— 42 —

Finite Metric Spaces and Partitions

42.1. Finite Metric Spaces

Definition 42.1.1. A ***metric space*** is a pair $(\mathcal{X}, \mathsf{d})$ where $\mathcal{X}$ is a set and $\mathsf{d} : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is a ***metric***, satisfying the following axioms:

(i) $\mathsf{d}(x, y) = 0 \iff x = y$,
(ii) $\mathsf{d}(x, y) = \mathsf{d}(y, x)$, and
(iii) $\mathsf{d}(x, y) + \mathsf{d}(y, z) \geq \mathsf{d}(x, z)$ (triangle inequality).

The plane, $\mathbb{R}^2$, with the regular Euclidean distance is a metric space.

Of special interest is the finite case, where $\mathcal{X}$ is an n-point set. Then, the function d can be specified by $\binom{n}{2}$ real numbers. Alternatively, one can think about $(\mathcal{X}, \mathsf{d})$ as a weighted complete graph, where positive weights are specified on the edges, and these weights comply with the triangle inequality.

Finite metric spaces rise naturally from (sparse) graphs. Indeed, let $\mathsf{G} = (\mathcal{X}, \mathsf{E})$ be an undirected weighted graph defined over $\mathcal{X}$, and let $\mathsf{d}_{\mathsf{G}}(x, y)$ be the length of the shortest path between x and y in G. It is easy to verify that $(\mathcal{X}, \mathsf{d}_G)$ is a finite metric space. As such if the graph G is sparse, it provides a compact representation to the finite space $(\mathcal{X}, \mathsf{d}_G)$.

Definition 42.1.2. Let $(\mathcal{X}, d)$ be an n-point metric space. We denote the ***open ball*** of radius r about $x \in \mathcal{X}$, by $\mathbf{b}(x, r) = \{y \in \mathcal{X} \mid \mathsf{d}(x, y) < r\}$.

Underling our discussion of metric spaces are algorithmic applications. The hardness of various computational problems depends heavily on the structure of the finite metric space. Thus, given a finite metric space, and a computational task, it is natural to try to map the given metric space into a new metric where the task at hand becomes easy.

Example 42.1.3. Computing the diameter of a point set is not trivial in two dimensions (if one wants near linear running time), but is easy in one dimension. Thus, if we could map points in two dimensions into points in

one dimension, such that the diameter is preserved, then computing the diameter becomes easy. This approach yields an efficient approximation algorithm, see Exercise 42.7.3 below.

Of course, this mapping from one metric space to another, is going to introduce error. Naturally, one would like to minimize the error introduced by such a mapping.

Definition 42.1.4. Let $(\mathcal{X}, \mathsf{d}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathsf{d}_{\mathcal{Y}})$ be two metric spaces. A mapping $f : \mathcal{X} \to \mathcal{Y}$ is an ***embedding***, and it is ***C-Lipschitz*** if $\mathsf{d}_Y(f(x), f(y)) \leq C \cdot \mathsf{d}_{\mathcal{X}}(x, y)$ for all $x, y \in \mathcal{X}$. The mapping f is ***K-bi-Lipschitz*** if there exists a $C > 0$ such that

$$CK^{-1} \cdot \mathsf{d}_{\mathcal{X}}(x, y) \leq \mathsf{d}_Y(f(x), f(y)) \leq C \cdot \mathsf{d}_{\mathcal{X}}(x, y),$$

for all $x, y \in \mathcal{X}$.

The least K for which f is K-bi-Lipschitz is the ***distortion*** of f, and is denoted $\operatorname{dist}(f)$. The least distortion with which $\mathcal{X}$ may be embedded in $\mathcal{Y}$ is denoted $c_{\mathcal{Y}}(\mathcal{X})$.

Informally, if $f : \mathcal{X} \to \mathcal{Y}$ has distortion K, then the distances in $\mathcal{X}$ and $f(\mathcal{X}) \subseteq \mathcal{Y}$ are the same up to a factor of K (one might need to scale up the distances by some constant C).

There are several powerful results about low distortion embeddings that would be presented:

(I) **Probabilistic trees**. Every finite metric can be randomly embedded into a tree such that the "expected" distortion for a specific pair of points is $O(\log n)$.

(II) **Bourgain embedding**. Any n-point metric space can be embedded into (finite dimensional) euclidean metric space with $O(\log n)$ distortion.

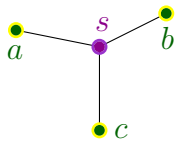
(III) **Johnson-Lindenstrauss lemma**. Any n-point set in Euclidean space with the regular Euclidean distance can be embedded into $\mathbb{R}^k$ with distortion $(1 + \varepsilon)$, where $k = O(\varepsilon^{-2} \log n)$.

42.2. Examples

What is distortion? When considering a mapping $f : \mathcal{X} \to \mathbb{R}^d$ of a metric space $(\mathcal{X}, \mathsf{d})$ to $\mathbb{R}^d$, it would useful to observe that since $\mathbb{R}^d$ can be

scaled, we can consider f to be an expansion (i.e., no distances shrink). Furthermore, we can assume that there is at least one pair of points $x, y \in \mathcal{X}$, such that $\mathsf{d}(x,y) = \|x-y\|$. As such, we have $\mathrm{dist}(f) = \max_{x,y} \frac{\|x-y\|}{\mathsf{d}(x,y)}$.

Why is distortion necessary? Consider the a graph $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ with one vertex s connected to three other vertices a, b, c, where the weights on the edges are all one (i.e., G is the star graph with three leafs). We claim that G can not be embedded into Euclidean space with distortion $\leq \sqrt{2}$. Indeed, consider the associated metric space $(\mathsf{V}, \mathsf{d}_G)$ and an (expansive) embedding $f : \mathsf{V} \to \mathbb{R}^d$.



Consider the triangle formed by $\triangle = a'b'c'$, where $a' = f(a), b' = f(b)$ and $c' = f(c)$. Next, consider the following quantity

$$\max(\|a'-s'\|, \|b'-s'\|, \|c'-s'\|),$$

which lower bounds the distortion of f. This quantity is minimized when $r = \|a'-s'\| = \|b'-s'\| = \|c'-s'\|$. Namely, s' is the center of the smallest enclosing circle of $\triangle$. However, r is minimized when all the edges of $\triangle$ are of equal length, and are of length $\mathsf{d}_G(a,b) = 2$. It follows that $\mathrm{dist}(f) \geq r \geq 2/\sqrt{3}$.

