15-859(J): Metric Embeddings and Methods **Lecturer:** Anupam Gupta
Topic: Introduction to Metrics **Date:** Sept 2, 2003
Scribe: Anupam Gupta

1 Distances and Metric Spaces

Given a set X of *points*, a *distance function* on X is a map $d : X \times X \to \mathbb{R}_+$ that is symmetric, and satisfies $d(i,i) = 0$ for all $i \in X$. The distance is said to be a *metric* if the *triangle inequality* holds, i.e.,

$$d(i,j) \leq d(i,k) + d(k,j) \qquad \forall i,j,k \in X.$$

It is often required that metrics should also satisfy the additional constraint that $d(i,j) = 0 \iff i = j$, and the distance is called a *pseudo-metric* or a *semi-metric* if this additional condition does not hold. We will be very lax with this distinction; however, you now know what these terms mean, just in case.

1.1 Finite Metrics and Graphs

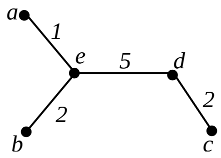
In this class, we shall usually study *finite metric spaces*, i.e., spaces (X,d) where X is a finite set of points; we will use n to denote the size of X. As described above, a metric can be specified by $\binom{n}{2}$ non-negative numbers, which give the distance between unordered pairs $\{i,j\} \in \binom{X}{2}$. E.g., the following matrix completely specifies a metric:

	a	b	c	d	e
a	0	3	8	6	1
b		0	9	7	2
c			0	2	7
d				0	5
e					0

It is often difficult to visualize metrics when specified thus, and hence we will use a natural correspondence between graphs and metrics. Given a graph G on n vertices endowed with lengths on the edges, one can get a natural metric d_G by setting, for every $i,j \in V(G)$, the distance $d_G(i,j)$ to be the length of the *shortest-path* between vertices i and j in G.

Conversely, given a metric (X,d), we can obtain a weighted graph $G(d)$ *representing* or *generating* the metric thus: we set X to be the vertices of the graph, add edges between all pairs of vertices, and set the length of the edge $\{i,j\}$ to be $d(i,j)$. It is trivial to see that the shortest path metric $d_{G(d)}$ is identical to the original metric d.

In fact, we may be able to get a "simpler" graph representing d by dropping any edge e such that the resulting metric $d_{G(d)-e}$ still equals d. We can continue removing edges until no more edges can be removed without changing the distances; the resulting minimal graph will be called the *critical graph* for the metric d. It can be checked that the critical graph for the metric above is



We will usually blur the distinction between finite metrics and graphs, unless there are good reasons not to do so.

1.2 Infinite Metric spaces

The most commonly used infinite metric spaces are obtained by equipping real space $\mathbb{R}^k$ with one of the so-called Minkowski norms ℓ_p for any $1 \leq p \leq \infty$. Given a point $x \in \mathbb{R}^k$, its ℓ_p length is given by

$$\|x\|_p = (\sum_{i=1}^{k} |x_i|^p)^{1/p} \qquad\qquad \text{for } 1 \leq p < \infty;$$

the special case of the ℓ_∞ norm specified by $\|x\|_\infty = \max_{i=1}^k \{|x_i|\}$. (All this can be extended to non-finite dimensions, but we will not encounter such situations in this class.) Given two points $x, y \in \mathbb{R}^k$, the ℓ_p-distance between them is naturally given by $\|x - y\|_p$.

Some of these spaces should be familiar to us: ℓ_2 is just Euclidean space, while ℓ_1 is real space endowed with the so-called Manhattan metric. It is often instructive to view the *unit balls* in these ℓ_p metrics; here are the balls for $p = 1, 2, \infty$:

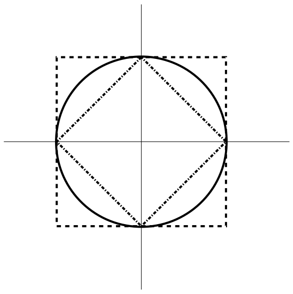


Figure 1.1: *Unit balls in ℓ_1, ℓ_2 and ℓ_∞.*

We call a finite metric (X, d) an ℓ_p-metric if there is a mapping $f : X \to \mathbb{R}^k$ (for some k) such that $\|f(i) - f(j)\|_p = d(x, y)$; by a cavalier abuse of notation, we will often denote this by $d \in \ell_p$. To denote the fact that the ℓ_p space had k dimensions, we will call the space ℓ_p^k as well.

