Algorithms, Geometry and Learning

Reading group

Paris Syminelakis

September 20, 2016

Contents

1 Advances in Metric Embedding Theory **5**

1 Definitions, Results and Applications 5

1.1 Applications 7

2 A general strategy of obtaining embeddings 9

2.1 Frechet Embeddings 9

2.2 Witnessing distances 9

2.3 General recipe 10

2.3.1 Entropic Approach - Bourgain's Theorem 10

2.3.2 Spatial Approach - Padded Decomposition 11

2.3.3 A hybrid approach: Measured descent 12

2.4 Beyond Frechet Embeddings 12

3 Uniformly Padded Probabilistic Partitions 15

3.1 The basic partitioning 16

3.2 The Probabilistic Decomposition 16

4 Main Embedding 17

5 Extensions 18

Chapter 1

Advances in Metric Embedding Theory

Definitions, Results and Applications

A finite metric space is given by a collection of points $\mathcal{X}$ and a collection of numbers (distances) $d : \mathcal{X} \times \mathcal{X}$ that satisfy the triangle inequality. The business of metric embeddings mainly concern of using the fact that the distances satisfy the triangle inequality to obtain a compressed "approximate" representation of the metric space. In generality, we need dimension $n-1$ to embed in L_∞ isometrically (in total $(n-1)\cdot n$ numbers), so the approximation will have to distort some distances if we are to obtain a more compact representation.

Definition 1.1.1 (Distortion)**:** An embedding $f : \mathcal{X} \to \mathcal{Y}$ has distortion $Q > 1$ if there exists a constant $c > 0$ such that $\forall u, v \in \mathcal{X}$

$$c \cdot d_{\mathcal{X}}(u, v) \leq d_{\mathcal{Y}}(f(u), f(v)) \leq cQ \cdot d_{\mathcal{X}}(u, v) \tag{1.1}$$

Compactly, we have $\mathrm{dist}(f) = \sup_{u,v\in\mathcal{X}} \mathrm{dist}_f(u, v)$, where $\mathrm{dist}_f(u, v) := \frac{d_{\mathcal{Y}}(f(u),f(v))}{d_{\mathcal{X}}(u,v)}$.

In particular, the spaces $\mathcal{Y}$ that we are mostly interested in are the ℓ_p^D *spaces* of some (small) dimension D, where we have an explicit representation of the points in terms of coordinates, and *Ultrametrics*, that have important algorithmic applications. Research in this area asks about what is the feasible (or best) combination of (Q, D, p) for different classes of metric spaces.

The above notion of distortion is worst case, however, in most cases we are interested that the distortion be *small on average* or that we get have better bounds for *most pairs of points*. We start by formalizing the average case.

Definition 1.1.2 (ℓ_q-distortion)**:** Given a distribution Π over $\binom{\mathcal{X}}{2}$ define for $1 \leq q \leq \infty$ the ℓ_q-distortion of f with respect to Π as:

$$\mathrm{dist}_q^{(\Pi)}(f) = \|\mathrm{dist}_f(u, v)\|_q^{(\Pi)} = \mathbb{E}_\Pi \left[\mathrm{dist}_f(u, v)^q\right]^{1/q} \tag{1.2}$$

Some special cases are when :

- *Average distortion:* for $q = 1$ and $\mathcal{U}$ the uniform distribution over all the pairs we have

$$\text{avgdist}(f) := \text{dist}_1^{\mathcal{U}} = \frac{1}{\binom{n}{2}} \sum_{u \neq v \in \mathcal{X}} \text{dist}_f(u, v) \tag{1.3}$$

- *Distortion:* for $q = \infty$ and $\mathcal{U}$ the uniform distribution over all the pairs we define

$$\text{dist}(f) := \text{dist}_\infty^{\mathcal{U}} = \sup_{u \neq v \in \mathcal{X}} \text{dist}_f(u, v) \tag{1.4}$$

In this talk, we are going to present techniques that allow us to get refined quantitative bounds on the distortion. To achieve this we first introduce a relaxed notion of embedding that will be useful in formalizing the improved guarantees that we get.

Definition 1.1.3 (Partial Embedding)**:** Given two metric spaces $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$, a *partial embedding* (PE) is a pair (f, G), where f is an embedding of $\mathcal{X}$ into $\mathcal{Y}$ and $G \subseteq \binom{\mathcal{X}}{2}$. The distortion is $\text{dist}(f, G) := \sup_{\{u,v\} \in G} \text{dist}_f(u, v)$.