This quantity is minimized when $r = \|a'-s'\| = \|b'-s'\| = \|c'-s'\|$. Namely, s' is the center of the smallest enclosing circle of $\triangle$. However, r is minimized when all the edges of $\triangle$ are of equal length and are of length $\mathsf{d}_\mathsf{G}(a,b) = 2$. Observe that the height of the equilateral triangle with sidelength 2 is $h = \sqrt{3}$, and the radius of its inscribing circle is $r = (2/3)h = 2/\sqrt{3}$; see Figure 42.1. As such, it follows that $\mathrm{dist}(f) \geq r = 2/\sqrt{3}$.

Note that the above argument is independent of the target dimension d. A packing argument shows that embedding the star graph with n leaves into $\mathbb{R}^d$ requires distortion $\Omega\big(n^{1/d}\big)$; see Exercise ??. It is known that $\Omega(\log n)$ distortion is necessary in the worst case when embedding a graph into Euclidean space (this is shown using expanders). A proof of distortion $\Omega(\log n / \log\log n)$ is sketched in the bibliographical notes.

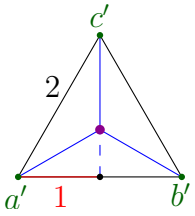


Figure 42.1: The equilateral triangle and its enclosing circle.

42.2.1. Hierarchical Tree Metrics

The following metric is quite useful in practice, and nicely demonstrate why algorithmically finite metric spaces are useful.

Definition 42.2.1. ***Hierarchically well-separated tree*** (HST) is a metric space defined on the leaves of a rooted tree T. To each vertex $u \in T$ there is associated a label $\Delta_u \geq 0$ such that $\Delta_u = 0$ if and only if u is a leaf of T. The labels are such that if a vertex u is a child of a vertex v then $\Delta_u \leq \Delta_v$. The distance between two leaves $x, y \in T$ is defined as $\Delta_{\mathrm{lca}(x,y)}$, where $\mathrm{lca}(x, y)$ is the least common ancestor of x and y in T.

A HST T is a ***k-HST*** if for a vertex $v \in T$, we have that $\Delta_v \leq \Delta_{\overline{\mathrm{p}}(v)}/k$, where $\overline{\mathrm{p}}(v)$ is the parent of v in T.

Note that a HST is a very limited metric. For example, consider the cycle $G = C_n$ of n vertices, with weight one on the edges, and consider an expansive embedding f of G into a HST HST. It is easy to verify, that there must be two consecutive nodes of the cycle, which are mapped to two different subtrees of the root r of HST. Since HST is expansive, it follows that $\Delta_r \geq n/2$. As such, $\mathrm{dist}(f) \geq n/2$. Namely, HSTs fail to faithfully represent even very simple metrics.

42.2.2. Clustering

One natural problem we might want to solve on a graph (i.e., finite metric space) $(\mathcal{X}, \mathsf{d})$ is to partition it into clusters. One such natural clustering is the ***k-median clustering***, where we would like to choose a set $C \subseteq \mathcal{X}$ of k centers, such that $\nu_C(\mathcal{X}, \mathsf{d}) = \sum_{\mathsf{u} \in \mathcal{X}} \mathsf{d}(\mathsf{u}, C)$ is minimized, where $\mathsf{d}(\mathsf{u}, C) = \min_{c \in C} \mathsf{d}(\mathsf{u}, c)$ is the distance of u to its closest center in C.

It is known that finding the optimal k-median clustering in a (general weighted) graph is NP-complete. As such, the best we can hope for is an approximation algorithm. However, if the structure of the finite metric space $(\mathcal{X}, \mathsf{d})$ is simple, then the problem can be solved efficiently. For example, if the points of $\mathcal{X}$ are on the real line (and the distance between a and b is just $|a - b|$), then k-median can be solved using dynamic programming.

Another interesting case is when the metric space $(\mathcal{X}, \mathsf{d})$ is a HST. Is not too hard to prove the following lemma. See Exercise 42.7.1.

Lemma 42.2.2. *Let $(\mathcal{X}, \mathsf{d})$ be a HST defined over n points, and let $k > 0$ be an integer. One can compute the optimal k-median clustering of $\mathcal{X}$ in $O(k^2 n)$ time.*

Thus, if we can embed a general graph G into a HST HST, with low distortion, then we could approximate the k-median clustering on G by clustering the resulting HST, and "importing" the resulting partition to the original space. The quality of approximation, would be bounded by the distortion of the embedding of G into HST.

42.3. Random Partitions

Let $(\mathcal{X}, d)$ be a finite metric space. Given a partition $P = \{C_1, \ldots, C_m\}$ of $\mathcal{X}$, we refer to the sets C_i as ***clusters***. We write $\mathcal{P}_{\mathcal{X}}$ for the set of all partitions of $\mathcal{X}$. For $x \in \mathcal{X}$ and a partition $P \in \mathcal{P}_{\mathcal{X}}$ we denote by $P(x)$ the unique cluster of P containing x. Finally, the set of all probability distributions on $\mathcal{P}_{\mathcal{X}}$ is denoted $\mathcal{D}_{\mathcal{X}}$.

The following partition scheme is due to [CKR04].

42.3.1. Constructing the partition

Consider a given metric space $(\mathcal{X}, \mathsf{d})$, where $\mathcal{X}$ is a set of n points.

Let $\Delta = 2^u$ be a prescribed parameter, which is the required diameter of the resulting clusters. Choose, uniformly at random, a permutation π of $\mathcal{X}$ and a random value $\alpha \in [1/4, 1/2]$. Let $R = \alpha\Delta$, and observe that it is uniformly distributed in the interval $[\Delta/4, \Delta/2]$.

