Advances in Metric Embedding Theory

Extended Abstract

Ittai Abraham
Hebrew University
ittaia@cs.huji.ac.il

Yair Bartal*
Hebrew University
yair@cs.huji.ac.il

Ofer Neiman†
Hebrew University
neiman@cs.huji.ac.il

ABSTRACT

Metric Embedding plays an important role in a vast range of application areas such as computer vision, computational biology, machine learning, networking, statistics, and mathematical psychology, to name a few.

The theory of metric embedding received much attention in recent years by mathematicians as well as computer scientists and has been applied in many algorithmic applications.

A cornerstone of the field is a celebrated theorem of Bourgain which states that every finite metric space on n points embeds in Euclidean space with $O(\log n)$ distortion.

Bourgain's result is best possible when considering the worst case distortion over all pairs of points in the metric space. Yet, it is possible that an embedding can do much better in terms of the *average distortion*.

Indeed, in most practical applications of metric embedding the main criteria for the quality of an embedding is its *average* distortion over all pairs.

In this paper we provide an embedding with *constant* average distortion for arbitrary metric spaces, while maintaining the same worst case bound provided by Bourgain's theorem.

In fact, our embedding possesses a much stronger property. We define the ℓ_q*-distortion* of a uniformly distributed pair of points. Our embedding achieves the best possible ℓ_q-distortion for all $1 \le q \le \infty$ *simultaneously.*

These results have several algorithmic implications, e.g. an $O(1)$ approximation for the unweighted uncapacitated quadratic assignment problem.

The results are based on novel embedding methods which improve on previous methods in another important aspect: the *dimension*.

The dimension of an embedding is of very high importance in particular in applications and much effort has been invested in analyzing it. However, no previous result improved the bound on the dimension which can be derived from Bourgain's embedding.

We prove that any metric space on n points embeds into L_p with distortion $O(\log n)$ in dimension $O(\log n)$. This provides an *optimal* bound on the dimension of the embedding.

Somewhat surprisingly, we show that a further small improvement is possible at a small price in the distortion, obtaining an embedding with distortion $O(\log^{1+\theta} n)$ in *optimal* dimension $O(\theta^{-1} \log n/ \log\log n)$, for any $\theta > 0$. It is worth noting that with the small loss in the distortion this improves upon the best known embedding of arbitrary spaces into *Euclidean* space, where dimension reduction is used.

Our techniques also allow to obtain the optimal distortion for embedding into L_p with nearly tight dimension. For any $1 \le p \le \infty$ and any $1 \le k \le p$, we give an embedding into L_p with distortion $O(\lceil \log n/k \rceil)$ in dimension $2^{O(k)} \log n$.

Underlying our results is a novel embedding method. Probabilistic metric decomposition techniques have played a central role in the field of finite metric embedding in recent years. Here we introduce a novel notion of probabilistic metric decompositions which comes particularly natural in the context of embedding. Our new methodology provides a *unified approach* to all known results on embedding of arbitrary metric spaces. Moreover, as described above, with some additional ideas they allow to get far stronger results. These metric decompositions seem of independent interest.[1]

Categories and Subject Descriptors: F.2.0 [Theory of Computation]: Analysis of Algorithms and Problem Complexity – General

General Terms: Algorithms, Theory

Keywords: Metric Embedding

1. INTRODUCTION

The theory of embeddings of finite metric spaces has attracted much attention in recent decades by several communities: mathematicians, researchers in theoretical Computer Science as well as researchers in the networking community and other applied fields of Computer Science.

The main objective of the field is to find embeddings of metric spaces into other more simple and structured spaces that have *low distortion*.

Given two metric spaces (X, d_X) and (Y, d_Y) an *injective* mapping $f : X \to Y$ is called an *embedding* of X into Y. An embedding is *non-contractive* if for any $u \ne v \in X$: $d_Y(f(u), f(v)) \ge d_X(u, v)$. The *distortion* of a non-

*Supported in part by a grant from the Israeli Science Foundation (195/02)

†Supported in part by a grant from the Israeli Science Foundation (195/02)

[1] The paper is based on the papers [2] and [10].

contractive embedding f is: $\text{dist}(f) = \sup_{u \neq v \in X} \text{dist}_f(u,v)$, where $\text{dist}_f(u,v) = \frac{d_Y(f(u),f(v))}{d_X(u,v)}$.

We say that X embeds in Y with distortion α if there exists an embedding of X into Y with distortion α.

In Computer Science, embeddings of finite metric spaces have played an important role, in recent years, in the development of algorithms. More general practical use of embeddings can be found in a vast range of application areas including computer vision, computational biology, machine learning, networking, statistics, and mathematical psychology to name a few.

From a mathematical perspective embeddings of finite metric spaces into normed spaces are considered natural non-linear analogues to the local theory of Banach spaces. The most classic fundamental question is that of embedding metric spaces into Hilbert Space.

Major effort has been put into investigating embeddings into L_p normed spaces (see the surveys [41, 30, 31] and the book [47] for an exposition of many of the known results).

The main cornerstone of the field has been the following theorem by Bourgain [15]:

THEOREM 1 (BOURGAIN). *For every n-point metric space there exists an embedding into Euclidean space with distortion $O(\log n)$.*

This theorem has been the basis on which the theory of embedding into finite metric spaces has been built. In [42] it is shown that Bourgain's embedding provides an embedding into L_p with distortion $O(\log n)$ and dimension $O(\log^2 n)$. In this paper we improve this result in two ways: we present an embedding with *average* distortion $O(1)$ and dimension $O(\log n)$.

1.1 Novel Embedding Methods

There are few general methods of embedding finite metric spaces that appear throughout the literature. One is indeed the method introduced in Bourgain's proof. This may be described as a Fréchet-style embedding where coordinates are defined as distances to randomly chosen sets in the space. Some examples of its use include [15, 42, 45, 46], essentially providing the best known bounds on embedding arbitrary metric spaces into L_p.

The other embedding method which has been extensively used in recent years, is based on *probabilistic partitions* of metric spaces [7] originally defined in the context of *probabilistic embedding* of metric spaces. Probabilistic partitions for arbitrary metric spaces were also given in [7] and similar constructions appeared in [43].

The probabilistic embedding of [7] (and later improvements in [8, 22, 9]) provide in particular embeddings into L_1 and serve as the first use of probabilistic partitions in the context of embeddings into normed spaces.

A major step was done in a paper by Rao [52] where he shows that a certain padding property of such partitions can be used to obtain embeddings into L_2. Informally, a probabilistic partition is *padded* if every ball of a certain radius depending on some *padding parameter* has a good chance of being contained in a cluster. Rao's embedding defines coordinates which may be described as the distance from a point to the edge of its cluster in the partition and the padding parameter provides a lower bound on this quantity (with some associated probability). While Rao's original proof was done in the context of embedding planar metrics, it has since been observed by many researchers that his methods are more general and in fact provide the first decomposition-based embedding into L_p. However, the resulting distortion bound still did not match those achievable by Bourgain's original techniques.

This gap has been recently closed by Krauthgamer et. al [35]. Their embedding, however, is already far more complex than the original decomposition-based argument by Rao. Their proof exploits special properties of the probabilistic partition of [22] which was also originally developed in the context of probabilistic embedding. This partition is based on an algorithm of [16] and further improvements by [21]. In particular, the main property of the probabilistic partition of [22] is that the padding parameter is defined separately at each point of the space and depends in a delicate fashion on the growth of the space in the local surrounding of that point.

This paper introduces *novel probabilistic partitions* with even more refined properties which allow stronger and more general results on embedding of finite metric spaces which could not be achieved using the previous methods. They also lead to relatively elegant embeddings and analysis.

Decomposition based embeddings also play a fundamental role in the recently developed metric Ramsey theory [11, 14]. In [13] it is shown that the standard Fréchet style embeddings do not allow similar results. One indication that our approach significantly differs from the previous embedding methods discussed above is that our new theorems crucially rely on the use of non-Fréchet embeddings.

The main idea is the construction of *uniformly padded* probabilistic partitions. That is the padding parameter is *uniform* over all points within a cluster. The key is that having this property allows partition-based embeddings to use the value of the padding parameter in the definition of the embedding in the most natural way. In particular, the most natural definition is to let a coordinate be the distance from a point to the edge of the cluster (as in [52]) multiplied by the inverse of the padding parameter. This provides an alternate embedding method with essentially similar benefits as the approach of [35].

We present a construction of uniformly padded probabilistic partitions which still posses intricate properties similar to those of [22]. The construction is mainly based on a decomposition lemma similar in spirit to a lemma which appeared in [9], which by itself is a generalization of the original probabilistic partitions of [7, 43]. However the proof that the new construction obeys the desired properties is quite technically involved and requires several new ideas that have not previously appeared.

We also give constructions of *uniformly padded hierarchical probabilistic partitions*. The idea is that these partitions are padded in a hierarchical manner – a much stronger demand than for only a single level partition. Although these are not strictly necessary for the proof of our main theorems they capture a *stronger* property of our partitions and play a central role in showing that arbitrary metric spaces embed in L_p with constant average distortion, while maintaining the best worst case distortion bounds. The embeddings in this paper demonstrate the versatility of these techniques and we expect that more applications will be found in the near future.

1.2 Low-Dimension Embeddings

Our new embeddings into L_p beat the previous embedding methods by achieving optimal dimension.

Recall that Bourgain proved that every n point metric space embeds into into L_p with $O(\log n)$ distortion. One of the most important parameters of the embedding into a normed space is the *dimension* of the embedding. This is of particular important in applications and has been the main object of study in the paper by Linial, London and Rabinovich [42] In particular, they ask: what is the dimension of the embedding in Theorem 1 ?

For embedding into Euclidean space, this can be answered by applying the Johnson and Lindenstrauss [32] dimension reduction lemma which states that any n-point metric space in L_2 can be embedded in Euclidean space of dimension $O(\log n)$ with constant distortion. This reduces the dimension in Bourgain's theorem to $O(\log n)$.

However, dimension reduction techniques *cannot* be used to generalize the low dimension bound to L_p for all p[2]. In particular, while every metric space embeds isometrically in L_∞ there are non-constant lower bounds on the distortion of embedding specific metric spaces into low dimensional L_∞ space [44].

This problem has been addressed by Linial, London, and Rabinovich [42] and separately by Matoušek [45] where they observe that the embedding given in Bourgain's proof of Theorem 1 can be used to bound the dimension of the embedding into L_p by $O(\log^2 n)$.

In this paper we prove the following:

THEOREM 2. *For any $1 \le p \le \infty$, every n-point metric space embeds in L_p with distortion $O(\log n)$ in dimension $O(\log n)$.*

In addition to the new embedding techniques discussed above the proof of Theorem 2 introduces a new trick of summing up the components of the embedding over all scales. This is in contrast to previous embeddings where such components were allocated separate coordinates. This saves us the extra logarithmic factor in dimension.

Moreover, with only a small price to pay in distortion, we provide an embedding into dimension $O(\log n/\log\log n)$:

THEOREM 3. *For any $1 \le p \le \infty$, and $\theta > 0$, every n-point metric space embeds in L_p with distortion $O(\log^{1+\theta} n)$ in dimension $O(\theta^{-1}\log n/\log\log n)$.*

The proof of Theorem 3 is considerably more involved and requires several more ideas, one of which, is that we make non-standard use of the padded decompositions in that we exploit a padding property with probability that may be very close to 1.

The bounds in theorems 2 and 3 are tight for the metric of an expander.

THEOREM 4. *For any fixed $1 \le p < \infty$ and any $\theta > 0$, if the metric of an n-node constant degree expander embeds into L_p with distortion $O(\log^{1+\theta} n)$ then the dimension of the embedding is $\Omega(\log n/\lceil\theta\log\log n\rceil)$.*

Matoušek extended Bouragin's proof to improve the distortion bound into L_p for large p to $O(\lceil\frac{\log n}{p}\rceil)$. He also showed this bound is tight [46]. The dimension obtained in Matoušek's analysis of the embedding into L_p is $e^{O(p)}\log^2 n$. Our methods extend to give the following improvement:

THEOREM 5. *For any $1 \le p \le \infty$ and any $1 \le k \le p$, every n-point metric space embeds in L_p with distortion $O(\lceil\log n/k\rceil)$ in dimension $e^{O(k)}\log n$.*

The bound on the dimension in Theorem 5 is nearly tight (up to lower order terms) as follows from volume arguments by Matoušek [44] (based on original methods of Bourgain [15]).

1.3 On the Average Distortion of Metric Embeddings

The $O(\log n)$ distortion guarantied by Bourgain's theorem is tight in the worst case. A somewhat weaker bound was already shown in Bourgain's paper and later Linial, London and Rabinovich [42] proved that embedding the metrics of constant-degree expander graphs into Euclidean space requires $\Omega(\log n)$ distortion.

Yet, this lower bound on the distortion is a *worst case* bound, i.e., it means that there *exists* a pair of points whose distortion is large. However, the *average case* is often more significant in terms of evaluating the quality of the embedding, in particular in relation to practical applications.

Formally, the *average distortion* of an embedding f is defined as: $\text{avgdist}(f) = \frac{1}{\binom{n}{2}}\sum_{u\neq v\in X}\text{dist}_f(u,v)$.

Indeed, in most real-world applications of metric embeddings *average distortion* and similar notions are used for evaluating the embedding's performance in practice, for example see [28, 29, 6, 27, 54, 55]. Moreover, in some cases it is desired that the average distortion would be small and the worst case distortion would still be reasonably bounded as well. While these papers provide some indication that such embeddings are possible in practice, the classic theory of metric embedding fails to address this natural question.

In particular, applying Bourgain's embedding to the metric of a constant-degree expander graph results in $\Omega(\log n)$ distortion for a *constant fraction* of the pairs[3].

In this paper we prove the following theorem which provides a qualitative strengthening of Bourgain's theorem:

THEOREM 6 (AVERAGE DISTORTION). *For every n-point metric space there exists an embedding into Euclidean space with distortion $O(\log n)$ and average distortion $O(1)$.*

In fact our results are even stronger. For $1 \le q \le \infty$, define the ℓ_q-*distortion* of an embedding f as:

$$\text{dist}_q(f) = \|\text{dist}_f(u,v)\|_q^{(\mathcal{U})} = \mathbb{E}[\text{dist}_f(u,v)^q]^{1/q},$$

where the expectation is taken according to the uniform distribution $\mathcal{U}$ over $\binom{X}{2}$. The classic notion of distortion is expressed by the ℓ_∞-distortion and the average distortion is expressed by the ℓ_1-distortion. Theorem 6 follows from the following:

THEOREM 7 (ℓ_q-DISTORTION). *For every n-point metric space (X,d) there exists an embedding f of X into Euclidean space such that for any $1 \le q \le \infty$, $\text{dist}_q(f) = O(\min\{q, \log n\})$.*

[2]For $1 \le p < 2$, a combination of lemmas of [32] and [23] can be used to obtain an embedding in dimension $O(\log n)$.

[3]Similar statements hold for the more recent metric embeddings of [52, 35] as well.

A variant of average distortion that is natural is what we call *distortion of average*: $\text{distavg}(f) = \frac{\sum_{u \neq v \in X} d_Y(f(u),f(v))}{\sum_{u \neq v \in X} d(u,v)}$, which can be naturally extended to its ℓ_q-normed extension termed *distortion of ℓ_q-norm*. Theorems 6 and 7 extend to those notions as well.

Besides $q = \infty$ and $q = 1$, the case of $q = 2$ provides a particularly natural measure. It is closely related to the notion of *stress* which is a standard measure in *multidimensional scaling* methods, invented by Kruskal [36] and later studied in many models and variants. Multidimensional scaling methods (see [37, 28]) are based on embedding of a metric representing the relations between entities into low dimensional space to allow feature extraction and are often used for indexing, clustering, nearest neighbor searching and visualization in many application areas including psychology and computational biology [29].

