

On Embedding of Finite Metric Spaces into Hilbert Space

Ittai Abraham*

Yair Bartal[†]
Hebrew University

Ofer Neiman[‡]

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 main criteria for the quality of an embedding is its average distortion over all pairs.

A celebrated theorem of Bourgain 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, is this the case for the *average distortion*?

Our main result is a strengthening of Bourgain's theorem providing an embedding with *constant* average distortion for arbitrary metric spaces. 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 $1 \leq q \leq \infty$ *simultaneously*.

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 $f : X \rightarrow Y$ is called an *embedding* of X into Y . An embedding is *non-contractive* if for any $u \neq v \in X$: $d_Y(f(u), f(v)) \geq d_X(u, v)$. For a non-contractive embedding the *distortion* of f is defined as $\text{dist}(f) = \sup_{u \neq v \in X} \text{dist}_f(u, v)$, where $\text{dist}_f(u, v) = \frac{d_Y(f(u), f(v))}{d_X(u, v)}$.

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. The main cornerstone of the field has been the following theorem by Bourgain [13]:

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

Bourgain also showed that this bound is nearly tight and later Linial, London and Rabinovich [33] prove 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 all real-world applications of metric embeddings *average distortion* and similar notions are used for evaluating the embedding's performance in practice, for example see [23, 24, 4, 22, 41, 42]. Moreover, in some

*email: ittaia@cs.huji.ac.il.

[†]email: yair@cs.huji.ac.il. Supported in part by a grant from the Israeli Science Foundation (195/02).

[‡]email: neiman@cs.huji.ac.il. Supported in part by a grant from the Israeli Science Foundation (195/02).

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 pairs¹.

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

Theorem 2 (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 $1 \leq q \leq \infty$, 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 2 is a corollary of the following theorem:

Theorem 3 (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 $1 \leq q \leq \infty$, $\text{dist}_q(f) = O(\min\{q, \log n\})$.*

Another variant of average distortion that is natural is what we call *distortion of average*: $\text{dist}_{\text{avg}}(f) = \frac{\sum_{u \neq v \in X} d_Y(f(u), f(v))}{\sum_{u \neq v \in X} d(u, v)}$, which can be naturally extended to its ℓ_q -normed extension termed *distortion of ℓ_q -norm*. Theorems 2 and 3 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 [28] and later studied in many models and variants. Multidimensional scaling methods (see [29, 23]) 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 [24].

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 [37] studied the notion of distortion of average² 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 [18]) that the work of Arora, Rao and Vazirani on Sparsest Cut [3] 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*³ (a notion defined by Gromov [21]) (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 [30] 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 distortion⁴.

Network embedding. Our work is largely motivated by a surge of interest in the networking community on performing *passive distance estimation* (see e.g. [19, 35, 32, 15, 41, 14]), 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 [15]). 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.

¹Similar statements hold for the more recent metric embeddings of [39, 27] as well.

²Usually this notion was called average distortion but the name is somewhat confusing.

³This notion is used here somewhat differently than its original purpose.

⁴The bound given in [30] is $O(\sqrt{\log n})$ which applies to a somewhat weaker notion.

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 [26]. 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 [26] define the notion of a $(1 - \epsilon)$ -partial embedding⁵ 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 [26]. In Abraham et. al. [1] it 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 [26] define a stronger notion of *scaling distortion*⁶. 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 [26], 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 [26, 1] is whether embeddings with small scaling distortion exist for embedding into Euclidean space. We give the following theorem⁷ which lies in the heart of the proof of Theorem 3:

Theorem 4. *For every finite metric space (X, d) , there exists an embedding of X into Euclidean space with scaling distortion $O(\log \frac{1}{\epsilon})$.*

Novel Techniques. While [1] certainly uses the state of the art methods in finite metric embedding, it appears all these techniques break when attempting to prove Theorem 4. Indeed, to prove the theorem and its generalizations we present novel embedding techniques.

Our embeddings are based on *probabilistic partitions* of metric spaces [6] originally defined in the context of *probabilistic embedding* of metric spaces and later used in the context of metric embedding in [7, 17, 8, 39, 27].

We make use of novel probabilistic partitions [9] with refined properties which allow stronger and more general results on embedding of finite metric spaces which cannot be achieved using any of the previous methods. These constructions as well as the embedding techniques based on them were developed in conjunction with this paper. Moreover, here we make a far more sophisticated use of these partitions.

Our embeddings into L_p are based on *uniformly padded hierarchical probabilistic partitions*. The properties of these partitions allow to define sophisticated embeddings in a natural way. We believe that the application of these techniques as presented here demonstrates their vast versatility and we expect that more applications will be found in the near future.

We stress that although the similarity in approach and techniques to [9] the proof of the main result in this paper is considerably more involved. This is not surprising 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 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 important 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 present a solution for another open problem from [26, 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* [36, 25]. We also extend our concepts to analyze Distance Oracles of Thorup and Zwick [44] providing results with strong relation to

⁵Called “embeddings with ϵ -slack” in [26].

⁶Called “gracefully degrading distortion” in [26].

⁷In fact in this theorem the definition of scaling distortion is even stronger. This is explained in detail in the appropriate section.

the questions addressed by [26]. 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.

The rest of the introduction provides a detailed description of the new concepts and results. We then provide the proof of Theorem 6 which contains the main technical contribution of this paper. The rest of the results are given in the detail in the appendix.

1.1 ℓ_q -Distortion and the Main Theorem

Given two metric spaces (X, d_X) and (Y, d_Y) an *injective* mapping $f : X \rightarrow Y$ 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* [21] 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 , $\text{dist}_f : \binom{X}{2} \rightarrow \mathbb{R}^+$, 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 *average distortion* 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 $\pi : \binom{X}{2} \rightarrow [0, 1]$, 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:⁸ $\hat{\Phi}(\Pi) = 2 \min\{\Phi(\Pi), \binom{n}{2}\}$

Theorem 5 (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 5 by showing an embedding with $\text{dist}_q^{*(\Pi)}(f) = O(\min\{q, (\log \lambda_X)^{1-\frac{1}{p}} (\log n)^{1/p}\} + \log \hat{\Phi}(\Pi))$.

The proof of Theorem 5 follows directly from results on embedding with scaling distortion, discussed in the next paragraph.

1.2 Partial Embedding, Scaling Distortion and Additional Results

Following [26] 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: $\text{dist}(f, G) = \sup_{\{u,v\} \in G} \text{dist}_f(u, v)$.

For $\epsilon \in [0, 1)$, a $(1 - \epsilon)$ -*partial embedding* is a partial embedding such that $|G| \geq (1 - \epsilon) \binom{n}{2}$.⁹

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 is a pair $(f, \hat{G}(\epsilon))$, where f is an embedding.¹⁰

Definition 7 (Scaling Distortion). Given two metric spaces (X, d_X) and (Y, d_Y) and a function $\alpha : [0, 1) \rightarrow \mathbb{R}^+$, we say that an embedding $f : X \rightarrow Y$ 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.¹¹

Definition 8 (Partial/Scaling Probabilistic 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* consist 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: $\text{dist}(\hat{\mathcal{F}}) = \sup_{\{u,v\} \in \hat{G}(\epsilon)} \mathbb{E}_{(f, \hat{G}(\epsilon)) \sim \hat{\mathcal{F}}}[\text{dist}_f(u, v)]$.

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 $f : X \rightarrow Y$ with scaling distortion α then for any distribution Π over $\binom{X}{2}$:¹²*

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

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

⁸The factor of 2 in the definition is placed solely for the sake of technical convenience.

⁹Note that the embedding is *strictly* partial only if $\epsilon \geq 1/\binom{n}{2}$.

¹⁰It 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.

¹¹Our upper bounds use this definition, while our lower bounds hold also for the non-coarsely case.

¹²Assuming the integral is defined. We note that lemma is stated using the integral for presentation reasons.

Combined with the following theorem we obtain Theorem 5. 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 6 (Scaling Distortion Theorem into L_p). *Let $1 \leq p \leq \infty$. For any n -point metric space (X, d) there exists an embedding $f : X \rightarrow L_p$ 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 7 (Probabilistic 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 $\binom{X}{2}$: $\text{dist}_q^{*(\Pi)}(\hat{\mathcal{F}}) = O(\min\{q, \log n\} + \log \hat{\Phi}(\Pi))$.*

For $q = 1$ and for a given fixed distribution Theorem 7 can be given a classic embedding (deterministic) version:

Theorem 8 (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 $\text{avgdist}^{(\Pi)}(f) = O(\log \hat{\Phi}(\Pi))$ and $\text{distavg}^{(\Pi)}(f') = O(\log \hat{\Phi}(\Pi))$.*

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 [6, 12] is $O(\frac{1}{\epsilon})$, the lower bound which follows from [1] by applying the $\Omega(n)$ lower bound on embedding into trees of [38] 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 [6]. This result is related both in spirit and techniques to recently developed metric Ramsey theorems [10, 12] where embeddings into ultrametrics and k -HSTs play a central role.

Theorem 9 (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 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 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* [31, 5, 33, 3, 2], *multi cut* [20], *minimum linear arrangement* [16, 40], *embedding in d -dimensional meshes* [16, 8], *multiple sequence alignment* [45] and *uncapacitated quadratic assignment* [36, 25].

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 [36]) and has been one of the main motivations for the work of Kleinberg and Tardos on metric labelling [25].

We also present a different use of our results for the problem of *min-sum k -clustering* [11].

1.4 Distance Oracles

Thorup and Zwick [44] 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 and a distance oracles scaling distortion. We present distance oracle analogues to our theorems on embeddings.

2 Proof of Main Result

In this section we prove Theorem 6. We make use of a new type of probabilistic partition described below.

2.1 Probabilistic Hierarchical Partitions

Definition 9 (Partition). Let (X, d) be a finite metric space. A partition P of X is a collection of disjoint sets $\mathcal{C}(P) = \{C_1, C_2, \dots, C_t\}$ 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 10 (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 11 (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 12 (Probabilistic 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 13 (Uniformly Padded Probabilistic Hierarchical Partition). Let $\hat{\mathcal{H}}$ be a probabilistic k -hierarchical partition. Given collection of functions $\eta = \{\eta_{P,i} : X \rightarrow [0, 1] | i \in I, P_i \in P, P \in \mathcal{H}\}$ 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} .

To state the main lemma we require the following additional notion:

Definition 14. 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 now present the main lemma on the existence of hierarchical partitions which are the main building block of our embedding. A variant of this lemma is given in [9].