A partial embedding thus gives guarantees only for a subset of the distances specified by G. Some important special cases are:

- $(1 - \epsilon)$- *partial embeddings*: are PE (f, G) for which $|G| \geq (1 - \epsilon)\binom{n}{2}$.
- *Coarsely* $(1 - \epsilon)$- *partial embeddings*: are PE where

$$\hat{G}(\epsilon) := \left\{ \{u, v\} \in \binom{\mathcal{X}}{2} \middle| \min\{|B(x, d_{xy})|, |B(y, d_{xy})|\} \geq \epsilon n/2 \right\} \tag{1.5}$$

The last definition aims to capture the fact that we expect to be able to estimate the distances better[1] for pair of points that involve *coarse features* of the space (the corresponding balls contain a significant fraction of points of the space). For us, the interesting case will be when a *single embedding* is simultaenously a coarsely-PE for all scales $\epsilon \in (0, 1)$ with controlled distortion.

Definition 1.1.4 (Scaling Distortion)**:** Given two metric spaces $(\mathcal{X}, d_\mathcal{X})$, $(\mathcal{Y}, d_\mathcal{Y})$, and a function $\alpha : (0, 1) \to \mathbb{R}^+$, we say that an embedding has scaling distortion α, if for any $\epsilon \in (0, 1)$, there is some set $G(\epsilon)$ such that $(f, G(\epsilon))$ is a $(1 - \epsilon)$-PE with distortion at most $\alpha(\epsilon)$. We say that f has *coarsely scaling distortion* if for every $\epsilon \in (0, 1)$, $G(\epsilon)$ can be taken to be equal to $\hat{G}(\epsilon)$.

The main technical component of the talk will be to show how we can obtain embeddings with *coarse* and *scaling distortion*. The main results are:

[1] We will see a bit later why this is true.

Theorem 1.1.5 (Coarsely Scaling Distortion)**.** *Let $1 \le p \le \infty$. For any n-point metric space (X, d) there exists an embedding $f : \mathcal{X} \to \ell_p$ with coarsely scaling distortion $O(\lceil (\log \frac{2}{\epsilon})/p \rceil)$ and dimension $e^{O(p)} \log n$.*

- *Matousek's theorem:* worst case distrotion $O(\lceil \log n/p \rceil)$ with dimension $e^{O(p)} \log^2 n$.

Theorem 1.1.6. *For any $1 \le p \le \infty$, any finite metric space $(\mathcal{X}, d)$ on n points and $\theta \ge 12/\log\log n$ there is an embedding $f : \mathcal{X} \to \ell_p^D$ with coarse scaling distortion $O(\log(2/\epsilon) \cdot \log^\theta n)$ where the dimension $D = O(\frac{\log n}{\theta \log\log n})$.*

- *Bourgain's theorem:* $\theta = \Theta(1/\log\log n)$ then $D = O(\log n)$ and worst case distortion $O(\log(n))$ improving dimension in Bourgain's theorem from $O(\log^2 n)$. In general, $O(n^{1/D} \log n)$ distortion is achievable for dimension $O(D)$.

Applications

Before, going into the proofs of the main results let us mention some of the applications of metric embeddings with scaling distortion.

Definition 1.1.7: The *aspect ratio* of a non-degenerate distribution Π over $\binom{\mathcal{X}}{2}$ with probability function $\pi \times \mathcal{X} \times \mathcal{X} \to [0, 1]$ is given by

$$\Phi(\Pi) = \frac{\max_{u \ne v \in \mathcal{X}} \pi(u, v)}{\min_{u \ne v \in \mathcal{X}} \pi(u, v)} \tag{1.6}$$

For an arbitrary distribution Π, we define its effective aspect ratio as $\hat{\Phi}(\Pi) := 2 \min\{\Phi(\Pi), \binom{n}{2}\}$.

Lemma 1.1.8 (ℓ_q-distortion from Scaling distortion)**.** *Let $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ be metric spaces. If there exists an embedding $f : \mathcal{X} \to \mathcal{Y}$ with scaling distortion α then for any distribution Π over $\binom{\mathcal{X}}{2}$:*

$$dist_q^{(\Pi)}(f) \le \left(2 \int_{\frac{1}{2}\binom{n}{2}^{-1} \hat{\Phi}(\Pi)}^{1} \alpha(x \hat{\Phi}(\Pi)^{-1})^q dx \right)^{1/q} + \alpha(\hat{\Phi}(\Pi)^{-1}) \tag{1.7}$$

Proof Sketch. We may assume w.l.o.g that $\Phi(\Pi) < \binom{n}{2}$. Let $G(\epsilon)$ be the $(1 - \epsilon)$ fraction of pairs with the smallest distortion. By definition of scaling distortion we have that $(f, G(\epsilon))$ is a $(1 - \epsilon)$-PE with distortion $\alpha(\epsilon)$, and therefore $\text{dist}_f(u, v) \le \alpha(\epsilon)$ for every $(u, v) \in G(\epsilon)$.

1. Define $\epsilon_i := 2^{-i} \hat{\Phi}(\Pi)^{-1}$ for $i = 1, \ldots, \lfloor \log \left(\binom{n}{2} \hat{\Phi}(\Pi)^{-1} \right) \rfloor$. Note that $\epsilon_i \ge \epsilon_{i+1}$ and hence that $G(\epsilon_i) \supseteq G(\epsilon_{i+1})$ by definition of the set $G(\epsilon)$.