Previous work on average distortion. Related notions to the ones studied in this paper have been considered before in several theoretical papers. Most notably, Yuri Rabinovich [50] studied the notion of distortion of average[4] motivated by its application to the Sparsest Cut problem. This however places the restriction that the embedding is Lipschitz or *non-expansive*. Other recent papers have address this version of distortion of average and its extension to weighted average. In particular, it has been recently shown (see for instance [20]) that the work of Arora, Rao and Vazirani on Sparsest Cut [5] can be rephrased as an embedding theorem using these notions.

In his paper, Rabinovich observes that for Lipschitz embeddings the lower bound of $\Omega(\log n)$ still holds. It is therefore *crucial* in our theorems that the embeddings are *co-Lipschitz*[5] (a notion defined by Gromov [26]) (and w.l.o.g *non-contractive*).

To the best of our knowledge the only paper addressing such embeddings prior to this work is by Lee, Mendel and Naor [38] where they seek to bound the *average distortion* of embedding n-point L_1 metrics into Euclidean space. However, even for this special case they do not give a constant bound on the average distortion[6].

Network embedding. Our work is largely motivated by a surge of interest in the networking community on performing *passive distance estimation* (see e.g. [24, 48, 40, 18, 54, 17]), assigning nodes with short labels in such a way that the network latency between nodes can be approximated efficiently by extracting information from the labels without the need to incur active network overhead. The motivation for such labelling schemes are many emerging large-scale decentralized applications that require *locality awareness*, the ability to know the relative distance between nodes. For example, in peer-to-peer networks, finding the nearest copy of a file may significantly reduce network load, or finding the nearest server in a distributed replicated application may improve response time. One promising approach for distance labelling is *network embedding* (see [18]). In this approach nodes are assigned coordinates in a low dimensional Euclidean space. The node coordinates form simple and efficient *distance labels*. Instead of repeatedly measuring the distance between nodes, these labels allow to extract an approximate measure of the latency between nodes. Hence these network coordinates can be used as an efficient building block for locality aware networks that significantly reduce network load.

As mentioned above the natural measure of efficiency in the networking research is how the embedding performs on average, where the notion of average distortion comes in several variations can be phrased in terms of the definitions given above. The phenomenon observed in measurements of network distances is that the average distortion of network embeddings was bounded by a small constant. Our work gives the *first* full theoretical explanation for this intriguing phenomenon.

Embedding with relaxed guaranties. The theoretical study of such phenomena was initiated by the work of Kleinberg, Slivkins and Wexler [34]. They mainly focus on the fact reported in the networking papers that the distortion of almost all pairwise distances is bounded by some small constant. In an attempt to provide theoretical justification for such phenomena [34] define the notion of a $(1-\epsilon)$-partial embedding[7] where the distortion is bounded for at least some $(1-\epsilon)$ fraction of the pairwise distances. They obtained some initial results for metrics which have constant doubling dimension [34]. In Abraham et. al. [1] is was shown that any finite metric space has a $(1-\epsilon)$-partial embedding into Euclidean space with $O(\log \frac{1}{\epsilon})$ distortion.

While this result is very appealing it has the disadvantage of lacking any promise for some fraction of the pairwise distances. This may be critical for applications - that is we really desire an embedding which in a sense does *"as well as possible"* for all distances. To question whether such an embedding exists [34] define a stronger notion of *scaling distortion*[8]. An embedding has scaling distortion of $\alpha(\epsilon)$ if it provides this bound on the distortion of a $(1-\epsilon)$ fraction of the pairwise distances, *for any* ϵ. In [34], such embeddings with $\alpha(\epsilon) = O(\log \frac{1}{\epsilon})$ were shown for metrics of bounded growth dimension, this was extended in [1] to metrics of bounded doubling dimension. In addition [1] give a rather simple probabilistic embedding with scaling distortion, implying an embedding into (high-dimensional) L_1.

The most important question arising from the work of [34, 1] is whether embeddings with small scaling distortion exist for embedding into Euclidean space. We give the following theorem[9] which lies in the heart of the proof of Theorem 7:

THEOREM 8. *For every finite metric space (X,d), there exists an embedding of X into Euclidean space with scaling distortion $O(\log \frac{1}{\epsilon})$.*

Techniques. While [1] certainly uses the state of the art methods in finite metric embedding, it appears these techniques break when attempting to prove Theorem 8. Indeed, to prove the theorem and its generalizations we present novel embedding techniques.

Our embeddings are based on the new decomposition based methods described in Section 1.1. Here we make a more sophisticated use of these techniques. This is not surprising

[4] Usually this notion was called average distortion but the name is somewhat confusing.

[5] This notion is used here somewhat differently than its original purpose.

[6] The bound given in [38] is $O(\sqrt{\log n})$ which applies to a somewhat weaker notion.

[7] Called "embeddings with ϵ-slack" in [34].

[8] Called "gracefully degrading distortion" in [34].

[9] In fact in this theorem the definition of scaling distortion is even stronger. This is explained in detail in the appropriate section.

given that here we desire to obtain distortions which depend solely on ϵ rather than on n. This requires clever ways of defining the embeddings so that the contribution would be limited as a function of ϵ. In particular, in our results for large p, in addition to the decomposition based embedding, a second component of our embedding uses a similar approach to that of Bourgain's original embedding. However using it in straightforward manner is impossible. It is here that we crucially rely on the hierarchical structure of our decompositions in order to do this in a way that will allow us to bound the contribution appropriately.

Additional Results and Applications. In addition to our main result, our paper contains several other contributions: we extend the results on average distortion to weighted averages. We show the bound is $O(\log \Phi)$ where Φ is the effective aspect ratio of the weight distribution. We also obtain average distortion results for embeddings into ultrametrics. In addition we provide a solution for another open problem from [34, 1] regarding partial embedding into trees.

Finally, we demonstrate some basic algorithmic applications of our theorems, mostly due to their extensions to general weighted averages. Among others is an application to *uncapacitated quadratic assignment* [49, 33]. We also extend our concepts to analyze Distance Oracles of Thorup and Zwick [56] providing results with strong relation to the questions addressed by [34]. We however feel that our current applications do not make full use of the strength of our theorems and techniques and it remains to be seen if such applications will arise.

1.3.1 ℓ_q-Distortion and the Main Theorem

Given two metric spaces (X, d_X) and (Y, d_Y) an *injective* mapping $f : X \to Y$ is called an *embedding* of X into Y. In what follows we define novel notions of distortion. In order to do that we start with the definition of the classic notion.

An embedding f is called c-co-Lipschitz [26] if for any $u \neq v \in X$: $d_Y(f(u), f(v)) \geq c \cdot d_X(u, v)$ and *non-contractive* if $c = 1$. In the context of this paper we will restrict attention to co-Lipschitz embeddings, which due to scaling may be further restricted to *non-contractive* embeddings. This has no difference for the classic notion of distortion but has a crucial role for the results presented in this paper. We will elaborate more on this issue in the sequel.

For a non-contractive embedding define the distortion function of f, $\text{dist}_f : \binom{X}{2} \to \mathbb{R}^+$, where for $u \neq v \in X$: $\text{dist}_f(u,v) = \frac{d_Y(f(u), f(v))}{d_X(u,v)}$. The distortion of f is defined as $\text{dist}(f) = \sup_{u \neq v \in X} \text{dist}_f(u, v)$.

Definition 1 (ℓ_q-Distortion). *Given a distribution* Π *over* $\binom{X}{2}$ *define for* $1 \leq q \leq \infty$ *the* ℓ_q-distortion *of* f *with respect to* Π*:*

$$\text{dist}_q^{(\Pi)}(f) = \|\text{dist}_f(u,v)\|_q^{(\Pi)} = \mathbb{E}_\Pi[\text{dist}_f(u,v)^q]^{1/q},$$

where $\|\cdot\|_q^{(\Pi)}$ *denotes the* normalized q norm *over the distribution* (Π)*, defined as in the equation above. Let* $\mathcal{U}$ *denote the uniform distribution over* $\binom{X}{2}$*. The* ℓ_q-distortion *of* f *is defined as:* $\text{dist}_q(f) = \text{dist}_q^{(\mathcal{U})}(f)$.

In particular the classic distortion may be viewed as the ℓ_∞-distortion: $\text{dist}(f) = \text{dist}_\infty(f)$. An important special case of ℓ_q-distortion is when $q = 1$:

Definition 2 (Average Distortion). *Given a distribution* Π *over* $\binom{X}{2}$ *define for* $1 \leq q \leq \infty$ *the* averagedistortion *of* f *with respect to* Π *is defined as:* $\text{avgdist}^{(\Pi)}(f) = \text{dist}_1^{(\Pi)}(f)$*, and the* average distortion *of* f *is given by:* $\text{avgdist}(f) = \text{dist}_1(f)$.

Another natural notion is the following:

Definition 3 (Distortion of ℓ_q-Norm). *Given a distribution* Π *over* $\binom{X}{2}$ *define the* distortion of ℓ_q-norm *of* f *with respect to* Π*:*

$$\text{distnorm}_q^{(\Pi)}(f) = \frac{\mathbb{E}_\Pi[d_Y(f(u), f(v))^q]^{1/q}}{\mathbb{E}_\Pi[d_X(u,v)^q]^{1/q}},$$

and let $\text{distnorm}_q(f) = \text{distnorm}_q^{(\mathcal{U})}(f)$.

Again, an important special case of distortion of ℓ_q-norm is when $q = 1$:

Definition 4 (Distortion of Average). *Given a distribution* Π *over* $\binom{X}{2}$ *define the* distortion of average *of* f *with respect to* Π *as:* $\text{distavg}^{(\Pi)}(f) = \text{distnorm}_1^{(\Pi)}(f)$ *and the* distortion of average *of* f *is given by:* $\text{distavg}(f) = \text{distnorm}_1(f)$.

For simplicity of the presentation of our main results we use the following notation:
$\text{dist}_q^{*(\Pi)}(f) = \max\{\text{dist}_q^{(\Pi)}(f), \text{distnorm}_q^{(\Pi)}(f)\}$, $\text{dist}_q^*(f) = \max\{\text{dist}_q(f), \text{distnorm}_q(f)\}$, and
$\text{avgdist}^*(f) = \max\{\text{avgdist}(f), \text{distavg}(f)\}$.

Definition 5. *A probability distribution* Π *over* $\binom{X}{2}$*, with probability function* $\pi : \binom{X}{2} \to [0,1]$*, is called* non-degenerate *if for every* $u \neq v \in X$*:* $\pi(u,v) > 0$*. The* aspect ratio *of a non-degenerate probability distribution* Π *is defined as:*

$$\Phi(\Pi) = \frac{\max_{u \neq v \in X} \pi(u,v)}{\min_{u \neq v \in X} \pi(u,v)}.$$

In particular $\Phi(\mathcal{U}) = 1$*. If* Π *is* not *non-degenerate then* $\Phi(\Pi) = \infty$.

For an arbitrary *probability distribution* Π *over* $\binom{X}{2}$*, define its* effective aspect ratio *as:*[10] $\hat{\Phi}(\Pi) = 2\min\{\Phi(\Pi), \binom{n}{2}\}$.

Theorem 9 (Embedding into L_p). *Let* (X, d) *an* n*-point metric space, and let* $1 \leq p \leq \infty$*. There exists an embedding* f *of* X *into* L_p *in dimension* $e^{O(p)} \log n$*, such that for every* $1 \leq q \leq \infty$*, and any distribution* Π *over* $\binom{X}{2}$*:* $\text{dist}_q^{*(\Pi)}(f) = O(\min\{q, \log n\}/p + \log \hat{\Phi}(\Pi))$*. In particular,* $\text{avgdist}^{*(\Pi)}(f) = O(\log \hat{\Phi}(\Pi))$*. Also:* $\text{dist}(f) = O(\lceil \log n/p \rceil)$*,* $\text{dist}_q^*(f) = O(\lceil q/p \rceil)$ *and* $\text{avgdist}^*(f) = O(1)$.

We show that all the bounds in the theorem above are tight.

In the full paper we also give a stronger version of the bounds for decomposable metrics. Recall that metric spaces (X, d) can be characterized by their decomposability parameter τ_X where it is known that $\tau_X = O(\log \lambda_X)$, where λ_X is the doubling constant of X, and for metrics of $K_{s,s}$-excluded minor graphs. $\tau_X = O(s^2)$. For metrics with a bounded decomposability parameter we extend Theorem 9 by showing

[10] The factor of 2 in the definition is placed solely for the sake of technical convenience.

an embedding with $\mathrm{dist}_q^{*(\Pi)}(f) = O(\min\{q, (\log \lambda_X)^{1-\frac{1}{p}} (\log n)^{1/p}\} + \log \hat{\Phi}(\Pi))$.

The proof of Theorem 9 follows directly from results on embedding with scaling distortion, discussed in the next paragraph.

1.3.2 Partial Embedding, Scaling Distortion and Additional Results

Following [34] we define:

Definition 6 (Partial Embedding). *Given two metric spaces (X, d_X) and (Y, d_Y), a* partial embedding *is a pair (f, G), where f is a non-contractive embedding of X into Y, and $G \subseteq \binom{X}{2}$. The distortion of (f, G) is defined as:* $\mathrm{dist}(f, G) = \sup_{\{u,v\} \in G} \mathrm{dist}_f(u, v)$.

For $\epsilon \in [0, 1)$, a $(1-\epsilon)$-partial embedding *is a partial embedding such that $|G| \geq (1-\epsilon)\binom{n}{2}$.*[11]

Next, we would like to define a special type of $(1-\epsilon)$-partial embeddings. For this aim we need a few more definitions. Let $r_\epsilon(x)$ denote the minimal radius r such that $|B(x,r)|/n \geq \epsilon$. Let $\hat{G}(\epsilon) = \{\{x, y\} \in \binom{X}{2} \mid d(x, y) \geq \max\{r_{\epsilon/2}(x), r_{\epsilon/2}(y)\}\}$.

A coarsely *$(1-\epsilon)$-partial embedding f is a partial embedding $(f, \hat{G}(\epsilon))$*[12].

Definition 7 (Scaling Distortion). *Given two metric spaces (X, d_X) and (Y, d_Y) and a function $\alpha : [0, 1) \to \mathbb{R}^+$, we say that an embedding $f : X \to Y$ has* scaling distortion *α if for any $\epsilon \in [0, 1)$, there is some set $G(\epsilon)$ such that $(f, G(\epsilon))$ is a $(1-\epsilon)$-partial embedding with distortion at most $\alpha(\epsilon)$. We say that f has* coarsely *scaling distortion if for every ϵ, $G(\epsilon) = \hat{G}(\epsilon)$.*

We can extend the notions of partial probabilistic embeddings and scaling distortion to probabilistic embeddings. For simplicity we will restrict to coarsely partial embeddings.[13]

Definition 8 (Partial/Scaling Prob. Embedding). *Given (X, d_X) and a set of metric spaces $\mathcal{S}$, for $\epsilon \in [0, 1)$, a* coarsely $(1-\epsilon)$-partial probabilistic embedding *consists of a distribution $\hat{\mathcal{F}}$ over a set $\mathcal{F}$ of* coarsely *$(1-\epsilon)$-partial embeddings from X into $Y \in \mathcal{S}$. The distortion of $\hat{\mathcal{F}}$ is defined as:* $\mathrm{dist}(\hat{\mathcal{F}}) = \sup_{\{u,v\} \in \hat{G}(\epsilon)} \mathbb{E}_{(f, \hat{G}(\epsilon)) \sim \hat{\mathcal{F}}}[\mathrm{dist}_f(u, v)]$.

The notion of scaling distortion is extended to probabilistic embedding in the obvious way.

We observe the following relation between partial embedding, scaling distortion and the ℓ_q-distortion.

Lemma 1 (Scaling Distortion vs. ℓ_q-Distortion). *Given an n-point metric space (X, d_X) and a metric space (Y, d_Y). If there exists an embedding $f : X \to Y$ with scaling distortion α then for any distribution Π over $\binom{X}{2}$:*[14]

$$\mathrm{dist}_q^{(\Pi)}(f) \leq \left(2 \int_{\frac{1}{2}\binom{n}{2}^{-1}\hat{\Phi}(\Pi)}^{1} \alpha(x \hat{\Phi}(\Pi)^{-1})^q dx \right)^{1/q} + \alpha(\hat{\Phi}(\Pi)^{-1}).$$

In the case of coarsely *scaling distortion this bound holds for $\mathrm{dist}_q^{*(\Pi)}(f)$.*

Combined with the following theorem we obtain Theorem 9. We note that when applying the lemma we use $\alpha(\epsilon) = O(\log \frac{1}{\epsilon})$ and the bounds in the theorem mentioned above follow from bounding the corresponding integral.

