$ , we have for all  $\sigma$ ,

$$(A.14) \quad \mathbb{E} [\exp(s \frac{2}{\sigma^2} [1 - \cos(\sigma \tau)])] \leq 1 + \frac{2}{\sigma^2} (1 - \exp(-\frac{1}{2}\sigma^2\ell^2)) s + \frac{4}{\sigma^4} (3 - 4 \exp(-\frac{1}{2}\sigma^2\ell^2) + \exp(-2\sigma^2\ell^2)) s^2.$$

As before define:

$$(A.15) \quad b(\sigma) := \frac{2}{\sigma^2} (1 - \exp(-\frac{1}{2}\sigma^2\ell^2))$$

$$(A.16) \quad a(\sigma) := b(\sigma)^2 ((1 + \exp(-\frac{1}{2}\sigma^2\ell^2))^2 + 2).$$

To complete the proof, suppose we can find an  $s_0 > 0$  such that

$$(A.17) \quad s_0 \leq \sigma_s^2/2.$$

$$(A.18) \quad -b(\sigma_t)s_0 - a(\sigma_t)s_0^2 + us_0 \geq \gamma \quad \text{for all } 1 \leq t \leq D.$$

for some constant  $\gamma > 0$ . Then, using the inequality  $1 + t \leq e^t$  and the preceding analysis, we would have the probability of  $\|\Theta(x) - \Theta(y)\|^2 \geq u$  being at most

$$(A.19) \quad \exp(-uD_{s_0}) \prod_{t=1}^D (1 + b(\sigma_t)s_0 + a(\sigma_t)s_0^2) \leq \exp \left( \sum_{t=1}^D (-us_0 + b(\sigma_t)s_0 + a(\sigma_t)s_0^2) \right) \leq \exp(-\gamma D).$$

Hence, we need to find an  $s_0$  for which (A.17) and (A.18) holds with  $u = (1 + \epsilon)G_r(\ell)^2 = (1 + \epsilon)b(\sigma_s)$  and  $\gamma = \frac{1}{24}\epsilon^2$ .

We show that choosing  $s_0$  such that the equality is attained in (A.8). That is, we set

$$(A.20) \quad s_0 = \frac{u - b(\sigma_s) - \sqrt{(b(\sigma_s) - u)^2 - 4a(\sigma_s)\gamma}}{2a(\sigma_s)}.$$

If this choice of  $s_0$  is positive, then (A.17) and (A.18) are automatically satisfied. For (A.17), note that since  $\ell^2 > \frac{\epsilon}{2}r^2 = \epsilon \frac{1}{\sigma_s^2}$  then

$$b(\sigma_s) \geq \frac{2}{\sigma_s^2} (1 - \exp(-\frac{1}{2}\sigma_s^2\ell^2)) \geq \frac{\epsilon}{2\sigma_s^2} \\ s_0 \leq \frac{\epsilon b(\sigma_s)}{2a(\sigma_s)} \leq \frac{\epsilon}{6b(\sigma_s)} < \sigma_s^2/2.$$

For (A.18), note that  $a$  and  $b$  are both decreasing functions of  $\sigma$  so we have

$$(A.21) \quad (u - b(\sigma_t))s_0 - a(\sigma_t)s_0^2 \geq (u - b(\sigma_s))s_0 - a(\sigma_s)s_0^2 = \gamma$$

All that remains is to verify that  $s_0$  is positive for the values of  $u$  and  $\gamma$  defined above. Note that  $s_0$  is positive as long as it is well define, i.e. if  $u - b(\sigma_{\mathbf{m}}) \geq 2\sqrt{\gamma a(\sigma_{\mathbf{m}})}$ .

Recall that  $u = (1 + \epsilon)b(\sigma_{\mathbf{m}})$  and  $\gamma = \frac{1}{24}\epsilon^2$ . Rearranging terms, we must show  $\epsilon b(\sigma_{\mathbf{m}}) \geq 2\epsilon\sqrt{\frac{1}{24}a(\sigma_{\mathbf{m}})}$  or that  $b(\sigma)^2 \geq 4\frac{1}{24}a(\sigma_{\mathbf{m}})$ , which clearly holds using the bound  $(1 + \exp(-\frac{1}{2}\sigma_{\mathbf{m}}^2\ell^2))^2 + 2 \leq 6$ .

## B Probabilistic Partitions Preliminaries

**B.1 Preliminaries** Consider a finite metric space  $(X, d)$  and let  $n = |X|$ . The *diameter* of  $X$  is denoted  $\text{diam}(X) = \max_{x,y \in X} d(x, y)$ . For a point  $x$  and  $r \geq 0$ , the ball at radius  $r$  around  $x$  is defined as  $B_X(x, r) = \{z \in X | d(x, z) \leq r\}$ . We omit the subscript  $X$  when it is clear from the context.

The following definitions are used in the context of partition-based embeddings into  $L_p$ :

DEFINITION B.1. The local growth rate of  $x \in X$  at radius  $r > 0$  for a given scale  $\gamma > 0$  is defined as

$$\rho(x, r, \gamma) = |B(x, r\gamma)|/|B(x, r/\gamma)|.$$

Given a subspace  $Z \subseteq X$ , the minimum local growth rate of  $Z$  at radius  $r > 0$  and scale  $\gamma > 0$  is defined as  $\rho(Z, r, \gamma) = \min_{x \in Z} \rho(x, r, \gamma)$ . The minimum local growth rate at radius  $r > 0$  and scale  $\gamma > 0$  is defined as  $\bar{\rho}(x, r, \gamma) = \rho(B(x, r), r, \gamma)$ .

The following simple fact about minimum local growth rate is useful:

CLAIM B.2. Let  $x, y \in X$ , let  $\gamma > 0$  and let  $r$  be such that  $2(1 + 1/\gamma)r < d(x, y) \leq (\gamma - 2 - 1/\gamma)r$ , then

$$\max\{\bar{\rho}(x, r, \gamma), \bar{\rho}(y, r, \gamma)\} \geq 2.$$

## B.2 Uniformly Padded Probabilistic Partitions

We start with describing the basic definition that captures the properties needed for the application for embeddings:

DEFINITION B.3. (PARTITION) Let  $(X, d)$  be a finite metric space. A partition  $P$  of  $X$  is a collection of disjoint set of clusters  $\mathcal{C}(P) = \{C_1, C_2, \dots, C_t\}$  such that  $X = \cup_j C_j$ . The sets  $C_j$  are called clusters. For  $x \in X$  we denote by  $P(x)$  the cluster containing  $x$ . Given  $\Delta > 0$ , a partition is  $\Delta$ -bounded if for all  $1 \leq j \leq t$ ,  $\text{diam}(C_j) \leq \Delta$ .

DEFINITION B.4. (UNIFORM FUNCTION) Given a partition  $P$  of a metric space  $(X, d)$ , a function  $f$  defined on  $X$  is called uniform with respect to  $P$  if for any  $x, y \in X$  such that  $P(x) = P(y)$  we have  $f(x) = f(y)$ .