Let us introduce two terms we will use often. Given metric spaces (X, d) and (X', d'), a map $f : X \to X'$ will be called an *embedding*, and a distance-preserving embedding will be called *isometric*. Before we continue, note that embeddings are a generic term for any map from a metric into another; transforming the metric (X, d) into its graphical representation $G(d)$ also gave us an isometric embedding.

1.3 Why embeddings?

While we may want to investigate the properties of metric spaces for their rich mathematical content, the study is rewarding from an algorithmic viewpoint as well. Here are two simple examples.

Suppose we wanted to solve the *traveling salesman problem (TSP)* for the 5-point metric above. (This is clearly a toy example!) We can run over all the possible tours, and output the cheapest; or we can use the embedding of the metric into the tree, and the fact that TSP can be solved optimally on trees to get a solution.

1.3.1 Furthest points in ℓ_1

A more substantial example is the following: suppose we are given a set X of n points in ℓ_1^k (with $n \gg 2^k$), and we want to figure out the *furthest pair* of points in this set. Clearly, a brute-force algorithm can do this in $O(kn^2)$ time; we now show how to do this in much less time using an isometric embedding from ℓ_1^k to $\ell_\infty^{2^k}$, where it is much easier to solve the furthest pair problem.

The Embedding: Consider $x \in \ell_1^k$, and let $\mathbb{B}_1$ be the unit ball in ℓ_1^k: then $\|x\|_1 = \min\{\lambda : x \in \lambda\,\mathbb{B}_1\}$. The normals to the faces of this unit ball are the vectors $\{-1, 1\}^k$, and if we were to project the point x down onto the $k' = 2^k$ vectors $\{-1, +1\}^k$, and the length $\|x\|_1$ would be the *largest* of these projections. I.e.,

$$\|x\|_1 = \max\{\, \langle x, y\rangle \mid y \in \{-1, +1\}^k \,\}. \tag{1.1}$$

This immediately suggests the following mapping: we create a coordinate $f_y : X \to \mathbb{R}$ for each vector $y \in \{-1, 1\}^k$, with $f_y(z) = \langle z, y\rangle$. The final map f is simply obtained by concatenating these coordinates together, i.e., $f = \oplus f_y$. That this is an isometric embedding follows the linearity of the map, which implies that $f(z) - f(z') = f(z - z')$, and (1.1) above. This transformation can be done in $O(n\,k\,2^k)$ time.

Solving the transformed problem: We claim that the furthest pair of points in $\ell_\infty^{k'}$ must be the furthest pair of points when projected down onto one of the k' dimensions. Indeed, the distance between the furthest pair in the set S is

$$\begin{aligned}
\max_{x,y \in S} \|x - y\|_\infty &= \max_{x,y \in S} \max_{i=1}^{k'} |x_i - y_i| \\
&= \max_{i=1}^{k'} \max_{x,y \in S} |x_i - y_i| \\
&= \max_{i=1}^{k'} \text{ furthest pair along the } i\text{-th coordinate .}
\end{aligned}$$

However, the problem of finding the furthest pair along any coordinate can be solved by finding the largest and the smallest value in that coordinate, and hence takes $2n$ time. Doing this for all the k' coordinates takes $O(k'n) = O(n\,2^k)$ time.