Lemma 2 (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 $\xi = \{\xi_{P,i} : X \rightarrow \{0, 1\} | P \in \mathcal{H}, i \in I\}$, 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 $P \in \mathcal{H}$, $0 < i \in I$, $P_i \in P$:

- $\sum_{j \leq i} \xi_{P,j}(x) \eta_{P,j}(x)^{-1} \leq 2^{11} \ln \left(\frac{n}{|B(x, \Delta_{i+4})|} \right) / \ln(1/\delta)$.
- If $\xi_{P,i}(x) = 1$ then: $\eta_{P,i}(x) \leq 2^{-8}$.
- If $\xi_{P,i}(x) = 0$ then: $\eta_{P,i}(x) \geq 2^{-8}$ and $\bar{\rho}(x, \Delta_{i-1}, \Gamma) < 1/\delta$.

2.2 The Proof

In this section we prove the following generalization of Theorem 6. The proof relies on Lemma 2.

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

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

Fix $t, 1 \leq t \leq D$. 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 2, 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 $x \in X, 0 < i \in I, \phi_i^{(t)} : X \rightarrow \mathbb{R}^+$, by $\phi_i^{(t)}(x) = \xi_{P_i(x)} \eta_{P_i(x)}^{-1}$.

Claim 3. *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 $\psi_i^{(t)} : X \rightarrow \mathbb{R}^+$ 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 $g_i^{(t)} : X \rightarrow \mathbb{R}^+$ is defined as: $g_i^{(t)}(x) = \min\{\phi_i^{(t)}(x) \cdot d(x, X \setminus P_i(x)), \Delta_i\}$. Define $\bar{g}_i^{(t)} : X \times X \rightarrow \mathbb{R}^+$ as follows: $\bar{g}_i^{(t)}(x, y) = \min\{\phi_i^{(t)}(x) \cdot d(x, y), \Delta_i\}$ (Note that $\bar{g}_i^{(t)}$ is nonsymmetric).

Claim 4. *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. 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 $g_i^{(t)}(x) - g_i^{(t)}(y) \leq \bar{g}_i^{(t)}(x, y)$. The bound $g_i^{(t)}(x) - g_i^{(t)}(y) \leq \Delta_i$ is immediate. To prove $g_i^{(t)}(x) - g_i^{(t)}(y) \leq \phi_i^{(t)}(x) \cdot d(x, y)$ 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 3 $\phi_i^{(t)}(y) = \phi_i^{(t)}(x)$ and therefore

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

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

Next, consider Case 2 where $P_i(x) \neq P_i(y)$. In this case we have that $d(x, X \setminus P_i(x)) \leq d(x, y)$ which implies that $\psi_i^{(t)}(x) \leq g_i^{(t)}(x) \leq \bar{g}_i^{(t)}(x, y)$. \square

Next, we define the function $\mu^{(t)}$, based on the embedding technique of Bourgain [13] and its generalization by Matoušek [34]. Let $T' = \lceil \log_s n \rceil$ and $K = \{k \in \mathbb{N} | 1 \leq k \leq T'\}$. 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:

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

We make the following two simple observations:

Claim 5. *For every $i \in I$: (1) For any $x \in X$: $|\{k | i \in I_k(x)\}| \leq 2$. (2) For every $k \in K$: the function $i \in I_k(x)$ is uniform with respect to P_i .*

We define $i_k : X \rightarrow I$, where $i_k(x) = \min\{i | i \in I_k(x)\}$. For each $k \in K$ we define a function $\mu_k^{(t)} : X \rightarrow \mathbb{R}^+$ and for $x \in X$ let $\mu^{(t)}(x) = \sum_{k \in K} \mu_k^{(t)}(x)$. Let $\Phi_0 = 2^8$. 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 $h_i^{(t)} : X \rightarrow \mathbb{R}^+$ is defined as: $h_i^{(t)}(x) = \min\{\Phi_0 \cdot d(x, X \setminus P_i(x)), \Delta_i\}$. Define $\bar{h}_i^{(t)} : X \times X \rightarrow \mathbb{R}^+$ 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 4:

Claim 6. *For any $k \in K$ and $x, y \in X$: $\mu_k^{(t)}(x) - \mu_k^{(t)}(y) \leq \bar{h}_{i_k(x)}^{(t)}(x, y)$.*

Proof. Let $i = i_k(x)$. We have two cases. In Case 1, assume $P_i(x) = P_i(y)$. Let $i' = i_k(y)$. By Claim 5 we have that $i \in I_k(y)$, implying $i' \leq i$. Since P is a hierarchical partition we have that $P_{i'}(x) = P_{i'}(y)$. Hence Claim 5 implies that $i' \in I_k(x)$, so that $i \leq i'$, which implies $i' = i$.

Since $h_i^{(t)}(x) \leq \Delta_i$ we have that $\mu_k^{(t)}(x) - \mu_k^{(t)}(y) \leq \Delta_i$. To prove $\mu_k^{(t)}(x) - \mu_k^{(t)}(y) \leq \Phi_0 \cdot d(x, y)$ consider the value of $\mu_k^{(t)}(y)$. If $\mu_k^{(t)}(y) = \frac{1}{4}d(y, A_k^{(t)})$ then $\mu_k^{(t)}(x) - \mu_k^{(t)}(y) \leq \frac{1}{4}(d(x, A_k^{(t)}) - d(y, A_k^{(t)})) \leq \frac{1}{4} \cdot d(x, y) \leq \Phi_0 \cdot d(x, y)$. Otherwise, if $\mu_k^{(t)}(y) = \Phi_0 \cdot d(y, X \setminus P_i(x))$ then

$$\mu_k^{(t)}(x) - \mu_k^{(t)}(y) \leq \Phi_0 \cdot (d(x, X \setminus P_i(x)) - d(y, X \setminus P_i(x))) \leq \Phi_0 \cdot d(x, y).$$

Finally, if $\mu_k^{(t)}(y) = \Delta_i$ then $\mu_k^{(t)}(x) - \mu_k^{(t)}(y) \leq \Delta_i - \Delta_i = 0$.

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

$$\mu_i^{(t)}(x) - \mu_i^{(t)}(y) \leq \mu_i^{(t)}(x) \leq h_i^{(t)}(x) \leq \bar{h}_i^{(t)}(x, y).$$

□

Lemma 7. *There exists a universal constant $C_1 > 0$ such that for any $\epsilon > 0$ and any $(x, y) \in \hat{G}(\epsilon)$:*

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

Proof. From Claim 4 we get

$$\sum_{0 < i \in I} (\psi_i^{(t)}(x) - \psi_i^{(t)}(y)) \leq \sum_{0 < i \in I} \bar{g}_i^{(t)}(x, y)$$

Define ℓ to be largest such that $\Delta_{\ell+4} \geq d(x, y) \geq \max\{r_{\epsilon/2}(x), r_{\epsilon/2}(y)\}$. If no such ℓ exists then let $\ell = 0$. By Lemma 2 we have

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

We also have that $\sum_{\ell < i \in I} \bar{g}_i^{(t)}(x, y) \leq \sum_{\ell < i \in I} \Delta_i \leq \Delta_\ell \leq 4^5 d(x, y)$.

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

From Claim 6 we get $\sum_{k \in K} (\mu_k^{(t)}(x) - \mu_k^{(t)}(y)) \leq \sum_{k \in K} \bar{h}_{i_k(x)}^{(t)}(x, y)$.

Let k' be the largest such that $s^{k'} \leq \epsilon n/2$. We have

$$\begin{aligned} \sum_{k' < k \in K} \bar{h}_{i_k(x)}^{(t)}(x, y) &\leq \sum_{k' < k \in K} \Phi_0 \cdot d(x, y) = 2^8 \cdot (\lceil \log_s n \rceil - \lfloor \log_s(\epsilon n/2) \rfloor) \cdot d(x, y) \\ &\leq 2^8 \cdot (\ln(2/\epsilon)/\kappa + 2) \cdot d(x, y). \end{aligned}$$

Now, if $k \leq k'$ and $i \in I_k(x)$ then for any $u \in P_i(x)$ we have $|B(x, 16\Delta_i)| \leq |B(u, 16\Delta_{i-1})| \leq s^k \leq \epsilon n/2$. It follows that $d(x, y) \geq r_{\epsilon/2}(x) \geq 16\Delta_i$. Denote the largest i satisfying the last inequality ℓ' . Using Claim 5 we get

$$\sum_{k' \geq k \in K} \bar{h}_{i_k(x)}^{(t)}(x, y) = \sum_{k' \geq k \in K} \Delta_{i_k(x)} \leq \sum_{\ell' \geq i \in I} \sum_{k \in K, i \in I_k(x)} \Delta_i \leq \sum_{\ell' \geq i \in I} 2\Delta_i \leq 4\Delta_{\ell'} \leq d(x, y)/4.$$

It follows that

$$|\mu^{(t)}(x) - \mu^{(t)}(y)| = \left| \sum_{k \in K} (\mu_k^{(t)}(x) - \mu_k^{(t)}(y)) \right| \leq 2^8 (\ln(2/\epsilon)/\kappa + 3) \cdot d(x, y).$$

Therefore

$$|f^{(t)}(x) - f^{(t)}(y)| = |\psi^{(t)}(x) + \mu^{(t)}(x) - \psi^{(t)}(y) - \mu^{(t)}(y)| \leq 2^{12} (\ln(2/\epsilon)/\kappa + 1) \cdot d(x, y).$$

□

Lemma 8. *There exists a universal constant $C_2 > 0$ such that for any $x, y \in X$, with probability at least $e^{-5\kappa}/4$:*

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

Proof. Let $0 < \ell \in I$ be such that $4\Delta_{\ell-1} \leq d(x, y) \leq 16\Delta_{\ell-1}$. 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 $\Pr[B(x, \eta_{P,\ell}(x)\Delta_\ell) \subseteq P_\ell(x)] \geq 1/s$. Therefore with probability at least $1/s$:

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

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

- $|f^{(t)}(x) - f^{(t)}(y) - (\psi_\ell^{(t)}(x) - \psi_\ell^{(t)}(y))| \geq \frac{1}{2}\Delta_\ell$. 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 $\psi_\ell^{(t)}(x) = \psi_\ell^{(t)}(y) = 0$.
- $|f^{(t)}(x) - f^{(t)}(y) - (\psi_\ell^{(t)}(x) - \psi_\ell^{(t)}(y))| \leq \frac{1}{2}\Delta_\ell$. Since $\text{diam}(P_\ell(x)) \leq \Delta_\ell < d(x, y)$ 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 $\psi_\ell^{(t)}(x) - \psi_\ell^{(t)}(y) \geq \Delta_\ell$.