2. Break up the pairs of edges in groups depending on the distortion f incurs. In particular, let $G_i := G(\epsilon_i) \setminus G(\epsilon_{i-1})$

3. Finally, using the facts that
   - $|G_i| = |G(\epsilon_i)| - |G(\epsilon_{i-1})| = [(1-\epsilon_i) - (1-\epsilon_{i-1})]\binom{n}{2} = \epsilon_i\binom{n}{2}$.
   - $\max_{u\neq v} \pi(u,v) = \hat{\Phi}(\Pi) \cdot \min_{u\neq} \pi(u,v)$
   - $\min_{u\neq v} \pi(u,v) \leq \frac{1}{\binom{n}{2}} \sum_{u\neq v} \pi(u,v)$
   - α is monotonically increasing

   we get the desired inequality.

□

In essence what this inequality says is that the ℓ_q distortion is $O\left(a(\hat{\Phi}(\Pi)^{-1})\right) = O\left(\log(\hat{\Phi}(\Pi))\right)$. When the aspect ratio is small this allows us to get constant average distortion and $O(\log n)$ worst case distortion. The following theorem summarizes the *algorithmic applications of scaling distortion.*

Theorem 1.1.9. *Let $\mathcal{X}$ be a metric space and $c : \binom{\mathcal{X}}{2} \to \mathbb{R}^+$ be a weight function, then:*

1. *There exists an embedding $f : X \to \ell_p$ such that* for any *weight function c:*
$$\sum_{u\neq v} c(u,v)\|f(u) - f(v)\|_p \leq O\left(\log(\hat{\Phi}(\Pi))\right) \cdot \sum_{u\neq v} c(u,v) d_{\mathcal{X}}(u,v) \tag{1.8}$$

2. *There is a distribution over ultrametrics and corresponding probabilistic embeddings $\mathcal{F}$ such that* for any *weight function c:*
$$\mathbb{E}_{f\sim\mathcal{F}}\left[\sum_{u\neq v} c(u,v) d(f(u),f(v))\right] \leq O\left(\log(\hat{\Phi}(\Pi))\right) \cdot \sum_{u\neq v} c(u,v) d_{\mathcal{X}}(u,v) \tag{1.9}$$

3. *For any fixed weight function c, there exists an ultrametric $(\mathcal{Y}, d_{\mathcal{Y}})$ and an embedding $f : \mathcal{X} \to \mathcal{Y}$ such that:*
$$\sum_{u\neq v} c(u,v) d_{\mathcal{Y}}(f(u),f(v)) \leq O\left(\log(\hat{\Phi}(\Pi))\right) \cdot \sum_{u\neq v} c(u,v) d_{\mathcal{X}}(u,v) \tag{1.10}$$

This gives us approximation algorithms for *Sparsest Cut, Multicut, Minimum Linear arrangement, Multiple Sequence Alignment, Metric Labeling and Min-sum k-clustering.* The way to obtain such approximations is:

1. Find a *metric relaxation* of the optimization problem that provides us with a metric space $(\mathcal{X}, d_{\mathcal{X}})$ for which $\sum_{u\neq v} c(u,v) d_{\mathcal{X}}(u,v)$ is a *lower bound* to the optimal value of the objective.

2. Find an *embedding f* either in ℓ_1 (or Ultrametric) and use a constant factor approximation algorithm for ℓ_1 (or ultrametrics).

3. Use the above theorem to bound the final approximation ratio!

The rest of this talk will focus on proving the main results.

A general strategy of obtaining embeddings

For the rest of the talk we assume that $\mathcal{X} = [n]$ and that $d := d_{\mathcal{X}}$ is an arbitrary metric. For simplicity assume we would like to obtain an embedding of the1 metric space $(\mathcal{X}, d)$ in ℓ_1 of dimension D. That means, we need to produce D coordinates for every point $x \in \mathcal{X}$ such that:

- Expansion: $\|f(x) - f(y)\| = \sum_i |f_i(x) - f_i(y)| \leq cQ \cdot d(x, y)$.
- Contraction: $\|f(x) - f(y)\| \geq c \cdot d(x, y)$.

Attempt #1 Observe, that we can trivially achieve this with $D = Q = n$ and $c = 1$ if $f_i(x) := d(x, i)$. This makes sure that there is always a coordinate $j \in [D]$ such that $|f_j(x) - f_j(y)| \geq d(x, y)$ and further that for any other coordinate $|f_{j'}(x) - f_{j'}(y)| \leq d(x, y)$ by the triangle inequality. However, this method leads to increased dimension and therefore increased distortion[2] as well!.

Question: Can we combine some coordinates together such that we still get the lower bound but with decreased dimension?