DEFINITION B.5. (PROBABILISTIC PARTITION) A probabilistic partition  $\hat{\mathcal{P}}$  of a finite metric space  $(X, d)$  is a distribution over a set  $\mathcal{P}$  of partitions of  $X$ . Given  $\Delta > 0$ ,  $\hat{\mathcal{P}}$  is  $\Delta$ -bounded if each  $P \in \mathcal{P}$  is  $\Delta$ -bounded.

DEFINITION B.6. (UNIFORMLY PADDED LOCAL PP) Given  $\Delta > 0$  and  $0 < \delta \leq 1$ , let  $\hat{\mathcal{P}}$  be a  $\Delta$ -bounded probabilistic partition of  $(X, d)$ . Given collection of functions  $\eta = \{\eta_P : X \rightarrow [0, 1] | P \in \mathcal{P}\}$  such that  $\eta_P$  is a uniform function with respect to  $P$ . We say that  $\hat{\mathcal{P}}$  is a  $(\eta, \delta)$ -uniformly padded local probabilistic partition if the event  $B(x, \eta_P(x)\Delta) \subseteq P(x)$  occurs with probability at least  $\delta$  and is independent of the structure of the partition outside  $B(x, 2\Delta)$ .

Formally for all  $C \subseteq X \setminus B(x, 2\Delta)$  and all partitions  $P'$  of  $C$ ,

$$\Pr[B(x, \eta_P(x)\Delta) \subseteq P(x) | P|_C = P'] \geq \delta$$

## B.3 Local Uniform Padding Lemma for Doubling Metrics

LEMMA B.7. (LOCAL UNIFORM PADDING LEMMA)

Let  $(X, d)$  be a  $\lambda$ -doubling finite metric space. Let  $0 < \Delta \leq \text{diam}(X)$ . Let  $\hat{\delta} \in (\lambda^{-2}, 1/2]$ , and let  $\Gamma = 64$ . There exists a  $\Delta$ -bounded probabilistic partition  $\hat{\mathcal{P}}$  of  $(X, d)$  and a collection of uniform functions  $\{\xi_P : X \rightarrow \{0, 1\} | P \in \mathcal{P}\}$  and  $\{\eta_P : X \rightarrow (0, 1/\ln(1/\hat{\delta})) | P \in \mathcal{P}\}$  such that for any  $\hat{\delta} \leq \delta \leq 1$ , and  $\eta^{(\delta)}$  defined by  $\eta_P^{(\delta)}(x) = \eta_P(x) \ln(1/\delta)$ , the probabilistic partition  $\hat{\mathcal{P}}$  is a  $(\eta^{(\delta)}, \delta)$ -uniformly padded local probabilistic partition; and the following conditions hold for any  $P \in \mathcal{P}$  and any  $x \in X$ :

- $\eta_P(x) \geq 2^{-9}/(\ln \lambda)$ .
- If  $\xi_P(x) = 1$  then:  $2^{-7}/\ln \rho(x, 4\Delta, \Gamma) \leq \eta_P(x) \leq 2^{-7}/\ln(1/\hat{\delta})$ .
- If  $\xi_P(x) = 0$  then:  $\eta_P(x) = 2^{-7}/\ln(1/\hat{\delta})$  and  $\bar{\rho}(x, 4\Delta, \Gamma) < 1/\hat{\delta}$ .

## C Embedding Distant Pairs

THEOREM 4.1 follows from the following theorem on local scaling embedding for doubling metrics.

Recall that  $X$  satisfies a weak growth rate condition (cf. [2]):  $\text{WGR}(\gamma)$  for some constant  $\gamma < 1$  if for every  $x \in X$  and  $r_1, r_2 > 0$ ,  $|B(x, r_2)| \leq |B(x, r_1)|^{(r_2/r_1)^\gamma}$ , and further assume  $\gamma < 0.2$ .

THEOREM C.1. Given a metric space  $(X, d)$  satisfying  $\text{WGR}(\gamma)$ . For any  $1 \leq p \leq \infty$ , and  $0 < \theta \leq 1$ , there exists an embedding of  $X$  into  $\ell_p^D$  in dimension  $D = O(\text{dim}(X)/\theta)$  and scaling distortion where the distortion for pairs  $x, y \in X$  and  $\hat{k}$  s.t.  $d(x, y) \leq \Delta_{\hat{k}}(x)$  is  $O(\log^{1+\theta} \hat{k}/\theta)$ .

The lower bound on the distortion guaranteed by THEOREM C.1 is a monotonic function of the distance from any particular point. This is stated in the following corollary:

COROLLARY C.1. Given a metric space  $(X, d)$  satisfying  $\text{WGR}(\gamma)$ . For any  $1 \leq p \leq \infty$ , and  $0 < \theta \leq 1$ , there exists an embedding  $f$  of  $X$  into  $\ell_p^D$  in dimension  $D = O(\text{dim}(X)/\theta)$  such that for any  $x, y \in X$  and  $\hat{k}$  s.t.  $d(x, y) \geq \Delta_{\hat{k}}(x)$  then  $\|f(x) - f(y)\|^p \geq \Delta_{\hat{k}}(x) \cdot \Omega(\theta/\log^{1+\theta} \hat{k})$ .

In the rest of this section we prove THEOREM C.1.

### C.1 Proof of THEOREM C.1 The Embedding.

Let  $\theta > 0$ . Let  $D = \lceil \frac{c \log \lambda}{\theta} \rceil$ , where  $c$  is a constant to be determined later. We will define an

embedding  $f : X \rightarrow l_p^D$  with scaling distortion where the distortion for pairs  $x, y \in X$  and  $\hat{k}$  s.t.  $d(x, y) \leq \Delta_{\hat{k}}(x)$  is  $O(\log^{1+\theta} \hat{k}/\theta)$ . We define  $f$  by defining for each  $1 \leq t \leq D$ , a function  $f^{(t)} : X \rightarrow \mathbb{R}^+$  and let  $f = D^{-1/p} \bigoplus_{1 \leq t \leq D} f^{(t)}$ .

In what follows we define the functions  $f^{(t)}$ . Let  $\Delta_0 = \text{diam}(X)$ ,  $I = \{i \in \mathbb{Z} \mid 1 \leq i \leq \log \Delta_0\}$ . For  $i \in I$  let  $\Delta_i = \Delta_0/4^i$ . For each  $0 < i \in I$  construct a  $\Delta_i$ -bounded uniformly padded probabilistic partition  $\hat{\mathcal{P}}_i$  (note that we have a different random choice for every  $t$ :  $\hat{\mathcal{P}}_i = \hat{\mathcal{P}}_i^{(t)}$  and we omit the superscript for simplicity of notation), as in [Lemma B.7](#) with parameter  $\Gamma = 64$ ,  $\hat{\delta} = 1/2$ . Fix some  $P_i \in \mathcal{P}_i$  for all  $i \in I$ . In the usual embedding via partitions scheme we obtain a lower bound for every pair  $x, y \in X$  from only one "critical" scale (which is approximately  $d(x, y)$ ). Here, we use the same idea, but since the cluster in the critical scale may contain too many points, we get contribution from two scales lower than the critical one, which is guaranteed to be small enough. For this reason we define a new function  $\bar{\xi}$  as follows, for each  $i \in I$ ,  $P \in \mathcal{H}$ :

$$\bar{\xi}_{P,i}(x) = \begin{cases} 1 & \rho(v(P_i(x)), 4\Delta_i, \Gamma^4) \geq 2 \\ \xi_{P,i}(x) & \text{otherwise} \end{cases}$$

where  $v(C)$  is the center of cluster  $C \in \mathcal{P}_i$ . It can be seen that the function  $\bar{\xi}$  is uniform as well.