We conclude that with probability at least $1/4s$: $|f^{(t)}(x) - f^{(t)}(y)| \geq \frac{1}{2}\Delta_\ell$.

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

It follows from Lemma 2 that $\max\{\bar{\rho}(x, \Delta_{\ell-1}, \Gamma), \bar{\rho}(y, \Delta_{\ell-1}, \Gamma)\} < s$. 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:

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

using that $d(x, x') \leq \Delta_{\ell-1}$ and $d(x, y') \leq d(x, y) + d(y, y') \leq 17\Delta_{\ell-1}$, 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 $s^{k-1} < |B(x, 32\Delta_{\ell-1})| \leq s^k$. 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 $d(x, u) \leq \Delta_\ell < \Delta_{\ell-1}$ it follows that $s^{k-2} < |B(u, 16\Delta_{\ell-1})| \leq s^k$. This implies that $\ell \in I_k(x)$ and therefore $i_k(x) \leq \ell$. As $\hat{\mathcal{H}}$ is (η, δ) -padded we have the following bound

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

Assume that this event occurs. Since P is hierarchical we get that for every $i \geq \ell$ $B(x, \eta_{P,\ell}(x)\Delta_\ell) \subseteq P_\ell(x) \subseteq P_i(x)$ and in particular this holds for $i = i_k(x)$. As $\xi_{P,\ell}(x) = 0$ we have that $\eta_{P,\ell}(x) \geq 2^{-8} = 1/\Phi_0$. Hence,

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

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

The following is a variant on the original argument in [13, 34]. Define the events: $\mathcal{A}_1 = B(y', \Delta_{\ell-1}/\Gamma) \cap A_k^{(t)} \neq \emptyset$, $\mathcal{A}_2 = B(x', \Delta_{\ell-1}/\Gamma) \cap A_k^{(t)} \neq \emptyset$ and $\mathcal{A}'_2 = [B(x, 32\Delta_{\ell-1}) \setminus B(y', \Delta_{\ell-1})] \cap A_k^{(t)} = \emptyset$. Then for $m \in \{1, 2\}$:

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

using $s \geq 2$. Observe that $d(x', y') \geq d(x, y) - 2\Delta_{\ell-1}/\Gamma \geq (1 - 2/\Gamma)\Delta_{\ell-1} > 2\Delta_{\ell-1}/\Gamma$, 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 $d(y, A_k^{(t)}) \leq d(y, y') + \Delta_{\ell-1}/\Gamma \leq (1 + 1/\Gamma)\Delta_{\ell-1}$. We distinguish between two cases:

- $|f^{(t)}(x) - f^{(t)}(y) - (\mu_k^{(t)}(x) - \mu_k^{(t)}(y))| \geq 3/2 \cdot \Delta_\ell$. In this case there is probability at least $s^{-2}/2$ that event \mathcal{A}_2 occurs, so that $|\mu_k^{(t)}(x) - \mu_k^{(t)}(y)| \leq \frac{1}{4} \max\{d(x, A_k^{(t)}), d(y, A_k^{(t)})\} \leq \frac{1}{4}(1 + 1/\Gamma)\Delta_{\ell-1} = (1 + 1/\Gamma)\Delta_\ell$, by bounding $d(x, A_k^{(t)})$ in the same way as done above for $d(y, A_k^{(t)})$. We therefore get with probability at least $s^{-2}/2$ that $|f^{(t)}(x) - f^{(t)}(y) \geq 3/2 \cdot \Delta_\ell - (1 + 1/\Gamma)\Delta_\ell \geq \Delta_\ell/4$.
- $|f^{(t)}(x) - f^{(t)}(y) - (\mu_\ell^{(t)}(x) - \mu_\ell^{(t)}(y))| \leq 3/2 \cdot \Delta_\ell$. In this case there is probability at least $1/4$ that event \mathcal{A}'_2 occurs. Observe that:

$$\begin{aligned} d(x, B(y', \Delta_{\ell-1}/\Gamma)) &\leq d(x, y) + d(y, y') + \Delta_{\ell-1}/\Gamma \\ &\leq 16\Delta_{\ell-1} + \Delta_{\ell-1} + \Delta_{\ell-1}/\Gamma \leq (17 + 1/\Gamma)\Delta_{\ell-1} < 32\Delta_{\ell-1}, \\ d(x, B(y', \Delta_{\ell-1}/\Gamma)) &\geq d(x, y) - d(y, y') - \Delta_{\ell-1}/\Gamma \\ &\geq 4\Delta_{\ell-1} - \Delta_{\ell-1} - \Delta_{\ell-1}/\Gamma \geq (3 - 1/\Gamma)\Delta_{\ell-1}, \end{aligned}$$

implying that $d(x, A_k^{(t)}) \geq d(x, B(y', \Delta_{\ell-1}/\Gamma)) \geq (3 - 1/\Gamma)\Delta_{\ell-1}$ and therefore $\mu_k^{(t)}(x) \geq \min\{\frac{1}{4}(3 - 1/\Gamma)\Delta_{\ell-1}, \Delta_\ell\} = (3 - 1/\Gamma)\Delta_\ell$. Since $\mu_k^{(t)}(y) \leq \frac{1}{4}d(y, A_k^{(t)}) \leq \frac{1}{4}(1 + 1/\Gamma)\Delta_{\ell-1} = (1 + 1/\Gamma)\Delta_\ell$ we obtain that: $\mu_k^{(t)}(x) - \mu_k^{(t)}(y) \geq (3 - 1/\Gamma)\Delta_\ell - (1 + 1/\Gamma)\Delta_\ell \geq (2 - 2/\Gamma)\Delta_\ell$. We therefore get with probability at least $1/4$ that $|f^{(t)}(x) - f^{(t)}(y)| \geq (2 - 2/\Gamma)\Delta_\ell - 3/2 \cdot \Delta_\ell \geq \Delta_\ell/4$.

We conclude that given \mathcal{A}_1 , with probability at least $s^{-2}/2$: $|f^{(t)}(x) - f^{(t)}(y)| \geq \Delta_\ell/4$.

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

Lemma 9. *There exists a universal constants $C'_1, C'_2 > 0$ such that w.h.p for any $\epsilon > 0$ and any $(x, y) \in \hat{G}(\epsilon)$:*

$$C'_2 \cdot d(x, y) \leq \|f(x) - f(y)\|_p \leq C'_1 (\ln(1/\epsilon)/\kappa + 1) \cdot d(x, y).$$

Proof. By definition $\|f(x) - f(y)\|_p^p = D^{-1} \sum_{1 \leq t \leq D} |f^{(t)}(x) - f^{(t)}(y)|^p$.

Lemma 7 implies that $\|f(x) - f(y)\|_p^p \leq (C_1(\ln(1/\epsilon)/\kappa + 1))^p d(x, y)^p$.

Using Lemma 8 and applying Chernoff bounds we get w.h.p for any $x, y \in X$:

$$\|f(x) - f(y)\|_p^p \geq \frac{1}{8}e^{-5\kappa} (C_2 d(x, y))^p \geq \frac{1}{8} (e^{-5} \cdot C_2 d(x, y))^p.$$

\square

References

- [1] I. Abraham, Y. Bartal, H. Chan, K. Dhamdhere, J. Kleiberg, A. Gupta, O. Neiman, and A. Slivkins. Metric embedding with relaxed guarantees.
- [2] S. Arora, J. R. Lee, and A. Naor. Euclidean distortion and the sparsest cut. In *STOC '05: Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, 2005.
- [3] Sanjeev Arora, Satish Rao, and Umesh Vazirani. Expander flows, geometric embeddings and graph partitioning. In *STOC '04: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 222–231. ACM Press, 2004.
- [4] Vassilis Athitsos and Stan Sclaroff. Database indexing methods for 3d hand pose estimation. In *Gesture Workshop*, pages 288–299, 2003.
- [5] Yonatan Aumann and Yuval Rabani. An $o(\log k)$ approximate min-cut max-flow theorem and approximation algorithm. *SIAM J. Comput.*, 27(1):291–301, 1998.
- [6] Y. Bartal. Probabilistic approximation of metric spaces and its algorithmic applications. In *37th Annual Symposium on Foundations of Computer Science (Burlington, VT, 1996)*, pages 184–193. IEEE Comput. Soc. Press, Los Alamitos, CA, 1996.
- [7] Y. Bartal. On approximating arbitrary metrics by tree metrics. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, pages 183–193, 1998.