Frechet Embeddings

One idea to combine coordinates while still having control is instead of a coordinate to indicate distance to a single point, to indicate distance to a set! Given a sequence of sets $W_1, \ldots, W_D$ we define the embedding:

$$f(x) := (f_1(x), \ldots, f_D(x)) = (d(x, W_1), \ldots, d(x, W_D)) \tag{1.11}$$

For each coordinate, we still have by triangle inequality (Exercise!) that $|f_i(x) - f_j(x)| = |d(x, W_i) - d(y, W_i)| \leq d(x, y)$. Thus, we see that we get immediately the following upper bound on the expansion: $\|f(x) - f(y)\| \leq D \cdot d(x, y)$.

Question: can we construct sets $W_1, \ldots, W_D$ such that for every pair x, y there is a set W_j such that $|d(x, W_j) - d(y, W_j)| \geq c \cdot d(x, y)$?

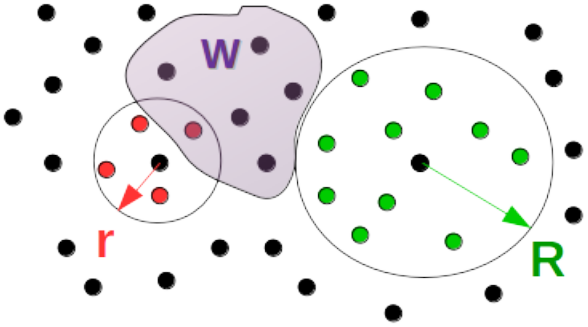
Witnessing distances

In order for $|d(x, W) - d(y, W)| \geq c \cdot d(x, y)$ to happen for a set W, we roughly need that:

- $d(x, W) \leq r_{xy} := a \cdot d(x, y) \Rightarrow B(x, r) \cap W = \emptyset$.
- $d(y, W) \geq R_{xy} := A \cdot d(x, y) \Rightarrow B(y, R) \cap W \neq \emptyset$.

for constants $A \geq a$ such that $(A - a) \geq c$. Thus, we need to produce sets $W_1, \ldots, W_D$ such that the above event happens for at least one set for all pairs!

[2] Observe that in ℓ_∞ this would give us an isometry!



General recipe

A common theme in all the different ways to construct embeddings is the following:

1. First obtain a partition $\mathcal{P} = \{P_k\}_{k \in I}$ of the $\binom{n}{2}$ distances (numbers). We will consider a different coordinate in the embedding for each part.

2. For each part $P_k \in \mathcal{P}$ construct a distribution $\mathcal{S}_k$ over sets $W_k \sim \mathcal{S}_k$ such that for any pair $(u, v) \in P_k$ with constant probability the following event happens

$$|d(x, W_k) - d(y, W_k)| \geq c \cdot d(x, y)$$

Depending on how we partition distances, and how we construct the functions $\{W_k\}_{k \in I}$ we get different embeddings.

Entropic Approach - Bourgain's Theorem

Bourgain's approach is to roughly:

1. partition pairs of points depending on the minimum size of the sets $|B(x, r_{xy})|, |B(y, r_{xy})|$, in particular,

$$P_k \approx \left\{ \{x, y\} \in \binom{\mathcal{X}}{2} \middle| \min\{|B(x, r_{xy})|, |B(y, r_{xy})|\} \approx 2^k \right\}$$

2. the (distribution over) sets W_k is constructed by sampling each point uniformly at random with probability $p_k = 2^{-k}$.

Dimension: since $|I| \approx \log n$ we need $O(\log n)$ dimensions to approximate distances in expectation. If we want guarantees with high probability, we need to sample from each distribution

Spatial Approach - Padded Decomposition

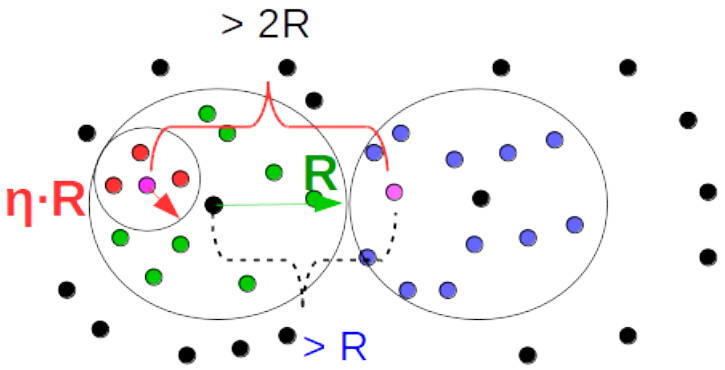
A different approach inspired by the works of Bartal, CKR and FRT is to:

1. partition pairs of points depending on the scale of the distance

$$P_k \approx \left\{ \{x, y\} \in \binom{\mathcal{X}}{2} \middle| \, 2^{k+1} < d(x, y) \leq 2^{k+2} \right\}$$

2. obtain a distribution over set W_k by doing the following:

   - We first obtain a *coarsening* of the metric space at scale k by partitioning the space into clusters $\mathcal{C} = \{C_1, \ldots, C_s\}$ compactly given by a function $C : \mathcal{X} \to \mathcal{C}$ such that
     - **Boundedness:** the diameter of clusters is bounded by 2^{k+1}.
     - **Padding property:** with constant probability $B(x, \eta \cdot 2^k) \subseteq C(x)$ for a padding parameter $\eta : \mathcal{X} \to (0, 1)$.