Let  $\varepsilon(\bar{k}) = \ln^{-\theta} \bar{k}$ ,  $\delta(\bar{k}) = 1 - \varepsilon(\bar{k})$ , and let  $\zeta(\bar{k}) = \ln^{1+2\theta} \bar{k}$ . We define the embedding by defining the coordinates for each  $x \in X$ . Define for  $x \in X$ ,  $0 < i \in I$ ,  $\hat{k}_i(x) = |B(v(P_i(x)), (4\Gamma + 1)\Delta_i)|$ . Define  $\phi_i^{(t)} : X \rightarrow \mathbb{R}^+$ , as:

$$\phi_i^{(t)}(x) = \frac{\bar{\xi}_{P_i}(x)}{\eta_{P_i}^{(\delta(\hat{k}_i(x)))}(x) \cdot \zeta(\hat{k}_i(x))}.$$

Let  $\{\sigma_i^{(t)}(C) \mid C \in P_i, 0 < i \in I\}$  be i.i.d random variables uniformly distributed in  $[0, 1]$ .

For each  $0 < i \in I$  we define a function  $f_i^{(t)} : X \rightarrow \mathbb{R}^+$  and for  $x \in X$ , let  $f^{(t)}(x) = \sum_{i \in I} f_i^{(t)}(x)$ .

The embedding is defined as follows: for each  $x \in X$ :

- For each  $0 < i \in I$ , let  $f_i^{(t)}(x) = \sigma_i^{(t)}(P_i(x)) \cdot g_i^{(t)}(x)$ , where  $g_i^{(t)} : X \rightarrow \mathbb{R}^+$  is defined as:  $g_i^{(t)}(x) = \min\{\phi_i^{(t)}(x) \cdot d(x, X \setminus P_i^{(t)}(x)), \Delta_i\}$ .

We have the following claims:

**CLAIM C.1.** For any  $x, y \in X$  and  $i \in I$  if  $P_i(x) = P_i(y)$  then  $\phi_i^{(t)}(x) = \phi_i^{(t)}(y)$ .

**CLAIM C.2.** There exists universal constant  $C_1$  such that for any  $x \in X$ ,  $1 \leq t \leq D$  we have  $\sum_{j \in I} \phi_j^{(t)}(x) \leq C_1/\theta$ .

*Proof.* Let  $b_i = \lfloor \ln |B(x, 4\Delta_i)| \rfloor$ . As  $d(v(P_i(x)), x) \leq \Delta_i$  we have that  $\log \hat{k}_i(x) = \log |B(v(P_i(x)), (4\Gamma + 1)\Delta_i)| \geq \log |B(x, 4\Gamma\Delta_i)| \geq b_i - 3$ .

$$\begin{aligned} & \sum_{j \in I} \phi_j(x) \\ &= \sum_{j \in I: \bar{\xi}_j(x)=1} \frac{\eta_j^{(\delta(\hat{k}_j(x)))}(x)^{-1}}{\zeta(\hat{k}_j(x))} \\ &\leq \sum_{j \in I: \bar{\xi}_j(x)=1} \frac{2^7 \ln \rho(x, 4\Delta_j, \Gamma)}{\zeta(\hat{k}_j(x)) \cdot \ln(\frac{1}{1-\varepsilon(\hat{k}_j(x))})} + \\ &\quad \sum_{j \in I: \bar{\xi}_j(x)=1, \xi_j(x)=0} \frac{2^7}{\zeta(\hat{k}_j(x)) \cdot \ln(\frac{1}{1-\varepsilon(\hat{k}_j(x))})} \\ &\leq 2^8 \sum_{j \in I: \bar{\xi}_j(x)=1} \frac{\rho(x, 4\Delta_j, \Gamma)}{\ln^{1+\theta} \hat{k}_j(x)} + 2^7 \sum_{j \in I} \frac{1}{\ln^{1+\theta} \hat{k}_j(x)} \\ &\leq 2^9 \sum_{j \in I: \bar{\xi}_j(x)=1} \frac{b_{j-3} - b_{j+2}}{(b_{j-3})^{1+\theta}} + 2^7 \sum_{h=1}^{\infty} \frac{1}{h^{1+\theta}} \\ &\leq 2^9 \sum_{j \in I} \sum_{h=b_{j+2}}^{b_{j-3}} \frac{1}{h^{1+\theta}} + O(1/\theta) \\ &\leq 2^{12} \sum_{h=1}^{\infty} \frac{1}{h^{1+\theta}} + O(1/\theta) = O(1/\theta). \end{aligned}$$

Define  $\bar{g}_i^{(t)} : X \times X \rightarrow \mathbb{R}^+$  as follows:  $\bar{g}_i^{(t)}(x, y) = \min\{\phi_i^{(t)}(x) \cdot d(x, y), \Delta_i\}$ . We have the following claim:

**CLAIM C.3.** For any  $0 < i \in I$  and  $x, y \in X$ :  $f_i^{(t)}(x) - f_i^{(t)}(y) \leq \bar{g}_i^{(t)}(x, y)$ .

**LEMMA C.4.** There exists a universal constant  $C_1 > 0$  such that for any  $x, y \in X$ :

$$\|f(x) - f(y)\|_p \leq (C_1/\theta) \cdot d(x, y).$$

*Proof.* From [Claim C.3](#) and [Claim C.2](#) we get

$$\begin{aligned} & \sum_{0 < i \in I} (f_i^{(t)}(x) - f_i^{(t)}(y)) \\ &\leq \sum_{0 < i \in I} \bar{g}_i^{(t)}(x, y) \leq \sum_{0 < i \in I} \phi_i^{(t)}(x) \cdot d(x, y) \\ &\leq (C_1/\theta) \cdot d(x, y). \end{aligned}$$

It follows that  $|f^{(t)}(x) - f^{(t)}(y)| = |\sum_{0 < i \in I} (f_i^{(t)}(x) - f_i^{(t)}(y))| \leq (C_1/\theta) \cdot d(x, y)$ , and therefore

$$\begin{aligned} \|f(x) - f(y)\|_p^p &= D^{-1} \sum_{1 \leq t \leq D} |f^{(t)}(x) - f^{(t)}(y)|^p \\ &\leq (C_1/\theta)^p d(x, y)^p. \end{aligned}$$