The partition is now defined as follows: A point $x \in \mathcal{X}$ is assigned to the cluster C_y of y, where y is the first point in the permutation in

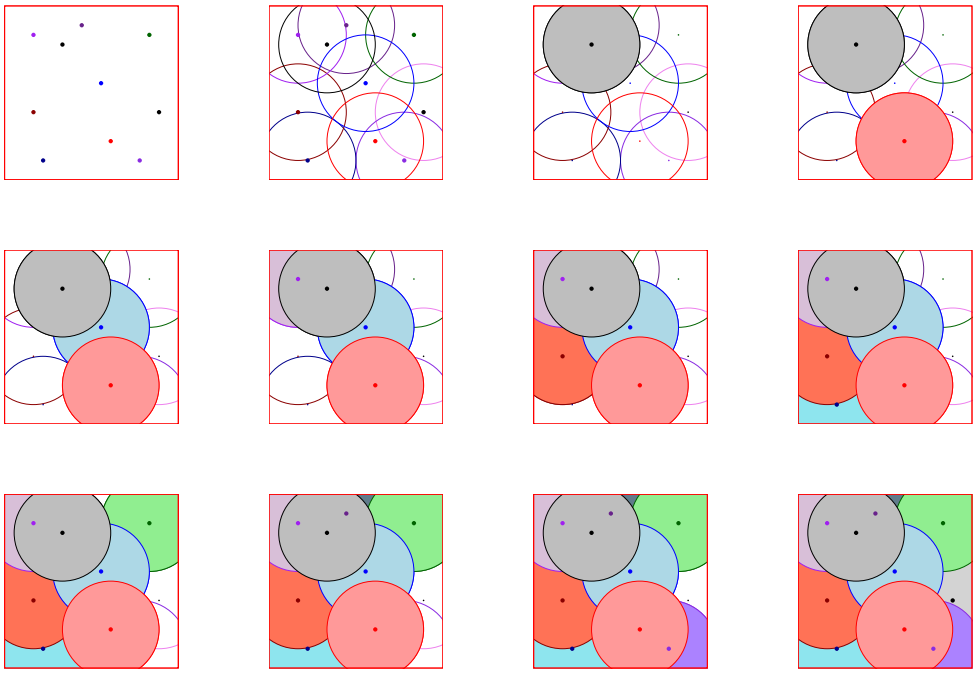


Figure 42.2: An example of the partition of a square (induced by a set of points) as described in Section 42.3.1.

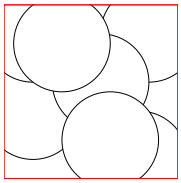


Figure 42.3: The resulting partition.

distance $\leq R$ from x. Formally,

$$C_y = \left\{ x \in \mathcal{X} \;\middle|\; \begin{array}{c} x \in \mathbf{b}(y,R) \text{ and } \pi(y) \leq \pi(z) \\ \text{for all } z \in \mathcal{X} \text{ with } x \in \mathbf{b}(z,R) \end{array} \right\}.$$

Let $P = \{C_y\}_{y \in \mathcal{X}}$ denote the resulting partition.

Here is a somewhat more intuitive explanation: Once we fix the radius of the clusters R, we start scooping out balls of radius R centered at the

points of the random permutation π. At the ith stage, we scoop out only the remaining mass at the ball centered at x_i of radius r, where x_i is the ith point in the random permutation.

42.3.2. Properties

The following lemma quantifies the probability of a (crystal) ball of radius t centered at a point x is fully contained in one of the clusters of the partition? (Otherwise, the crystal ball is of course broken.)

Lemma 42.3.1. *Let $(\mathcal{X}, d)$ be a finite metric space, $\Delta = 2^u$ a prescribed parameter, and let P be the partition of $\mathcal{X}$ generated by the above random partition. Then the following holds:*

(i) For any $C \in P$, we have $\operatorname{diam}(C) \leq \Delta$.

(ii) Let x be any point of $\mathcal{X}$, and t a parameter $\leq \Delta/8$. Then,

$$\mathbb{P}[\mathbf{b}(x,t) \not\subseteq P(x)] \leq \frac{8t}{\Delta} \ln \frac{b}{a},$$

where $a = |\mathbf{b}(x, \Delta/8)|$, and $b = |\mathbf{b}(x, \Delta)|$.

Proof: Since $C_y \subseteq \mathbf{b}(y, R)$, we have that $\operatorname{diam}(C_y) \leq \Delta$, and thus the first claim holds.

Let U be the set of points of $\mathbf{b}(x, \Delta)$, such that $w \in U$ iff $\mathbf{b}(w, R) \cap \mathbf{b}(x, t) \neq \emptyset$. Arrange the points of U in increasing distance from x, and let $w_1, \ldots, w_{b'}$ denote the resulting order, where $b' = |U|$. Let $I_k = [d(x, w_k) - t, d(x, w_k) + t]$ and write $\mathcal{E}_k$ for the event that w_k is the first point in π such that $\mathbf{b}(x,t) \cap C_{w_k} \neq \emptyset$, and yet $\mathbf{b}(x,t) \not\subseteq C_{w_k}$. Note that if $w_k \in \mathbf{b}(x, \Delta/8)$, then $\mathbb{P}[\mathcal{E}_k] = 0$ since $\mathbf{b}(x,t) \subseteq \mathbf{b}(x, \Delta/8) \subseteq \mathbf{b}(w_k, \Delta/4) \subseteq \mathbf{b}(w_k, R)$.

In particular, $w_1, \ldots, w_a \in \mathbf{b}(x, \Delta/8)$ and as such $\mathbb{P}[\mathcal{E}_1] = \cdots = \mathbb{P}[\mathcal{E}_a] = 0$. Also, note that if $\mathsf{d}(x, w_k) < R - t$ then $\mathbf{b}(w_k, R)$ contains $\mathbf{b}(x,t)$ and as such $\mathcal{E}_k$ can not happen. Similarly, if $\mathsf{d}(x, w_k) > R + t$ then $\mathbf{b}(w_k, R) \cap \mathbf{b}(x,t) = \emptyset$ and $\mathcal{E}_k$ can not happen. As such, if $\mathcal{E}_k$ happen then $R - t \leq \mathsf{d}(x, w_k) \leq R + t$. Namely, if $\mathcal{E}_k$ happen then $R \in I_k$. Namely, $\mathbb{P}[\mathcal{E}_k] = \mathbb{P}[\mathcal{E}_k \cap (R \in I_k)] = \mathbb{P}[R \in I_k] \cdot \mathbb{P}[\mathcal{E}_k \mid R \in I_k]$. Now, R is uniformly distributed in the interval $[\Delta/4, \Delta/2]$, and I_k is an interval of length $2t$. Thus, $\mathbb{P}[R \in I_k] \leq 2t/(\Delta/4) = 8t/\Delta$.