   - The final set W_k is constructed by including each cluster $C_1, \ldots, C_s$ with probability $1/2$.

   Since, the points belong in different clusters then with constant probability exactly one of them will not belong to W_k in which case we have that:

$$|d(x, W_k) - d(y, W_k)| \geq d(x, \mathcal{X} \setminus C(x)) \geq \eta \cdot R \approx \eta \cdot d(x, y)$$

Dimension: let Δ be the aspect ratio[3] of the space. Based on this approach we need $O(\log \Delta)$ dimensions to approximate all distances in expectation. If we need guarantees with high probability, then the dimension becomes $O(\log \Delta \log n)$.

[3] Maximum over minimum distance between distinct points in $\mathcal{X}$.

A hybrid approach: Measured descent

The last approach although more sophisticated than Bourgain's approach results in worse guarantees whenever $\Delta \gg n$. The reason being, that if there are many different scales in the metric space, we will be adding coordinates that are not useful for most pairs of points. In other words, although there are $\log \Delta$ *spatial scales* there can be only $\log n$ *entropic scales* as in Bourgain's approach. Krauthgamer, Lee, Mendel and Naor proposed the following approach:

1. Partition the distances again according to the entropic scale t of the distance.

2. For each point $x \in \mathcal{X}$ and entropic scale $t \in \log n$ we consider the spatial scale $K(x,t)$ for which the entropic scale is achieved, i.e., $K(x,t)$ is such that $|B(x, 2^{K(x,t)})| \approx 2^t$. To construct the set W_t we:

   - Let $\tilde{W}_1, \ldots, \tilde{W}_{\log \Delta}$ be sets generated using the padded decomposition from the previous section.
   - For each $t \in [\log_2 n]$ we define the set $W_t := \{x \in \mathcal{X} : x \in \tilde{W}_{K(x,t)}\}$.

   That, is locally we use different scales for a specific entropy scale. This works because of the locality property of the padded decomposition and smoothness property of $K(x,t)$ (cannot change too much locally).

To make the connection with Bourgain's approach, here we roughly partition the space into $n/2^t$ "local" clusters of cardinality 2^t and then pick each one with probability $1/2$. So, we perform an entropic coarsening of the metric space and then sample uniformly at random.

Dimension: using the entropy scale to partition the distances allows us to approximate distances in expectation using dimension $O(\log n)$ and $O(\log^2 n$ if we require with high probability. The importance of this method compared to Bourgain's is that for different metric spaces we can get improved distortion by using the geometry in a non-trivial way as captured by the existence of a padding decomposition with

Beyond Frechet Embeddings

The previous considerations illustrate three important points:

- **Expansion:** depends mostly on the *dimension* and for ℓ_1 is typically proportional to the dimension! Roughly this means that if we can reduce the amount of irrelevant information per pair of points we can improve our expansion/dimension.

  Question: can we reduce the dimension? Besides, in the analysis we always end up adding the different components together, so perhaps a one dimensional embedding might be possible!

  Consider a one dimensional embedding of the form $\phi(x) = \sum_{k \in \mathbb{Z}} \phi_k(x)$.

– By triangle inequality we have the same upper bound as when we had separate dimensions $|\phi(x) - \phi(y)| \leq \sum_{k\in\mathbb{Z}} |\phi_k(x) - \phi_k(y)|$

- **Contraction:** depends on the *padding parameter* $\eta : \mathcal{X} \to (0, 1)$ of the padded decomposition. In fact to show the lower bound, we use the fact that with constant probability $d(x, \mathcal{X} \setminus C(x)) \geq \eta \cdot d(x, y)$ instead of directly bounding $d(x, W)$.

  Question: Perhaps it is easier to work with $d(x, W_k(x))$, where $W_k(x) := \mathcal{X} \setminus C(x)$ instead of having a specific set $W_k(x) := W_k$ for all $x \in \mathcal{X}$? How can we get upper bounds of the same kind?

  Consider functions $\tilde{\phi}_k(x) = \sigma_k(x) \cdot \min\{d(x, W_k(x)), R_k\}$ where $\sigma_k(x) := \sigma_k(C(x))$ where $\{\sigma_k(C_i)\}_{i\leq s}$ are i.i.d Bernouli(1/2) random variables.