LEMMA C.5. *There exists a universal constant  $C_2 > 0$  such that with constant probability for any  $x, y \in X$  s.t.  $d(x, y) \leq \Delta_{\hat{k}}(x)$ :*

$$\|f(x) - f(y)\|_p \geq C_2 \ln^{-1-3\theta} \hat{k} \cdot d(x, y).$$

*Proof.* We will prove that with constant probability for every  $x, y \in X$  s.t.  $d(x, y) \leq \Delta_{\hat{k}}(x)$ , there exists a set  $T(x, y) \subseteq \{1, \dots, D\}$  of size at least  $D/2$  such that for any  $t \in T(x, y)$ :

$$(C.22) \quad |f^{(t)}(x) - f^{(t)}(y)| \geq 2^{-6} \ln^{-1-3\theta} \hat{k} \cdot d(x, y).$$

The theorem follows directly:

$$\begin{aligned} & \|f(x) - f(y)\|_p^p \\ &= D^{-1} \sum_{1 \leq t \leq D} |f^{(t)}(x) - f^{(t)}(y)|^p \\ &\geq D^{-1} \sum_{t \in T(x, y)} |f^{(t)}(x) - f^{(t)}(y)|^p \\ &\geq D^{-1} |T(x, y)| \cdot \left(2^{-6} \ln^{-1-3\theta} \hat{k} \cdot d(x, y)\right)^p \\ &\geq \frac{1}{2} \left(2^{-6} \ln^{-1-3\theta} \hat{k} \cdot d(x, y)\right)^p. \end{aligned}$$

The proof of (C.22) uses a set of nets of the space. For any  $0 < i \in I$ , and  $1 \leq k = 2^j \leq n$ , let  $N_i^k$  be a  $\frac{\theta \cdot \varepsilon(k) \Delta_i}{16C_1 \zeta(4k)}$ -net of  $X$ . Let

$$(C.23) \quad \begin{aligned} M = \{ & (i, k, u, v) \mid i \in I, u, v \in N_i^k, \\ & 3\Delta_{i-4} \leq d(u, v) \leq 17\Delta_{i-4}, \\ & k \leq \min\{\hat{k}_i(u), \hat{k}_i(v)\} < 2k\}. \end{aligned}$$

Given an embedding  $f$  define a function  $T : M \rightarrow 2^{[D]}$  such that for  $t \in [D]$ :

$$t \in T(i, k, u, v) \Leftrightarrow |f^{(t)}(u) - f^{(t)}(v)| \geq \frac{1}{2} \frac{\varepsilon(k)}{\zeta(4k)} \Delta_i.$$

For all  $(i, k, u, v) \in M$ , let  $\mathcal{E}_{(i, k, u, v)}$  be the event  $|T(i, k, u, v)| \geq D/2$ .

Define the event  $\mathcal{E} = \bigcap_{(i, k, u, v) \in M} \mathcal{E}_{(i, k, u, v)}$  that captures the case that all triplets in  $M$  have the desired property. The main technical lemma is that  $\mathcal{E}$  occurs with non-zero probability:

LEMMA C.6.  $\Pr[\mathcal{E}] > 0$ .

Let us first show that if the event  $\mathcal{E}$  took place, then the lower bound follows. Let  $x, y \in X$ , and let  $0 < i \in I$  be such that  $4\Delta_{i-4} \leq d(x, y) < 16\Delta_{i-4}$ .

Consider  $u, v \in N_i$  satisfying  $d(x, u) = d(x, N_i^k)$  and  $d(y, v) = d(y, N_i^k)$ , then  $d(u, v) \leq d(x, y) + d(u, x) +$

$$d(y, v) \leq 16\Delta_{i-4} + 2\frac{\Delta_i}{C_1} \leq 17\Delta_{i-4} \text{ and } d(u, v) \geq d(x, y) - d(x, u) - d(y, v) \geq 4\Delta_{i-4} - 2\frac{\Delta_i}{C_1} \geq 3\Delta_{i-4}.$$

Let  $k$  be such that  $k \leq \min\{\hat{k}_i(u), \hat{k}_i(v)\} < 2k$ . By the definition of  $M$  it follows that  $(i, k, u, v) \in M$ . It also holds that  $k \leq |B(v, P_i(u)), (4\Gamma + 1)\Delta_i| \leq |B(x, 4\Delta_{i-4})| \leq |B(x, d(x, y))| \leq \hat{k}$ .

The next lemma shows that since  $x, y$  are very close to  $u, v$  respectively, then by the triangle inequality the embedding  $f$  of  $x, y$  cannot differ by much from that of  $u, v$  (respectively).

LEMMA C.7. *Let  $x, y \in X$ , let  $i$  be such that  $4\Delta_{i-4} \leq d(x, y) \leq 16\Delta_{i-4}$ , and  $u, v \in N_i^k$  satisfying  $d(x, u) = d(x, N_i^k)$  and  $d(y, v) = d(y, N_i^k)$ .*

*Given  $\mathcal{E}$ , for any  $t \in T(i, k, u, v)$ :*

$$|f^{(t)}(x) - f^{(t)}(y)| \geq \frac{1}{4} \frac{\varepsilon(\hat{k})}{\zeta(4\hat{k})} \Delta_i.$$

*Proof.* Since  $N_i^k$  is  $\frac{\theta \cdot \varepsilon(k) \Delta_i}{16C_1 \zeta(4k)}$ -net, then  $d(x, u) \leq \frac{\theta \cdot \varepsilon(k) \Delta_i}{16C_1 \zeta(4k)}$ . By Lemma C.4  $|f^{(t)}(x) - f^{(t)}(u)| \leq (C_1/\theta) \cdot d(x, u) \leq \frac{1}{16} \frac{\varepsilon(k)}{\zeta(4k)} \Delta_i$ , and similarly  $|f^{(t)}(y) - f^{(t)}(v)| \leq \frac{1}{16} \frac{\varepsilon(k)}{\zeta(4k)} \Delta_i$ . Then

$$\begin{aligned} & |f^{(t)}(x) - f^{(t)}(y)| \\ &= |f^{(t)}(x) - f^{(t)}(u) + f^{(t)}(u) - f^{(t)}(v) + f^{(t)}(v) - f^{(t)}(y)| \\ &\geq |f^{(t)}(u) - f^{(t)}(v)| - |f^{(t)}(x) - f^{(t)}(u)| - |f^{(t)}(y) - f^{(t)}(v)| \\ &\geq \frac{1}{2} \frac{\varepsilon(k)}{\zeta(4k)} \Delta_i - 2 \frac{1}{16} \frac{\varepsilon(k)}{\zeta(4k)} \Delta_i \geq \frac{1}{4} \frac{\varepsilon(k)}{\zeta(4k)} \Delta_i \geq \frac{1}{4} \frac{\varepsilon(\hat{k})}{\zeta(4\hat{k})} \Delta_i. \end{aligned}$$

Let  $\kappa(k) = \lceil \log \log(4k) \rceil$ . Let  $(i, k, u, v) \in M$  and  $t \in [D]$ . Define  $\mathcal{F}_{(i, k, u, v, t)}$  be the event that:

$$\left| \sum_{0 < j \leq i + \kappa(k)} (f_j^{(t)}(u) - f_j^{(t)}(v)) \right| \geq \frac{\varepsilon(k)}{\zeta(4k)} \Delta_i.$$

Let  $\hat{\mathcal{E}}_{(i, k, u, v)}$  be the event that  $|\{t | \mathcal{F}_{(i, k, u, v, t)}\}| \geq D/2$ .