- [8] Y. Bartal. Graph decomposition lemmas and their role in metric embedding methods. In *12th Annual European Symposium on Algorithms*, pages 89–97, 2004.
- [9] Y. Bartal. On embedding finite metric spaces in low-dimensional normed spaces, 2005. Submitted.
- [10] Y. Bartal, B. Bollobás, and M. Mendel. Ramsey-type theorems for metric spaces with applications to online problems, 2002. To appear in Special issue of Journal of Computer and System Science.
- [11] Yair Bartal, Moses Charikar, and Danny Raz. Approximating min-sum k-clustering in metric spaces. In *STOC '01: Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 11–20. ACM Press, 2001.
- [12] Yair Bartal, Nathan Linial, Manor Mendel, and Assaf Naor. On metric ramsey-type phenomena. *Annals Math*, 2003. To appear.
- [13] J. Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel J. Math.*, 52(1-2):46–52, 1985.
- [14] Manuel Costa, Miguel Castro, Antony I. T. Rowstron, and Peter B. Key. Pic: Practical internet coordinates for distance estimation. In *24th International Conference on Distributed Computing Systems*, pages 178–187, 2004.
- [15] Russ Cox, Frank Dabek, M. Frans Kaashoek, Jinyang Li, and Robert Morris. Practical, distributed network coordinates. *ACM SIGCOMM Computer Communication Review*, 34(1):113–118, 2004.
- [16] Guy Even, Joseph Seffi Naor, Satish Rao, and Baruch Schieber. Divide-and-conquer approximation algorithms via spreading metrics. *J. ACM*, 47(4):585–616, 2000.
- [17] Jittat Fakcharoenphol, Satish Rao, and Kunal Talwar. A tight bound on approximating arbitrary metrics by tree metrics. In *STOC '03: Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 448–455. ACM Press, 2003.
- [18] U. Feige, M. T. Hajiaghayi, and J. R. Lee. Improved approximation algorithms for minimum-weight vertex separators. In *Annual ACM Symposium on Theory of Computing*, pages 563–572, 2005.
- [19] Paul Francis, Sugih Jamin, Cheng Jin, Yixin Jin, Danny Raz, Yuval Shavitt, and Lixia Zhang. Idmaps: a global internet host distance estimation service. *IEEE/ACM Trans. Netw.*, 9(5):525–540, 2001.
- [20] Naveen Garg, Vijay V. Vazirani, and Mihalis Yannakakis. Approximate max-flow min-(multi)cut theorems and their applications. In *ACM Symposium on Theory of Computing*, pages 698–707, 1993.
- [21] Mikhael Gromov. Filling Riemannian manifolds. *J. Differential Geom.*, 18(1):1–147, 1983.
- [22] Eran Halperin, Jeremy Buhler, Richard M. Karp, Robert Krauthgamer, and B. Westover. Detecting protein sequence conservation via metric embeddings. In *ISMB (Supplement of Bioinformatics)*, pages 122–129, 2003.
- [23] G. Hjaltason and H. Samet. Contractive embedding methods for similarity searching in metric spaces, 2000.
- [24] Gabriela Hristescu and Martin Farach-Colton. Cofe: A scalable method for feature extraction from complex objects. In *DaWaK*, pages 358–371, 2000.
- [25] Jon Kleinberg and Eva Tardos. Approximation algorithms for classification problems with pairwise relationships: metric labeling and markov random fields. *J. ACM*, 49(5):616–639, 2002.
- [26] Jon M. Kleinberg, Aleksandrs Slivkins, and Tom Wexler. Triangulation and embedding using small sets of beacons. In *FOCS*, pages 444–453, 2004.
- [27] R. Krauthgamer, J. R. Lee, M. Mendel, and A. Naor. Measured descent: A new embedding method for finite metrics. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 434–443. IEEE, October 2004.
- [28] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

- [29] Joseph B. Kruskal and Myron Wish. *Multidimensional Scaling*. M. Sage Publications, CA, 1978.
- [30] James R. Lee, Manor Mendel, and Assaf Naor. Metric structures in ℓ_1 : Dimension, snowflakes, and average distortion. In *LATIN*, pages 401–412, 2004.
- [31] Frank Thomson Leighton and Satish Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *J. ACM*, 46(6):787–832, 1999.
- [32] Hyuk Lim, Jennifer C. Hou, and Chong-Ho Choi. Constructing internet coordinate system based on delay measurement. In *3rd ACM SIGCOMM Conference on Internet Measurement*, pages 129–142, 2003.
- [33] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.
- [34] J. Matoušek. Note on bi-lipschitz embeddings into low-dimensional euclidean spaces. *Comment. Math. Univ. Carolinae*, 31:589–600, 1990.
- [35] T. S. Eugene Ng and Hui Zhang. Predicting internet network distance with coordinates-based approaches. In *21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, pages 178–187, 2002.
- [36] P. Pardalos, F. Rendl, and H. Wolkowicz. The quadratic assignment problem: a survey and recent developments. In P. Pardalos and H. Wolkowicz, editors, *Quadratic assignment and related problems (New Brunswick, NJ, 1993)*, pages 1–42. Amer. Math. Soc., Providence, RI, 1994.
- [37] Yuri Rabinovich. On average distortion of embedding metrics into the line and into ℓ_1 . In *STOC '03: Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 456–462. ACM Press, 2003.
- [38] Yuri Rabinovich and Ran Raz. Lower bounds on the distortion of embedding finite metric spaces in graphs. *Discrete & Computational Geometry*, 19(1):79–94, 1998.
- [39] S. Rao. Small distortion and volume preserving embeddings for planar and Euclidean metrics. In *Proceedings of the Fifteenth Annual Symposium on Computational Geometry*, pages 300–306, New York, 1999. ACM.
- [40] S. Rao and A. Richa. New approximation techniques for some ordering problems. In *SODA*, pages 211–219, 1998.
- [41] Yuval Shavitt and Tomer Tanel. Big-bang simulation for embedding network distances in euclidean space. *IEEE/ACM Trans. Netw.*, 12(6):993–1006, 2004.
- [42] Liying Tang and Mark Crovella. Geometric exploration of the landmark selection problem. In *PAM*, pages 63–72, 2004.
- [43] M. Thorup and U. Zwick. Compact routing schemes. In *13th Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA)*, pages 1–10. ACM Press, July 2001.
- [44] Mikkel Thorup and Uri Zwick. Approximate distance oracles. *J. ACM*, 52(1):1–24, 2005.
- [45] Bang Ye Wu, Giuseppe Lancia, Vineet Bafna, Kun-Mao Chao, R. Ravi, and Chuan Yi Tang. A polynomial time approximation scheme for minimum routing cost spanning trees. In *SODA '98: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pages 21–32. Society for Industrial and Applied Mathematics, 1998.

Contents

1	Introduction	1
1.1	ℓ_q -Distortion and the Main Theorem	4
1.2	Partial Embedding, Scaling Distortion and Additional Results	5
1.3	Algorithmic Applications	6
1.4	Distance Oracles	6
2	Proof of Main Result	7
2.1	Probabilistic Hierarchical Partitions	7
2.2	The Proof	8
A	Partial Embedding, Scaling Distortion and the ℓ_q-Distortion	15
A.1	Distortion of ℓ_q -Norm for Fixed q	17
B	Embedding into Ultrametrics	18
C	Applications	19
C.1	Sparsest cut	20
C.2	Multi cut	20
C.3	Minimum Linear Arrangement	21
C.4	Multiple sequence alignment	21
C.5	Uncapacitated quadratic assignment	22
C.6	Min-sum k -clustering	22
D	Distance Oracles	23
D.1	Distance oracles with scaling distortion	23
D.2	Partial distance oracles	24
E	Constructing Hierarchical Probabilistic Partitions	25

A 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 10 (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 $1 \leq q \leq \infty$, 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 $1 \leq q \leq \infty$ 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 $\text{dist}_f(u, v) \geq \alpha(\epsilon)$. Therefore:

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

□

The following lemma in particular implies constant average distortion for any scaling embedding with distortion $O(\log(1/\epsilon))$. However, the argument extends to the weighted average case to provide $O(\log \Phi)$ distortion, where Φ is the effective aspect ratio of the weight function.

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 $f : X \rightarrow Y$ with scaling distortion α then for any distribution Π over $\binom{X}{2}$.¹³*

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

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

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

Define for each $\epsilon \in [0, 1)$ the set $G(\epsilon)$ of the $(1 - \epsilon)\binom{n}{2}$ pairs u, v of smallest distortion $\text{dist}_f(u, v)$ over all pairs in $\binom{X}{2}$. Since f is a $(1 - \epsilon)$ -partial embedding for any $\epsilon \in [0, 1)$ we have that for each $\{u, v\} \in G(\epsilon)$, $\text{dist}_f(u, v) \leq \alpha(\epsilon)$. Let $G_i = G(2^{-i}\hat{\Phi}(\Pi)^{-1}) \setminus G(2^{-(i-1)}\hat{\Phi}(\Pi)^{-1})$. Since α is a monotonic non-increasing function, it follows that

$$\begin{aligned} \mathbb{E}_{\Pi}[\text{dist}_f(u, v)^q] &= \sum_{u \neq v \in X} \pi(u, v) \text{dist}_f(u, v)^q \\ &\leq \sum_{\{u, v\} \in G(\hat{\Phi}(\Pi)^{-1})} \pi(u, v) \alpha(\hat{\Phi}(\Pi)^{-1})^q + \\ &\quad \sum_{i=1}^{\lfloor \log(\binom{n}{2}\hat{\Phi}(\Pi)^{-1}) \rfloor} \sum_{\{u, v\} \in G_i} \pi(u, v) \alpha(2^{-i}\hat{\Phi}(\Pi)^{-1})^q \\ &\leq \sum_{u \neq v \in X} \pi(u, v) \cdot \alpha(\hat{\Phi}(\Pi)^{-1})^q + \\ &\quad \sum_{i=1}^{\lfloor \log(\binom{n}{2}\hat{\Phi}(\Pi)^{-1}) \rfloor} |G_i| \cdot \left(\frac{\hat{\Phi}(\Pi)}{\binom{n}{2}} \sum_{u \neq v \in X} \pi(u, v) \right) \cdot \alpha(2^{-i}\hat{\Phi}(\Pi)^{-1})^q \\ &\leq \alpha(\hat{\Phi}(\Pi)^{-1})^q + \sum_{i=1}^{\lfloor \log(\binom{n}{2}\hat{\Phi}(\Pi)^{-1}) \rfloor} 2^{-i} \cdot \alpha(2^{-i}\hat{\Phi}(\Pi)^{-1})^q \\ &\leq \alpha(\hat{\Phi}(\Pi)^{-1})^q + \left(2 \int_{\frac{1}{2}\binom{n}{2}^{-1}\hat{\Phi}(\Pi)}^1 \alpha(x\hat{\Phi}(\Pi)^{-1})^q dx \right). \end{aligned}$$

¹³Assuming the integral is defined. We note that lemma is stated using the integral for presentation reasons.

□

The following somewhat more sophisticated lemma in particular implies constant *distortion of average* or any scaling embedding with distortion $O(\log(1/\epsilon))$. Again, the argument extends to the weighted average case to provide $O(\log \Phi)$ distortion.

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

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

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

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

For $\epsilon \in [0, 1)$ recall that $\hat{G}(\epsilon) = \{\{x, y\} \in \binom{X}{2} \mid d(x, y) \geq \max\{r_{\epsilon/2}(x), r_{\epsilon/2}(y)\}\}$. Since (f, \hat{G}) is a $(1 - \epsilon)$ -partial embedding for any $\epsilon \in [0, 1)$ we have that for each $\{u, v\} \in \hat{G}(\epsilon)$, $\text{dist}_f(u, v) \leq \alpha(\epsilon)$. 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:

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

To prove this fix some $u \in X$. Let $S = \{v \mid \{u, v\} \notin \hat{G}(2^{-(i-1)}\hat{\Phi}(\Pi)^{-1})\}$. 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 $d(u, v) \leq d(u, v')$. It follows that:

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

¹⁴Assuming the integral is defined.

