

CS265/CME309: Randomized Algorithms and Probabilistic Analysis

Lecture #7: Metric Embeddings

Gregory Valiant*, updated by Mary Wootters

August 18, 2020

1 Introduction

Definition 1. A metric space is a pair (X, d) , where X is a set, and $d : X \times X \rightarrow \mathbb{R}$ is the distance function (also referred to as the metric), which satisfies the following conditions:

- For all $x, y \in X$ $d(x, x) = 0$ and if $x \neq y$, $d(x, y) = d(y, x) > 0$.
- Triangle inequality: for all $x, y, z \in X$, $d(x, y) \leq d(x, z) + d(z, y)$.

The following examples illustrate some common metric spaces:

Example 2. Consider the ℓ_p metric space (\mathbb{R}^k, d_p) , with the distance function

$$d_p(x, y) = \left(\sum_{i=1}^k |x_i - y_i|^p \right)^{1/p}.$$

In the case that $p = 2$, this is Euclidean distance, in the case that $p = 1$ this is sometimes referred to as “Manhattan distance”. For $p < 1$, this is not a valid metric space because the triangle inequality no longer holds for $k \geq 2$, as can be seen by considering the points $x = (0, 0)$, $y = (1, 1)$, $z = (0, 1)$.

Example 3. Given a graph on n points, with positive edge weights, the corresponding graph metric (V, d) is defined by the vertex set V , with distance function $d(x, y)$ being the length of the shortest-path between x and y in the graph. Note that any finite metric (any metric (X, d) where X is a set of finite cardinality) can be represented as the graph metric corresponding to the complete graph over vertex set X , with the edge weights defined to be the corresponding distances.

*©2019, Gregory Valiant. Not to be sold, published, or distributed without the authors’ consent.

2 Low-distortion embeddings

ℓ_p metric spaces, especially ℓ_1, ℓ_2 and ℓ_∞ are especially nice to deal with, largely because there are many algorithms which leverage the geometry of these spaces to efficiently solve problems. These metrics also allow one to easily visualize the relationships between the points. This prompts the fundamental question: to what extent can one *embed* one metric (X, d) into another (usually “nicer” metric) (Y, d') in such a way so that most of the distances are essentially preserved?

We might first hope for *exact* preservation of distances: that is, given metric spaces (X, d_X) and (Y, d_Y) , we might like a map $f : X \rightarrow Y$ so that $d_X(x, x') = d_Y(f(x), f(x'))$ for all $x, x' \in X$. Such a map is called an *isometric embedding*. Sometimes these exist! For example, you can check that, given an n point metric space $X = \{x_1, \dots, x_n\}$ with metric d , the map $f(x) = (d(x, x_1), d(x, x_2), \dots, d(x, x_n))$ is an isometric embedding of (X, d) into (\mathbb{R}^n, d_∞) .

However, isometric embeddings don't always exist. For example, consider the metric space on four points $\{x, y, z, w\}$ so that $d(x, y) = d(y, z) = d(z, w) = d(w, x) = 1$, and $d(x, z) = d(y, w) = 2$. That is, this is the graph metric space where the underlying graph is a square. I bet you can't embed this in ℓ_2 ! (Seriously, try it).

This raises the question of whether we can *approximately* preserve distances in an embedding. This question can be formalized via the notion of the *distortion* of an embedding:

Definition 4. *Given two metric spaces (X, d) and (Y, d') , and some function $f : X \rightarrow Y$, we say that f is an embedding of (X, d) into (Y, d') with distortion $\alpha \geq 1$ if there is a scaling factor β such that for all $x, y \in X$, it holds that $\beta \cdot d(x, y) \leq d'(f(x), f(y)) \leq \alpha\beta \cdot d(x, y)$. And we say that (X, d) can be embedded into (Y, d') with distortion α if such a function f exists.*

3 Bourgain's Embedding

Perhaps the most important result in the area of metric embeddings is the following theorem, proved by Bourgain in 1985 [?] (actually, he proved a slightly weaker result, giving an embedding into a higher-dimensional space, the result below was made explicit by Linial, London, and Rabinovich in 1995 [?]):