CLAIM C.8. *For all  $(i, k, u, v) \in M$ ,  $\hat{\mathcal{E}}_{(i, k, u, v)}$  implies  $\mathcal{E}_{(i, u, v)}$ .*

*Proof.* Let  $S = \{t | \mathcal{F}_{(i, k, u, v, t)}\}$ . Then for  $t \in S$ :  $|\sum_{0 < j \leq i + \kappa(k)} f_j^{(t)}(u) - f_j^{(t)}(v)| \geq \frac{\varepsilon(k)}{\zeta(4k)} \Delta_i$ , from Claim C.3 it follows that  $|\sum_{j > i + \kappa(k)} f_j^{(t)}(u) - f_j^{(t)}(v)| \leq \sum_{j > i + \kappa(k)} \Delta_j \leq \frac{1}{2} \frac{\varepsilon(k)}{\zeta(4k)} \Delta_i$ , which implies that  $|f^{(t)}(u) - f^{(t)}(v)| = |\sum_{j \in I} f_j^{(t)}(u) - f_j^{(t)}(v)| \geq \frac{1}{2} \frac{\varepsilon(k)}{\zeta(4k)} \Delta_i$ .

LEMMA C.9. (LOVASZ LOCAL LEMMA - GENERAL CASE) *Let  $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n$  be events in some probability space.*

Let  $G(V, E)$  be a directed graph on  $n$  vertices, each vertex corresponds to an event. Let  $c : V \rightarrow [m]$  be a rating function of events, such that if  $(\mathcal{A}_i, \mathcal{A}_j) \in E$  then  $c(\mathcal{A}_i) \leq c(\mathcal{A}_j)$ . Assume that for all  $i = 1, \dots, n$  there exists  $x_i \in [0, 1)$  such that

$$\Pr \left[ \mathcal{A}_i \mid \bigwedge_{j \in Q} \neg \mathcal{A}_j \right] \leq x_i \prod_{j: (i,j) \in E} (1 - x_j),$$

for all  $Q \subseteq \{j : (\mathcal{A}_i, \mathcal{A}_j) \notin E \wedge c(\mathcal{A}_i) \geq c(\mathcal{A}_j)\}$ , then

$$\Pr \left[ \bigwedge_{i=1}^n \neg \mathcal{A}_i \right] > 0$$

Define a graph  $G = (V, E)$ , where  $V = \{\hat{\mathcal{E}}_{(i,k,u,v)} \mid (i,k,u,v) \in M\}$ , and the rating of a vertex  $c(\hat{\mathcal{E}}_{(i,k,u,v)}) = i$ . Let  $x_{(i,k,u,v)} = \lambda^{-60 \ln(\frac{2 \ln k}{\theta})}$ .

Define that  $(\hat{\mathcal{E}}_{(i,k,u,v)}, \hat{\mathcal{E}}_{(i',k',u',v')}) \in E$  iff  $d(\{u,v\}, \{u',v'\}) \leq 4\Delta_i$ , and  $i' \leq i + \kappa(k)$ , and  $\frac{1}{3} \leq \frac{\log \log(4k')}{\log \log(4k)} \leq 3$ .

CLAIM C.10. Let  $\hat{\mathcal{E}}_{(i,k,u,v)} \in V$ , then the number of edges  $(\hat{\mathcal{E}}_{(i,k,u,v)}, \hat{\mathcal{E}}_{(i',k',u',v')}) \in E$  is at most  $\lambda^{20 \ln(\frac{2 \ln k}{\theta})}$ .

*Proof.* We bound the number of pairs  $u', v' \in N_{i'}^k$  such that  $(\hat{\mathcal{E}}_{(i,k,u,v)}, \hat{\mathcal{E}}_{(i',k',u',v')}) \in E$  for  $i \leq i' \leq i + \kappa(k)$  and  $\frac{1}{3} \leq \frac{\log \log(4k')}{\log \log(4k)} \leq 3$ .

Assume w.l.o.g  $d(u, u') \leq 4\Delta_i$ , since  $d(u', v') \leq 17\Delta_{i-4}$  we have  $u', v' \in B = B(u, 40\Delta_{i-4})$ . The number of pairs can be bounded by  $|N_{i'}^k \cap B|^2$ . There is at most point from the net  $N_{i'}^k$  in every ball of radius  $r = \frac{\theta \cdot \varepsilon(k)^3}{16C_1(\zeta(4k))^3} \Delta_{i+\kappa(k)}$ . Since  $(X, d)$  is  $\lambda$ -doubling, the ball  $B$  can be covered by  $\lambda^{\log(40\Delta_{i-4}/r)}$  balls of radius  $r$ . Now,  $\log(40\Delta_{i-4}/r) \leq 8 \ln \ln k + 18 + \log(1/\theta)$ . It conclude that the number of possible pairs is bounded above by  $\lambda^{20 \ln(\frac{2 \ln k}{\theta})}$ .

The construction of the graph is based on the proposition that vertices that do not have an edge are either farther than  $\approx \Delta_i$  apart or have different scales and hence do not change each other's bound on their success probability.

LEMMA C.11.

$$\Pr \left[ \neg \hat{\mathcal{E}}_{(i,k,u,v)} \mid \bigwedge_{(i',k',u',v') \in Q} \hat{\mathcal{E}}_{(i',k',u',v')} \right] \leq \lambda^{-61 \ln(\frac{2 \ln k}{\theta})},$$

for all

$$Q \subseteq \left\{ (i', k', u', v') \mid i \geq i' \wedge \left( \hat{\mathcal{E}}_{(i,k,u,v)}, \hat{\mathcal{E}}_{(i',k',u',v')} \right) \notin E \right\}$$

Before we prove this lemma, let us see that it implies Lemma C.6.