Since α is a monotonic non-increasing function, it follows that

$$\begin{aligned}
\mathbb{E}_\Pi[d_Y(f(u), f(v))^q] &= \sum_{u \neq v \in X} \pi(u, v) d_Y(f(u), f(v))^q \\
&= \sum_{u \neq v \in X} \pi(u, v) d_X(u, v)^q \text{dist}_f(u, v)^q \\
&\leq \sum_{\{u, v\} \in \hat{G}(\hat{\Phi}(\Pi)^{-1})} \pi(u, v) d_X(u, v)^q \alpha(\hat{\Phi}(\Pi)^{-1})^q + \\
&\quad \sum_{i=1}^{\lfloor \log(\binom{n}{2} \hat{\Phi}(\Pi)^{-1}) \rfloor} \sum_{\{u, v\} \in \hat{G}_i} \pi(u, v) d_X(u, v)^q \alpha(2^{-i} \hat{\Phi}(\Pi)^{-1})^q \\
&\leq \sum_{u \neq v \in X} \pi(u, v) d_X(u, v)^q \cdot \alpha(\hat{\Phi}(\Pi)^{-1})^q + \\
&\quad \sum_{i=1}^{\lfloor \log(\binom{n}{2} \hat{\Phi}(\Pi)^{-1}) \rfloor} \sum_{\{u, v\} \in \hat{G}_i} d_X(u, v)^q \cdot \hat{\Phi}(\Pi) \cdot \min_{w \neq z \in X} \pi(w, z) \cdot \alpha(2^{-i} \hat{\Phi}(\Pi)^{-1})^q \\
&\leq \sum_{u \neq v \in X} \pi(u, v) d_X(u, v)^q \cdot \alpha(\hat{\Phi}(\Pi)^{-1})^q + \\
&\quad \sum_{i=1}^{\lfloor \log(\binom{n}{2} \hat{\Phi}(\Pi)^{-1}) \rfloor} \sum_{u \neq v \in X} 2^{-i} d_X(u, v)^q \cdot \min_{w \neq z \in X} \pi(w, z) \cdot \alpha(2^{-i} \hat{\Phi}(\Pi)^{-1})^q \\
&\leq \sum_{u \neq v \in X} \pi(u, v) d_X(u, v)^q \cdot \alpha(\hat{\Phi}(\Pi)^{-1})^q + \\
&\quad \sum_{i=1}^{\lfloor \log(\binom{n}{2} \hat{\Phi}(\Pi)^{-1}) \rfloor} \sum_{u \neq v \in X} \pi(u, v) d_X(u, v)^q \cdot 2^{-i} \cdot \alpha(2^{-i} \hat{\Phi}(\Pi)^{-1})^q \\
&\leq \mathbb{E}_\Pi[d_X(u, v)^q] \cdot \left[\alpha(\hat{\Phi}(\Pi)^{-1})^q + \left(2 \int_{\frac{1}{2} \binom{n}{2}^{-1} \hat{\Phi}(\Pi)}^1 \alpha(x \hat{\Phi}(\Pi)^{-1})^q dx \right) \right].
\end{aligned}$$

□

A.1 Distortion of ℓ_q -Norm for Fixed q

Lemma 1. *Let $1 \leq q \leq \infty$. For any finite metric space (X, d) , there exists an embedding f from X into a star metric such that for any non-degenerate distribution Π : $\text{distanorm}_q^{(\Pi)}(f) \leq 2^{1/q} (2^q - 1)^{1/q} \hat{\Phi}(\Pi)^{1/q}$. In particular: $\text{distanorm}_q(f) \leq 2^{1/q} (2^q - 1)^{1/q} \leq \sqrt{6}$.*

Proof. Let $w \in X$ be the point that minimizes $(\sum_{x \in X} d(w, x)^q)^{1/q}$. Let $Y = X \cup \{r\}$. Define a star metric (Y, d')

where r is the center and for every $x \in X$: $d'(r, x) = d(w, x)$. Thus $d'(x, y) = d(w, x) + d(w, y)$. Then

$$\begin{aligned}
\mathbb{E}_\Pi[d'(u, v)^q] &= \sum_{u \neq v \in X} \pi(u, v) d'(u, v)^q \leq \sum_{u \neq v \in X} \pi(u, v) (d(u, w) + d(w, v))^q \\
&\leq (2^q - 1) \sum_{u \neq v \in X} \pi(u, v) (d(u, w)^q + d(w, v)^q) \\
&\leq (2^q - 1) \sum_{u \neq v \in X} \left(\Phi(\Pi) \min_{s \neq t \in X} \pi(s, t) \right) \cdot (d(u, w)^q + d(w, v)^q) \\
&= (2^q - 1) \cdot \Phi(\Pi) \min_{s \neq t \in X} \pi(s, t) \cdot \frac{n-1}{2} \left(\sum_{u \in X} d(u, w)^q + \sum_{v \in X} d(w, v)^q \right) \\
&\leq (2^q - 1) \cdot \Phi(\Pi) \cdot (n-1) \cdot \frac{1}{n} \sum_{z \in X} \sum_{u \in X} \min_{s \neq t \in X} \pi(s, t) \cdot d(u, z)^q \\
&\leq 2(2^q - 1) \cdot \Phi(\Pi) \cdot \sum_{u \neq v \in X} \pi(u, v) \cdot d(u, v)^q \\
&= 2(2^q - 1) \cdot \Phi(\Pi) \cdot \mathbb{E}_\Pi[d(u, v)^q].
\end{aligned}$$

□

B Embedding into Ultrametrics

In this section we show $(1 - \epsilon)$ partial embedding with distortion $\log(1/\sqrt{\epsilon})$ into ultrametrics.

Definition 15. An ultrametric U (or 1-HST) is a metric space (U, d_U) whose elements are the leaves of a rooted tree T . Each $v \in T$ is associated a label $\Delta(v) \geq 0$ such that if $u \in T$ is a descendant of v then $\Delta(u) \leq \Delta(v)$ and $\Delta(u) = 0$ iff $u \in U$ is a leaf. The distance between leaves $x, y \in U$ is defined as $d_U(x, y) = \Delta(\text{lca}(x, y))$ where $\text{lca}(x, y)$ is the least common ancestor of x and y in T .

Theorem 9. 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 $O(\frac{1}{\sqrt{\epsilon}})$.

Proof. The proof is by induction on the size of X (the base case is where $|X| = 1$ and is trivial). Assume the claim is true for any metric space with less than $|X|$ points. Denote $\Lambda = \text{diam}(X)$. Let $u, v \in X$ such that $d_X(u, v) = \Lambda$, and assume $|B(u, \frac{\Lambda}{3})| \leq \frac{n}{2}$ (otherwise switch the roles of u and v).

Let $I = \{1, 2, 3, \dots, \lceil \sqrt{5/\epsilon} \rceil\}$, for every $i \in I$ let $A_i = B(u, i\sqrt{\epsilon/5}\frac{\Lambda}{3})$, $\alpha_i = \frac{|A_i|}{n}$, $S_i = A_{i+1} \setminus A_i$.

There are two cases to consider. *Case 1:* $\alpha_1 \leq \epsilon$. In this case label the root of U with Λ , and it will have two children: one is a leaf formed by the singleton $\{u\}$ and the other is the tree formed recursively by $X \setminus \{u\}$.

Let $v \in X \setminus A_i$ then $d_X(u, v) \geq \sqrt{\epsilon/5}\frac{\Lambda}{3}$. Since $d_U(u, v) = \Lambda$ then distortion is $O(\frac{1}{\sqrt{\epsilon}})$. Hence, only pairs (u, w) where $w \in A_1$ may need to be discarded, and using the induction hypothesis the total number of distorted distances is at most:

$$\epsilon(n-1) + \epsilon \binom{n-1}{2} = \epsilon \frac{n^2 - n}{2} = \epsilon \binom{n}{2}$$

as required.

Case 2: $\alpha_1 > \epsilon$. Notice that for all $i \in I$ we have $\alpha_i \leq \frac{1}{2}$ because $|B(u, \frac{\Lambda}{3})| \leq \frac{n}{2}$.

Claim 1. There exists $i \in I$ such that $|S_i|^2 \leq \epsilon \alpha_i n^2$

Proof. Seeking a contradiction, assume that for all $i \in I$, $|S_i|^2 = (\alpha_{i+1} - \alpha_i)^2 n^2 > \epsilon \alpha_i n^2$ and hence $\alpha_{i+1} - \alpha_i > \sqrt{\epsilon \alpha_i}$. Under this assumption $\alpha_i \geq \frac{1}{9} \epsilon i^2$ for all $i \in I$, and the proof is by induction.

For the base case $\alpha_1 > \epsilon > \frac{1}{9} \epsilon$. For the induction step, assume $\alpha_i \geq \frac{1}{9} \epsilon i^2$, then

$$\begin{aligned}\alpha_{i+1} - \alpha_i &\geq \sqrt{\epsilon\alpha_i} \\ \alpha_{i+1} &\geq \frac{1}{9}\epsilon i^2 + \sqrt{\frac{1}{9}\epsilon^2 i^2} = \frac{1}{9}\epsilon(i^2 + 3i) \geq \frac{1}{9}\epsilon(i+1)^2\end{aligned}$$

Therefore $\alpha_{\sqrt{5/\epsilon}} \geq \frac{1}{9}\epsilon(\sqrt{5/\epsilon})^2 > \frac{1}{2}$ and contradiction follows. This completes the proof of Claim 1 \square

Let $i \in I$ be an index such that $|S_i| \leq \epsilon\alpha_i n^2$. We partition X to $X_1 = B\left(u, (i + \frac{1}{2})\sqrt{\epsilon/5}\frac{\Lambda}{3}\right)$ and $X_2 = X \setminus X_1$ (we divide the shell S_i in half). Label the root of U with Λ and create two children: one is the tree formed recursively by X_1 and the other is the tree formed recursively by X_2 .

Since $|X_1| \geq \alpha_i n$ and $|X_2| \geq \frac{n}{2}$ we get

$$\binom{|S_i|}{2} \leq \frac{|S_i|^2}{2} \leq \epsilon|X_1||X_2|$$

Let $C = \{(u, v) \mid u \in X_1, v \in X_2\}$ and $D = \{(u, v) \mid u \in X_1 \cap S_i, v \in X_2 \cap S_i\}$. So for any $(u, v) \in C \setminus D$, $d(u, v) \geq \min\{d_X(X_1, X_2 \setminus S_i), d_X(X_1 \setminus S_i, X_2)\} \geq \sqrt{(\epsilon/5)}\Lambda/6$. Hence only distances $(u, v) \in D$ may be distorted by more than $O(\frac{1}{\sqrt{\epsilon}})$, and there are at most $\binom{|S_i|}{2}$ such distances.