Next, to bound $\mathbb{P}[\mathcal{E}_k \mid R \in I_k]$, we observe that $w_1, \ldots, w_{k-1}$ are closer to x than w_k and their distance to $\mathbf{b}(x,t)$ is smaller than R. Thus, if any of

them appear before w_k in π then $\mathcal{E}_k$ does not happen. Thus, $\mathbb{P}[\mathcal{E}_k \mid R \in I_k]$ is bounded by the probability that w_k is the first to appear in π out of $w_1, \ldots, w_k$. But this probability is $1/k$, and thus $\mathbb{P}[\mathcal{E}_k \mid R \in I_k] \leq 1/k$.

We are now ready for the kill. Indeed,

$$\begin{aligned}
\mathbb{P}[\mathbf{b}(x,t) \not\subseteq P(x)] &= \sum_{k=1}^{b'} \mathbb{P}[\mathcal{E}_k] \\
&= \sum_{k=a+1}^{b'} \mathbb{P}[\mathcal{E}_k] \\
&= \sum_{k=a+1}^{b'} \mathbb{P}[R \in I_k] \cdot \mathbb{P}[\mathcal{E}_k \mid R \in I_k] \\
&\leq \sum_{k=a+1}^{b'} \frac{8t}{\Delta} \cdot \frac{1}{k} \leq \frac{8t}{\Delta} \ln \frac{b'}{a} \leq \frac{8t}{\Delta} \ln \frac{b}{a},
\end{aligned}$$

since $\sum_{k=a+1}^{b} \frac{1}{k} \leq \int_a^b \frac{dx}{x} = \ln \frac{b}{a}$ and $b' \leq b$.

42.4. Probabilistic embedding into trees

In this section, given n-point finite metric $(\mathcal{X}, \mathsf{d})$. we would like to embed it into a HST. As mentioned above, one can verify that for any embedding into HST, the distortion in the worst case is $\Omega(n)$. Thus, we define a randomized algorithm that embed $(\mathcal{X}, d)$ into a tree. Let T be the resulting tree, and consider two points $x, y \in \mathcal{X}$. Consider the ***random variable*** $\mathsf{d}_T(x,y)$. We constructed the tree T such that distances never shrink; i.e. $\mathsf{d}(x,y) \leq \mathsf{d}_T(x,y)$. The ***probabilistic distortion*** of this embedding is $\max_{x,y} \mathbb{E}\left[\frac{\mathsf{d}_T(x,y)}{\mathsf{d}(x,y)}\right]$. Somewhat surprisingly, one can find such an embedding with logarithmic probabilistic distortion.

Theorem 42.4.1. *Given n-point metric $(\mathcal{X}, d)$ one can randomly embed it into a 2-HST with probabilistic distortion $\leq 24 \ln n$.*

Proof: The construction is recursive. Let $\operatorname{diam}(P)$, and compute a random partition of $\mathcal{X}$ with cluster diameter $\operatorname{diam}(P)/2$, using the construction of Section 42.3.1. We recursively construct a 2-HST for each cluster, and

hang the resulting clusters on the root node v, which is marked by $\Delta_v = \operatorname{diam}(P)$. Clearly, the resulting tree is a 2-HST.

For a node $v \in T$, let $\mathcal{X}(v)$ be the set of points of $\mathcal{X}$ contained in the subtree of v.

For the analysis, assume $\operatorname{diam}(P) = 1$, and consider two points $x, y \in \mathcal{X}$. We consider a node $v \in T$ to be in level i if $\operatorname{level}(v) = \lceil \lg \Delta_v \rceil = i$. The two points x and y correspond to two leaves in T, and let $\hat{u}$ be the least common ancestor of x and y in t. We have $\mathsf{d}_T(x, y) \leq 2^{\operatorname{level}(v)}$. Furthermore, note that along a path the levels are strictly monotonically increasing.

Being more conservative, let w be the first ancestor of x, such that $\mathbf{b} = \mathbf{b}(x, \mathsf{d}(x, y))$ is not completely contained in $\mathcal{X}(u_1), \ldots, \mathcal{X}(u_m)$, where $u_1, \ldots, u_m$ are the children of w. Clearly, $\operatorname{level}(w) > \operatorname{level}(\hat{u})$. Thus, $\mathsf{d}_T(x, y) \leq 2^{\operatorname{level}(w)}$.

Consider the path σ from the root of T to x, and let $\mathcal{E}_i$ be the event that $\mathbf{b}$ is not fully contained in $\mathcal{X}(v_i)$, where v_i is the node of σ of level i (if such a node exists). Furthermore, let Y_i be the indicator variable which is 1 if $\mathcal{E}_i$ is the first to happened out of the sequence of events $\mathcal{E}_0, \mathcal{E}_{-1}, \ldots$. Clearly, $\mathsf{d}_T(x, y) \leq \sum Y_i 2^i$.

Let $t = \mathsf{d}(x, y)$ and $j = \lfloor \lg \mathsf{d}(x, y) \rfloor$, and $n_i = \left| \mathbf{b}(x, 2^i) \right|$ for $i = 0, \ldots, -\infty$. We have

$$\begin{aligned}
\mathbb{E}[\mathsf{d}_T(x, y)] &\leq \sum_{i=j}^{0} \mathbb{E}[Y_i]\, 2^i \leq \sum_{i=j}^{0} 2^i \, \mathbb{P}\left[\mathcal{E}_i \cap \overline{\mathcal{E}_{i-1}} \cap \overline{\mathcal{E}_{i-1}} \cdots \overline{\mathcal{E}_0}\right] \\
&\leq \sum_{i=j}^{0} 2^i \cdot \frac{8t}{2^i} \ln \frac{n_i}{n_{i-3}},
\end{aligned}$$

by Lemma 42.3.1. Thus,

$$\mathbb{E}[\mathsf{d}_T(x, y)] \leq 8t \ln \left(\prod_{i=j}^{0} \frac{n_i}{n_{i-3}} \right) \leq 8t \ln(n_0 \cdot n_1 \cdot n_2) \leq 24t \ln n.$$

It thus follows, that the expected distortion for x and y is $\leq 24 \ln n$. 

42.4.1. Application: approximation algorithm for k-median clustering

Let $(\mathcal{X}, \mathsf{d})$ be a n-point metric space, and let k be an integer number. We would like to compute the optimal k-median clustering. Number, find a subset $C_{\text{opt}} \subseteq \mathcal{X}$, such that $\nu_{C_{\text{opt}}}(\mathcal{X}, \mathsf{d})$ is minimized, see Section 42.2.2. To this end, we randomly embed $(\mathcal{X}, \mathsf{d})$ into a HST HST using Theorem 42.4.1. Next, using Lemma 42.2.2, we compute the optimal k-median clustering of HST. Let C be the set of centers computed. We return C together with the partition of $\mathcal{X}$ it induces as the required clustering.