Theorem 10 (Scaling Distortion Theorem into L_p). *Let $1 \leq p \leq \infty$. For any n-point metric space (X, d) there exists an embedding $f : X \to L_p$ with coarsely scaling distortion $O(\lceil (\log \frac{1}{\epsilon})/p \rceil)$ and dimension $e^{O(p)} \log n$.*

For metrics with a decomposability parameter τ_X the distortion improves to: $O(\min\{\tau_X^{1-1/p} (\log \frac{1}{\epsilon})^{1/p}, \log \frac{1}{\epsilon}\})$.

Applying the lemma on the probabilistic embedding into ultrametrics with scaling distortion $O(\log \frac{1}{\epsilon})$ of [1] we obtain:

Theorem 11 (Prob. Embedding into Ultrametrics). *Let (X, d) an n-point metric space. There exists a probabilistic embedding $\hat{\mathcal{F}}$ of X into ultrametrics, such that for every $1 \leq q \leq \infty$, and any distribution Π over $\binom{X}{2}$: $\mathrm{dist}_q^{*(\Pi)}(\hat{\mathcal{F}}) = O(\min\{q, \log n\} + \log \hat{\Phi}(\Pi))$.*

For $q = 1$ and for a given fixed distribution Theorem 11 can be given a classic embedding (deterministic) version:

Theorem 12 (Embedding into Ultrametrics). *Given an arbitrary fixed distribution Π over $\binom{X}{2}$, for any finite metric space (X, d) there exists embeddings f, f' into ultrametrics, such that $\mathrm{avgdist}^{(\Pi)}(f) = O(\log \hat{\Phi}(\Pi))$ and $\mathrm{distavg}^{(\Pi)}(f') = O(\log \hat{\Phi}(\Pi))$.*

The results in [1] leave as an open question the distortion of partial embedding into an ultrametric. While the upper bound which follows from [1] by applying [7, 14] is $O(\frac{1}{\epsilon})$, the lower bound which follows from [1] by applying the $\Omega(n)$ lower bound on embedding into trees of [51] is only $\Omega(\frac{1}{\sqrt{\epsilon}})$.

We show that the correct answer to this question is the latter bound which is achievable by embedding into *ultrametrics*. In fact we can obtain embeddings into low-degree k-HSTs [7]. This result is related both in spirit and techniques to recently developed metric Ramsey theorems [11, 14] where embeddings into ultrametrics and k-HSTs play a central role.

Theorem 13 (Partial Embedding into Ultrametrics). *For every n-point metric space (X, d) and any $\epsilon \in (0, 1)$ there exists a $(1-\epsilon)$ partial embedding into an ultrametric with distortion $O(\frac{1}{\sqrt{\epsilon}})$.*

1.3.3 Algorithmic Applications

We demonstrate some basic applications of our main theorems. We must stress however that our current applications do not use the full strength of these theorems. Most of our applications are based on the bound given on the *distortion*

[11] Note that the embedding is *strictly* partial only if $\epsilon \geq 1/\binom{n}{2}$.

[12] It is elementary to verify that indeed this defines a $(1-\epsilon)$-partial embedding. We also note that in most of the proofs we can use a min rather than max in the definition of $\hat{G}(\epsilon)$. However, this definition seems more natural and of more general applicability.

[13] Our upper bounds use this definition, while our lower bounds hold also for the non-coarsely case.

[14] Assuming the integral is defined. We note that lemma is stated using the integral for presentation reasons.

of average for general distributions of embeddings f into L_p and into ultrametrics with $\text{distavg}^{(\Pi)}(f) = O(\log \hat{\Phi}(\Pi))$. In some of these applications it is crucial that the result holds for all such distributions Π This is useful for problems which are defined with respect to weights $c(u, v)$ in a graph or in a metric space, where the solution involves minimizing the sum over distances weighted according to c. This is common for many optimization problem either as part of the objective function or alternatively it may come up in the linear programming relaxation of the problem. These weights can be normalized to define the distribution Π. Using this paradigm we obtain $O(\log \hat{\Phi}(c))$ approximation algorithms, improving on the general bound which depends on n in the case that $\hat{\Phi}(c)$ is small. This is the *first* result of this nature.

We are able to obtain such results for the following group of problems: *general sparsest cut* [39, 4, 42, 5, 3], *multi cut* [25], *minimum linear arrangement* [19, 53], *embedding in d-dimensional meshes* [19, 9], *multiple sequence alignment* [57] and *uncapacitated quadratic assignment* [49, 33].

We would like to emphasize that the notion of bounded weights is in particular natural in the last application mentioned above. The problem of *uncapacitated quadratic assignment* is one of the most basic problems in operations research (see the survey [49]) and has been one of the main motivations for the work of Kleinberg and Tardos on metric labelling [33].

We also give a different use of our results for the problem of *min-sum k-clustering* [12].

1.3.4 Distance Oracles

Thorup and Zwick [56] study the problem of creating *distance oracles* for a given metric space. A distance oracle is a space efficient data structure which allows efficient queries for the approximate distance between pairs of points.

They give a distance oracle of space $O(kn^{1+1/k})$, query time of $O(k)$ and *worst case* distortion (also called stretch) of $2k - 1$. They also show that this is nearly best possible in terms of the space-distortion tradeoff.

We extend the new notions of distortion in the context of distance oracles. In particular, we can define the ℓ_q-distortion of a distance oracle. Of particular interest are the average distortion and distortion of average notion. We also define partial distance oracles, distance oracle scaling distortion, and extend our results to distance labels and distributed labeled compact routing schemes in a similar fashion. Our main result is the following strengthening of [56]:

THEOREM 14. *Let (X, d) be a finite metric space. Let $k = O(\ln n)$ be a parameter. The metric space can be preprocessed in polynomial time, producing a data structure of $O(n^{1+1/k} \log n)$ size, such that distance queries can be answered in $O(k)$ time. The distance oracle has worst case distortion $2k - 1$. Given any distribution Π, its average distortion (and distortion of average) with respect to Π is $O(\log \hat{\Phi}(\Pi))$. In particular the average distortion (and distortion of average) is $O(1)$.*

1.4 Organization of the Paper

In Section 3 we define the new probabilistic partitions. The constructions are described in Section 3.1. In Section 4 we present the proof of Theorem 2, providing an embedding in $O(\log n)$ dimension. We also give its extension to scaling distortion which implies $O(1)$ average distortion, as stated in Theorem 6. Section 5 includes the proof of Theorem 3, providing better bounds on the dimension. Section 6 provides better scaling distortion embeddings for large p, as stated in Theorem 5. Finally, we describe the relations between scaling distortion, partial embedding and the ℓ_q-distortion in Section 7.

2. 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 form the context.

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

DEFINITION 9. *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 of $x \in X$ at radius $r > 0$ and scale $\gamma > 0$ is defined as $\bar{\rho}(x, r, \gamma) = \rho(B(x, r), r, \gamma)$.

We make use of the following simple fact (the proof is omitted).

CLAIM 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*

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

3. PROBABILISTIC PARTITIONS

DEFINITION 10 (PARTITION). *Let (X, d) be a finite metric space. A partition P of X is a collection of disjoint sets $\mathcal{C}(P) = \{C_1, C_2, \ldots, C_t\}$ such that $X = \cup_j C_j$. The sets $C_j \subseteq X$ are called clusters. For $x \in X$ we denote by $P(x)$ the cluster containing x. Given $\Delta > 0$, a partition is Δ-*bounded *if for all $1 \leq j \leq t$, $\text{diam}(C_j) \leq \Delta$.*

DEFINITION 11 (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 12 (HIERARCHICAL PARTITION). *Fix some integer $L > 0$. Let $I = \{0 \leq i \leq L | i \in \mathbb{Z}\}$. A* hierarchical partition *P of a finite metric space (X, d) is a hierarchical collection of partitions $\{P_i\}_{i \in I}$ where P_0 consists of a single cluster equal to X and for any $0 < i \in I$ and $x \in X$, $P_i(x) \subseteq P_{i-1}(x)$. Given $k > 1$, let $L = \lceil \log_k(\text{diam}(X)) \rceil$ and set $\Delta_0 = \text{diam}(X)$, and for each $0 < i \in I$, $\Delta_i = \Delta_{i-1}/k$. We say that P is k-hierarchical if for each $i \in I$, $P_i \in P$, P_i is Δ_i-bounded.*

DEFINITION 13 (PROB. HIERARCHICAL PARTITION). *A* probabilistic k-hierarchical partition *$\hat{\mathcal{H}}$ of a finite metric space (X, d) consists of a probability distribution over a set $\mathcal{H}$ of k-hierarchical partitions.*

A collection of functions defined on X, $f = \{f_{P,i} | P \in \mathcal{H}, i \in I\}$ is uniform *with respect to $\mathcal{H}$ if for every $P \in \mathcal{H}$ and $i \in I$, $f_{P,i}$ is uniform with respect to P_i.*

Definition 14 (Uniformly Padded PHP). *Let $\hat{\mathcal{H}}$ be a probabilistic k-hierarchical partition. Given collection of functions $\eta = \{\eta_{P,i} : X \to [0,1] | i \in I, P_i \in P, P \in \mathcal{H}\}$ and $\delta \in (0,1]$, $\hat{\mathcal{H}}$ is called* (η, δ)-padded *if the following condition holds for all $i \in I$ and for any $x \in X$:*

$$\Pr[B(x, \eta_{P,i}(x)\Delta_i) \subseteq P_i(x)] \geq \delta.$$

We say $\hat{\mathcal{H}}$ is uniformly padded *if η is uniform with respect to $\mathcal{H}$.*

We now present the main lemma on the existence of hierarchical partitions which are the main building block of our embeddings.

Lemma 3 (Hierarchical Uniform Padding Lemma). *Let $\Gamma = 64$. Let $\delta \in (0, \frac{1}{2}]$. Given a finite metric space (X,d), there exists a probabilistic 4-hierarchical partition $\hat{\mathcal{H}}$ of (X,d) and a uniform collection of functions $\xi = \{\xi_{P,i} : X \to \{0,1\} | P \in \mathcal{H}, i \in I\}$, such that for the collection of functions η, defined below, we have that $\hat{\mathcal{H}}$ is (η,δ)-uniformly padded, and the following properties hold for any $P \in \mathcal{H}$, $0 < i \in I$, $P_i \in P$:*

- $\sum_{j \leq i} \xi_{P,j}(x) \eta_{P,j}(x)^{-1} \leq 2^{10} \ln \left(\frac{n}{|B(x, \Delta_{i+4})|} \right) / \ln(1/\delta)$.
- *If $\xi_{P,i}(x) = 1$ then: $\eta_{P,i}(x) \leq 2^{-7}$.*
- *If $\xi_{P,i}(x) = 0$ then: $\eta_{P,i}(x) \geq 2^{-7}$ and $\bar{\rho}(x, \Delta_{i-1}, \Gamma) < 1/\delta$.*

The proof of Lemma 3 is described below. The rest of the paper is based only on the statement of the lemma.

3.1 The Constructions

In this section we provide details on the proof of Lemma 3.

The main building block in the proof of the lemma is a lemma about uniformly padded probabilistic partitions.

Definition 15 (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 Δ-bounded if each $P \in \mathcal{P}$ is Δ-bounded.*

Definition 16 (Uniformly Padded PP). *Given $\Delta > 0$, let $\hat{\mathcal{P}}$ be a Δ-bounded probabilistic partition of (X,d). Given collection of functions $\eta = \{\eta_P : X \to [0,1] | P \in \mathcal{P}\}$ and $\delta \in (0,1]$, is called* (η,δ)-padded *if the following condition holds for any $x \in X$:*

$$\Pr[B(x, \eta_P(x)\Delta) \subseteq P(x)] \geq \delta.$$

We say $\hat{\mathcal{P}}$ is uniformly padded *if for any $P \in \mathcal{P}$ the function η_P is uniform with respect to P.*

Lemma 4 (Uniform Padding Lemma). *Let (X,d) be a finite metric space. Let $Z \subseteq X$. Let $\bar{\Delta}$ be such that $\mathrm{diam}(Z) \leq \bar{\Delta}$. Let Δ be such that $\Delta \leq \bar{\Delta}/4$ and let Γ be such that $\Gamma \geq 4\bar{\Delta}/\Delta$. Let $\hat{\delta} \in (0, \frac{1}{2}]$. There exists a Δ-bounded probabilistic partition $\hat{\mathcal{P}}$ of (Z,d) and a collection of uniform functions $\{\xi_P : X \to \{0,1\} | P \in \mathcal{P}\}$ and $\{\hat{\eta}_P : X \to \{0, 1/\ln(1/\hat{\delta})\} | P \in \mathcal{P}\}$, such that for any $\hat{\delta} \leq \delta \leq 1$, and $\eta^{(\delta)}$ defined by $\eta_P^{(\delta)}(x) = \hat{\eta}_P(x) \ln(1/\delta)$, the probabilistic partition $\hat{\mathcal{P}}$ is $(\eta^{(\delta)}, \delta)$-uniformly padded, and the following conditions hold for any $P \in \mathcal{P}$ and any $x \in Z$:*

- *If $\xi_P(x) = 1$ then: $2^{-6}/\ln \rho(x, \bar{\Delta}, \Gamma) \leq \hat{\eta}_P(x) \leq 2^{-6}/\ln(1/\hat{\delta})$.*
- *If $\xi_P(x) = 0$ then: $\hat{\eta}_P^{(\delta)}(x) = 2^{-6}/\ln(1/\hat{\delta})$ and $\bar{\rho}(x, \bar{\Delta}, \Gamma) < 1/\hat{\delta}$.*

The proof of Lemma 4 is based on a the following technical lemma, non-trivial generalization of arguments of [7, 9], which lies in the heart of the constructions.

Given a finite metric space (X,d), and a subspace $Z \subseteq X$, we define a decomposition of Z into $(S, \bar{S})$, where $S \subseteq Z$. For some given parameter $0 < \Delta < \mathrm{diam}(Z)$, the decomposition creates a cluster S of diameter in the range $[\Delta/2, \Delta]$ as follows. For short we use the notation $A \bowtie (S, \bar{S})$ to abbreviate $A \cap S \neq \emptyset$ and $A \cap \bar{S} \neq \emptyset$.

Lemma 5 (Probabilistic Decomposition). *Let (X,d) be a metric space and $Z \subseteq X$. let $\chi \geq 2$ be a parameter. Given $0 < \Delta < \mathrm{diam}(Z)$ and a center point $v \in Z$, there exists a probability distribution over partitions $(S, \bar{S})$ of Z such that $S = B_Z(v,r)$, and r is chosen from a probability distribution in the interval $[\Delta/4, \Delta/2]$, such that for any $\theta \in (0,1)$ satisfying $\theta \geq \chi^{-1}$, let $\eta = \frac{1}{16} \ln(1/\theta)/\ln \chi$ then for any $x \in Z$, the following holds:*

$$\Pr[B_Z(x, \eta\Delta) \bowtie (S, \bar{S})] \leq (1-\theta) \left[\Pr[B_Z(x, \eta\Delta) \nsubseteq \bar{S}] + 2\chi^{-2} \right].$$

Proof. Let $R = \Delta/8$. Choose a radius r in the interval $[2R, 4R]$ according to the distribution $p(r) = (\frac{\chi^2}{1-\chi^{-2}}) \frac{\ln \chi}{R} \chi^{-\frac{r}{R}}$. Now define a partition $(S, \bar{S})$ where $S = B_Z(v,r)$ and $\bar{S} = Z \setminus S$.