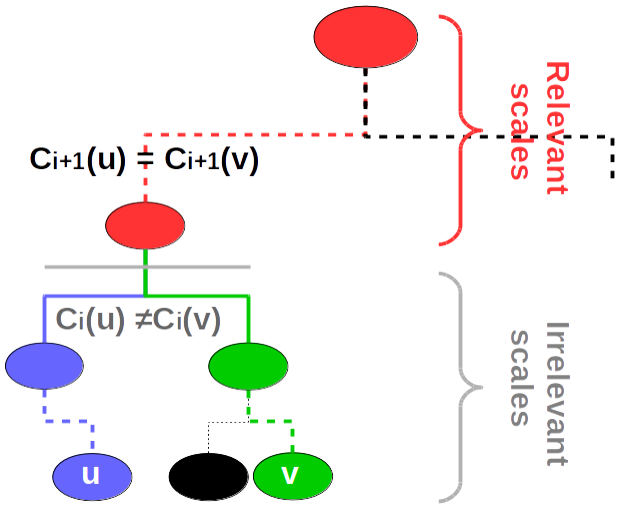
  – The random variables σ are used to simulate the effect that cluster sampling had in the construction of sets W_k. In particular, we get as before that with constant probability $|\phi_k(x) - \phi_k(y)| \geq \eta \cdot d(x, y)$ for at least one $k \in \mathbb{Z}$.

  – if $C(x) \neq C(y)$ then $|\phi_k(x) - \phi_k(y)| \leq R_k$ (irrelevant scales).

  – if $C(x) = C(y)$ then $W_k(x) = W_k(y)$ and $|d(x, W_k(x)) - d(y, W_k(y))| \leq d(x, y)$.

- **Entropic scale:** since for each point we need to somehow encode the distances to other $n - 1$ points, we have seen that the right way to look at the problem is based on entropy. To bound the distortion given that we have witness "coordinates" (terms in the sum) for every scale, we need to make sure that we do not increase the sum too much.

  Question: How can we make sure that we only add terms that carry additional new information about some pair of vertices?



Assume that $d(x,y) \approx 2^{k+2}$ (k-spatial scale), i.e. that k is the first integer for which $C_k(x) \neq C_k(y)$. Then, let $\mathcal{R}$ be the set of scales $\geq k$ for which, x, y are in the same cluster and $\phi_k(x) \neq \phi_k(y)$. We have:

$$\begin{aligned} |\phi(x) - \phi(y)| &\leq \sum_{\ell<k} |\phi_\ell(x) - \phi_\ell(y)| + \sum_{\ell \geq k} |\phi_\ell(x) - \phi_\ell(y)| && (1.12) \\ &\leq \sum_{k<\ell} (|\phi_\ell(x)| + |\phi_\ell(y)|) + \sum_{\ell \geq k} 1 \cdot d(x,y) && (1.13) \\ &\leq 2 \sum_{k<\ell} R_\ell + |\mathcal{R}| \cdot d(x,y) && (1.14) \end{aligned}$$

- *Geometric decay:* if we set $R_\ell = \gamma 2^{\ell+2}$ then the first term is $\approx c \cdot d(x,y)$ for some small constant c.
- Thus, if we can reduce $|\mathcal{R}|$ for most pairs of distances we can obtain better distortion. In other words, we need to include coordinates only when (locally) a significant number of pairs are *separated*.
- To be able to have a lower bound on this *entropic gain*: (a) We keep track of the *minimum gain per cluster* that provides a lower bound on all points in the cluster. (b) Only add coordinates when this lower bound is *large enough*.

The final form of the embedding is given by:

$$\phi_k(x) = \sigma_k(x) \cdot \min \{ w_k(x) \cdot d(x, \mathcal{X}), R_k \}$$

- *Contraction:* with constant probabbility $\exists k, |\phi_k(x) - \phi_k(y)| \geq \eta_k(x) w_k(x) d(x,y)$.
- *Expansion:* the first term of irrelevant scales does not change. However, we need to take some care for the upper bound:

$$\sum_{\ell \geq k} |\phi_\ell(x) - \phi_\ell(y)|$$

  - *Uniformity:* we assume that $w_k(x)$ is constant for all $x \in C_k(x)$.

$$\sum_{\ell \geq k} |\phi_\ell(x) - \phi_\ell(y)| \leq \left[\sum_{k \geq \ell} w_k(x) \right] d(x,y)$$

To bound the distortion we have:

$$dist_\phi(x,y) = O\left(\frac{\sum_{\ell \geq k} w_k(x)}{w_k(x) \eta_k(x)} \right)$$

There are two final ideas:

- *Telescoping:* to control the numerator we define weights such that we have some sort of telescoping.

- *Padding:* We need to pick a padding parameter such that the "padding" event happens with constant probability.

These, constraints are met with the choice of weights:

$$w_k(x) = \mathbb{I}_{\{\text{Local Expansion of C(x) is large}\}} \cdot \eta_k(x)^{-1}$$

$$\eta_k(x) \approx \ln\left(\frac{|B(x, \gamma_1\Delta_k)|}{|B(x, \gamma_2\Delta_k)|}\right)$$

Thus, to complete the analysis outlined above we need to construct a probabilistic partition such that:

- Lower bound on the growth ratio for each point in each cluster.
- Uniformity of the weights in the embedding, i.e. use a single point to define the local growth ratio.
- Local padding is given in terms of the logarithm of the growth ratio, i.e., there must be a constant probability of success, when we shrink the ball by that much.