Apply Lemma C.9 to the graph  $G$  we defined. Using Claim C.10 we can bound the number of edges  $(\hat{\mathcal{E}}_{(i,k,u,v)}, \hat{\mathcal{E}}_{(i',k',u',v')}) \in E$  is at most  $d = \lambda^{20 \ln(\frac{2 \ln k}{\theta})}$ . Recall that  $x_{(i,k,u,v)} = \lambda^{-60 \ln(\frac{2 \ln k}{\theta})}$ . Also it follows that  $x_{(i',k',u',v')} = \lambda^{-60 \ln(\frac{2 \ln k'}{\theta})} \leq \lambda^{-20 \ln(\frac{2 \ln k}{\theta})}$ . Therefore the probability bound in Lemma C.11 satisfies the first condition of Lemma C.9  $\lambda^{-61 \ln(\frac{2 \ln k}{\theta})} \leq \lambda^{-60 \ln(\frac{2 \ln k}{\theta})} (1 - \lambda^{-20 \ln(\frac{2 \ln k}{\theta})})^d$ . Therefore  $\Pr[\mathcal{E}] = \Pr[\bigwedge_{(i,k,u,v) \in M} \hat{\mathcal{E}}_{(i,k,u,v)}] > 0$ , which concludes the proof of Lemma C.6.

C.1.1 Proof of Lemma C.11 In what follows we use of the following simple technical claim.

CLAIM C.12. Let  $A, B \in \mathbb{R}^+$  and let  $\alpha, \beta$  be i.i.d random variables uniformly distributed in  $[0, 1]$ . Then for any  $C \in \mathbb{R}$  and  $\varepsilon > 0$ :

$$\Pr[|C + A\alpha - B\beta| < \varepsilon \cdot \max\{A, B\}] < 2\varepsilon.$$

*Proof.* Assume wlog  $A \geq B$ . Consider the condition  $|C + A\alpha - B\beta| < \varepsilon \cdot \max\{A, B\} = \varepsilon A$ . If  $C - B\beta \geq 0$  then it implies  $\alpha < \varepsilon$ . Otherwise  $|\alpha - \frac{B\beta - C}{A}| < \varepsilon$ .

CLAIM C.13. Let  $(i, k, u, v) \in M$ ,  $t \in [D]$ , then  $\Pr[\mathcal{F}_{(i,k,u,v,t)}] \geq 1 - 3\varepsilon(k)$ .

*Proof.* Set  $\varepsilon = \varepsilon(k)$  and  $\delta = 1 - \varepsilon$ . Consider some  $(i, k, u, v) \in M$ . Then  $3\Delta_{i-4} \leq d(u, v) \leq 17\Delta_{i-4}$ . By Claim B.2 we have that  $\max\{\bar{\rho}(u, \Delta_{i-4}, \Gamma), \bar{\rho}(v, \Delta_{i-4}, \Gamma)\} \geq 2$ . Assume w.l.o.g that  $\bar{\rho}(u, \Delta_{i-4}, \Gamma) \geq 2$ . It follows that also  $\rho(v(P_i(u)), 4\Delta_i, \Gamma^4) \geq 2$  from Lemma B.7 that  $\bar{\xi}_{P^{(t)}, i}(u) = 1$  which implies that  $\phi_i^{(t)}(u) = \frac{\eta_{P^{(t)}, i}^{(\delta(k_i(u)))}(u)^{-1}}{\zeta(k_i(u))}$ . As  $k_i(u) \geq k$  we have that  $\phi_i^{(t)}(u) \geq \frac{\eta_{P^{(t)}, i}^{(\delta)}(u)^{-1}}{\zeta(k_i(u))}$ . As  $\hat{\mathcal{H}}^{(t)}$  is  $(\eta^{(\delta)}, 1 - \varepsilon)$ -padded we have the following bound

$$\Pr[B(u, \eta_{P^{(t)}, i}^{(\delta)}(u) \Delta_i) \subseteq P_i^{(t)}(u)] \geq 1 - \varepsilon.$$

Therefore with probability at least  $1 - \varepsilon$ :

$$(C.24) \quad g_i^{(t)}(u) \geq \phi_i^{(t)}(u) \cdot d(u, X \setminus P_i^{(t)}(u)) \geq \frac{\Delta_i}{\zeta(\hat{k}_i(u))}.$$

If  $\hat{k}_i(u) \leq 4k$  then  $g_i^{(t)}(u) \geq \frac{\Delta_i}{\zeta(4k)}$ . Otherwise

it must be the case that  $\hat{k}_i(v) \leq 2k$ . It follows that  $\rho(v(P_i(u)), 4\Delta_i, \Gamma^4) \geq 2$  and thus  $\bar{\xi}_{P^{(t)}, i}(v) = 1$ ,

and hence by analogues argument to the one above we get that  $g_i^{(t)}(v) \geq \frac{\Delta_i}{\zeta(4k)}$ . We conclude that  $\max\{g_i^{(t)}(u), g_i^{(t)}(v)\} \geq \frac{\Delta_i}{\zeta(4k)}$ .

Let  $\mathcal{A}$  denote the event that (C.24) occurs.

Recall that we are interested in the expression:  $|\sum_{0 < j \leq i + \kappa(k)} (f_j^{(t)}(u) - f_j^{(t)}(v))|$  and

$$f_i^{(t)}(u) - f_i^{(t)}(v) = \sigma_i^{(t)}(P_i^{(t)}(u)) \cdot g_i^{(t)}(u) - \sigma_i^{(t)}(P_i^{(t)}(v)) \cdot g_i^{(t)}(v).$$

Define  $A = g_i^{(t)}(u)$ ,  $B = g_i^{(t)}(v)$ ,  $\alpha = \sigma_i^{(t)}(P_i^{(t)}(u))$ ,  $\beta = \sigma_i^{(t)}(P_i^{(t)}(v))$  and  $C = \sum_{i \neq j \leq i + \kappa(k)} (f_j^{(t)}(u) - f_j^{(t)}(v))$ . Since  $\text{diam}(P_i^{(t)}(u)) \leq \Delta_i < d(u, v)$  we have that  $P_i^{(t)}(v) \neq P_i^{(t)}(u)$ . Thus  $\alpha$  and  $\beta$  are independent random variables uniformly distributed in  $[0, 1]$ , hence we can apply claim C.12 and using (C.24) we have:

$$\begin{aligned} \Pr\left[ \sum_{0 < j \leq i + \kappa(k)} (f_j^{(t)}(u) - f_j^{(t)}(v)) < \varepsilon \frac{\Delta_i}{\zeta(4k)} \mid \mathcal{A} \right] \\ = \Pr[C + A\alpha - B\beta] < \varepsilon \cdot \max\{A, B\} \mid \mathcal{A} < 2\varepsilon. \end{aligned}$$

Therefore with probability at least  $1 - 3\varepsilon(k)$ :

$$(C.25) \quad |f^{(t)}(u) - f^{(t)}(v)| \geq \frac{\varepsilon(k)}{\zeta(4k)} \Delta_i.$$

CLAIM C.14. *Let  $(i, k, u, v) \in M$ ,  $t \in [D]$ , then*

$$\Pr \left[ \neg \mathcal{F}_{(i,k,u,v,t)} \mid \bigwedge_{(i',k',u',v') \in Q} \hat{\mathcal{E}}_{(i',k',u',v')} \right] \leq 3\varepsilon(k),$$

for all  $Q \subseteq \{(i', k', u', v') \in M \mid i \geq i' \wedge (\hat{\mathcal{E}}_{(i,k,u,v)}, \hat{\mathcal{E}}_{(i',k',u',v')}) \notin E\}$ .