Let y and z be the nearest and farthest points to v in $B_Z(x, \eta\Delta)$, respectively. We have:

$$\Pr[B_Z(x, \eta\Delta) \bowtie (S, \bar{S})] = \int_{d(v,y)}^{d(v,z)} p(r) dr = (\tfrac{\chi^2}{1-\chi^{-2}}) \chi^{-\frac{d(v,y)}{R}} (1 - \chi^{-\frac{d(v,z)-d(v,y)}{R}}) \leq (\tfrac{\chi^2}{1-\chi^{-2}}) \chi^{-\frac{d(v,y)}{R}} (1-\theta), \quad (1)$$

which follows since:

$$\frac{d(v,z) - d(v,y)}{R} \leq \frac{d(y,z)}{R} \leq \frac{2\eta\Delta}{R} = 16\eta = \ln_\chi(1/\theta).$$

$$\Pr[B_Z(x, \eta\Delta) \nsubseteq \bar{S}] = \int_{d(v,y)}^{4R} p(r) dr = (\tfrac{\chi^2}{1-\chi^{-2}})(\chi^{-\frac{d(v,y)}{R}} - \chi^{-4}). \quad (2)$$

Therefore we have:

$$\Pr[B_Z(x, \eta\Delta) \bowtie (S, \bar{S})] - (1-\theta) \cdot \Pr[B_Z(x, \eta\Delta) \nsubseteq \bar{S}] \leq (1-\theta)(\tfrac{\chi^2}{1-\chi^{-2}}) \chi^{-4} \leq (1-\theta) \cdot 2\chi^{-2},$$

where in the last inequality we have used the assumption that $\chi > 2$. This completes the proof of the lemma. □

We are now ready to prove Lemma 4:

Proof of Lemma 4. We generate a probabilistic partition $\hat{\mathcal{P}}$ of Z by invoking the probabilistic decomposition

Lemma 5 iteratively. Define the partition P of Z into clusters by generating a sequence of clusters: $C_1, C_2, \dots C_t$. Notice that we are generating a distribution over partitions and therefore t and the generated clusters are random variables.

Let $Z_1 = Z$. The clusters are created as follows: For $j \geq 1$, let v_j be a random variable equal to the point $x \in Z_j$ minimizing $\rho(x, \bar{\Delta}, \Gamma)$. Define the random variables $\hat{\chi}_j = \rho(v_j, \bar{\Delta}, \Gamma) = \rho(Z_j, \bar{\Delta}, \Gamma)$ and $\chi_j = \max\{\hat{\chi}_j, 2/\hat{\delta}^{1/2}\}$.

Throughout the analysis fix some $\delta \geq \hat{\delta}$ and let $\theta = \delta^{1/2}$. Note that $\theta \geq 2\chi_j^{-1}$ as required.

Invoke Lemma 5 to construct a probabilistic decomposition for Z_j with $v = v_j$ as center and the parameter $\chi = \chi_j$ defined above. Recall that $\eta_j = 2^{-4}\ln(1/\theta)/\ln \chi_j = 2^{-5}\ln(1/\delta)/\ln \chi_j$. This results in a partition $(S_{v_j}, \bar{S}_{v_j})$ where v_j is the center of $S_{v_j} = B_{Z_j}(v_j, r_j)$, for $r_j \in [\Delta/4, \Delta/2]$ chosen from the distribution defined in Lemma 5. Set $C_j = S_{v_j}$, $Z_{j+1} = Z_j \setminus S_{v_j}$. For every $x \in S_{v_j}$ (i.e., $P(x) = S_{v_j}$), define: $\hat{\eta}_P(x) = 2^{-6}/\max\{\ln \hat{\chi}_j, \ln(1/\hat{\delta})\}$ and let $\eta_P^{(\delta)}(x) = \hat{\eta}_P(x)\ln(1/\delta)$ (it is easy to verify that $\eta_P^{(\delta)}(x) \leq \eta_j$). If $\hat{\chi}_j \geq 1/\hat{\delta}$ let $\xi_P(x) = 1$ otherwise $\xi_P(x) = 0$. Clearly, ξ_P and η_P are uniform functions with respect to P. Continue the process while $Z_{j+1} \neq \emptyset$, otherwise $t = j$.

We now prove the properties in the lemma for some $x \in Z$. We will consider the distribution over partitions of Z into clusters $C_1, C_2, \dots C_t$ as defined above. For $1 \leq m \leq t$, define the events:

$$\begin{aligned} \mathcal{Z}_m &= \{\forall j, 1 \leq j < m,\ B_Z(x, \eta_j\Delta) \subseteq Z_{j+1}\}, \\ \mathcal{E}_m &= \{\exists j,\ m \leq j < t \text{ s.t. } B_Z(x, \eta_j\Delta) \bowtie (S_{v_j}, \bar{S}_{v_j}) | \mathcal{Z}_m\}. \end{aligned}$$

Also let $T = B(x, 2\Delta)$. We prove the following inductive claim: For every $1 \leq m \leq t$:

$$\Pr[\mathcal{E}_m] \leq (1-\theta)(1 + \theta\mathbb{E}[\sum_{j \geq m, v_j \in T} \chi_j^{-1} | \mathcal{Z}_m]). \qquad (3)$$

Note that $\Pr[\mathcal{E}_t] = 0$. Assume the claim holds for $m+1$ and we will prove for m. Define the events:

$$\begin{aligned} \mathcal{F}_m &= \{B_Z(x, \eta_m\Delta) \bowtie (S_{v_m}, \bar{S}_{v_m}) | \mathcal{Z}_m\}, \\ \mathcal{G}_m &= \{B_Z(x, \eta_m\Delta) \subseteq \bar{S}_{v_m} | \mathcal{Z}_m\} = \{\mathcal{Z}_{m+1} | \mathcal{Z}_m\}. \end{aligned}$$

First we bound $\Pr[\mathcal{F}_m]$. Assume first a particular choice of the clusters $C_1, \dots C_{m-1}$ such that event $\mathcal{Z}_m$ occurs. Call this specific event $\mathcal{A}$, then given that $\mathcal{A}$ occurred the center v_m of C_m is now determined deterministically, and so is the value of χ_j. Notice that since $r_j \leq \Delta$, if $B_Z(x, \eta_m\Delta) \bowtie (S_{v_m}, \bar{S}_{v_m})$ then $d(v_m, x) \leq \Delta + \eta_m\Delta \leq 2\Delta$, and thus $v_m \in T$. Now, applying Lemma 5 we get

$$\begin{aligned} &\Pr[B_Z(x, \eta_m\Delta) \bowtie (S_{v_m}, \bar{S}_{v_m}) | \mathcal{A}] \leq \\ &\quad (1-\theta)(\Pr[B_Z(x, \eta_m\Delta) \nsubseteq \bar{S}_{v_m} | \mathcal{A}] + \mathbf{1}_T(v_m)\theta\chi_m^{-1}. \end{aligned}$$

It follows that

$$\Pr[\mathcal{F}_m] \leq (1-\theta)(\Pr[\bar{\mathcal{G}}_m] + \theta\mathbb{E}[\mathbf{1}_T(v_m)\chi_m | \mathcal{Z}_m]).$$

Using the induction hypothesis we prove the inductive claim:

$$\begin{aligned} \Pr[\mathcal{E}_m] &\leq \Pr[\mathcal{F}_m] + \Pr[\mathcal{G}_m]\Pr[\mathcal{E}_{m+1}] \\ &\leq (1-\theta)(\Pr[\bar{\mathcal{G}}_m] + \theta\mathbb{E}[\mathbf{1}_T(v_m)\chi_m^{-1} | \mathcal{Z}_m]) + \\ &\quad \Pr[\mathcal{G}_m] \cdot (1-\theta)(1 + \theta\mathbb{E}[\sum_{j \geq m+1, v_j \in T} \chi_j^{-1} | \mathcal{Z}_{m+1}]) \\ &\leq (1-\theta)(1 + \theta\mathbb{E}[\sum_{j \geq m, v_j \in T} \chi_j^{-1} | \mathcal{Z}_m]), \end{aligned}$$

Now consider a fixed choice of partition P. Observe that for all $v_j \in T$, $d(v_j, x) \leq 2\Delta$, and so we get $B(v_j, \bar{\Delta}/\Gamma) \subseteq B(x, 3\Delta)$. On other hand $B(v_j, \bar{\Delta}\Gamma) \supseteq B(x, 3\Delta)$. Also since the radius of the ball defining S_{v_j} is at least $\Delta/4$ we have that for any $j \neq j', d(v_j, v_j') > \Delta/4 \geq \bar{\Delta}/\Gamma$, so that $B(v_j, \bar{\Delta}/\Gamma) \cap B(v_{j'}, \bar{\Delta}/\Gamma) = \emptyset$. Hence, we get:

$$\begin{aligned} &\sum_{j \geq m, v_j \in T} \chi_j^{-1} \leq \sum_{j \geq m, v_j \in T} \hat{\chi}_j^{-1} = \\ &\quad \sum_{j \geq m, v_j \in T} \frac{|B(v_j, \bar{\Delta}/\Gamma)|}{|B(v_j, \bar{\Delta}\Gamma)|} \leq \sum_{j \geq m, v_j \in T} \frac{|B(v_j, \bar{\Delta}/\Gamma)|}{|B(x, 3\Delta)|} \leq 1. \end{aligned}$$

For $x \in X$, if $P(x) = S_{v_j}$ then by definition $\eta_P^{(\delta)}(x) \leq \eta_j$ and so $B(x, (\eta_P^{(\delta)}(x))\Delta) \subseteq B(x, \eta_j\Delta)$. We conclude from the claim (3) for $m = 1$ that:

$$\begin{aligned} &\Pr[B(x, (\eta_P^{(\delta)}(x))\Delta) \nsubseteq P(x)] = \Pr[\mathcal{E}_1] \leq \\ &\quad (1-\theta)(1 + \theta \cdot \mathbb{E}[\sum_{j \geq m, v_j \in T} \chi_j^{-1}]) \leq (1-\theta)(1+\theta) = 1 - \delta. \end{aligned}$$

It follows that $\hat{\mathcal{P}}$ is uniformly padded. Finally, we show the properties states in the lemma. Recall that $\hat{\eta}_P(x) = 2^{-6}/\max\{\ln \hat{\chi}_j, \ln(1/\hat{\delta})\}$. By the choice of v_j, $\hat{\chi}_j = \rho(Z_j, \bar{\Delta}, \Gamma)$. As $x \in Z_j$ we have that $\hat{\chi}_j \leq \rho(x, \bar{\Delta}, \Gamma)$. If follows that if $\xi_P(x) = 1$ then $\hat{\chi}_j \geq 1/\hat{\delta}$ and therefore $\hat{\eta}_P(x) \geq 2^{-6}/\ln \hat{\chi}_j \geq 2^{-6}\ln(1/\delta)/\ln \rho(x, \bar{\Delta}, \Gamma)$.

On the other hand if $\xi_P(x) = 0$ then $\hat{\chi}_j < 1/\hat{\delta}$ and $\hat{\eta}_P(x) = 2^{-6}/\ln(1/\hat{\delta})$. Since $\mathrm{diam}(Z_j) \leq \mathrm{diam}(Z) \leq \bar{\Delta}$ we have that $Z_j \subseteq B(x, \bar{\Delta})$, and therefore $\bar{\rho}(x, \bar{\Delta}, \Gamma) \leq \hat{\chi}_j < 1/\hat{\delta}$. □

Using this lemma we can prove the following lemma on uniformly padded hierarchical probabilistic partitions from which Lemma 3 is derived.

Lemma 6 (Hierarchical Uniform Padding Lemma). *Let $\Gamma = 64$. Let $\hat{\delta} \in (0, \frac{1}{2}]$. Given a finite metric space (X, d), there exists a probabilistic 4-hierarchical partition $\hat{\mathcal{H}}$ of (X, d) and uniform collections of functions $\xi = \{\xi_{P,i} : X \to \{0,1\} | P \in \mathcal{H}, i \in I\}$ and $\hat{\eta} = \{\hat{\eta}_{P,i} : X \to \{0, 1/\ln(1/\hat{\delta})\} | P \in \mathcal{H}, i \in I\}$, such that for any $\hat{\delta} \leq \delta \leq 1$ and $\eta^{(\delta)}$ defined by $\eta^{(\delta)}(x) = \hat{\eta}^{(\delta)}(x)\ln(1/\delta)$, we have that $\hat{\mathcal{H}}$ is $(\eta^{(\delta)}, \delta)$-uniformly padded, and the following properties hold:*

- $$\sum_{j \leq i} \xi_{P,j}(x)\eta_{P,j}^{(\delta)}(x)^{-1} \leq 2^{10}\ln\left(\frac{n}{|B(x, \Delta_{i+4})|}\right)/\ln(1/\delta).$$

and for any $P \in \mathcal{H}$, $0 < i \in I$, $P_i \in P$:

- *If $\xi_{P,i}(x) = 1$ then: $\hat{\eta}_{P,i}(x) \leq 2^{-7}/\ln(1/\hat{\delta})$.*
- *If $\xi_{P,i}(x) = 0$ then: $\hat{\eta}_{P,i}(x) \geq 2^{-7}/\ln(1/\hat{\delta})$ and $\bar{\rho}(x, \Delta_{i-1}, \Gamma) < 1/\hat{\delta}$.*

Proof. We create a random hierarchical partition P. By definition P_0 consists of a single cluster equal to X. Set for all $x \in X$, $\Delta_0 = \mathrm{diam}(X)$, $\hat{\eta}_{P,0}(x) = 1/\ln(1/\hat{\delta})$, $\xi_{P,0}(x) = 0$. For each $i \in \mathbb{Z}$ we set $\Delta_i = 4^{-i}\Delta_0$. The rest of the levels of the partition are created by invoking iteratively Lemma 4. For $0 < i \in I$, assume we have created clusters in P_{i-1}. Set $\bar{\Delta} = \Delta_{i-1}$. Now, for each cluster $S \in P_{i-1}$, invoke Lemma 4 to create a Δ_i-bounded probabilistic partition $\mathcal{Q}[S]$ of (S, d). Let $Q[S]$ be the generated partition. Set $P_i = Q[S]$. Let ${\xi'}_{Q[S]}$, ${\hat{\eta}'}_{Q[S]}$ be the uniform functions defined in Lemma 4. Recall that for $\delta' \geq \hat{\delta}$ we have that $\mathcal{Q}[S]$ is $({\eta'}^{(\delta')}, \delta')$-uniformly padded, where ${\eta'}^{(\delta')}_{Q[S]}(x) = {\hat{\eta}'}_{Q[S]}(x) \ln(1/\delta')$. Define $\hat{\eta}_{P,i}(x) = \min\{\frac{1}{2} \cdot {\hat{\eta}'}_{Q[S]}(x), 2 \cdot \hat{\eta}_{P,i-1}(x)\}$ and let $\eta^{(\delta)}_{P,i}(x) = \hat{\eta}_{P,i}(x)\ln(1/\delta)$. If it is the case that $\hat{\eta}_{P,i}(x) = \frac{1}{2} \cdot {\hat{\eta}'}_{Q[S]}(x)$ and also ${\xi'}_{Q[S]}(x) = 0$ then set $\xi_{P,i}(x) = 0$, otherwise $\xi_{P,i}(x) = 1$.

Setting $\delta' = \delta^{1/2} \geq \hat{\delta}$, observe that by definition: $\eta^{(\delta)}_{P,i}(x) = \min\{{\eta'}^{(\delta')}_{Q[S]}(x), 2\eta^{(\delta)}_{P,i-1}(x)\}$.

Note, that for $i \in I$, $x, y \in X$ such that $P_i(x) = P_i(y)$, it follows by induction that $\hat{\eta}_{P,i}(x) = \hat{\eta}_{P,i}(y)$ (and hence $\eta^{(\delta)}_{P,i}(x) = \eta^{(\delta)}_{P,i}(y)$) and $\xi_{P,i}(x) = \xi_{P,i}(y)$, by using the fact that $\hat{\eta}'$ and ξ' are uniform functions with respect to $Q[S]$, where $S = P_{i-1}(x) = P_{i-1}(y)$.