Using the induction hypothesis we conclude that the total number of distorted distances is at most

$$\begin{aligned}\epsilon \binom{|X_1|}{2} + \epsilon \binom{|X_2|}{2} + \binom{|S_i|}{2} &\leq \frac{1}{2} (|X_1|^2 - |X_1| + |X_2|^2 - |X_2| + 2|X_1||X_2|) \\ &\leq \frac{1}{2} (|X_1| + |X_2|)(|X_1| + |X_2| - 1) \\ &= \epsilon \binom{n}{2}\end{aligned}$$

\square

C 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 alternatively it may come up in the linear programming relaxation of the problem.

Then these weights are can be normalized to define the distribution Π where $\pi(u, v) = \frac{c(u, v)}{\sum_{x, y} c(x, y)}$ so that the goal translates into minimizing the *expected distance* according to the distribution Π . We can now use our results to construct embeddings with small *distortion of average* provided in Theorem 5, Theorem 7 and Theorem 8. Thus we get 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 (Theorems 5 and Theorem 7).

Define $\Phi(c) = \Phi(\Pi)$ and $\hat{\Phi}(c) = \hat{\Phi}(\Pi)$. Note that if for all $u \neq v$, $c(u, v) > 0$ then $\Phi(c) = \frac{\max_{u,v} c(u, v)}{\min_{u,v} c(u, v)}$.

Using this paradigm we obtain $O(\log \hat{\Phi}(c)) = O(\min\{\log(\Phi(c)), \log n\})$ approximation algorithms.

This lemma below summarizes the specific propositions which will be useful in most of the applications in the sequel:

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

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

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

2. There is 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 :

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

3. For any given weight function c , there exists an ultrametric (Y, d_Y) and an embedding $f : X \rightarrow Y$ such that

$$\sum_{\{u,v\} \in \binom{X}{2}} c(u,v) d_Y(f(u), f(v)) \leq O(\log \hat{\Phi}(c)) \sum_{\{u,v\} \in \binom{X}{2}} c(u,v) d_X(u,v)$$

C.1 Sparsest cut

We show an approximation for the sparsest cut problem for complete weighted graphs, i.e., for the following problem:

Given a complete graph $G(V, E)$ with capacities $c(u, v) : E \rightarrow \mathbb{R}_+$ and demands $D(u, v) : E \rightarrow \mathbb{R}_+$. Define the weight of a cut (S, \bar{S}) as

$$\frac{\sum_{u \in S, v \in \bar{S}} c(u, v)}{\sum_{u \in S, v \in \bar{S}} D(u, v)}$$

We seek a subset $S \subseteq V$ minimizing the weight of the cut.

The uniform demand case of the problem was first given an approximation algorithm of $O(\log n)$ by Leighton and Rao [31]. For the general case $O(\log k)$ approximation algorithms were given by Aumann and Rabani [5] and London, Linial and Rabinovich [33] (where k is the number of demands), via embeddings into L_1 of Bourgain. Recently Arora, Rao and Vazirani improved the uniform case bound to $O(\sqrt{\log n})$ and subsequently Arora, Lee and Naor gave an $O(\sqrt{\log n} \log \log n)$ approximation for the general demand case based on embedding of negative-type metrics into L_1 .

We show an $O(\log \hat{\Phi}(c))$ approximation. We apply the method of [33]: build the following linear program:

$$\begin{aligned} & \min_{\tau} \sum_{u,v} c(u,v) \tau(u,v) \\ & \text{subject to: } \sum_{u,v} D(u,v) \tau(u,v) \geq 1 \\ & \tau \text{ satisfies triangle inequality} \\ & \tau \geq 0 \end{aligned}$$

If the solution would yield a cut metric it would be the optimal solution. We solve the relaxed program for all metrics, obtaining a metric (V, τ) , then embed (V, τ) into l_1 , using f of Lemma 2. Since the embedding is non-contractive $\tau(u, v) \leq \|f(u) - f(v)\|_1$, hence

$$\frac{\sum_{u,v} c(u,v) \|f(u) - f(v)\|_1}{\sum_{u,v} D(u,v) \|f(u) - f(v)\|_1} \leq O(\log \hat{\Phi}(c)) \frac{\sum_{u,v} c(u,v) \tau(u,v)}{\sum_{u,v} D(u,v) \tau(u,v)}$$

Following [33], we can obtain a cut that provides a $O(\log \hat{\Phi}(c))$ approximation.

C.2 Multi cut

The multi cut problem is: given a complete graph $G(V, E)$ with weights $c(u, v) : E \rightarrow \mathbb{R}_+$, and k set of pairs $(s_i, t_i) \subseteq V \times V$ $i = 1, \dots, k$ find a minimal weight subset $E' \subseteq E$, such that removing every edge in E' disconnects every pair (s_i, t_i) .

The best approximation algorithm for this problem due to Garg, Vazirani and Yannakakis [20] has performance $O(\log k)$.

We show a $O(\log \hat{\Phi}(c))$ approximation. We slightly change the methods of [20], create a linear program:

$$\begin{aligned} & \min_{\tau} \sum_{(u,v) \in \binom{V}{2}} c(u,v)\tau(u,v) \\ \text{subject to: } & \forall i,j \sum_{(u,v) \in p_i^j} \tau(u,v) \geq 1 \\ & \tau \text{ satisfies triangle inequality} \\ & \tau \geq 0 \end{aligned}$$

where p_i^j is the j -th path from s_i to t_i . Now solve the relaxed version obtaining metric space (V, τ) . Using (3.) of Lemma 2 we get an embedding $f : V \rightarrow Y$ into an HST (Y, d_Y) . We use this metric to partition the graph instead of the region growing method introduced by [20].

Hence $\sum_{(u,v) \in \binom{V}{2}} c(u,v)d_Y(u,v) \leq O(\log \hat{\Phi}(c)) \sum_{(u,v) \in \binom{V}{2}} c(u,v)\tau(u,v)$. We build a multi cut E' : for every pair (s_i, t_i) find their $\text{lca}(s_i, t_i) = r_i$, and create two clusters containing all the vertices under each child: insert into E' all the edges between the points in each subtree and the rest of the graph. Since we have the constraint that $\sum_{(u,v) \in p_i^j} \tau(u,v) \geq 1$, we get from the fact that f is non-contractive that $\Delta(r_i) = d_Y(s_i, t_i) \geq 1$. It follows that if an edge $(u, v) \in E'$ then $d(u, v) \geq 1$. It follows that

$$\sum_{(u,v) \in E'} c(u,v) \leq \sum_{(u,v) \in \binom{V}{2}} c(u,v)d_Y(u,v) \leq O(\log \hat{\Phi}(c))OPT$$

C.3 Minimum Linear Arrangement

The same idea can be used in the minimum linear arrangement problem, where we have an undirected graph $G(V, E)$ with capacities $c(e)$ for every $e \in E$, we wish to find a one to one arrangement of vertices $h : V \rightarrow \{1, \dots, |V|\}$, minimizing the total edge length: $\sum_{(u,v) \in E} c(u,v)|h(u) - h(v)|$.

This problem was first given an $O(\log n \log \log n)$ approximation by Even, Naor, Rao and Schieber [16], which was subsequently improved by Rao and Richa [40] to $O(\log n)$.

As shown in [16], this can be done using the following LP:

$$\begin{aligned} & \min \sum_{u \neq v \in V} c(u,v)d(u,v) \\ \text{s.t. } & \forall U \subseteq V, \quad \forall v \in U : \sum_{u \in U} d(u,v) \geq \frac{1}{4}(|U|^2 - 1) \\ & \forall (u,v) : d(u,v) \geq 0 \end{aligned}$$

which is proven there to be a lower bound to the optimal solution. Even et. al [16] use this LP formulation to define a *spreading metric* which they use to recursively solve the problem in a divide-and-conquer approach. Their method can be in fact viewed as an embedding into an ultrametric (HST) (the argument is similar to the one given for the special case of the *multi cut* problem) and so by using assertion (3.) of Lemma 2 we obtain an $O(\log \hat{\Phi}(c))$ approximation.

The problem of embedding in d -dimensional meshes is basically an expansion of h to d dimensions, and can be solved in the same manner.

C.4 Multiple sequence alignment

Multiple sequence alignments are important tools in highlighting similar patterns in a set of genetic or molecular sequence.

Given n strings over a small character set, the goal is to insert gaps in each string as to minimize the total number of different characters between all pairs of strings, when the cost of gap is considered 0.

In their paper, [45] showed an approximation algorithm for the generalized version, where each pair of string has an importance parameter $c(u, v)$, they phrased the problem as finding a minimum *communication cost spanning tree*, i.e. finding a tree that minimizes $\sum_{u,v} c(u,v)d(u,v)$, where d is the edit distance. They apply probabilistic embedding into trees to bound the cost of such a tree. This gives an approximation ratio of $O(\log n)$.

Using Lemma 2 we get an $O(\log \hat{\Phi}(c))$ approximation.

C.5 Uncapacitated quadratic assignment

The uncapacitated quadratic assignment problem is one of the most studied problems in operations research (see the survey [36]) and is once of the main applications of metric labelling [25]. Given three $n \times n$ input matrices C, D, F , such that C is symmetric with 0 in the diagonal, D is a metric and all matrices are non-negative. The objective is to minimize

$$\min_{\sigma \in \mathcal{S}_n} \sum_{i,j} C(i,j) D(\sigma(i), \sigma(j)) + \sum_i F(i, \sigma(i))$$

where \mathcal{S}_n is the set of all permutations over n elements.

One of the major applications of uncapacitated quadratic assignment is in location theory: where $C(i, j)$ is the material flow from facility i to j , $D(\sigma(i), \sigma(j))$ is their distance after locating them and $F(i, \sigma(i))$ is the cost for positioning facility i at location $\sigma(i)$.

Unlike the previous applications here C is not a fixed weight function on the metric D , but the actual weights depends on σ which is determined by the algorithm. Hence we require the probabilistic result (1) of Lemma 2 which is oblivious to the weight function C .

Kleinberg and Tardos [25] gave an approximation algorithm based on probabilistic embedding into ultrametrics. They give an $O(1)$ approximation algorithm for an ultrametric (they in fact use a 3-HST). This implies an $O(\log k)$ approximation for general metrics, where k is the number of labels.