Uniformly Padded Probabilistic Partitions

Definition 1.3.1 (Local growth rate)**:** The local growth rate of $x \in \mathcal{X}$ at radius r for given scales $\gamma_1, \gamma_2 > 0$ is defined as

$$\rho(x, r, \gamma_1, \gamma_2) := \frac{|B(x, r\gamma_1)|}{|B(x, r\gamma_2)|} \tag{1.15}$$

Given $Z \subseteq \mathcal{X}$ we define $\rho(Z, r, \gamma_1, \gamma_2)$ as $\min_{x \in Z} \rho(x, r, \gamma_1, \gamma_2)$. The *minimum* local growth rate of x at radius r and scales γ_1, γ_2 is defined as $\bar{\rho}(x, r, \gamma_1, \gamma_2) := \rho(B(x, r), r, \gamma_1, \gamma_2)$.

Definition 1.3.2 (Uniformity)**:** Given a partition $\mathcal{C}$ of a metric space, a function f is called *uniform* with respect to $\mathcal{C}$ if for any x, y such that $C(x) = C(y)$ we have $f(x) = f(y)$[4].

Definition 1.3.3 (Uniformly padded Local PP)**:** Given $\Delta > 0$ and $0 < \delta \leq 1$, let $\hat{\mathcal{P}}$ be a Δ-bounded probabilistic partition of $(\mathcal{X}, d)$. Given a collection of functions $\eta = \{\eta_C : \mathcal{X} | C \in \hat{\mathcal{P}}\}$, we say that $\hat{\mathcal{P}}$ is locally (η, δ)-*locally padded* if the event $B(x, \eta(x)\Delta) \subseteq C(x)$ occurs with probability at least δ regardless of the structure of the partition outside $B(x, 2\Delta)$. We say that $\hat{\mathcal{P}}$ is strongly $(\eta, \hat{\delta})$-locally padded if for any $\hat{\delta} \leq \delta \leq 1$, $\hat{\mathcal{P}}$ is $(\eta \ln(1/\delta), \delta)$-padded. We say that $\hat{P}$ is (η, δ)-*uniformly* locally padded if η is uniform with respect to $\hat{\mathcal{P}}$.

Our goal is to construct Uniformly Padded Probabilistic Partitions where η provides a uniform lower bound on the growth ratio of the points in each cluster. To achieve this, we first define a procedure that given a specific center, gives us the local padding property.

[4]The same definition extends to probabilistic partitions if it holds for every partition in its support.

The basic partitioning

Definition 1.3.4: Given sets $A, B, C \subset \mathcal{X}$ we denote by $A \bowtie (B, C)$ the property that $A \cap B \neq \emptyset$ and $A \cap C \neq \emptyset$.

Using this definition we may express the event that padding doesn't happen at x as $B(x, \eta\Delta) \bowtie (S, \bar{S})$. We consider that the decomposition failed when the ball $B(x, \eta\Delta)$ is partially cut by the decomposition.

Lemma 1.3.5. *For any metric space (Z, d) point $v \in Z$, real parameters $\chi \geq 2$, $\Delta > 0$, let r be a random variable sampled from a truncated exponential density function with parameter $\kappa = 8\ln(\chi)/\Delta$*

$$f(r) = \begin{cases} \frac{\chi^2}{1-\chi^{-2}} \kappa e^{-\kappa r}, & r \in [\Delta/4, \Delta/2] \\ 0, & \text{otherwise} \end{cases}$$

If $S = B(v, r)$ and $\bar{S} = Z \setminus S$ then for any $\theta \in [\chi^{-1}, 1)$ and any $x \in Z$:

$$\Pr\left[B(x, \eta \cdot \Delta) \bowtie (S, \bar{S})\right] \leq (1-\theta)\left(\Pr[B(x, \eta \cdot \Delta) \subsetneq \bar{S}] + \frac{2\theta}{\chi}\right)$$

where $\eta = 2^{-4}\ln(1/\theta)/\ln(\chi)$.

The proof of this lemma is a simple exercise in calculus.

The Probabilistic Decomposition

Let (X, d) be a metric space. To generate the probabilistic decomposition we invoke the basic partitioning iteratively on a carefully selected sequence of centers that guarantees the desired lower bounds on the growth ratio. First we deterministically assign centers $v_1, \ldots, v_s$ and parameters $\chi_1, \ldots, \chi_s$ to be determined shortly. Let $Z_1 = X$ and $j = 1$. Conduct the following iterative process

1. Let $v_j \in Z_j$ be the point minimizing $\hat{\chi}_j = \rho(x, 2\Delta, \gamma_1, \gamma_2)$ over all $x \in Z_j$ (Lower bound growth rate).
2. Set $\chi_j = \max\{2/\delta^{1/2}, \hat{\chi}_j\}$.
3. $Z_{j+1} = Z_j \setminus B(v_j, \Delta/4)$.
4. Set $j = j + 1$. If $Z_j \neq \emptyset$ return to step 1.