Theorem 42.4.2. *Let $(\mathcal{X}, \mathsf{d})$ be a n-point metric space. One can compute in polynomial time a k-median clustering of $\mathcal{X}$ which has expected price $O(\alpha \log n)$, where α is the price of the optimal k-median clustering of $(\mathcal{X}, \mathsf{d})$.*

Proof: The algorithm is described above, and the fact that its running time is polynomial can be easily be verified. To prove the bound on the quality of the clustering, for any point $p \in \mathcal{X}$, let $\operatorname{cen}(p)$ denote the closest point in C_{opt} to p according to d, where C_{opt} is the set of k-medians in the optimal clustering. Let C be the set of k-medians returned by the algorithm, and let HST be the HST used by the algorithm. We have

$$\begin{aligned}
\beta = \nu_C(\mathcal{X}, \mathsf{d}) &\leq \nu_C(\mathcal{X}, \mathsf{d}_{\mathsf{HST}}) \\
&\leq \nu_{C_{\text{opt}}}(\mathcal{X}, \mathsf{d}_{\mathsf{HST}}) \leq \sum_{p \in \mathcal{X}} \mathsf{d}_{\mathsf{HST}}(p, C_{\text{opt}}) \\
&\leq \sum_{p \in \mathcal{X}} \mathsf{d}_{\mathsf{HST}}(p, \operatorname{cen}(p)).
\end{aligned}$$

Thus, in expectation we have

$$\begin{aligned}
\mathbb{E}[\beta] &= \mathbb{E}\Big[\sum_{p \in \mathcal{X}} \mathsf{d}_{\mathsf{HST}}(p, \operatorname{cen}(p))\Big] \\
&= \sum_{p \in \mathcal{X}} \mathbb{E}[\mathsf{d}_{\mathsf{HST}}(p, \operatorname{cen}(p))] \\
&= \sum_{p \in \mathcal{X}} O(\mathsf{d}(p, \operatorname{cen}(p)) \log n) \\
&= O\Big((\log n) \sum_{p \in \mathcal{X}} \mathsf{d}(p, \operatorname{cen}(p))\Big) = O\Big(\nu_{C_{\text{opt}}}(\mathcal{X}, \mathsf{d}) \log n\Big),
\end{aligned}$$

by linearity of expectation and Theorem 42.4.1. 

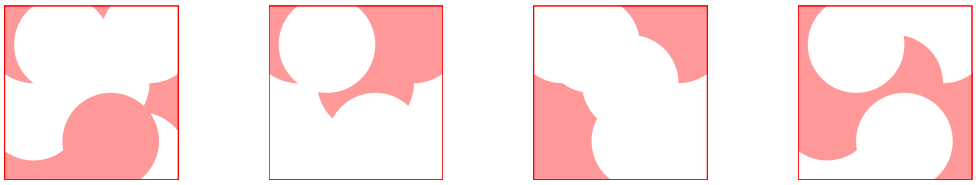


Figure 42.4: Examples of the sets resulting from the partition of Figure 42.2 and taking clusters into a set with probability 1/2.

42.5. Embedding any metric space into Euclidean space

Lemma 42.5.1. *Let $(\mathcal{X}, \mathsf{d})$ be a metric, and let $Y \subset \mathcal{X}$. Consider the mapping $f : \mathcal{X} \to \mathbb{R}$, where $f(x) = \mathsf{d}(x, Y) = \min_{y \in Y} \mathsf{d}(x, y)$. Then for any $x, y \in \mathcal{X}$, we have $|f(x) - f(y)| \leq \mathsf{d}(x, y)$. Namely f is nonexpansive.*

Proof: Indeed, let x' and y' be the closet points of Y, to x and y, respectively. Observe that

$$f(x) = \mathsf{d}(x, x') \leq \mathsf{d}(x, y') \leq \mathsf{d}(x, y) + \mathsf{d}(y, y') = \mathsf{d}(x, y) + f(y)$$

by the triangle inequality. Thus, $f(x) - f(y) \leq \mathsf{d}(x, y)$. By symmetry, we have $f(y) - f(x) \leq \mathsf{d}(x, y)$. Thus, $|f(x) - f(y)| \leq \mathsf{d}(x, y)$.

42.5.1. The bounded spread case

Let $(\mathcal{X}, \mathsf{d})$ be a n-point metric. The ***spread*** of $\mathcal{X}$, denoted by

$$\Phi(\mathcal{X}) = \frac{\operatorname{diam}(\mathcal{X})}{\min_{x,y \in \mathcal{X}, x \neq y} \mathsf{d}(x, y)},$$

is the ratio between the diameter of $\mathcal{X}$ and the distance between the closest pair of points.

Theorem 42.5.2. *Given a n-point metric $\mathcal{Y} = (\mathcal{X}, d)$, with spread Φ, one can embed it into Euclidean space $\mathbb{R}^k$ with distortion $O(\sqrt{\ln \Phi} \ln n)$, where $k = O(\ln \Phi \ln n)$.*

Proof: Assume that $\operatorname{diam}(\mathcal{Y}) = \Phi$ (i.e., the smallest distance in $\mathcal{Y}$ is 1), and let $r_i = 2^{i-2}$, for $i = 1, \ldots, \alpha$, where $\alpha = \lceil \lg \Phi \rceil$. Let $P_{i,j}$ be a random partition of P with diameter r_i, using Theorem 42.4.1, for $i = 1, \ldots, \alpha$ and $j = 1, \ldots, \beta$, where $\beta = \lceil c \log n \rceil$ and c is a large enough constant to be determined shortly.