We prove by induction on i that P_i is $(\eta^{(\delta)}, \delta)$-uniformly padded for all $\delta \geq \hat{\delta}$. Assume it holds for $i-1$ and we will prove for i. Now fix some appropriate value of δ. Let $B_i = B(x, \eta^{(\delta)}_{P,i}(x)\Delta_i)$. We have:

$$\Pr[B_i \subseteq P_i(x)] = \Pr[B_i \subseteq P_{i-1}(x)] \cdot \Pr[B_i \subseteq P_i(x) | B_i \subseteq P_{i-1}(x)]. \quad (4)$$

As $\eta^{(\delta)}_{P,i}(x) \leq {\eta'}^{(\delta')}_{Q[P_{i-1}(x)]}(x)$ we have $B(x, {\eta'}^{(\delta')}_{Q[P_{i-1}(x)]}(x)\Delta_i) \supseteq B_i$. It follows that if $B_i \subseteq P_{i-1}(x)$ then it also holds that $B_i \subseteq B_{P_{i-1}}(x, {\eta'}^{(\delta')}_{Q[P_{i-1}(x)]}(x)\Delta_i)$. Using Lemma 4 we have $\Pr[B_i \subseteq P_i(x) | B_i \subseteq P_{i-1}(x)] \geq \delta'$.

Next observe that by definition $\eta^{(\delta)}_{P,i}(x) \leq 2\eta^{(\delta)}_{P,i-1}(x)$ and $\eta^{(\delta)}_{P,i-1}(x) = 2\eta^{(\delta^{1/2})}_{P,i-1}(x)$. Since $\Delta_i = \Delta_{i-1}/4$ we get that $\eta^{(\delta)}_{P,i}(x)\Delta_i \leq \eta^{(\delta')}_{P,i-1}(x)\Delta_{i-1}$ for $\delta' = \delta^{1/2}$. We therefore obtain that $B_i \subseteq B(x, \eta^{(\delta')}_{P,i-1}(x)\Delta_{i-1})$. Using the induction hypothesis we get $\Pr[B_i \subseteq P_{i-1}(x)] \geq \delta'$. We conclude from (4) above that the inductive claim holds: $\Pr[B_i \subseteq P_i(x)] \geq \delta' \cdot \delta' = \delta$.

This completes the proof that $\mathcal{H}$ is $(\eta^{(\delta)}, \delta)$-uniformly padded.

We now turn to prove the properties stated in the lemma. Consider some $i \in I$ and $x \in X$. The second property holds as $\hat{\eta}_{P,i}(x) \leq \frac{1}{2}\hat{\eta}'_{Q[P_{i-1}(x)]}(x) \leq 2^{-7}/\ln(1/\hat{\delta})$, using Lemma 4. Let us prove the third property. By definition if $\xi_{P,i}(x) = 0$ then $\hat{\eta}_{P,i}(x) = \frac{1}{2}\hat{\eta}'_{Q[P_{i-1}(x)]}(x)$ and ${\xi'}_{Q[P_{i-1}(x)]}(x) = 0$. Using Lemma 4 we have that $\hat{\eta}_{P,i}(x) \geq 2^{-7}/\ln(1/\hat{\delta})$ and that $\bar{\rho}(x, \Delta_{i-1}, \Gamma) < 1/\hat{\delta}$.

It remains to prove the first property of the lemma. Define $\psi_{P,i}(x) = 2^{-7} \cdot \xi_{P,i}(x)\hat{\eta}_{P,i}(x)^{-1}$. It is easy to derive the following recursion: $\psi_{P,i}(x) \leq \ln \rho(x, \Delta_{i-1}, \Gamma) + \psi_{P,i-1}(x)/2$. A simple induction on t shows that for any $0 \leq t < i$: $\sum_{t<j\leq i} \psi_{P,j}(x) \leq 2\sum_{t<j\leq i} \ln \rho(x, \Delta_{j-1}, \Gamma) + (1-2^{t-i})\psi_{P,t}(x)$. Now observe that as $\Gamma = 64$, and that for any $j \in I$:

$$\ln \rho(x, \Delta_j, \Gamma) = \ln\left(\frac{|B(x, \Delta_j\Gamma)|}{|B(x, \Delta_j/\Gamma)|}\right) = \sum_{h=-4}^{3} \ln\left(\frac{|B(x, 4\Delta_{j+h})|}{|B(x, \Delta_{j+h})|}\right).$$

It follows that

$$\sum_{0<j\leq i} \psi_{P,j}(x) \leq 2\sum_{0<j\leq i} \ln \rho(x, \Delta_{j-1}, \Gamma) = 2\sum_{0<j\leq i}\sum_{h=-4}^{3} \ln\left(\frac{|B(x, 4\Delta_{j+h})|}{|B(x, \Delta_{j+h})|}\right) = 2\sum_{h=-4}^{3}\sum_{0<j\leq i} \ln\left(\frac{|B(x, 4\Delta_{j+h})|}{|B(x, \Delta_{j+h})|}\right) = 8\ln\left(\frac{n}{|B(x, \Delta_{i+4})|}\right).$$

This completes the proof of the first property of the lemma. □

4. THE MAIN THEOREM

In this section we prove Theorem 2 and its scaling distortion version: every n-point metric space embeds in L_p with scaling distortion $O(\log \frac{1}{\epsilon})$ and dimension $O(\log n)$. In particular, this theorem implies $O(1)$ average distortion (Theorem 6).

The start with the basic theorem of $O(\log n)$ distortion and dimension. The proof relies on Lemma 3.

In this section we construct an embedding into L_p. Let $D = \Theta(\ln n)$. We will define an embedding $f : X \to l_p^D$ with distortion $O(\ln n)$. We define f by defining for each $1 \leq t \leq D$, a function $f^{(t)} : X \to \mathbb{R}^+$ and let $f = D^{-1/p}\bigoplus_{1\leq t\leq D} f^{(t)}$.

Fix t, $1 \leq t \leq D$. In what follows we define $f^{(t)}$. We construct a uniformly $(\eta, 1/2)$-padded probabilistic 4-hierarchical partition $\hat{\mathcal{H}}$ as in Lemma 3, and let ξ be as defined in the lemma. Now fix a hierarchical partition $P \in \mathcal{H}$. We define the embedding by defining the coordinates for each $x \in X$. Define for $x \in X$, $0 < i \in I$, $\phi_i^{(t)} : X \to \mathbb{R}^+$, by $\phi_i^{(t)}(x) = \xi_{P,i}(x)\eta_{P,i}(x)^{-1}$.

Lemma 3 ensures that ξ and η are uniform functions with respect to $\mathcal{H}$ so we have:

Claim 7. *For any $x, y \in X$ and $i \in I$ if $P_i(x) = P_i(y)$ then $\phi_i^{(t)}(x) = \phi_i^{(t)}(y)$.*

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

Let $\{\sigma_i^{(t)}(C) | C \in P_i, 0 < i \in I\}$ be i.i.d symmetric $\{0,1\}$-valued Bernoulli random variables. The embedding is defined as follows: for each $x \in X$:

- For each $0 < i \in I$, let $\psi_i^{(t)}(x) = \sigma_i^{(t)}(P_i(x)) \cdot \phi_i^{(t)}(x) \cdot d(x, X \setminus P_i(x))$.

Claim 8. *For any $0 < i \in I$ and $x, y \in X$: $\psi_i^{(t)}(x) - \psi_i^{(t)}(y) \leq \phi_i^{(t)}(x) \cdot d(x, y)$.*

Proof. We have two cases. In Case 1, assume $P_i(x) = P_i(y)$. By Claim 7 $\phi_i^{(t)}(y) = \phi_i^{(t)}(x)$. It follows that

$$\psi_i^{(t)}(x) - \psi_i^{(t)}(y) = \sigma_i^{(t)}(P_i(x)) \cdot \phi_i^{(t)}(x) \cdot (d(x, X \setminus P_i(x)) - d(y, X \setminus P_i(x))) \leq \phi_i^{(t)}(x) \cdot d(x, y).$$

Next, consider Case 2 where $P_i(x) \neq P_i(y)$. In this case we have that $d(x, X \setminus P_i(x)) \leq d(x, y)$ which implies that $\psi_i^{(t)}(x) - \psi_i^{(t)}(y) \leq \psi_i^{(t)}(x) \leq \phi_i^{(t)}(x) \cdot d(x, y)$. □

LEMMA 9. *There exists a universal constant $C_1 > 0$ such that for any $x, y \in X$:*

$$|f^{(t)}(x) - f^{(t)}(y)| \le C_1 \ln n \cdot d(x,y).$$

PROOF. From Claim 8 and using Lemma 3 we get

$$\begin{aligned}\sum_{0<i\in I} (\psi_i^{(t)}(x) - \psi_i^{(t)}(y)) &\le \sum_{0<i\in I} \phi_i^{(t)}(x) \cdot d(x,y) \\ &\le 2^{11} \ln n \cdot d(x,y).\end{aligned}$$

It follows that $|f^{(t)}(x)-f^{(t)}(y)| = |\sum_{0<i\in I}(\psi_i^{(t)}(x)-\psi_i^{(t)}(y))| \le 2^{11} \ln n \cdot d(x,y)$. □

LEMMA 10. *There exists a universal constant $C_2 > 0$ such that for any $x, y \in X$, with probability at least 1/8:*

$$|f^{(t)}(x) - f^{(t)}(y)| \ge C_2 \cdot d(x,y).$$

PROOF. Let $0 < \ell \in I$ be such that $4\Delta_{\ell-1} \le d(x,y) \le 16\Delta_{\ell-1}$. By Claim 2 we have $\max\{\bar\rho(x, \Delta_{\ell-1}, \Gamma), \bar\rho(y, \Delta_{\ell-1}, \Gamma)\} \ge 2$. Assume w.l.o.g that $\bar\rho(x, \Delta_{\ell-1}, \Gamma) \ge 2$. It follows from Lemma 3 that $\xi_{P,\ell}(x) = 1$ which implies that $\phi_\ell^{(t)}(x) = \eta_{P,\ell}(x)^{-1}$. As $\hat{\mathcal{H}}$ is $(\eta, 1/2)$-padded we have the following bound

$$\Pr[B(x, \eta_{P,\ell}(x)\Delta_\ell) \subseteq P_\ell(x)] \ge 1/2.$$

Therefore with probability at least 1/2:

$$\phi_\ell^{(t)}(x) \cdot d(x, X \setminus P_\ell(x)) \ge \phi_\ell^{(t)}(x) \cdot \eta_{P,\ell}(x)\Delta_\ell \ge \Delta_\ell.$$

Assume that this event occurs. We distinguish between two cases:

- $|\sum_{0<i\neq\ell}(\psi_i^{(t)}(x) - \psi_i^{(t)}(y))| \ge \frac{1}{2}\Delta_\ell$. In this case there is probability at least 1/4 that $\sigma_\ell^{(t)}(P_\ell(x)) = \sigma_\ell^{(t)}(P_\ell(y)) = 0$, so that $\psi_\ell^{(t)}(x) = \psi_\ell^{(t)}(y) = 0$.
- $|\sum_{0<i\neq\ell}(\psi_i^{(t)}(x) - \psi_i^{(t)}(y))| \le \frac{1}{2}\Delta_\ell$.
  Since $\operatorname{diam}(P_\ell(x)) \le \Delta_\ell < d(x,y)$ we have that $P_\ell(y) \neq P_\ell(x)$. We get that there is probability 1/4 that $\sigma_\ell^{(t)}(P_\ell(x)) = 1$ and $\sigma_\ell^{(t)}(P_\ell(y)) = 0$ so that $\psi_\ell^{(t)}(x) - \psi_\ell^{(t)}(y) \ge \Delta_\ell$.

We conclude that with probability at least 1/8:

$$|\sum_{0<i\in I} (\psi_i^{(t)}(x) - \psi_i^{(t)}(y))| \ge \frac{1}{2}\Delta_\ell.$$

It follows that with probability at least 1/8:

$$|f^{(t)}(x) - f^{(t)}(y)| = |\sum_{0<i\in I} (\psi_i^{(t)}(x) - \psi_i^{(t)}(y))| \ge 2^{-7} d(x,y).$$

□

LEMMA 11. *There exist universal constants $C_1', C_2' > 0$ such that w.h.p for any $x, y \in X$:*

$$C_2' \cdot d(x,y) \le \|f(x) - f(y)\|_p \le C_1' \ln n \cdot d(x,y).$$

PROOF. By definition $\|f(x) - f(y)\|_p^p = D^{-1}\sum_{1\le t\le D} |f^{(t)}(x) - f^{(t)}(y)|^p$. Lemma 9 implies that $\|f(x)-f(y)\|_p^p \le (C_1 \ln n)^p \, d(x,y)^p$. Using Lemma 10 and applying Chernoff bounds we get w.h.p for any $x, y \in X$: $\|f(x)-f(y)\|_p^p \ge \frac{1}{16}(C_2 d(x,y))^p$. □

4.1 Scaling Distortion Embedding

The proof for scaling distortion version of the theorem follows the same structure above. The embedding is modified using the following definition:

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

It is easy to see that the new definition does not affect the lower bound part of the proof (Lemma 10). The upper bound is described below:

Define $\bar g_i^{(t)} : X \times X \to \mathbb{R}^+$ as follows: $\bar g_i^{(t)}(x,y) = \min\{\phi_i^{(t)}(x) \cdot d(x,y), \Delta_i\}$ (Note that $\bar g_i^{(t)}$ is nonsymmetric).

CLAIM 12. *For any $0 < i \in I$ and $x, y \in X$:*

$$\psi_i^{(t)}(x) - \psi_i^{(t)}(y) \le \bar g_i^{(t)}(x,y).$$

PROOF. We have two cases. In Case 1, assume $P_i(x) = P_i(y)$. It follows that

$$\psi_i^{(t)}(x) - \psi_i^{(t)}(y) = \sigma_i^{(t)}(P_i(x)) \cdot (g_i^{(t)}(x) - g_i^{(t)}(y)).$$

We will show that $g_i^{(t)}(x) - g_i^{(t)}(y) \le \bar g_i^{(t)}(x,y)$. The bound $g_i^{(t)}(x) - g_i^{(t)}(y) \le \Delta_i$ is immediate. To prove $g_i^{(t)}(x) - g_i^{(t)}(y) \le \phi_i^{(t)}(x) \cdot d(x,y)$ consider the value of $g_i^{(t)}(y)$. Assume first $g_i^{(t)}(y) = \phi_i^{(t)}(y) \cdot d(y, X \setminus P_i(x))$. By Claim 7 $\phi_i^{(t)}(y) = \phi_i^{(t)}(x)$ and therefore

$$\begin{aligned}g_i^{(t)}(x) - g_i^{(t)}(y) &\le \phi_i^{(t)}(x) \cdot (d(x, X \setminus P_i(x)) - d(y, X \setminus P_i(x))) \\ &\le \phi_i^{(t)}(x) \cdot d(x,y).\end{aligned}$$

In the second case $g_i^{(t)}(y) = \Delta_i$ and therefore $g_i^{(t)}(x) - g_i^{(t)}(y) \le \Delta_i - \Delta_i = 0$, proving the claim in this case.

Next, consider Case 2 where $P_i(x) \neq P_i(y)$. In this case we have that $d(x, X \setminus P_i(x)) \le d(x,y)$ which implies that

$$\psi_i^{(t)}(x) - \psi_i^{(t)}(y) \le \psi_i^{(t)}(x) \le g_i^{(t)}(x) \le \bar g_i^{(t)}(x,y).$$

□

LEMMA 13. *There exists a universal constant $C_1 > 0$ such that for any $\epsilon > 0$ and any $(x,y) \in \hat G(\epsilon)$:*

$$|f^{(t)}(x) - f^{(t)}(y)| \le C_1 \ln(1/\epsilon) \cdot d(x,y).$$

PROOF. From Claim 12 we get

$$\sum_{0<i\in I} (\psi_i^{(t)}(x) - \psi_i^{(t)}(y)) \le \sum_{0<i\in I} \bar g_i^{(t)}(x,y).$$

Now, define ℓ to be largest such that $\Delta_{\ell+4} \ge d(x,y) \ge \max\{r_{\epsilon/2}(x), r_{\epsilon/2}(y)\}$. If no such ℓ exists then let $\ell = 0$.