The above process determines the centers and the local parameters of the partition. We know show how to obtain the probabilistic partition and uniform weights for the embedding. Let $Z_1 = X$. For $j = 1, \ldots, s$:

1. Let $(S_{v_j}, \bar{S}_{v_j})$ be the partition created by $S_{v_j} = B_{Z_j}(v_j, r)$ and $\bar{S}_{v_j} = Z_j \setminus S_{v_j}$ where r is distributed as before with parameter $\kappa = 8\ln(\chi_j)/\Delta$.

2. Set $C_j = S_{v_j}$ and $Z_{j+1} = \bar{S}_{v_j}$.

3. For all $x \in C_j$ let $\eta_C(x) = 2^{-6}/\max\{\ln \hat{\chi}_j, \ln(1/\hat{\delta})\}$.

4. If $\hat{\chi}_j \geq 1/\hat{\delta}$ then $\xi_C(x) = 1$ otherwise $\xi_C(x) = 0$.

Where $\theta = \delta^{1/2}$ for some fixed $\delta \geq \hat{\delta}$. We have the following lemma.

Lemma 1.3.6. *Let $0 < \Delta \leq diam(Z)$. Let $\hat{\delta} \in (0, 1/2], \gamma_1 \geq 2, \gamma_2 \leq 2^{-4}$. There exists a Δ-bounded probabilistic partition $\hat{\mathcal{P}}$ of (Z, d) and a collection of uniform functions $\{\xi_C : Z \to \{0,1\} | C \in \hat{\mathcal{P}}\}$ and $\{\eta_C : Z \to (0,1] | C \in \hat{\mathcal{P}}\}$ such that the probabilistic partition is a strong $(\eta, \hat{\delta})$-uniformly locally padded PP and the following conditions hold for any $C \in \hat{\mathcal{P}}$ and any $x \in Z$:*

- *If $\xi_C(x) = 1$ then: $\frac{2^{-6}}{\ln \rho(x, 2\Delta, \gamma_1, \gamma_2)} \leq \eta_C(x) \leq \frac{2^{-6}}{\ln(1/\hat{\delta})}$.*

- *If $\xi_C(x) = 0$ then: $\eta_C(x) = \frac{2^{-6}}{\ln(1/\hat{\delta})}$ and $\bar{\rho}(x, 2\Delta, \gamma_1, \gamma_2) < 1/\hat{\delta}$.*

Main Embedding

- $\Delta_0 := \mathrm{diam}(X)$ and $\Delta_i = \left(\frac{\zeta}{8}\right)^i \Delta_0$ for $i \in \mathbb{N}$

- For all $i \in \mathbb{N}$ let C_i be a strong $(\eta_i, \hat{\delta})$- uniformly locally padded PP with parameters $\gamma_1 = 8/\zeta$, $\gamma_2 = 1/16$, $\hat{\delta} = 1/2$, $\Delta = \Delta_i$ and $Z = X$ as given by the previous lemma.

- Let $\sigma_i(x)$, $\xi_i(x)$, $\eta_i(x)$ be uniform functions with respect to C_i

Lemma 1.4.1. *Let (X,d) be a finite metric space on n points and $0 < \zeta \leq 1/8$, then there exists a distribution $\mathcal{D}$ over functions $f : X \to \mathbb{R}$ such that for al $u, v \in X$:*

1. *For all $f \in supp(\mathcal{D})$,*

$$|f(u) - f(v)| \leq C \left\lceil \ln\left(\frac{n}{|B(u, d(u,v))|}\right)\right\rceil \cdot d(u,v)$$

2. $\Pr_{f \sim \mathcal{D}}[|f(u) - f(v)| \geq \zeta^3 \cdot d(u,v)/C] \geq 1 - \zeta$,

where $C > 0$ is a universal constant.

Using this lemma and Chernoff bounds we can prove our main result.

Extensions

Theorem 1.5.1 (Scaling Distortion for Doubling Spaces)**.** *For an doubling metric space $(\mathcal{X}, d)$ there exists an embedding $f : X \to \ell_p^D$ with coarse scaling distortion $O(\log^{26}(1/\epsilon))$ where $D = O(\dim(\mathcal{X}) \log \dim(X))$.*

Theorem 1.5.2 (Scaling Distortion for Decomposable Metrics)**.** *Let $1 \leq p \leq \infty$. For any n point τ-decomposable metric space $(\mathcal{X}, d)$ there exists an embedding with coarse scaling distortion $O\left(\min\left\{(1/\tau)^{1-1/p}\log(2/\epsilon)^{1/p}, \log(2\epsilon)\right\}\right)$.*

Theorem 1.5.3 (Ultrametrics)**.** *For any n-point metric space $(\mathcal{X}, d)$ there exists an embedding into a distribution over ultrametrics with coarse scaling distortion $O(\log(2/\epsilon))$.*