*Proof.* If  $i' + \kappa(k') < i$ , then event  $\hat{\mathcal{E}}_{(i',k',u',v')}$  depend on events  $\mathcal{F}_{(i',k',u',v',t')}$ , and these events depend only on the choice of partition for scales at most  $i$ . Hence the padding probability for  $u, v$  in scale  $i$  and the choice of  $\sigma_i$  is independent of these events.

Otherwise, if  $i - \kappa(k') \leq i' \leq i$ , let  $(i', k', u', v') \in M$  such that  $(\hat{\mathcal{E}}_{(i,k,u,v)}, \hat{\mathcal{E}}_{(i',k',u',v')}) \notin E$ . By the construction of  $G$  there are two cases. If  $u', v' \notin B(u, 4\Delta_{i'})$  and  $u', v' \notin B(v, 4\Delta_{i'})$  then  $u', v'$  are far from  $u, v$  and they fall into different clusters in every possible partition of scale  $i$ . From Lemma B.7, the padding of  $u, v$  in scale  $i$  depends only on the local neighborhoods,  $B(u, 2\Delta_i) \cup B(v, 2\Delta_i)$ , which are disjoint from those of  $u', v'$ . The second case is that  $d(\{u, v\}, \{u', v'\}) \leq 4\Delta_i$ . Recall that  $k' \leq k_{i'}(u') = |B(v(P_{i'}(u')), (4\Gamma + 1)\Delta_{i'})|$

and  $k \geq \frac{1}{2}k_i(u) = \frac{1}{2}|B(v(P_i(u)), (4\Gamma + 1)\Delta_i)|$ . We have  $d(v(P_{i'}(u')), v(P_i(u))) \leq d(v(P_{i'}(u'), u') + d(u', u) + d(u, v(P_i(u))) \leq \Delta_{i'} + 4\Delta_i + \Delta_i \leq 6\Delta_{i'}$  and therefore  $k' \leq |B(v(P_i(u)), 2(4\Gamma + 1)\Delta_{i'})|$ . It follows from the WGR( $\gamma$ ) assumption that  $k' \leq 2k^{4\gamma\kappa(k')}$  implying  $\log \log(4k') \leq \log \log(4k) + 2\gamma\kappa(k') \leq \log \log(4k) + 3\gamma \log \log(4k')$ , and therefore  $\frac{\log \log(4k')}{\log \log(4k)} \leq 1/(1 - 3\gamma) \leq 3$  assuming  $\gamma < 0.2$ . A similar bound can be derived in the reverse direction which yields a contradiction.

By Claim C.13 there is probability  $\geq 1 - 3\varepsilon(k)$  to succeed, no matter what happened in scales  $\neq i$  or “far away” in scale  $i$ .

We now prove Lemma C.11. By Claim C.14 the probability a single coordinate  $t$  fails is at most  $3\varepsilon(k)$ . It follows from Chernoff bounds that the probability that more than  $D/2$  coordinates fail is bounded above by:

$$(C.26) \quad \Pr \left[ \neg \hat{\mathcal{E}}_{(i,k,u,v)} \mid \bigwedge_{(i',k',u',v') \in Q} \hat{\mathcal{E}}_{(i',k',u',v')} \right] \leq (6e(3\varepsilon(k)))^{D/2} \leq \lambda^{-\frac{c}{8} \ln(\frac{2 \ln k}{\theta})}.$$

Setting  $c$  large enough implies that (C.26) is at most  $\lambda^{-61 \ln(\frac{2 \ln k}{\theta})}$ , as required.

## D Proof of Theorem 5.2

The general argument is similar to that of Section 3.2. We shall highlight the main differences. We use the following property of the Nash device (corollary of Lemma 3.2):

LEMMA D.1. *Consider the Nash device  $\Theta$  with  $\sigma_1 = \dots = \sigma_D = \sigma$ , and let  $r = \sqrt{2}/\sigma$ . Let  $0 < \epsilon < \frac{1}{2}$  and  $x$  and  $y \in \mathfrak{R}^U$ . Then*

$$(1 - \epsilon) \leq \left( \frac{\|\Theta(x) - \Theta(y)\|}{G_r(\|x - y\|)} \right)^2 \leq (1 + \epsilon)$$

with probability exceeding  $1 - 2 \exp(-\frac{D\epsilon^2}{24})$ .

We also replace the cardinality based probabilistic partitions with the more standard diameter bounded probabilistic partitions. The following lemmas are easily obtained from the work of [2]:

LEMMA D.2. (LPPP) *Let  $(X, d)$  be a finite metric space. Let  $\Delta > 0$  and  $k \in \mathbb{N}$  such that for each  $x \in X$ ,  $\Delta \leq \Delta_k(x)$ . There exists a  $\Delta$ -bounded probabilistic partition  $\hat{\mathcal{P}}$  of  $(X, d)$  such that  $\hat{\mathcal{P}}$  is  $(\eta^{(\delta)}, \delta)$ -locally padded for  $\eta^{(\delta)} = 2^{-11} \cdot \ln(1/\delta)/\ln k$ , for each  $\delta \in (1/k, 1]$ .*

LEMMA D.3. *The construction in Lemma 2.4 holds with respect to the probabilistic partitions in Lemma D.2.*

We define the embedding in the same way as in [Section 3.2](#) except we replace the cardinality based partitions with those of [Lemma D.3](#). We also let  $\sigma = 2^{12}(\ln k)/\epsilon \cdot \Delta^{-1}$ ,  $\hat{\tau} = \frac{\tau}{2\sqrt{2}}$ , and  $\hat{\sigma} = \hat{\tau}^{-1}\sigma$ . Let  $\Delta^* = \sigma^{-1}/2$  and  $r = \sqrt{2}/\hat{\sigma} = \tau\Delta^*$ . That is to say,  $\sigma$  is now constant for all clusters  $C \in P$ . The amplitude  $A^{(t)}$  is scaled down by a factor of  $\hat{\tau}$  as follows:  $A^{(t)}(x) = \hat{\tau} \cdot \min\{d(x, X \setminus C), \sigma^{-1}\}$ , and let

$$\Phi^{(t)}(x) = A^{(t)}(x)\hat{\varphi}^{(t)}(x)$$

where,

$$\hat{\varphi}^{(t)}(x) = \hat{\sigma}\varphi^{(t)}(x) = \begin{bmatrix} \cos(\hat{\sigma}\omega^{(t)}(x)'x) \\ \sin(\hat{\sigma}\omega^{(t)}(x)'x) \end{bmatrix}.$$

We have the following version of [Lemma 3.3](#)