By Lemma 3 we have

$$\begin{aligned}\sum_{0<i\le\ell} \bar g_i^{(t)}(x,y) &\le \sum_{0<i\le\ell} \phi_i^{(t)}(x) \cdot d(x,y) \\ &\le 2^{11} \cdot \ln\left(\frac{n}{|B(x, \Delta_{\ell+4})|}\right) \cdot d(x,y) \\ &\le (2^{11} \ln(2/\epsilon)) \cdot d(x,y).\end{aligned}$$

We also have that

$$\sum_{\ell<i\in I} \bar g_i^{(t)}(x,y) \le \sum_{\ell<i\in I} \Delta_i \le \Delta_\ell \le 4^5 d(x,y).$$

It follows that

$$|f^{(t)}(x)-f^{(t)}(y)| = |\sum_{0<i\in I}(\psi_i^{(t)}(x)-\psi_i^{(t)}(y))| \leq (2^{11}\ln(2/\epsilon)+4^5)\cdot d(x,y).$$

□

5. TIGHT BOUNDS ON THE DIMENSION

In this section we provide an embedding in optimal dimension, proving Theorem 3. The proof relies on the following slightly strengthened version of Lemma 3. We first require a claim based on a simple technique from [7].

CLAIM 14. *Given a finite metric space (X,d) and let $\epsilon > 0$, there exists a metric space $(X^{(\epsilon)}, d^{(\epsilon)})$ (where $X^{(\epsilon)}$ is a quotient of X) and a mapping $\nu^{(\epsilon)}: X \to X^{(\epsilon)}$ such that:*

- *For any $x,y\in X$: if $d(x,y)\leq\epsilon$ then $\nu^{(\epsilon)}(x)=\nu^{(\epsilon)}(y)$.*
- *For any $x,y \in X$: $d^{(\epsilon)}(\nu^{(\epsilon)}(x),\nu^{(\epsilon)}(y)) \leq d(x,y) \leq d^{(\epsilon)}(\nu^{(\epsilon)}(x),\nu^{(\epsilon)}(y)) + 2\epsilon n$.*

The proof is left for the full version. We obtain the following lemma (the proof is omitted).

LEMMA 15. *Let $c\geq 2$. Given a finite metric space (X,d), there exists a probabilistic 4-hierarchical partition $\hat{\mathcal{H}}$ of (X,d) and uniform collections of functions $\xi = \{\xi_{P,i} : X \to \{0,1\} | P\in\mathcal{H}, i\in I\}$ and $\hat{\eta} = \{\hat{\eta}_{P,i} : X \to \{0, 1/\ln(1/\hat{\delta})\} | P\in\mathcal{H}, i\in I\}$, such that for any $\hat{\delta}\leq\delta\leq 1$ and $\eta^{(\delta)}$ defined by:*

- *$\eta^{(\delta)}_{P,i}(x) = \max\{\hat{\eta}^{(\delta)}_{P,i}(x)\ln(1/\delta), 1/(cn)\}$,*

we have that $\hat{\mathcal{H}}$ is $(\eta^{(\delta)},\delta)$-uniformly padded, and $\hat{\eta}$ satisfies the same properties as in Lemma 6.

The Embedding. Let $\theta>0$. Let $D = \lceil 8\cdot\theta^{-1}\ln n/\ln\ln n\rceil$. We will define an embedding $f: X\to l_p^D$ with distortion $O(\ln^{1+2\theta} n)$. We define f by defining for each $1\leq t\leq D$, a function $f^{(t)}: X\to\mathbb{R}^+$ and let $f = D^{-1/p}\bigoplus_{1\leq t\leq D} f^{(t)}$.

In what follows we define the functions $f^{(t)}$.

Let $\varepsilon = \ln^{-\theta} n$ (we may assume $\varepsilon\leq 1/2$). We construct a uniformly $(\eta, 1-\varepsilon)$-padded probabilistic 4-hierarchical partition $\bar{\mathcal{H}}^{(t)}$ by applying Lemma 15 with $\hat{\delta}=1/2$ and $\delta = 1-\varepsilon$, let ξ be as defined in the lemma. Now fix a hierarchical partition $P^{(t)}\in\mathcal{H}^{(t)}$. We define the embedding by defining the coordinates for each $x\in X$. Define for $x\in X$, $0<i\in I$, $\phi_i^{(t)}: X\to\mathbb{R}^+$, by $\phi_i^{(t)}(x) = \xi_{P^{(t)},i}(x)\eta_{P^{(t)},i}(x)^{-1}$.

As in Section 4 we have Claim 7.

Let $\{\sigma_m^{(t)} | 1\leq t\leq D, 1\leq m\leq n\}$ be i.i.d random variables uniformly distributed in $[0,1]$. Let $\{\bar{\sigma}_i^{(t)} | 1\leq t\leq D, 0\leq i< 2c\log n\}$ be i.i.d symmetric $\{0,1\}$-valued Bernoulli random variables, and extend it into a periodic sequence so that for $i\in\mathbb{N}$, $\bar{\sigma}_i^{(t)} = \bar{\sigma}^{(t)}_{i \bmod \lceil 2c\log n\rceil}$, where c is determined later in the proof.

Let $C_1,\dots C_{n_i}$ be the clusters of $P_i^{(t)}$. Define for $C_j\in P_i^{(t)}$: $\hat{\sigma}_i^{(t)}(C_j) = \sigma_j^{(t)}\bar{\sigma}_i^{(t)}$.

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

To define the embedding we need to make use of some parameters in the construction of the probabilistic partition as described in Lemma 15. Denote $\epsilon_i = \Delta_i/(cn)$ and let $\nu_i^{(\epsilon_i)}$ be the mapping as in Claim 14. The embedding is defined as follows: for each $x\in X$:

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

We have the corresponding variant of Claim 12 from Section 4.1 . Define $\bar{g}_i^{(t)}: X\times X\to\mathbb{R}^+$ as follows: $\bar{g}_i^{(t)}(x,y) = \min\{\phi_i^{(t)}(x)\cdot d(x,y),\Delta_i\}$.

CLAIM 16. *For any $0<i\in I$ and $x,y\in X$: $\psi_i^{(t)}(x)-\psi_i^{(t)}(y)\leq\bar{g}_i^{(t)}(x,y)$.*

PROOF. The proof is the same as in Claim 12, by using $d(\nu_i^{(\epsilon_i)}(x),\nu_i^{(\epsilon_i)}(y))\leq d(x,y)$. □

LEMMA 17. *There exists a universal constant $C_1>0$ such that for any $x,y\in X$:*

$$\|f(x)-f(y)\|_p\leq C_1\ln^{1+\theta} n\cdot d(x,y).$$

PROOF. From Claim 16 and using Lemma 3 we get

$$\sum_{0<i\in I}(\psi_i^{(t)}(x)-\psi_i^{(t)}(y))\leq\sum_{0<i\in I}\bar{g}_i^{(t)}(x,y)\leq$$
$$\sum_{0<i\in I}\phi_i^{(t)}(x)\cdot d(x,y)\leq 2^{10}\ln n/\ln(\frac{1}{1-\varepsilon})\cdot d(x,y)\leq$$
$$2^{11}\ln n/\varepsilon\cdot d(x,y) = 2^{11}\ln^{1+\theta} n\cdot d(x,y).$$

It follows that $|f^{(t)}(x)-f^{(t)}(y)| = |\sum_{0<i\in I}(\psi_i^{(t)}(x)-\psi_i^{(t)}(y))| \leq 2^{11}\ln^{1+\theta} n\cdot d(x,y)$,and therefore

$$\|f(x)-f(y)\|_p^p =$$
$$D^{-1}\sum_{1\leq t\leq D}|f^{(t)}(x)-f^{(t)}(y)|^p\leq\left(C_1\ln^{1+\theta} n\right)^p d(x,y)^p.$$

□

The next Lemma makes use of the following simple technical claim.

CLAIM 18. *Let $A,B\in\mathbb{R}^+$ and let α,β be i.i.d random variables uniformly distributed in $[0,1]$. Then for any $\varepsilon>0$:*

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

LEMMA 19. *There exists a universal constant $C_2>0$ such that with constant probability for any $x,y\in X$:*

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

PROOF. We will prove that with constant probability for every $x,y\in X$, there exists a set $T(x,y)\subseteq\{1,\dots,D\}$ of size at least $D/16$ such that for any $t\in T(x,y)$:

$$|f^{(t)}(x)-f^{(t)}(y)| \geq 2^{-9}\varepsilon\cdot d(x,y). \quad (5)$$

The theorem follows directly:

$$\|f(x)-f(y)\|_p^p = D^{-1}\sum_{1\leq t\leq D}|f^{(t)}(x)-f^{(t)}(y)|^p \geq D^{-1}|T(x,y)|\cdot\left(2^{-9}\varepsilon\cdot d(x,y)\right)^p \geq \frac{1}{16}\left(2^{-9}\log^{-\theta} n\cdot d(x,y)\right)^p.$$

The proof follows the general principle of Lemma 10. In order to do this we need to make more delicate analysis of the probability that certain events occur.

Fix some $x, y \in X$. Let $0 < \ell = \ell(x,y) \in I$ be such that $4\Delta_{\ell-1} \le d(x,y) \le 16\Delta_{\ell-1}$. By Claim 2 we have that $\max\{\bar\rho(x,\Delta_{\ell-1},\Gamma),\bar\rho(y,\Delta_{\ell-1},\Gamma)\} \ge 2$. Assume w.l.o.g that $\bar\rho(x,\Delta_{\ell-1},\Gamma) \ge 2$. It follows from Lemma 3 that $\xi_{P^{(t)},\ell}(x) = 1$ which implies that $\phi_\ell^{(t)}(x) = \eta_{P^{(t)},\ell}(x)^{-1}$. As $\hat{\mathcal{H}}^{(t)}$ is $(\eta, 1-\varepsilon)$-padded we have the following bound

$$\Pr[B(x,\eta_{P^{(t)},\ell}(x)\Delta_\ell) \subseteq P_\ell^{(t)}(x)] \ge 1-\varepsilon.$$

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

$$\begin{aligned}&\phi_\ell^{(t)}(x)\cdot d(\nu_\ell^{(\epsilon_\ell)}(x), X\setminus P_\ell^{(t)}(x)) \ge \\ &\quad \phi_\ell^{(t)}(x)\cdot(d(x,X\setminus P_\ell^{(t)}(x)) - d(x,\nu_\ell^{(\epsilon_\ell)}(x)) \ge \\ &\quad \phi_\ell^{(t)}(x)\cdot(\eta_{P^{(t)},\ell}(x) - \frac{1}{cn})\Delta_\ell \ge \frac{1}{2}\Delta_\ell, \end{aligned} \tag{6}$$

for an appropriate choice of c as $\phi_\ell^{(t)}(x) \le c\ln n/2$. It follows that also $g_\ell^{(t)}(x) \ge \frac{1}{2}\Delta_\ell$.

Let $\mathcal{A}$ denote the event that (6) occurs. Recall that:

$$\begin{aligned}&|\psi_\ell^{(t)}(x) - \psi_\ell^{(t)}(y)| = \\ &\quad |\hat\sigma_\ell^{(t)}(P_\ell^{(t)}(x))\cdot g_\ell^{(t)}(x) - \hat\sigma_\ell^{(t)}(P_\ell^{(t)}(y))\cdot g_\ell^{(t)}(y)|.\end{aligned}$$

Since $\operatorname{diam}(P_\ell^{(t)}(x)) \le \Delta_\ell < d(x,y)$ we have that $P_\ell^{(t)}(y) \ne P_\ell^{(t)}(x)$. It follows that there are $a \ne b$ such that $\hat\sigma_\ell^{(t)}(P_\ell^{(t)}(x)) = \bar\sigma_\ell^{(t)}\cdot\sigma_a^{(t)}$ and $\hat\sigma_\ell^{(t)}(P_\ell^{(t)}(y)) = \bar\sigma_\ell^{(t)}\cdot\sigma_b^{(t)}$. Hence,

$$|\psi_\ell^{(t)}(x) - \psi_\ell^{(t)}(y)| = \bar\sigma_\ell^{(t)}\cdot|\sigma_a^{(t)}\cdot g_\ell^{(t)}(x) - \sigma_b^{(t)}\cdot g_\ell^{(t)}(y)|.$$

By Claim 18 and using (6) we have:

$$\Pr[|\sigma_a^{(t)}\cdot g_\ell^{(t)}(x) - \sigma_b^{(t)}\cdot g_\ell^{(t)}(y)|] < \varepsilon\cdot\frac{1}{2}\Delta_\ell|\mathcal{A}] < 2\varepsilon.$$

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

$$|\psi_\ell^{(t)}(x) - \psi_\ell^{(t)}(y)| \quad\ge\quad \bar\sigma_\ell^{(t)}\cdot\frac{1}{2}\varepsilon\cdot\Delta_\ell. \tag{7}$$

We are now ready to prove (5). Specifically, we will prove that with constant probability for every $x,y \in X$ there exists $T(x,y)$ of size at least $D/16$ such that for any $t \in T(x,y)$:

$$\begin{aligned}&|f^{(t)}(x) - f^{(t)}(y)| = \\ &\quad |\sum_{0<i\in I}(\psi_i^{(t)}(x) - \psi_i^{(t)}(y))| \ge \frac{1}{8}\varepsilon\cdot\Delta_{\ell(x,y)}.\end{aligned} \tag{8}$$

Inequality (5) now follows as $\Delta_{\ell(x,y)} \ge 4^{-3}\cdot d(x,y)$.

For each $x,y \in X$ define $I(x,y) = \{0 < i \in I | \ell(x,y) - \log(cn) \le i < \ell(x,y) + \log(cn)\}$ and let $I^*(x,y) = I(x,y)\setminus\{\ell(x,y)\}$. We will first show that with constant probability for every $x,y \in X$ there exists $T(x,y)$ of size at least $D/16$ such that for any $t \in T(x,y)$:

$$|\sum_{i\in I(x,y)}(\psi_i^{(t)}(x) - \psi_i^{(t)}(y))| \quad\ge\quad \frac{1}{4}\varepsilon\cdot\Delta_{\ell(x,y)}. \tag{9}$$

Consider the set $T_L(x,y) = \{t \in \{1,\ldots,D\}|$ $|\sum_{i\in I^*(x,y)}(\psi_i^{(t)}(x) - \psi_i^{(t)}(y))| \ge \frac{1}{4}\varepsilon\cdot\Delta_{\ell(x,y)}\}$. We partition the pairs of points in X as follows. Let $W = \{\{x,y\} \in \binom{X}{2} | |T_L(x,y)| \ge D/2\}$ and $\bar W = \{\{x,y\} \in \binom{X}{2} | |T_L(x,y)| < D/2\}$. Let $W_l = \{\{x,y\} \in W | \ell(x,y) = l\}$ and $\bar W_l = \{\{x,y\} \in \bar W | \ell(x,y) = l\}$. The analysis proceeds according to this partition:

- $\{x,y\} \in W_l$. If for some $t \in T_L(x,y)$, $\bar\sigma_l^{(t)} = 0$ then $|\sum_{i\in I(x,y)}(\psi_i^{(t)}(x) - \psi_i^{(t)}(y))| \ge \frac{1}{4}\varepsilon\cdot\Delta_l$. Using Chernoff bounds, we have that the probability that this fails to hold for at least $D/8$ values of t is at most $2^{-D/8}$.

- $\{x,y\} \in \bar W_l$. If for some $t \notin T_L(x,y)$, $\bar\sigma_l^{(t)} = 1$ then according to inequality (7) with probability at least $1-3\varepsilon$, $|\sum_{i\in I(x,y)}(\psi_i^{(t)}(x) - \psi_i^{(t)}(y))| \ge \frac{1}{4}\varepsilon\cdot\Delta_l$. Using Chernoff bounds, we have that the probability that $\bar\sigma_l^{(t)} = 1$ fails to hold for at least $D/8$ values of t is at most $2^{-D/8}$.