As uncapacitated quadratic assignment is a special case of metric labelling it can be solved in the same manner, yielding a $O(\log \hat{\Phi}(C))$ approximation ratio by applying result (1) of Lemma 2 together with the $O(1)$ approximation for ultrametrics of [25].

C.6 Min-sum k -clustering

Recall the min-sum k -clustering problem, where one has to partition a graph H to k clusters C_1, \dots, C_k as to minimize

$$\sum_{i=1}^k \sum_{u,v \in C_i} d_H(u, v)$$

[11] showed that using probabilistic embedding into HST and solving the problem there by dynamic programming yields a constant approximation factor in the HST and therefore a factor of $O\left(\frac{1}{\epsilon}(\log n)^{1+\epsilon}\right)$ in H , with running time $n^{O(1/\epsilon)}$. Let $\Phi = \Phi(d)$.

Lemma 3. *For a graph H equipped with the shortest path metric, if $\Phi \leq 2^{\frac{\log n}{\log \log n}}$, there is a polynomial time $O(\log(k\Phi))$ approximation for min-sum k -clustering problem.*

Denote by OPT the optimum solution for the problem with clusters C_i^{OPT} , and OPT_T the optimum solution for a family of HST \mathcal{T} with clusters $C_i^{OPT_T}$. Also denote ALG for the result of [11] algorithm with clusters $C_i^{ALG_T}$.

We know that by using probabilistic $(1 - \epsilon)$ partial embedding H into \mathcal{T} , edges $e \in G$ are expanded by $O(\log \frac{1}{\epsilon})$ and for $e \notin G$ the maximum expansion is Φ (no distance is contracted), therefore choosing $\epsilon = \frac{1}{k\Phi}$ yields:

$$\begin{aligned} \mathbb{E}[ALG] &= \sum_{T \in \mathcal{T}} \Pr[T] \sum_{i=1}^k \sum_{u,v \in C_i^{ALG_T}} d_H(u, v) \leq \sum_{T \in \mathcal{T}} \Pr[T] \sum_{i=1}^k \sum_{u,v \in C_i^{ALG_T}} d_T(u, v) \\ &\leq O(1) \sum_{T \in \mathcal{T}} \Pr[T] \sum_{i=1}^k \sum_{u,v \in C_i^{OPT_T}} d_T(u, v) \leq O(1) \sum_{T \in \mathcal{T}} \Pr[T] \sum_{i=1}^k \sum_{u,v \in C_i^{OPT_T}} d_T(u, v) \\ &\leq O(1) \left(\sum_{i=1}^k \sum_{u,v \in C_i^{OPT_T} \cap G} \sum_{T \in \mathcal{T}} \Pr[T] d_T(u, v) + \sum_{i=1}^k \sum_{u,v \in C_i^{OPT_T} \setminus G} \sum_{T \in \mathcal{T}} \Pr[T] d_T(u, v) \right) \\ &\leq O(1) \left(\sum_{i=1}^k \sum_{u,v \in C_i^{OPT_T} \cap G} O(\log(1/\epsilon)) d_H(u, v) + \sum_{i=1}^k \sum_{u,v \in C_i^{OPT_T} \setminus G} \Phi \right) \\ &\leq O((\log(1/\epsilon)) OPT + k\epsilon n^2 \Phi) = O(\log(k\Phi)) OPT + n^2/k = O(\log(k\Phi)) OPT \end{aligned}$$

the last equation follows from the fact that $\frac{n^2}{k} \leq OPT$ (assuming we scaled the distances such that $\min_{u \neq v \in H} d_H(u, v) \geq 1$).

Let the clusters be of size a_1, \dots, a_k , naturally $\sum_{i=1}^k a_i = n$, and there are $\sum_{i=1}^k a_i^2$ edges inside clusters. Let $b = (1, 1, \dots, 1) \in \mathbb{R}^k$. From Cauchy-Schwartz we get

$$\left(\sum_{i=1}^k a_i \right)^2 = (a \times b)^2 \leq \|a\|^2 \|b\|^2 = \sum_i (a_i^2) k$$

therefore $\sum_i (a_i^2) \geq \frac{n^2}{k}$, meaning $OPT \geq \frac{n^2}{k}$.

The running time of the algorithm is $O(\log n)^{O(\log \Phi)} = n^{O(1)}$, since the number of levels in the HST is relatively small $O(\log \Phi)$. (see [11] for details).

D Distance Oracles

A distance oracle for a metric space (X, d) , $|X| = n$ is a data structure that given any pair returns an estimate of their distance. In this section we study scaling distance oracles and partial distance oracles.

D.1 Distance oracles with scaling distortion

Given a distance oracle with $O(n^{1/k})$ bits, the worst case stretch can indeed be $2k - 1$ for some pairs in some graphs. However we prove the existence of distance oracles with a *scaling stretch* property. For these distance oracles, the average stretch over all pairs is only $O(1)$.

We repeat the same preprocessing and distance query algorithm of Thorup and Zwick [44] with sampling probability $3n^{-1/k} \ln n$ for the first set and $n^{-1/k}$ thereafter.

```

Given  $(X, d)$  and parameter  $k$ :
 $A_0 := X$ ;  $A_k = \emptyset$ ;
for  $i = 1$  to  $k - 1$ 
  let  $A_i$  contain each element of  $A_{i-1}$ ,
  independently with probability  $\begin{cases} 3n^{-1/k} \ln n & i = 1 \\ n^{-1/k} & i > 1 \end{cases}$ ;
for every  $x \in X$ 
  for  $i = 0$  to  $k - 1$ 
    let  $p_i(x)$  be the nearest node in  $A_i$ ,
    so  $d(x, A_i) = d(x, p_i(x))$ ;
    let  $B_i(x) := \{y \in A_i \setminus A_{i+1} \mid d(x, y) < d(x, A_{i+1})\}$ ;

```

Figure 1: Preprocessing algorithm.

```

Given  $x, y \in X$ :
 $z := x$ ;  $i := 0$ ;
while  $z \notin B_i(y)$ 
   $i := i + 1$ ;
   $(x, y) := (y, x)$ ;
   $z := p_i(x)$ ;
return  $d(x, z) + d(z, y)$ ;

```

Figure 2: Distance query algorithm.

Theorem 11. *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 coarsely scaling distortion bounded by $\left(2 \left\lceil \frac{\log(2/\epsilon)k}{\log n} \right\rceil + 1\right)$.*

Proof. Fix $\epsilon \in (0, 1)$, and $x, y \in \hat{G}(\epsilon)$. Let j be the integer such that $n^{j/k} \leq \epsilon n/2 < n^{(j+1)/k}$. We prove by induction that at the end of the ℓ th iteration of the while loop of the distance query algorithm:

1. $d(x, z) \leq d(x, y) \max\{1, \ell - j\}$
2. $d(z, y) \leq d(x, y) \max\{2, \ell - j + 1\}$.

Observe that

$$\Pr[B(x, r_{n^{(i-k)/k}}(x)) \cap A_i = \emptyset] \leq (1 - n^{-i/k} 3 \ln n)^{n^{i/k}} \leq n^{-3}$$

for all $x \in X$ and $i \in \{0, 1, 2, \dots, k-1\}$. Hence with high probability (1.) holds for any $\ell < j$ since $d(x, p_{\ell+1}(x)) \leq r_{\epsilon/2}(x) \leq d(x, y)$ and (2.) follows from (1.) and the triangle inequality. For $\ell \geq j$, from the induction hypothesis, at the beginning of the ℓ th iteration, $d(z', y) \leq d(x, y) \max\{1, \ell - j\}$, where $z' = p_\ell(x)$, $z' \in A_\ell$. Since $z' \notin B_\ell(y)$ then after the swap (the line $(x, y) := (y, x)$) we have

$$d(x, z) = d(x, p_{\ell+1}(x)) \leq d(x, y) \max\{1, \ell - j\}$$

and $d(z, y) \leq d(x, y) \max\{2, \ell - j + 1\}$ follows from the triangle inequality. This completes the inductive argument. Since $p_{k-1}(x) \in A_{k-1} = B_{k-1}(y)$ then $\ell \leq k-1$ and therefore the stretch of the response is bounded by $2(k-j) - 1 \leq 2 \left\lceil \frac{\log(2/\epsilon)k}{\log n} \right\rceil + 1$. \square

We note that a similar argument showing scaling stretch can be given for variation of Thorup and Zwick's compact routing scheme [43].

D.2 Partial distance oracles

We construct a distance oracle with linear memory that guarantees stretch to $1 - \epsilon$ fraction of the pairs.

Theorem 11. *Let (X, d) be a finite metric space. Let $0 < \epsilon < 1$ be a parameter. Let $k \leq O(\log \frac{1}{\epsilon})$. The metric space can be preprocessed in polynomial time, producing data structure such that distance queries can be answered in $O(k)$ time. With either one of the following properties:*

1. *Either with $O\left(n \log \frac{1}{\epsilon} + k \left(\frac{\log \frac{1}{\epsilon}}{\epsilon}\right)^{1+1/k}\right)$ size, and stretch $6k - 1$ for some set $G \subseteq \binom{X}{2}$, $|G| \geq (1 - \epsilon) \binom{n}{2}$.*
2. *Or, with $O\left(n \log n \log \frac{1}{\epsilon} + k \log n \left(\frac{\log \frac{1}{\epsilon}}{\epsilon}\right)^{1+1/k}\right)$ size, and stretch $6k - 1$ for the set $\hat{G}(\epsilon)$.*

Proof. We begin with a proof of (1.). Let $b = \lceil (8/\epsilon) \ln(16/\epsilon) \rceil$. Let B be a set of b beacons chosen uniformly at random. Construct a distance oracle of [44] on the subspace (B, d) with parameter $k \leq \log b$ yielding stretch $2k - 1$ and using $O(kb^{1+1/k})$ storage. For every $x \in X$ let $p(x)$ be the closest node in B . The resulting data structure's size is $O(n \log b) + O(kb^{1+1/k}) = O(n \log b + kb^{1+1/k})$. Queries are processed as follows: given two nodes $x, y \in X$ let r be the response of the distance oracle on the beacons $p(x), p(y)$ then return $d(x, p(x)) + r + d(p(y), y)$.