For each cluster of $P_{i,j}$ randomly toss a coin, and let $V_{i,j}$ be the all the points of $\mathcal{X}$ that belong to clusters in $P_{i,j}$ that got 'T' in their coin toss. For a point $u \in x$, let

$$f_{i,j}(x) = \mathsf{d}(x, \mathcal{X} \setminus V_{i,j}) = \min_{v \in \mathcal{X} \setminus V_{i,j}} \mathsf{d}(x, v),$$

for $i = 0, \ldots, m$ and $j = 1, \ldots, \beta$. Let $F : \mathcal{X} \to \mathbb{R}^{(m+1)\cdot\beta}$ be the embedding, such that

$$F(x) = (\underbrace{f_{0,1}(x), f_{0,2}(x), \ldots, f_{0,\beta}(x)}_{\text{first n resolution block}}, f_{1,1}(x), f_{1,2}(x), \ldots, f_{1,\beta}(x), \ldots, f_{m,1}(x), f_{m,2}(x), \ldots, f_{\alpha,\beta}(x)).$$

Next, consider two points $x, y \in \mathcal{X}$, with distance $\phi = \mathsf{d}(x, y)$. Let k be an integer such that $r_u \leq \phi/2 \leq r_{u+1}$. Clearly, in any partition of $P_{u,1}, \ldots, P_{u,\beta}$ the points x and y belong to different clusters. Furthermore, with probability half $x \in V_{u,j}$ and $y \notin V_{u,j}$ or $x \notin V_{u,j}$ and $y \in V_{u,j}$, for $1 \leq j \leq \beta$.

Let $\mathcal{E}_j$ denote the event that $\mathbf{b}(x, \rho) \subseteq V_{u,j}$ and $y \notin V_{u,j}$, for $j = 1, \ldots, \beta$, where $\rho = \phi/(64 \ln n)$. By Lemma 42.3.1, we have

$$\mathbb{P}\big[\mathbf{b}(x, \rho) \not\subseteq P_{u,j}(x)\big] \leq \frac{8\rho}{r_u} \ln n \leq \frac{\phi}{8 r_u} \leq 1/2.$$

Thus,

$$\begin{aligned}\mathbb{P}[\mathcal{E}_j] &= \mathbb{P}\big[\big(\mathbf{b}(x, \rho) \subseteq P_{u,j}(x)\big) \cap \big(x \in V_{u,j}\big) \cap \big(y \notin V_{u,j}\big)\big] \\ &= \mathbb{P}\big[\mathbf{b}(x, \rho) \subseteq P_{u,j}(x)\big] \cdot \mathbb{P}\big[x \in V_{u,j}\big] \cdot \mathbb{P}\big[y \notin V_{u,j}\big] \geq 1/8,\end{aligned}$$

since those three events are independent. Notice, that if $\mathcal{E}_j$ happens, than $f_{u,j}(x) \geq \rho$ and $f_{u,j}(y) = 0$.

Let X_j be an indicator variable which is 1 if $\mathcal{E}_j$ happens, for $j = 1, \ldots, \beta$. Let $Z = \sum_j X_j$, and we have $\mu = \mathbb{E}[Z] = \mathbb{E}\big[\sum_j X_j\big] \geq \beta/8$. Thus, the

probability that only $\beta/16$ of $\mathcal{E}_1, \ldots, \mathcal{E}_\beta$ happens, is $\mathbb{P}[Z < (1-1/2)\,\mathbb{E}[Z]]$. By the Chernoff inequality, we have

$$\mathbb{P}[Z < (1-1/2)\,\mathbb{E}[Z]] \leq \exp(-\mu 1/(2\cdot 2^2)) = \exp(-\beta/64) \leq 1/n^{10},$$

if we set $c = 640$.

Thus, with high probability

$$\begin{aligned}\|F(x) - F(y)\| &\geq \sqrt{\sum_{j=1}^{\beta} \big(f_{u,j}(x) - f_{u,j}(y)\big)^2} \geq \sqrt{\rho^2 \frac{\beta}{16}} \\ &= \sqrt{\beta}\frac{\rho}{4} = \phi \cdot \frac{\sqrt{\beta}}{256 \ln n}.\end{aligned}$$

On the other hand, $\big|f_{i,j}(x) - f_{i,j}(y)\big| \leq \mathsf{d}(x,y) = \phi \leq 64\rho \ln n$. Thus,

$$\|F(x) - F(y)\| \leq \sqrt{\alpha\beta(64\rho \ln n)^2} \leq 64\sqrt{\alpha\beta}\rho \ln n = \sqrt{\alpha\beta}\cdot\phi.$$

Thus, setting $G(x) = F(x)\frac{256 \ln n}{\sqrt{\beta}}$, we get a mapping that maps two points of distance ϕ from each other to two points with distance in the range $\Big[\phi, \phi\cdot\sqrt{\alpha\beta}\cdot\frac{256\ln n}{\sqrt{\beta}}\Big]$. Namely, $G(\cdot)$ is an embedding with distortion $O(\sqrt{\alpha}\ln n) = O(\sqrt{\ln \Phi}\ln n)$.

The probability that G fails on one of the pairs, is smaller than $(1/n^{10})\cdot\binom{n}{2} < 1/n^8$. In particular, we can check the distortion of G for all $\binom{n}{2}$ pairs, and if any of them fail (i.e., the distortion is too big), we restart the process.

42.5.2. The unbounded spread case

Our next task, is to extend Theorem 42.5.2 to the case of unbounded spread. Indeed, let $(\mathcal{X}, d)$ be a n-point metric, such that $\operatorname{diam}(\mathcal{X}) \leq 1/2$. Again, we look on the different resolutions $r_1, r_2, \ldots$, where $r_i = 1/2^{i-1}$. For each one of those resolutions r_i, we can embed this resolution into β coordinates, as done for the bounded case. Then we concatenate the coordinates together.

There are two problems with this approach: (i) the number of resulting coordinates is infinite, and (ii) a pair x, y, might be distorted a "lot" because it contributes to all resolutions, not only to its "relevant" resolutions.

Both problems can be overcome with careful tinkering. Indeed, for a resolution r_i, we are going to modify the metric, so that it ignores short distances (i.e., distances $\leq r_i/n^2$). Formally, for each resolution r_i, let $G_i = (\mathcal{X}, \widehat{E_i})$ be the graph where two points x and y are connected if $\mathsf{d}(x,y) \leq r_i/n^2$. Consider a connected component $C \in G_i$. For any two points $x, y \in C$, we have $\mathsf{d}(x,y) \leq n(r_i/n^2) \leq r_i/n$. Let $\mathcal{X}_i$ be the set of connected components of G_i, and define the distances between two connected components $C, C' \in \mathcal{X}_i$, to be $\mathsf{d}_i(C,C') = \mathsf{d}(C,C') = \min_{c\in C, c'\in C'} \mathsf{d}(c,c')$.