As there are $O(\log n)$ independent values for $\bar\sigma_l^{(t)}$ we get that with constant probability for every x,y there are at least $D/8$ values of t for which there is probability at least $1-3\varepsilon$ that $|\sum_{i\in I(x,y)}(\psi_i^{(t)}(x) - \psi_i^{(t)}(y))| \ge \frac{1}{4}\varepsilon\cdot\Delta_l$ holds. By Chernoff bounds, the probability that this fails for more than $D/16$ values of t is at most $(6e\varepsilon)^{-D/16} = O(n^{-2})$. This argument implies claim (9).

Now, notice that for $i < \ell(x,y) - \log(cn)$, $d(x,y) \le \Delta_i/cn$. It follows that $\nu_i^{(\epsilon_i)}(x) = \nu_i^{(\epsilon_i)}(y)$ and as $\eta_{P^{(t)},i}(x) \ge 1/cn$ we have $P_i^{(t)}(x) = P_i^{(t)}(y)$ implying $\phi_i^{(t)}(x) = \phi_i^{(t)}(y)$ so that:

$$\begin{aligned}&g_i^{(t)}(x) = \min\{\phi_i^{(t)}(x)\cdot d(\nu_i^{(\epsilon_i)}(x), X\setminus P_i^{(t)}(x)),\Delta_i\} = \\ &\quad \min\{\phi_i^{(t)}(y)\cdot d(\nu_i^{(\epsilon_i)}(y), X\setminus P_i^{(t)}(y)),\Delta_i\} = g_i^{(t)}(y),\end{aligned}$$

and

$$\begin{aligned}&|\psi_i^{(t)}(x) - \psi_i^{(t)}(y)| = \\ &\quad |\hat\sigma_i^{(t)}(P_i^{(t)}(x))\cdot g_i^{(t)}(x) - \hat\sigma_i^{(t)}(P_i^{(t)}(y))\cdot g_i^{(t)}(y)| = 0.\end{aligned}$$

On the other hand we have

$$\begin{aligned}&\left|\sum_{i\ge\ell+\log(cn)}(\psi_i^{(t)}(x) - \psi_i^{(t)}(y))\right| \le \\ &\quad \sum_{i\ge\ell+\log(cn)}\max\{g_i^{(t)}(x), g_i^{(t)}(y)\} \le \\ &\quad \sum_{i\ge\ell+\log(cn)}\Delta_i \le \frac{\Delta_\ell}{cn} \le \frac{1}{8}\varepsilon\cdot\Delta_\ell,\end{aligned}$$

where $\ell = \ell(x,y)$, for an appropriate choice of c. From claim (9) and the above inequalities we now easily derive claim (8) which complete the proof. □

6. BETTER EMBEDDINGS INTO L_P

In this section we prove the following generalization of Theorem 10.

THEOREM 15. *Let $1 \le p \le \infty$ and let $1 \le \kappa \le p$. For any n-point metric space (X,d) there exists an embedding $f : X \to L_p$ with coarsely scaling distortion $O(\lceil(\log\frac{1}{\epsilon})/\kappa\rceil)$ and dimension $e^{O(\kappa)}\log n$.*

The proof relies on Lemma 3.

Let $1 \le \kappa \le p$. Let $s = e^\kappa$. Let $D = e^{\Theta(\kappa)}\ln n$. We will define an embedding $f : X \to l_p^D$, by defining for each $1 \le t \le D$, function $f^{(t)}, \psi^{(t)}, \mu^{(t)} : X \to \mathbb{R}^+$ and let $f^{(t)} = \psi^{(t)} + \mu^{(t)}$ and $f = D^{-1/p}\bigoplus_{1\le t\le D} f^{(t)}$.

Fix t, $1 \le t \le D$. In what follows we define $\psi^{(t)}$. We construct a uniformly $(\eta, 1/s)$-padded probabilistic 4-hierarchical partition $\bar{\mathcal{H}}$ as in Lemma 3, and let ξ be as defined in the lemma. Now fix a hierarchical partition $P \in \mathcal{H}$. We define the embedding by defining the coordinates for each $x \in X$. Define for $x \in X$, $0 < i \in I$, $\phi_i^{(t)} : X \to \mathbb{R}^+$, by $\phi_i^{(t)}(x) = \xi_{P,i}(x)\eta_{P,i}(x)^{-1}$.

For each $0 < i \in I$ we define a function $\psi_i^{(t)} : X \to \mathbb{R}^+$ and for $x \in X$, let $\psi^{(t)}(x) = \sum_{i\in I} \psi_i^{(t)}(x)$. Let $\{\sigma_i^{(t)}(C)|C \in P_i, 0 < i \in I\}$ be i.i.d symmetric $\{0,1\}$-valued Bernoulli random variables. For each $x \in X$: For each $0 < i \in I$, let $\psi_i^{(t)}(x) = \sigma_i^{(t)}(P_i(x)) \cdot g_i^{(t)}(x)$, where $g_i^{(t)} : X \to \mathbb{R}^+$ is defined as: $g_i^{(t)}(x) = \min\{\phi_i^{(t)}(x) \cdot d(x, X \setminus P_i(x)), \Delta_i\}$.

Next, we define the function $\mu^{(t)}$, based on the embedding technique of Bourgain [15] and its generalization by Matoušek [45]. Let $T' = \lceil \log_s n \rceil$ and $K = \{k \in \mathbb{N}|1 \le k \le T'\}$. For each $k \in K$ define a randomly chosen subset $A_k^{(t)} \subseteq X$, with each point of X included in $A_k^{(t)}$ independently with probability s^{-k}. For each $k \in K$ and $x \in X$, define:

$$I_k(x) = \{i \in I|\forall u \in P_i(x),\ s^{k-2} < |B(u, 16\Delta_{i-1})| \le s^k\}.$$

We make the following two simple observations:

Claim 20. *For every $i \in I$: (1) For any $x \in X$: $|\{k|i \in I_k(x)\}| \le 2$. (2) For every $k \in K$: the function $i \in I_k(x)$ is uniform with respect to P_i.*

We define $i_k : X \to I$, where $i_k(x) = 0$ if $I_k(x) = \emptyset$ and $i_k(x) = \min\{i|i \in I_k(x)\}$ otherwise. For each $k \in K$ we define a function $\mu_k^{(t)} : X \to \mathbb{R}^+$ and for $x \in X$ let $\mu^{(t)}(x) = \sum_{k\in K} \mu_k^{(t)}(x)$. Let $\Phi_0 = 2^7$. The function $\mu_k^{(t)}$ is defined as follows: for each $x \in X$: For each $k \in K$, let $\mu_k^{(t)}(x) = \min\{\frac{1}{4}d(x, A_k^{(t)}), h_{i_k(x)}^{(t)}(x)\}$, where $h_i^{(t)} : X \to \mathbb{R}^+$ is defined as: $h_i^{(t)}(x) = \min\{\Phi_0 \cdot d(x, X \setminus P_i(x)), \Delta_i\}$. Define $\bar{h}_i^{(t)} : X \times X \to \mathbb{R}^+$ as follows: $\bar{h}_i^{(t)}(x, y) = \min\{\Phi_0 \cdot d(x, y), \Delta_i\}$ (Note that $\bar{h}_i^{(t)}$ is nonsymmetric). We have the following analogue of Claim 8:

Claim 21. *For any $k \in K$ and $x, y \in X$: $\mu_k^{(t)}(x) - \mu_k^{(t)}(y) \le \bar{h}_{i_k(x)}^{(t)}(x, y)$.*

Lemma 22. *There exists a universal constant $C_1 > 0$ such that for any $\epsilon > 0$ and any $(x, y) \in \hat{G}(\epsilon)$:*

$$|f^{(t)}(x) - f^{(t)}(y)| \le C_1 (\ln(1/\epsilon)/\kappa + 1) \cdot d(x, y).$$

The proof is left for the full version of the paper.

Lemma 23. *There exists a universal constant $C_2 > 0$ such that for any $x, y \in X$, with probability at least $e^{-5\kappa}/4$:*

$$|f^{(t)}(x) - f^{(t)}(y)| \ge C_2 \cdot d(x, y).$$

Proof. Let $0 < \ell \in I$ be such that $4\Delta_{\ell-1} \le d(x, y) \le 16\Delta_{\ell-1}$. We distinguish between the following two cases:

- **Case 1:** Either $\xi_{P,\ell}(x) = 1$ or $\xi_{P,\ell}(y) = 1$.

Assume w.l.o.g that $\xi_{P,\ell}(x) = 1$. It follows that $\phi_\ell^{(t)}(x) = \eta_{P,\ell}(x)^{-1}$. As $\hat{\mathcal{H}}$ is (η, δ)-padded we have the following bound $\Pr[B(x, \eta_{P,\ell}(x)\Delta_\ell) \subseteq P_\ell(x)] \ge 1/s$.Therefore with probability at least $1/s$:

$$\phi_\ell^{(t)}(x) \cdot d(x, X \setminus P_\ell(x)) \ge \phi_\ell^{(t)}(x) \cdot \eta_{P,\ell}(x)\Delta_\ell \ge \Delta_\ell.$$

Assume that this event occurs.

We distinguish between two cases according to whether $|f^{(t)}(x) - f^{(t)}(y) - (\psi_\ell^{(t)}(x) - \psi_\ell^{(t)}(y))| \ge \frac{1}{2}\Delta_\ell$ and show that with probability at least $1/4s$: $|f^{(t)}(x) - f^{(t)}(y)| \ge \frac{1}{2}\Delta_\ell$.

- **Case 2:** $\xi_{P,\ell}(x) = \xi_{P,\ell}(y) = 0$

From Lemma 3, $\max\{\bar{\rho}(x, \Delta_{\ell-1}, \Gamma), \bar{\rho}(y, \Delta_{\ell-1}, \Gamma)\} < s$. Let $x' \in B(x, \Delta_{\ell-1})$ and $y' \in B(y, \Delta_{\ell-1})$ such that $\rho(x', \Delta_{\ell-1}, \Gamma) = \bar{\rho}(x, \Delta_{\ell-1}, \Gamma)$ and $\rho(y', \Delta_{\ell-1}, \Gamma) = \bar{\rho}(y, \Delta_{\ell-1}, \Gamma)$. For $z \in \{x', y'\}$ we have:

$$s > \frac{|B(z, \Gamma\Delta_{\ell-1})|}{|B(z, \Delta_{\ell-1}/\Gamma)|} \ge \frac{|B(x, 32\Delta_{\ell-1})|}{|B(z, \Delta_{\ell-1}/\Gamma)|},$$

using that $d(x, x') \le \Delta_{\ell-1}$ and $d(x, y') \le d(x, y) + d(y, y') \le 17\Delta_{\ell-1}$, and $\Gamma = 64$, so that $B(x, 32\Delta_{\ell-1}) \subseteq B(z, \Gamma\Delta_{\ell-1})$. Let $k \in K$ be such that $s^{k-1} < |B(x, 32\Delta_{\ell-1})| \le s^k$. We deduce that for $z \in \{x', y'\}$, $|B(z, \Delta_{\ell-1}/\Gamma)| > s^{k-2}$. Consider an arbitrary point $u \in P_\ell(x)$ as $d(x, u) \le \Delta_\ell < \Delta_{\ell-1}$ it follows that $s^{k-2} < |B(u, 16\Delta_{\ell-1})| \le s^k$. This implies that $\ell \in I_k(x)$ and therefore $i_k(x) \le \ell$. As $\hat{\mathcal{H}}$ is (η, δ)-padded we have the following bound

$$\Pr[B(x, \eta_{P,\ell}(x)\Delta_\ell) \subseteq P_\ell(x)] \ge 1/s.$$

Assume that this event occurs. Since P is hierarchical we get that for every $i \le \ell$ $B(x, \eta_{P,\ell}(x)\Delta_\ell) \subseteq P_\ell(x) \subseteq P_i(x)$ and in particular this holds for $i = i_k(x)$. As $\xi_{P,\ell}(x) = 0$ we have that $\eta_{P,\ell}(x) \ge 2^{-7} = 1/\Phi_0$. Hence,

$$\Phi_0 \cdot d(x, X \setminus P_i(x)) \ge \Phi_0 \cdot \eta_{P,\ell}(x)\Delta_\ell \ge \Delta_\ell.$$

Implying: $\mu_k^{(t)}(x) = \min\{\frac{1}{4}d(x, A_k^{(t)}), \Phi_0 \cdot d(x, X \setminus P_i(x)), \Delta_i\} \ge \min\{\frac{1}{4}d(x, A_k^{(t)}), \Delta_\ell\}$.

The following is a variant on the original argument in [15, 45]. Define the events: $\mathcal{A}_1 = B(y', \Delta_{\ell-1}/\Gamma) \cap A_k^{(t)} \ne \emptyset$, $\mathcal{A}_2 = B(x', \Delta_{\ell-1}/\Gamma) \cap A_k^{(t)} \ne \emptyset$ and $\mathcal{A}_2' = [B(x, 32\Delta_{\ell-1}) \setminus B(y', \Delta_{\ell-1})] \cap A_k^{(t)} = \emptyset$. Then for $m \in \{1, 2\}$:

$$\begin{aligned} \Pr[\mathcal{A}_m] &\ge 1 - \left(1 - s^{-k}\right)^{s^{k-2}} \ge 1 - e^{-s^{-k} \cdot s^{k-2}} \\ &\ge 1 - e^{-s^{-2}} \ge s^{-2}/2, \\ \Pr[\mathcal{A}_2'] &\ge \left(1 - s^{-k}\right)^{s^k} \ge 1/4, \end{aligned}$$

using $s \ge 2$. Observe that $d(x', y') \ge d(x, y) - 2\Delta_{\ell-1}/\Gamma \ge (1 - 2/\Gamma)\Delta_{\ell-1} > 2\Delta_{\ell-1}/\Gamma$, implying $B(y', \Delta_{\ell-1}/\Gamma) \cap B(x', \Delta_{\ell-1}/\Gamma) = \emptyset$. It follows that event $\mathcal{A}_1$ is independent of either event $\mathcal{A}_2$ or $\mathcal{A}_2'$.

Assume event $\mathcal{A}_1$ occurs. It follows that $d(y, A_k^{(t)}) \le d(y, y') + \Delta_{\ell-1}/\Gamma \le (1 + 1/\Gamma)\Delta_{\ell-1}$.

We distinguish between two cases according to whether $|f^{(t)}(x) - f^{(t)}(y) - (\mu_k^{(t)}(x) - \mu_k^{(t)}(y))| \ge 3/2 \cdot \Delta_\ell$ and show that given $\mathcal{A}_1$, with probability at least $s^{-2}/2$: $|f^{(t)}(x) - f^{(t)}(y)| \ge \Delta_\ell/4$. This concludes case 2.

It follows that with probability at least $s^{-5}/4$: $|f^{(t)}(x) - f^{(t)}(y)| \ge \frac{1}{4}\Delta_\ell \ge \frac{1}{4}4^{-3}d(x, y) = 2^{-8}d(x, y)$. □

Lemma 24. *There exist universal constants $C_1', C_2' > 0$ such that w.h.p for any $\epsilon > 0$ and any $(x, y) \in \hat{G}(\epsilon)$:*

$$C_2' \cdot d(x, y) \le \|f(x) - f(y)\|_p \le C_1' (\ln(1/\epsilon)/\kappa + 1) \cdot d(x, y).$$

7. PARTIAL EMBEDDING, SCALING DISTORTION AND THE ℓ_Q-DISTORTION

The following lemma states that lower bounds on the l_q-distortion follow from lower bound on $(1-\epsilon)$-partial embeddings. Applying this on the lower bound results from [1] we obtain the tightness of our bounds.