Theorem 1. *Given any finite metric (X, d) with $|X| = n$, there exists an embedding of (X, d) into \mathbb{R}^k under the ℓ_1 distance metric (actually, any ℓ_p metric), where $k = O(\log^2 n)$, and the distortion of the embedding is $O(\log n)$.*

Remark 5. *You might wonder if the parameters in the theorem above are the best possible. It turns out that the distortion of $O(\log n)$ is the best possible, while the target dimension can be improved to $O(\log n)$. See [1] for such an improvement.*

The proof of this theorem is constructive: we will give a randomized construction of the embedding, and prove that with high probability over the random choices of the construction, the resulting embedding will have the desired distortion. We begin by defining the randomized embedding, and then provide some intuition for this choice of embedding, and then prove the theorem.

The embedding will be defined in terms of $k = c \log^2 n$ random sets, $S_{i,j} \subset X$ for $i \in \{1, 2, \dots, \log n\}$ and $j \in \{1, \dots, c \log n\}$ for some constant c that we will choose later. The set

$S_{i,j}$ is chosen by including each point in X independently with probability $\frac{1}{2^i}$. Given the sets, the mapping $f : X \rightarrow \mathbb{R}^k$ is defined by

$$f(x) = (d(x, S_{1,1}), d(x, S_{1,2}), \dots, d(x, S_{\log n, c \log n})),$$

where $d(x, S) = \min_{y \in S} d(x, y)$.

The intuition for this embedding is as follows: suppose we could pick a subset $S \subset X$ such that for every pair of points x, y we could guarantee that there is some point $z \in S$ such that $d(x, z) \leq d(x, y)/4$, but no point in S has distance less than $d(x, y)/2$ from y . In that case, we could simply define the embedding into \mathbb{R} by $f(x) = d(x, S)$, and such an embedding would have constant distortion, and we would be done. Obviously no single set S can have this property; the intuition for the above construction is that if we randomly pick a number of sets, $S_{i,j}$, with the sets having a large range of different sizes, then we can argue that, with very high probability, for every pair of points x, y at least one of these sets will have the desired property with respect to the pair x, y . To ensure that there is an appropriate sized set for each pair x, y , we need a logarithmic number of different sizes, i.e. sizes roughly $2, 4, 8, \dots, n$. To ensure that each pair is satisfied with sufficiently high probability so as to allow us to perform a union bound over the n^2 pairs, we will need to have $O(\log n)$ redundancy for each size. Thus the $O(\log^2 n)$ sets, and the geometrically decreasing expected sizes of the sets. Because we can only expect a small number of the different sized sets to be relevant for a given pair x, y , we will lose a logarithmic factor in the distortion. The following proof formalizes this intuition.

Proof of Theorem 1. We begin by showing that for all x, y $\|f(x) - f(y)\|_1 \leq k \cdot d(x, y)$. For the coordinate corresponding to set $S_{i,j}$, assume without loss of generality that $d(x, S_{i,j}) \geq d(y, S_{i,j})$. Let $z \in S_{i,j}$ satisfy $d(y, S_{i,j}) = d(y, z)$, and note that $|d(x, S_{i,j}) - d(y, S_{i,j})| = d(x, S_{i,j}) - d(y, z) \leq d(x, z) - d(y, z) \leq d(x, y)$, where the last inequality holds by the triangle inequality. Hence the contribution to the ℓ_1 distance from each coordinate is at most $d(x, y)$, yielding the desired claim.

We now consider the meat of the proof—proving that $\|f(x) - f(y)\|_1 \geq \frac{k}{b \log n} d(x, y)$, for some constant b , with high probability over the randomness of the construction of the sets $S_{i,j}$. The proof will proceed by showing that with high probability, for a given pair x, y , the claim holds. We will then apply a union bound over all $O(n^2)$ such pairs.

Given a pair $x, y \in X$, define a set of distances $0 = \delta_0 < \delta_1 < \dots < \delta_t$. Let $B(x, \delta) = \{z : d(x, z) \leq \delta\}$ denote the closed ball of radius δ centered at x , $B^-(x, \delta) = \{z : d(x, z) < \delta\}$ denote the open ball of radius δ centered at x , and define δ_i to be the smallest δ such that both $|B(x, \delta)| \geq 2^i$ and $|B(y, \delta)| \geq 2^i$. Let t be the maximum i for which $\delta_i < d(x, y)/3$, and let $\delta_{t+1} = d(x, y)/3$.