It is easy to verify that $(\mathcal{X}_i, \mathsf{d}_i)$ is a metric space (see Exercise 42.7.2). Furthermore, we can naturally embed $(\mathcal{X}, \mathsf{d})$ into $(\mathcal{X}_i, \mathsf{d}_i)$ by mapping a point $x \in \mathcal{X}$ to its connected components in $\mathcal{X}_i$. Essentially $(\mathcal{X}_i, \mathsf{d}_i)$ is a snapped version of the metric $(\mathcal{X}, d)$, with the advantage that $\Phi((\mathcal{X}, \mathsf{d}_i)) = O(n^2)$. We now embed $\mathcal{X}_i$ into $\beta = O(\log n)$ coordinates. Next, for any point of $\mathcal{X}$ we embed it into those β coordinates, by using the embedding of its connected component in $\mathcal{X}_i$. Let E_i be the embedding for resolution r_i. Namely, $E_i(x) = (f_{i,1}(x), f_{i,2}(x), \ldots, f_{i,\beta}(x))$, where $f_{i,j}(x) = \min(\mathsf{d}_i(x, \mathcal{X} \setminus V_{i,j}), 2r_i)$. The resulting embedding is $F(x) = \oplus E_i(x) = (E_1(x), E_2(x), \ldots,)$.

Since we slightly modified the definition of $f_{i,j}(\cdot)$, we have to show that $f_{i,j}(\cdot)$ is nonexpansive. Indeed, consider two points $x, y \in \mathcal{X}_i$, and observe that

$$\left|f_{i,j}(x) - f_{i,j}(y)\right| \leq \left|\mathsf{d}_i(x, V_{i,j}) - \mathsf{d}_i(y, V_{i,j})\right| \leq \mathsf{d}_i(x,y) \leq \mathsf{d}(x,y),$$

as a simple case analysis[42.1] shows.

For a pair $x, y \in \mathcal{X}$, and let $\phi = \mathsf{d}(x,y)$. To see that $F(\cdot)$ is the required embedding (up to scaling), observe that, by the same argumentation of Theorem 42.5.2, we have that with high probability

$$\|F(x) - F(y)\| \geq \phi \cdot \frac{\sqrt{\beta}}{256 \ln n}.$$

To get an upper bound on this distance, observe that for i such that

[42.1] Indeed, if $f_{i,j}(x) < \mathsf{d}_i(x, V_{i,j})$ and $f_{i,j}(y) < \mathsf{d}_i(x, V_{i,j})$ then $f_{i,j}(x) = 2r_i$ and $f_{i,j}(y) = 2r_i$, which implies the above inequality. If $f_{i,j}(x) = \mathsf{d}_i(x, V_{i,j})$ and $f_{i,j}(y) = \mathsf{d}_i(x, V_{i,j})$ then the inequality trivially holds. The other option is handled in a similar fashion.

$r_i > \phi n^2$, we have $E_i(x) = E_i(y)$. Thus,

$$\begin{aligned}
\|F(x) - F(y)\|^2 &= \sum_i \|E_i(x) - E_i(y)\|^2 \\
&= \sum_{i, r_i < \phi n^2} \|E_i(x) - E_i(y)\|^2 \\
&= \sum_{i, \phi/n^2 < r_i < \phi n^2} \|E_i(x) - E_i(y)\|^2 \\
&\quad + \sum_{i, r_i < \phi/n^2} \|E_i(x) - E_i(y)\|^2 \\
&= \beta\phi^2 \lg(n^4) + \sum_{i, r_i < \phi/n^2} (2r_i)^2 \beta \\
&\leq 4\beta\phi^2 \lg n + \frac{4\phi^2\beta}{n^4} \leq 5\beta\phi^2 \lg n.
\end{aligned}$$

Thus, $\|F(x) - F(y)\| \leq \phi\sqrt{5\beta \lg n}$. We conclude, that with high probability, $F(\cdot)$ is an embedding of $\mathcal{X}$ into Euclidean space with distortion $\left(\phi\sqrt{5\beta \lg n}\right) / \left(\phi \cdot \frac{\sqrt{\beta}}{256 \ln n}\right) = O(\log^{3/2} n)$.

We still have to handle the infinite number of coordinates problem. However, the above proof shows that we care about a resolution r_i (i.e., it contributes to the estimates in the above proof) only if there is a pair x and y such that $r_i/n^2 \leq \mathsf{d}(x, y) \leq r_i n^2$. Thus, for every pair of distances there are $O(\log n)$ relevant resolutions. Thus, there are at most $\eta = O(n^2 \beta \log n) = O(n^2 \log^2 n)$ relevant coordinates, and we can ignore all the other coordinates. Next, consider the affine subspace h that spans $F(P)$. Clearly, it is $n-1$ dimensional, and consider the projection $G : \mathbb{R}^\eta \to \mathbb{R}^{n-1}$ that projects a point to its closest point in h. Clearly, $G(F(\cdot))$ is an embedding with the same distortion for P, and the target space is of dimension $n - 1$.

Note, that all this process succeeds with high probability. If it fails, we try again. We conclude:

Theorem 42.5.3 (Low quality Bourgain theorem). *Given a n-point metric M, one can embed it into Euclidean space of dimension $n-1$, such that the distortion of the embedding is at most $O(\log^{3/2} n)$.*

Using the Johnson-Lindenstrauss lemma, the dimension can be further reduced to $O(\log n)$. Being more careful in the proof, it is possible to reduce the dimension to $O(\log n)$ directly.

42.6. Bibliographical notes

The partitions we use are due to Calinescu *et al.* [CKR04]. The idea of embedding into spanning trees is due to Alon *et al.* [AKPW95], which showed that one can get a probabilistic distortion of $2^{O(\sqrt{\log n \log\log n})}$. Yair Bartal realized that by allowing trees with additional vertices, one can get a considerably better result. In particular, he showed [Bar96] that probabilistic embedding into trees can be done with polylogarithmic average distortion. He later improved the distortion to $O(\log n \log\log n)$ in [Bar98]. Improving this result was an open question, culminating in the work of Fakcharoenphol *et al.* [FRT04] which achieve the optimal $O(\log n)$ distortion.

Our proof of Lemma 42.3.1 (which is originally from [FRT04]) is taken from [KLMN05]. The proof of Theorem 42.5.3 is by Gupta [Gup00].

A good exposition of metric spaces is available in Matoušek [Mat02].