LEMMA 25 (PARTIAL EMBEDDING VS. ℓ_q-DISTORTION). *Let Y be a target metric space, let $\mathcal{X}$ be a family of metric spaces. If for any $\epsilon \in [0,1)$, there is a lower bound of $\alpha(\epsilon)$ on the distortion of $(1-\epsilon)$ partial embedding of metric spaces in $\mathcal{X}$ into Y, then for any $1 \le q \le \infty$, there is a lower bound of $\frac{1}{2}\alpha(2^{-q})$ on the ℓ_q-distortion of embedding metric spaces in $\mathcal{X}$ into Y.*

PROOF. For any $1 \le q \le \infty$ set $\epsilon = 2^{-q}$ and let $X \in \mathcal{X}$ be a metric space such that any $(1-\epsilon)$ partial embedding into Y has distortion at least $\alpha(\epsilon)$. Now, let f be an embedding of X into Y. It follows that there are at least $\epsilon\binom{n}{2}$ pairs $\{u,v\} \in \binom{X}{2}$ such that $\mathrm{dist}_f(u,v) \ge \alpha(\epsilon)$. Therefore:

$$\left(\mathbb{E}\left[\mathrm{dist}_f(u,v)^q\right]\right)^{1/q} \ge \left(\epsilon\alpha(\epsilon)^q\right)^{1/q} \ge \left(2^{-q}\alpha\left(2^{-q}\right)^q\right)^{1/q} = \frac{1}{2}\alpha\left(2^{-q}\right).$$

□

In what follows we prove Lemma 1. We give here the proof for the *distortion of average*. The proof for the *average distortion* is somewhat simpler.

PROOF. We may restrict to the case $\Phi(\Pi) \le \binom{n}{2}$. Otherwise $\hat{\Phi}(\Pi) > \binom{n}{2}$ and therefore $\mathrm{distnorm}_q^{(\Pi)}(f) \le \mathrm{dist}(f) \le \alpha(\hat{\Phi}(\Pi)^{-1})$. Recall that

$$\mathrm{distnorm}_q^{(\Pi)}(f) = \frac{\mathbb{E}_\Pi[d_Y(f(u),f(v))^q]^{1/q}}{\mathbb{E}_\Pi[d_X(u,v)^q]^{1/q}}.$$

For $\epsilon \in [0,1)$ recall that $\hat{G}(\epsilon) = \{\{x,y\} \in \binom{X}{2} | d(x,y) \ge \max\{r_{\epsilon/2}(x), r_{\epsilon/2}(y)\}\}$. Since $(f,\hat{G})$ is a $(1-\epsilon)$-partial embedding for any $\epsilon \in [0,1)$ we have that for each $\{u,v\} \in \hat{G}(\epsilon)$, $\mathrm{dist}_f(u,v) \le \alpha(\epsilon)$. Let $\hat{G}_i = \hat{G}(2^{-i}\hat{\Phi}(\Pi)^{-1}) \setminus \hat{G}(2^{-(i-1)}\hat{\Phi}(\Pi)^{-1})$. We first need to prove the following property:

$$\sum_{\{u,v\}\in\hat{G}_i} d_X(u,v)^q \le 2^{-i}\hat{\Phi}(\Pi)^{-1} \sum_{u\ne v\in X} d_X(u,v)^q. \quad (10)$$

To prove this fix some $u \in X$. Let $S = \{v | \{u,v\} \notin \hat{G}(2^{-(i-1)}\hat{\Phi}(\Pi)^{-1})\}$. Then $S = B(u, r_{2^{-i}\hat{\Phi}(\Pi)^{-1}}(u))$. Thus, $|S|/\binom{n}{2}2^{-i}\hat{\Phi}(\Pi)^{-1}$ and for each $v \in S$, $v' \in \bar{S}$ we have $d(u,v) \le d(u,v')$. It follows that:

$$\begin{aligned}\sum_{v;u\ne v\in X} d_X(u,v)^q &= \sum_{v\in S} d_X(u,v)^q + \sum_{v\in\bar{S}} d_X(u,v)^q \\ &\ge |S|\cdot\frac{\sum_{v\in S} d_X(u,v)^q}{|S|} + |\bar{S}|\cdot\frac{\sum_{v\in S} d_X(u,v)^q}{|S|} \\ &= \frac{\binom{n}{2}}{|S|}\sum_{v\in S} d_X(u,v)^q.\end{aligned}$$

Denote $T = \lfloor \log\left(\binom{n}{2}\hat{\Phi}(\Pi)^{-1}\right)\rfloor$. Since α is a monotonic non-increasing function, it follows that

$$\begin{aligned}&\mathbb{E}_\Pi[d_Y(f(u),f(v))^q] = \sum_{u\ne v\in X}\pi(u,v)d_Y(f(u),f(v))^q = \\ &\sum_{u\ne v\in X}\pi(u,v)d_X(u,v)^q\mathrm{dist}_f(u,v)^q \\ &\le \sum_{\{u,v\}\in\hat{G}(\hat{\Phi}(\Pi)^{-1})}\pi(u,v)d_X(u,v)^q\alpha(\hat{\Phi}(\Pi)^{-1})^q + \\ &\sum_{i=1}^{T}\sum_{\{u,v\}\in\hat{G}_i}\pi(u,v)d_X(u,v)^q\alpha(2^{-i}\hat{\Phi}(\Pi)^{-1})^q \\ &\le \sum_{u\ne v\in X}\pi(u,v)d_X(u,v)^q\cdot\alpha(\hat{\Phi}(\Pi)^{-1})^q + \min_{w\ne z\in X}\pi(w,z)\cdot \\ &\sum_{i=1}^{T}\sum_{\{u,v\}\in\hat{G}_i} d_X(u,v)^q\cdot\hat{\Phi}(\Pi)\cdot\alpha(2^{-i}\hat{\Phi}(\Pi)^{-1})^q.\end{aligned}$$

Now using (10) we get that

$$\begin{aligned}&\mathbb{E}_\Pi[d_Y(f(u),f(v))^q] \\ &\quad\le \sum_{u\ne v\in X}\pi(u,v)d_X(u,v)^q\cdot\alpha(\hat{\Phi}(\Pi)^{-1})^q + \\ &\quad\sum_{i=1}^{T}\sum_{u\ne v\in X}\pi(u,v)d_X(u,v)^q\cdot 2^{-i}\cdot\alpha(2^{-i}\hat{\Phi}(\Pi)^{-1})^q \\ &\quad\le \left[\alpha(\hat{\Phi}(\Pi)^{-1})^q + \left(2\int_{\frac{1}{2}\binom{n}{2}^{-1}\hat{\Phi}(\Pi)}^{1}\alpha(x\hat{\Phi}(\Pi)^{-1})^q dx\right)\right] \\ &\quad\cdot\mathbb{E}_\Pi[d_X(u,v)^q].\end{aligned}$$

□

8. ALGORITHMIC APPLICATIONS

Consider an optimization problem defined with respect to weights $c(u,v)$ in a graph or in a metric space, where the solution involves minimizing the sum over distances weighted according to c: $\sum_{u,v} c(u,v)d(u,v)$. It is common for many optimization problem that such a term appears either in the objective function or in its LP relaxation.

Define $\hat{\Phi}(c) = \min\{n, \frac{\max_{u,v} c(u,v)}{\min_{u,v} c(u,v)}\}$. The lemma below, based on Theorem 9 and Theorem 11, summarizes the propositions which yields $O(\log\hat{\Phi}(c))$ approximation algorithms for the optimization problems mentioned in Section 1.3.3.

LEMMA 26. *Let X be a metric space, with a weight function on the pairs $c: \binom{X}{2} \to \mathbb{R}_+$. Then:*

1. *There exists an embedding $f : X \to L_p$ such that for any weight function c:*

$$\frac{\sum_{\{u,v\}\in\binom{X}{2}} c(u,v)\|f(u)-f(v)\|_p}{\sum_{\{u,v\}\in\binom{X}{2}} c(u,v)d_X(u,v)} \le O(\log\hat{\Phi}(c)).$$

2. *There exist a set of ultrametrics $\mathcal{S}$ and a probabilistic embedding $\hat{\mathcal{F}}$ of X into $\mathcal{S}$ such that for any weight function c:*

$$\frac{\mathbb{E}_{f\sim\hat{\mathcal{F}}}\left[\sum_{\{u,v\}\in\binom{X}{2}} c(u,v)d_Y(f(u),f(v))\right]}{\sum_{\{u,v\}\in\binom{X}{2}} c(u,v)d_X(u,v)} \le O(\log\hat{\Phi}(c)).$$

9. REFERENCES

[1] I. Abraham, Y. Bartal, H. Chan, K. Dhamdhere, J. Kleiberg A. Gupta, O. Neiman, and A. Slivkins. Metric embedding with relaxed guarantees. FOCS 05, pages 83–100.

[2] I. Abraham, Y. Bartal, and O. Neiman. On Embedding of Finite Metric Spaces into Hilbert Space. Tech. Report, 2006.

[3] S. Arora, J. R. Lee, and A. Naor. Euclidean distortion and the sparsest cut. STOC 05, pages 553-562.

[4] Y. Aumann and Y. Rabani. An o(log k) approximate min-cut max-flow theorem and approximation algorithm. *SIAM J. Comput.*, 27(1):291–301, 1998.

[5] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. STOC 04, pages 222–231.

[6] V. Athitsos and S. Sclaroff. Database indexing methods for 3d hand pose estimation. In *Gesture Workshop*, pages 288–299, 2003.

[7] Y. Bartal. Probabilistic approximation of metric spaces and its algorithmic applications. FOCS 96, pages 184–193.

[8] Y. Bartal. On approximating arbitrary metrics by tree metrics. STOC 98, pages 183–193.

[9] Y. Bartal. Graph decomposition lemmas and their role in metric embedding methods. ESA 04, pages 89–97.

[10] Y. Bartal. Embedding Finite Metric Spaces in Low Dimension. Tech. Report, 2006.

[11] Y. Bartal, B. Bollobás, and M. Mendel. Ramsey-type theorems for metric spaces with applications to online problems. In Special issue of Journal of Computer and System Science, 2002.

[12] Y. Bartal, M. Charikar, and D. Raz. Approximating min-sum k-clustering in metric spaces. STOC 01, pages 11–20.

[13] Y. Bartal, N. Linial, M. Mendel, and A. Naor. Limitations to frechet embedding of metric spaces. *Israel J. Math.*, 151: 111-124, 2006.

[14] Y. Bartal, N. Linial, M. Mendel, and A. Naor. On metric ramsey-type phenomena. *Annals Math*, 162(2): 643-709, 2005.

[15] J. Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel J. Math.*, 52(1-2):46–52, 1985.

[16] G. Calinescu, H. J. Karloff, and Y. Rabani. Approximation algorithms for the 0-extension problem. SODA 01, pages 8–16.

[17] M. Costa, M. Castro, A. I. T. Rowstron, and P. B. Key. Pic: Practical internet coordinates for distance estimation. ICDCS 04, pages 178–187.

[18] R. Cox, F. Dabek, M. F. Kaashoek, J. Li, and R. Morris. Practical, distributed network coordinates. SIGCOMM 04, 34(1):113–118.

[19] G. Even, J. S. Naor, S. Rao, and B. Schieber. Divide-and-conquer approximation algorithms via spreading metrics. *J. ACM*, 47(4):585–616, 2000.

[20] U. Feige, M. T. Hajiaghayi, and J. R. Lee. Improved approximation algorithms for minimum-weight vertex separators. STOC 05, pages 563–572.

[21] J. Fakcharoenphol, C. Harrelson, S. Rao, and K. Talwar. An improved approximation algorithm for the 0-extension problem. SODA 03, pages 257–265.

[22] J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. STOC 03, pages 448–455.

[23] T. Figiel, J. Lindenstrauss, and V. D. Milman. The dimension of almost spherical sections of convex bodies. *Acta Math.*, 139(1-2):53–94, 1977.

[24] P. Francis, S. Jamin, C. Jin, Y Jin, D. Raz, Y. Shavitt, and L. Zhang. Idmaps: a global internet host distance estimation service. *IEEE/ACM Trans. Netw.*, 9(5):525–540, 2001.

[25] N. Garg, V. V. Vazirani, and M. Yannakakis. Approximate max-flow min-(multi)cut theorems and their applications. STOC 93, pages 698–707.

[26] M. Gromov. Filling Riemannian manifolds. *J. Differential Geom.*, 18(1):1–147, 1983.

[27] E. Halperin, J. Buhler, R. M. Karp, R. Krauthgamer, and B. Westover. Detecting protein sequence conservation via metric embeddings. ISMB 03, pages 122–129.

[28] G. Hjaltason and H. Samet. Contractive embedding methods for similarity searching in metric spaces, Tech. Report, 2000.

[29] G. Hristescu and M. Farach-Colton. Cofe: A scalable method for feature extraction from complex objects. DaWaK 00, pages 358–371.

[30] P. Indyk. Algorithmic applications of low-distortion geometric embeddings. FOCS 01, pages 10–33.

[31] P. Indyk and J. Matoušek. Low-distortion embeddings of finite metric spaces. Handbook of Discrete and Computational Geometry, 2004.

[32] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability*, pages 189–206. Amer. Math. Soc., 1984.

[33] J. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: metric labeling and markov random fields. *J. ACM*, 49(5):616–639, 2002.

[34] J. Kleinberg, A. Slivkins, and T. Wexler. Triangulation and embedding using small sets of beacons. FOCS 04, pages 444–453.

[35] R. Krauthgamer, J. R. Lee, M. Mendel, and A. Naor. Measured descent: A new embedding method for finite metrics. FOCS 04, pages 434–443.

[36] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[37] J. B. Kruskal and M. Wish. *Multidimensional Scaling*. M. Sage Publications, CA, 1978.

[38] J. R. Lee, M. Mendel, and A. Naor. Metric structures in l1: Dimension, snowflakes, and average distortion. LATIN 04, pages 401–412.

[39] T. Leighton and S. Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *J. ACM*, 46(6):787–832, 1999.

[40] H. Lim, J C. Hou, and C. Choi. Constructing internet coordinate system based on delay measurement. SIGCOMM 03, pages 129–142.

[41] N. Linial. Finite metric spaces- combinatorics, geometry and algorithms. ICM 2002.

[42] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

[43] N. Linial and M. Saks. Decomposing graphs into regions of small diameter. SODA 91, pages 320–330.

[44] J. Matoušek. On the distortion required for embedding finite metric spaces into normed spaces. Israel Journal of Mathematics, 93:333–344, 1996.

[45] J. Matoušek. Note on bi-lipschitz embeddings into low-dimensional euclidean spaces. *Comment. Math. Univ. Carolinae*, 31:589–600, 1990.

[46] J. Matoušek. On embedding expanders into l_p spaces. *Israel J. Math.*, 102:189–197, 1997.

[47] J. Matoušek. *Lectures on discrete geometry*. Springer-Verlag, New York, 2002.

[48] T. S. Eugene Ng and H. Zhang. Predicting internet network distance with coordinates-based approaches. INFOCOM 02, pages 178–187.

[49] P. Pardalos, F. Rendl, and H. Wolkowicz. The quadratic assignment problem: a survey and recent developments. In *Quadratic assignment and related problems*, pages 1–42. Amer. Math. Soc., 1994.

[50] Y. Rabinovich. On average distortion of embedding metrics into the line and into l1. STOC 03, pages 456–462.

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

[52] S. Rao. Small distortion and volume preserving embeddings for planar and Euclidean metrics. SoCG 99, pages 300–306.

[53] S. Rao and A. Richa. New approximation techniques for some ordering problems. SODA 98, pages 211–219.

[54] Y. Shavitt and T. Tankel. Big-bang simulation for embedding network distances in euclidean space. *IEEE/ACM Trans. Netw.*, 12(6):993–1006, 2004.

[55] L. Tang and M. Crovella. Geometric exploration of the landmark selection problem. In *PAM*, pages 63–72, 2004.

[56] M. Thorup and U. Zwick. Approximate distance oracles. *J. ACM*, 52(1):1–24, 2005.

[57] B. Wu, G. Lancia, V. Bafna, K. Chao, R. Ravi, and C. Tang. A polynomial time approximation scheme for minimum routing cost spanning trees. SODA 98, pages 21–32.