1.4 The Universal Space: ℓ_∞

We just saw that the metric space ℓ_1^k isometrically embeds into $\ell_\infty^{2^k}$. In fact, a stronger result can be shown: *any* metric space can be embedded isometrically into ℓ_∞. We will show the finite version of this:

Theorem 1.1 (Fréchet) *Any n point metric space (X, d) can be embedded into ℓ_∞.*

Proof: For each point i in X, let us define a coordinate $\phi_i : X \to \mathbb{R}_+$ thus: $f_i(j) = d(i, j)$. We claim this is an isometric embedding of d into ℓ_∞. Indeed,

$$\begin{aligned} \|f(x) - f(y)\|_\infty &= \max_{i \in V} |f_i(x) - f_i(y)| \\ &= \max_{i \in V} |d(i, x) - d(i, y)| \qquad (1.2) \\ &\leq d(x, y). \end{aligned}$$

However, the value of the expression (1.2) is also *at least* $d(x, y)$ (obtained by setting $i = y$), and hence the embedding is isometric. ■

Exercise 1.2 *Show that the number of coordinates can be reduced to $n-1$. Show that, for any tree metric, the number can be reduced to $O(\log n)$.*

In contrast, other metric spaces, say ℓ_2, are not universal in this sense. E.g., consider the metric generated by the 3-star $K_{1,3}$. A simple argument shows that this cannot be embedded into ℓ_2 isometrically.

2 Distorted Embeddings

It is very rare to find cases where isometric embeddings can be found between two spaces of interest, and hence we often have to allow the mappings to alter distances in some (hopefully restricted) fashion. Even this may be useful: suppose we were given a metric which is "close" to being a tree metric, then we would hope that the optimal TSP tour on the tree would have a length close to that of an optimal TSP tour on the original metric space.

There are many notions of "close" (and for the TSP example, most of these definitions of close would suffice); most of the course will focus on the following notions. Given two metrics (X, d) and (X', d') and a map $f : X \to X'$, the *contraction* of f is the maximum factor by which distances are shrunk, i.e.,

$$\max_{x,y \in X} \frac{d(x, y)}{d'(f(x), f(y))},$$

the *expansion* or *stretch* of f is the maximum factor by which distances are stretched:

$$\max_{x,y \in X} \frac{d'(f(x), f(y))}{d(x, y)},$$

and the *distortion* of f, denoted by $\|f\|_{\text{dist}}$, is the product of the distortion and the expansion. An isometric embedding has distortion 1, and this is the lowest it can get. (Right?)

Another equivalent definition is the following: the distortion of f, denoted by $\|f\|_{\text{dist}}$, is the smallest value $\alpha \geq 1$ for which there exists an $r > 0$ such that for all $x, y \in X$,

$$r\, d(x, y) \leq d'(f(x), f(y)) \leq \alpha\, r\, d(x, y).$$

(Check.) A very useful property of distortion is that it is invariant under scaling; i.e., replacing d and d' by, say $10\, d$ and $\frac{1}{3}\, d'$ does not change $\|f\|_{\text{dist}}$. Hence, in many arguments, we will feel free to rescale embeddings to make arguments easier.

To continue the toy TSP example, if we could embed the given metric into a tree with distortion α, then the optimal solution to the TSP on this tree would be within α of the optimal TSP tour on the input metric; hence if $\alpha < 1.5$, we would get an improved approximation algorithm. We will see more compelling algorithmic applications of low distortion embeddings later in the course.

3 The main results

The main results in low-distortion embeddings can be classified into three main groups based on the basic techniques involved in obtaining them.

3.1 Fréchet-type embeddings

These results embed a metric (X, d) into ℓ_p spaces, and fall into the following general framework. Let us define a mapping $f_S : X \to \mathbb{R}$ for each subset of the vertices $S \subseteq X$ thus:

$$f_S(x) = d(x, S) = \min_{s \in S} d(x, s).$$

A *Fréchet-type embedding* is a map $f : X \to \mathbb{R}^k$, each of whose coordinates is a scaled one of these mappings above. Formally, we will denote this as:

$$f(x) = \bigoplus_{S \subseteq X} \beta_S \, f_S,$$

for $\beta_S \in \mathbb{R}$. Indeed, to get Fréchet's mapping of metrics into ℓ_∞ from Section 3.1, we can set $\beta_S = 1 \iff |S| = 1$. The main result in this area was proved by Bourgain (1985), who showed the following theorem:

Theorem 3.1 (Bourgain (1985)) *Given a metric (X, d), the map f obtained by setting*

$$\beta_S = \frac{1}{|S| \binom{n}{|S|}}$$

is an $O(\log n)$-distortion embedding of d into ℓ_1.