Embedding into spanning trees. The above embeds the graph into a Steiner tree. A more useful representation, would be a random embedding into a spanning tree. Surprisingly, this can be done, as shown by Emek *et al.* [EEST08]. This was improved to $O(\log n \cdot \log\log n \cdot (\log\log\log n)^3)$[42.1] by Abraham *et al.* [ABN08a, ABN08b].

Alternative proof of the tree embedding result. Interestingly, if one does not care about the optimal distortion, one can get similar result (for embedding into probabilistic trees), by first embedding the metric into Euclidean space, then reduce the dimension by the Johnson-Lindenstrauss lemma, and finally, construct an HST by constructing a quadtree over the points. The "trick" is to randomly translate the quadtree. It is easy to verify that this yields $O(\log^4 n)$ distortion. See the survey by Indyk [Ind01] for more details. This random shifting of quadtrees is a powerful technique that was used in getting several result, and it is a crucial ingredient in Arora [Aro98] approximation algorithm for Euclidean TSP.

[42.1] Truely a polyglot of logs.

42.7. Exercises

Exercise 42.7.1 (Clustering for HST). Let $(\mathcal{X}, \mathsf{d})$ be a HST defined over n points, and let $k > 0$ be an integer. Provide an algorithm that computes the optimal k-median clustering of $\mathcal{X}$ in $O(k^2 n)$ time.

[Transform the HST into a tree where every node has only two children. Next, run a dynamic programming algorithm on this tree.]

Exercise 42.7.2 (Partition induced metric).

(a) Give a counter example to the following claim: Let $(\mathcal{X}, \mathsf{d})$ be a metric space, and let P be a partition of $\mathcal{X}$. Then, the pair (P, d') is a metric, where $\mathsf{d}'(C, C') = \mathsf{d}(C, C') = \min_{x \in C, y \in C'} \mathsf{d}(x, y)$ and $C, C' \in P$.

(b) Let $(\mathcal{X}, \mathsf{d})$ be a n-point metric space, and consider the set $U = \left\{ i \;\middle|\; 2^i \leq \mathsf{d}(x, y) \leq 2^{i+1}, \text{ for } x, y \in \mathcal{X} \right\}$. Prove that

$$|U| = O(n).$$

Namely, there are only n different resolutions that "matter" for a finite metric space.

Exercise 42.7.3 (Computing the diameter via embeddings).

(a) (h:1) Let ℓ be a line in the plane, and consider the embedding $f : \mathbb{R}^2 \to \ell$, which is the projection of the plane into ℓ. Prove that f is 1-Lipschitz, but it is not K-bi-Lipschitz for any constant K.

(b) (h:3) Prove that one can find a family of projections $\mathcal{F}$ of size $O(1/\sqrt{\varepsilon})$, such that for any two points $x, y \in \mathbb{R}^2$, for one of the projections $f \in \mathcal{F}$ we have $\mathsf{d}(f(x), f(y)) \geq (1 - \varepsilon)\mathsf{d}(x, y)$.

(c) (h:1) Given a set P of n in the plane, given a $O(n/\sqrt{\varepsilon})$ time algorithm that outputs two points $x, y \in P$, such that $\mathsf{d}(x, y) \geq (1 - \varepsilon)\mathrm{diam}(P)$, where $\mathrm{diam}(P) = \max_{z, w \in P} \mathsf{d}(z, w)$ is the diameter of P.

(d) (h:2) Given P, show how to extract, in $O(n)$ time, a set $Q \subseteq P$ of size $O(\varepsilon^{-2})$, such that $\mathrm{diam}(Q) \geq (1 - \varepsilon/2)\mathrm{diam}(P)$. (Hint: Construct a grid of appropriate resolution.)

In particular, give an $(1 - \varepsilon)$-approximation algorithm to the diameter of P that works in $O(n + \varepsilon^{-2.5})$ time. (There are slightly faster approximation algorithms known for approximating the diameter.)

Acknowledgments

The presentation in this write-up follows closely the insightful suggestions of Manor Mendel.

References

[ABN08a] I. Abraham, Y. Bartal, and O. Neiman. Nearly tight low stretch spanning trees. *Proc. 49th Annu. IEEE Sympos. Found. Comput. Sci.* (FOCS), 781–790, 2008.

[ABN08b] I. Abraham, Y. Bartal, and O. Neiman. Nearly tight low stretch spanning trees. *CoRR*, abs/0808.2017, 2008. arXiv: 0808.2017.

[AKPW95] N. Alon, R. M. Karp, D. Peleg, and D. West. A graph-theoretic game and its application to the k-server problem. *SIAM J. Comput.*, 24(1): 78–100, 1995.

[Aro98] S. Arora. Polynomial time approximation schemes for Euclidean TSP and other geometric problems. *J. Assoc. Comput. Mach.*, 45(5): 753–782, 1998.

[Bar96] Y. Bartal. Probabilistic approximations of metric space and its algorithmic application. *Proc. 37th Annu. IEEE Sympos. Found. Comput. Sci.* (FOCS), 183–193, 1996.

[Bar98] Y. Bartal. On approximating arbitrary metrics by tree metrics. *Proc. 30th Annu. ACM Sympos. Theory Comput.* (STOC), 161–168, 1998.

[CKR04] G. Călinescu, H. J. Karloff, and Y. Rabani. Approximation algorithms for the 0-extension problem. *SIAM J. Comput.*, 34(2): 358–372, 2004.

[EEST08] M. Elkin, Y. Emek, D. A. Spielman, and S. Teng. Lower-stretch spanning trees. *SIAM J. Comput.*, 38(2): 608–628, 2008.

[FRT04] J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. *J. Comput. Sys. Sci.*, 69(3): 485–497, 2004.

[Gup00] A. Gupta. *Embeddings of Finite Metrics*. PhD thesis. University of California, Berkeley, 2000.

[Ind01] P. Indyk. Algorithmic applications of low-distortion geometric embeddings. *Proc. 42nd Annu. IEEE Sympos. Found. Comput. Sci.* (FOCS), Tutorial. 10–31, 2001.

[KLMN05] R. Krauthgamer, J. R. Lee, M. Mendel, and A. Naor. Measured descent: a new embedding method for finite metric spaces. *Geom. funct. anal. (GAFA)*, 15(4): 839–858, 2005.

[Mat02] J. Matoušek. *Lectures on Discrete Geometry*. Vol. 212. Grad. Text in Math. Springer, 2002.