First observe that for $i, j \leq t+1$, $B(x, \delta_i) \cap B(y, \delta_j) = \emptyset$. Suppose that some set $S \subset X$ has the property that $S \cap B(x, \delta_i) \neq \emptyset$, but $S \cap B^-(y, \delta_{i+1}) = \emptyset$. Then $d(x, S) \leq \delta_i$, and $d(y, S) \geq \delta_{i+1}$, and hence $|d(x, S) - d(y, S)| \geq \delta_{i+1} - \delta_i$. We now show that for each of our random sets $S_{i,j}$, there is a constant chance that it will have this property for δ_{i+1} , and hence there will be a significant contribution to the ℓ_1 distance between $f(x)$ and $f(y)$ from the corresponding coordinates.

Note that from the definition of δ_{i+1} , it must be the case that either $|B^-(y, \delta_{i+1})| < 2^{i+1}$, or $|B^-(x, \delta_{i+1})| < 2^{i+1}$, otherwise we would have picked a smaller value for δ_{i+1} . In the following, assume that $|B^-(y, \delta_{i+1})| < 2^{i+1}$; a symmetric argument would apply in the case that $|B^-(x, \delta_{i+1})| < 2^{i+1}$. Leveraging the fact that the event that $S_{i,j}$ intersects $B(x, \delta_i)$ is independent of the event that $S_{i,j}$ intersects $B^-(y, \delta_{i+1})$ —since these balls are disjoint, and the events that each element of either ball is in the set is independent with probability $1/2^i$ —we have the probability that

$S_{i,j} \cap B(x, \delta_i) \neq \emptyset$ and $S_{i,j} \cap B^-(y, \delta_{i+1}) = \emptyset$ is the product of the probabilities of each of these separate events. Because there are at least 2^i points in $B(x, \delta_i)$ the probability that at least one of them is in $S_{i,j}$ is at least $1 - (1 - 1/2^i)^{2^i} \geq 1 - 1/e > 0.6$. On the other hand, by assumption, there are at most 2^{i+1} points in $B^-(y, \delta_{i+1})$, and the probability that they all avoid being in set $S_{i,j}$ is at least $(1 - 1/2^i)^{2^{i+1}} \geq 1/2^4$. Hence for each i , the probability that $S_{i,j}$ contributes at least $\delta_{i+1} - \delta_i$ to the ℓ_1 distance between $f(x)$ and $f(y)$ is bounded below by a constant $> 1/2^5$.

To finish the proof, we use a Chernoff bound to show that for a given i , the probability that fewer than a $1/2^6$ fraction of the $c \log n$ sets $S_{i,j}$ contribute at least $\delta_{i+1} - \delta_i$ to the distance is at most $e^{-\frac{c \log n}{2^8}}$. (Here we used the fact that the expected number of such sets is $c \log n / 2^5$, and each of these events is independent, and the probability that such a sum of independent indicators does not exceed half its expectation is bounded by $e^{-\mu/8}$.) Setting $c = 3 \cdot 2^8$, this probability is at most $1/n^3$, and hence via a union bound over the $\log n$ i 's and the $O(n^2)$ pairs x, y , it holds that with high probability, for every pair x, y ,

$$\|f(x) - f(y)\|_1 = \sum_{i,j} |d(x, S_{i,j}) - d(y, S_{i,j})| \geq \frac{c \log n}{2^6} \sum_{i=0}^t (\delta_{i+1} - \delta_i) \geq \frac{k}{2^6 \log n} d(x, y)/3,$$

where the last inequality followed from the fact that this sum telescopes and $\delta_{t+1} = d(x, y)/3$, and $\delta_0 = 0$.

This embedding also works for other ℓ_p distance metrics aside from just ℓ_1 , (in fact, Bourgain originally showed it for ℓ_2), and we might see this in class and/or on a problem set. \square

References

- [1] Ittai Abraham, Yair Bartal, and Ofer Neiman. Advances in metric embedding theory. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 271–286, 2006.