Observe that from triangle inequality the response is at least $d(x, y)$. Let \mathcal{E}_x for any $x \in X$ be the event

$$\mathcal{E}_x = \{d(x, B) > r_{\epsilon/8}(x)\}.$$

Then $\Pr[\mathcal{E}_x] \leq (1 - 8/\epsilon)^b \leq \epsilon/16$ and so by Markov inequality, $\Pr[|\{\mathcal{E}_x \mid x \in X\}| \leq \epsilon n/8] \geq 1/2$. In such a case let

$$\mathcal{G} = \{(x, y) \in \binom{X}{2} \mid \neg \mathcal{E}_x \wedge \neg \mathcal{E}_y \wedge d(x, y) \geq \max\{r_{\epsilon/8}(x), r_{\epsilon/8}(y)\}\}.$$

We bound the size of \mathcal{G} . At most $\epsilon n/8$ nodes are removed due to \mathcal{E}_x occurring, and from each remaining node at most $\epsilon n/8$ distances are discarded so $|\mathcal{G}| \geq \binom{X}{2} - \epsilon n^2/4 \geq (1 - \epsilon) \binom{X}{2}$. For $(x, y) \in \mathcal{G}$, we have $d(p(x), p(y)) \leq 3d(x, y)$ so from the distance oracle $r \leq (6k - 3)d(x, y)$ and in addition $\max\{d(x, p(x)), d(y, p(y))\} \leq d(x, y)$ so the stretch is bounded by $6k - 1$.

The proof of (2.) is a slight modification of the above procedure. Let $m = \lceil 3 \ln n \rceil$. Let B_1, \dots, B_m be sets each containing b beacons chosen uniformly at random. Let DO_i be the distance oracle on (B_i, d) . For every $x \in X$ we store $p_1(x), \dots, p_m(x)$ where $p_i(x)$ is the closest node in B_i . The resulting data structure's size is $O(n \log b \ln n) + O(kb^{1+1/k} \ln n) = O(n \log b \ln n + kb^{1+1/k} \ln n)$. Queries are processed as follows: given two nodes $x, y \in X$ let r_i be the response of the distance oracle DO_i on the beacons $p_i(x), p_i(y)$ then return $\min_{1 \leq i \leq m} d(x, p_i(x)) + r_i + d(p_i(y), y)$.

For every $(x, y) \in \binom{X}{2}$, $1 \leq i \leq m$ define the event $\mathcal{E}_{x,y}^i = \{d(x, B_i) > r_{\epsilon/8}(x) \vee d(y, B_i) > r_{\epsilon/8}(y)\}$. Since beacons are chosen independently, then for every $(x, y) \in \hat{G}(\epsilon/4)$, with high probability, there exists $1 \leq i \leq m$ such that $\neg \mathcal{E}_{x,y}^i$. Hence, in a similar argument as in (1.) above, the stretch of $d(x, p_i(x)) + r_i + d(p_i(y), y)$ is at most $6k - 1$. □

E Constructing Hierarchical Probabilistic Partitions

In this section we provide details on the proof of Lemma 2.

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

Definition 16 (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 17 (Uniformly Padded Probabilistic Partition). Given $\Delta > 0$, let $\hat{\mathcal{P}}$ be a Δ -bounded probabilistic partition of (X, d) . Given collection of functions $\eta = \{\eta_P : X \rightarrow [0, 1] | P \in \mathcal{P}\}$ 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 $\text{diam}(Z) \leq \bar{\Delta}$. 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 $\{\xi_P : X \rightarrow \{0, 1\} | P \in \mathcal{P}\}$ and $\{\hat{\eta}_P : X \rightarrow \{0, 1/\ln(1/\hat{\delta})\} | P \in \mathcal{P}\}$, such that for any $\hat{\delta} \leq \delta \leq 1$, and $\eta^{(\delta)}$ defined by $\eta_P^{(\delta)}(x) = \hat{\eta}_P(x) \ln(1/\delta)$, 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: $2^{-7}/\ln \rho(x, \bar{\Delta}, \Gamma) \leq \hat{\eta}_P(x) \leq 2^{-7}/\ln(1/\hat{\delta})$.
- If $\xi_P(x) = 0$ then: $\hat{\eta}_P^{(\delta)}(x) = 2^{-7}/\ln(1/\hat{\delta})$ and $\bar{\rho}(x, \bar{\Delta}, \Gamma) < 1/\hat{\delta}$.

The proof of Lemma 4 is a non-trivial generalization of arguments of [6, 8] and is given in [9].

Using this lemma we can prove the following lemma on uniformly padded hierarchical probabilistic partitions¹⁵ from which Lemma 2 is derived.

Lemma 5 (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 $\xi = \{\xi_{P,i} : X \rightarrow \{0, 1\} | P \in \mathcal{H}, i \in I\}$ and $\hat{\eta} = \{\hat{\eta}_{P,i} : X \rightarrow \{0, 1/\ln(1/\hat{\delta})\} | P \in \mathcal{H}, i \in I\}$, 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:

•

$$\sum_{j \leq i} \xi_{P,j}(x) \eta_{P,j}^{(\delta)}(x)^{-1} \leq 2^{11} \ln \left(\frac{n}{|B(x, \Delta_{i+4})|} \right) / \ln(1/\delta).$$

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

- If $\xi_{P,i}(x) = 1$ then: $\hat{\eta}_{P,i}(x) \leq 2^{-8}/\ln(1/\hat{\delta})$.
- If $\xi_{P,i}(x) = 0$ then: $\hat{\eta}_{P,i}(x) \geq 2^{-8}/\ln(1/\hat{\delta})$ and $\bar{\rho}(x, \Delta_{i-1}, \Gamma) < 1/\hat{\delta}$.

¹⁵a variant of this lemma appears also in [9]

Proof. We create a random hierarchical partition P . By definition P_0 consists of a single cluster equal to X . Set for all $x \in X$, $\Delta_0 = \text{diam}(X)$, $\hat{\eta}_{P,0}(x) = 1/\ln(1/\hat{\delta})$, $\xi_{P,0}(x) = 0$. 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 $Q[S]$ 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 $Q[S]$ is $(\eta'^{(\delta')}, \delta')$ -uniformly padded, where $\eta'^{(\delta')}(x) = \hat{\eta}'_{Q[S]}(x) \ln(1/\delta')$. 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 $\eta_{P,i}^{(\delta)}(x) = \hat{\eta}_{P,i}(x) \ln(1/\delta)$. 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: $\eta_{P,i}^{(\delta)}(x) = \min\{\eta'^{(\delta')}(x), 2\eta_{P,i-1}^{(\delta)}(x)\}$.

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 $\eta_{P,i}^{(\delta)}(x) = \eta_{P,i}^{(\delta)}(y)$) 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 Property 3 of the lemma holds 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 $B_i = B(x, \eta_{P,i}^{(\delta)}(x)\Delta_i)$. We have:

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

As $\eta_{P,i}^{(\delta)}(x) \leq \eta'^{(\delta')}(x)$ we have $B_i \subseteq B(x, \eta'^{(\delta')}(x)\Delta_i)$. It follows that if $B_i \subseteq P_{i-1}(x)$ then $B_i \subseteq B_{P_{i-1}}(x, \eta'^{(\delta')}(x)\Delta_i)$. 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 $\eta_{P,i}^{(\delta)}(x) \leq 2\eta_{P,i-1}^{(\delta)}(x)$. A simple induction on i shows that $\eta_{P,i-1}^{(\delta)}(x) \leq 2\eta_{P,i-1}^{(\delta^{1/2})}(x)$ for any $\delta \geq \hat{\delta}$. Since $\Delta_i = \Delta_{i-1}/4$ we get that $\eta_{P,i}^{(\delta)}(x)\Delta_i \leq \eta_{P,i-1}^{(\delta')}(x)\Delta_{i-1}$ for $\delta' = \delta^{1/2}$. We therefore obtain that $B_i \subseteq B(x, \eta_{P,i-1}^{(\delta')}(x)\Delta_{i-1})$. Using the induction hypothesis we get $\Pr[B_i \subseteq P_{i-1}(x)] \geq \delta'$. We conclude from (1) 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 $\hat{\eta}_{P,i}(x) \leq \frac{1}{2}\hat{\eta}'_{Q[P_{i-1}(x)]}(x) \leq 2^{-8}/\ln(1/\hat{\delta})$, 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 $\hat{\eta}_{P,i}(x) = 2^{-8}/\ln(1/\hat{\delta})$ and that $\bar{\rho}(x, \Delta_{i-1}, \Gamma) < 1/\hat{\delta}$.

It remains to prove the first property of the lemma. Define $\psi_{P,i}(x) = 2^{-8} \cdot \xi_{P,i}(x)\hat{\eta}_{P,i}(x)^{-1}$. We claim that the following recursion holds: $\psi_{P,i}(x) \leq \max\{\ln \rho(x, \Delta_{i-1}, \Gamma), \psi_{P,i-1}(x)/2\}$. By definition if $\xi_{P,i}(x) = 1$ then one of the two following cases is possible. In the first case $\hat{\eta}_{P,i}(x) = 2\hat{\eta}_{P,i-1}(x) \leq \frac{1}{2}\hat{\eta}'_{Q[P_{i-1}(x)]}(x)$. It follows from Lemma 4 that $\hat{\eta}_{P,i-1}(x) \leq \frac{1}{2}\hat{\eta}'_{Q[P_{i-1}(x)]}(x) \leq \frac{1}{2}2^{-8}/\ln(1/\hat{\delta}) = 2^{-9}/\ln(1/\hat{\delta})$. By the second property just proved above we deduce that $\xi_{P,i-1}(x) = 1$. Therefore $\psi_{P,i}(x) = 2^{-8} \cdot \xi_{P,i}(x)\hat{\eta}_{P,i}(x)^{-1} = 2^{-8} \cdot \xi_{P,i-1}(x)\hat{\eta}_{P,i-1}(x)^{-1}/2 = \psi_{P,i-1}(x)/2$. In the second case $\hat{\eta}_{P,i}(x) = \frac{1}{2}\hat{\eta}'_{Q[P_{i-1}(x)]}(x)$ and $\xi'_{Q[P_{i-1}(x)]}(x) = 1$. Now, using Lemma 4 we get that $\hat{\eta}_{P,i}(x) \geq 2^{-8}/\ln \rho(x, \Delta_{i-1}, \Gamma)$, and hence $\psi_{P,i}(x) \leq \ln \rho(x, \Delta_{i-1}, \Gamma)$.

We conclude that the following recursion holds: $\psi_{P,i}(x) \leq \ln \rho(x, \Delta_{i-1}, \Gamma) + \psi_{P,i-1}(x)/2$. 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

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

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

□