LEMMA D.4. *Let  $x, y \in X$ . Then,*

- (i.) *If  $P^{(t)}(x) \neq P^{(t)}(y)$ ,  $\|\Phi^{(t)}(x) - \Phi^{(t)}(y)\| \leq 2\hat{\tau}\|x - y\|$ .*
- (ii.) *If  $P^{(t)}(x) \neq P^{(t)}(y)$ ,  $d(x, X \setminus P^{(t)}(x)) \geq 2\sigma^{-1}$  and  $d(y, X \setminus P^{(t)}(y)) \geq 2\sigma^{-1}$ , then  $\|\Phi^{(t)}(x) - \Phi^{(t)}(y)\| \leq \hat{\tau}\|x - y\|$ .*
- (iii.) *If  $P^{(t)}(x) = P^{(t)}(y)$ ,  $\|\Phi^{(t)}(x) - \Phi^{(t)}(y)\|^2 \leq \hat{\tau}^2\|x - y\|^2 + \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2$ .*
- (iv.) *If  $P^{(t)}(x) = P^{(t)}(y)$ ,  $\sigma^{-1} \leq d(x, X \setminus P^{(t)}(x))$  and  $\sigma^{-1} \leq d(y, X \setminus P^{(t)}(y))$ , then  $\|\Phi^{(t)}(x) - \Phi^{(t)}(y)\| = \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|$ .*

The proof of [Theorem 5.2](#) now follows from the properties in [Lemma D.4](#) using an argument similar to that of [Section 3.2](#).

The following inequality replaces [\(3.5\)](#):

$$(D.27) \quad \|\Phi(x) - \Phi(y)\|^2 \leq \frac{1}{D} \cdot |T_=(x, y)| \cdot \frac{\sum_{t \in T_=(x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2}{|T_=(x, y)|} + \frac{1}{D} \cdot |T_{\neq}(x, y)| \cdot \tau^2\|x - y\|^2 + 4\epsilon\tau^2\|x - y\|^2.$$

Inequality [\(3.6\)](#) remains as before. For  $x, y$  such that:  $\|x - y\| \leq \Delta^*$  we have:

$$(D.28) \quad \|\Phi(x) - \Phi(y)\|^2 \geq (1 - 4\epsilon) \cdot \frac{\sum_{t \in T_=(x, y) \cap T_o(x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2}{|T_=(x, y) \cap T_o(x, y)|}.$$

As in [Section 3.2](#) we define:

$$L(x, y) = \frac{\sum_{t \in T_=(x, y) \cap T_o(x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2}{|T_=(x, y) \cap T_o(x, y)|},$$

$$U(x, y) = \frac{\sum_{t \in T_=(x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2}{|T_=(x, y)|}.$$

We define the following events for pairs. Let  $A_U(x, y)$  be the event that  $U(x, y) > (1 + \epsilon)\|x - y\|^2$ . For pairs  $x, y$  that are close neighbors, that is:  $\|x - y\| \leq \Delta^*$ , define  $A_L(x, y)$  to be the event that:  $L(x, y) < (1 - \epsilon)G_r(\|x - y\|)^2$  and  $A_{U'}(x, y)$  to be the event that:  $U(x, y) > (1 + \epsilon)G_r(\|x - y\|)^2$ . Let  $A(x, y) = A_L(x, y) \vee A_{U'}(x, y) \vee A_U(x, y)$ . If  $x, y$  are not close neighbors then  $A(x, y) = A_U(x, y)$ . The rest of the argument utilizes the Lovász Local Lemma to prove that there is positive probability that none of the events  $A(x, y)$  occurs.

We create a dependency graph  $G_A$  whose vertices are the events  $A(x, y)$  as in [Section 3.1](#). Let  $d_{G_A}$  denote its maximum degree. Using the condition that  $\Delta \leq \Delta_k(x)$  for all  $x \in X$  we obtain the bound  $d_{G_A} \leq \binom{k}{2}$ .

As before, by part (a) of [Lemma 3.2](#) the probability that  $U(x, y) > (1 + \epsilon)\|x - y\|^2$  is at most  $e^{-D(\epsilon^2/4 - \epsilon^3/6)} \leq k^{-2}/4$ . Hence, for pairs  $x, y$  that are not close neighbors this implies that the probability that event  $A(x, y)$  occurs is at most  $1/(e(\binom{k}{2} + 1)) \leq 1/(e \cdot d_{G_A} + 1)$ .

Consider now pairs  $x, y$  that are close neighbors:  $\|x - y\| \leq \Delta^*$ . By [Lemma D.1](#) that the probability that either  $A_L(x, y)$  or  $A_{U'}(x, y)$  hold is at most  $2e^{-D\epsilon^2/24} \leq k^{-2}/4$ . Hence the probability the event  $A(x, y)$  occurs is at most  $k^{-2}/2 < 1/(e \cdot d_{G_A} + 1)$ . This complete the proof that the conditions of the Local Lemma are satisfied.

As before, by [\(D.27\)](#) we get that for any pair  $x, y \in X$ :  $\|\Phi(x) - \Phi(y)\|^2 \leq (1 + 5\epsilon)\|x - y\|^2$ .

Now consider close neighbors  $x, y$  such that  $\|x - y\| \leq \Delta_k^*$ . By [\(3.6\)](#) we have:

$$\begin{aligned} \|\Phi(x) - \Phi(y)\|^2 &\geq (1 - 4\epsilon)L(x, y) \\ &\geq (1 - 4\epsilon)(1 - \epsilon)G_r(\|x - y\|)^2. \end{aligned}$$

Recall that in [Section 3.1](#) it is shown that for close neighbors:  $|T_=(x, y) \cap T_o(x, y)| \geq (1 - 4\epsilon)D$ . It follows that  $|T_{\neq}(x, y)| \leq 4\epsilon D$  and so by [\(D.27\)](#) we have

$$(D.29) \quad \|\Phi(x) - \Phi(y)\|^2 \leq \frac{1}{D} \cdot |T_=(x, y)| \cdot \frac{\sum_{t \in T_=(x, y)} \|\varphi^{(t)}(x) - \varphi^{(t)}(y)\|^2}{|T_=(x, y)|} + 8\epsilon\tau^2\|x - y\|^2.$$

It follows that when  $\|x - y\| \leq \Delta^* = r/\tau$  it holds that  $G_r(\|x - y\|) \geq \sqrt{1 - 1/e} \cdot \tau\|x - y\|$  (for  $\tau < 1$ ), using [Claim 3.5](#), and hence using [Lemma D.1](#) we obtain:

$$(D.30) \quad \begin{aligned} \|\Phi(x) - \Phi(y)\|^2 &\leq (1 + \epsilon)G_r(\|x - y\|)^2 + \frac{8\epsilon}{e - 1}\epsilon G_r(\|x - y\|)^2 \\ &\leq (1 + 14\epsilon)G_r(\|x - y\|)^2. \end{aligned}$$

Finally, for property (c) of [Theorem 5.2](#) note that it follows directly from the definition of  $\Phi$  that for all  $x \in X$ :  $\|\Phi(x)\| \leq \hat{\sigma}^{-1} \leq r$ .