This result is tight for ℓ_1, as was shown by Linial, London and Rabinovich (1995). They extended a theorem of Leighton and Rao (1989) to show that constant-degree expanders require a distortion of $\Omega(\log n)$ to embed into ℓ_1.

In 1995, Matoušek sharpened Theorem 3.1 to show that a suitable choice of β_S allows us to embed any n-point metric into ℓ_p with distortion $O((\log n)/p)$. Furthermore, he showed a matching lower bound of $\Omega((\log n)/p)$ on the distortion into ℓ_p, using the same constant-degree expanders.

Theorem 3.2 *Given a metric and $p \in [1, \infty)$, one can embed d into ℓ_p^k with $O((\log n)/p)$ distortion.*

The embeddings can be modified by random sampling ideas to use a small number of dimensions:

Theorem 3.3 *Given a metric and $p \in [1, \infty)$, one can embed d into ℓ_p^k with $k = O(\log^2 n)$ dimensions and $O(\log n)$ distortion.*

3.2 Flattening Lemmas

Perhaps the most surprising result is the last one, which reduces the dimensions required for representing n-point Euclidean point sets, if inter-point distances are allowed to be slightly distorted. The theorem is exceedingly simple to state:

Theorem 3.4 (Johnson and Lindenstrauss, 1984) *Given $X \subseteq \mathbb{R}^k$, there is an embedding $f : X \to \mathbb{R}^{O(\log n/\epsilon^2)}$ with distortion $(1+\epsilon)$.*

In fact, the proof of this result also shows that such a map can be obtained (with high probability) by just projecting the point set on a *random* $k' = O(\log n/\epsilon^2)$-dimensional subspace, and then scaling up the resulting distances by $\sqrt{k/k'}$. (This should be contrasted with the observation that we require $n-1$ dimensions to embed the n-point uniform metric into ℓ_2.)

We will see some (fairly simple) proofs of this theorem later in the course, as well as many applications of this powerful theorem; let us allude to two of the applications here:

3.2.1 Near-Neighbor Searching

Given a set X of n points in ℓ_2^k, one wants to preprocess these points and create a data structure capable of answering queries of the form: given a query point $q \in \ell_2^k$, return a point $p \in X$ that is "close" to q. Ideally, we would like to obtain a point p^* that is *closest* to q, but since that may require a large cost in both storage space and query time, often it is sufficient to return a point that is a $(1+\epsilon)$-near neighbor. Formally, if the closest point is p^*, we demand that the algorithm return a point p such that $d(p,q) \leq (1+\epsilon)\, d(p^*,q)$. The flattening lemma allows us to replace the term that depends on k, the number of dimensions of the point set, by $O(\log n/\epsilon^2)$. More details later in the course.

3.2.2 Learning Mixtures of Gaussians

Suppose we are given access to a random source which outputs a point in $\mathbb{R}^k$ thus: it has t Gaussians, with the Gaussian i having center at the point $\mu_i \in \mathbb{R}^k$, a covariance matrix Σ_i, and a *weight* w_i. The random source picks one of the t Gaussians with probability proportional to the weight w_i, and then outputs a point from the distribution $N^k(\mu_i, \Sigma_i)$. Our objective is to learn the parameters μ_i, Σ_i, w_i. In the same vein as the idea above, one could imagine solving the problem in low dimensions and calculating the number of sample points s required, and then arguing that only $O(\log s)$ dimensions would be needed.

However, it turns out that $O(\log t)$ dimensions suffice, where t is the number of Gaussians: this number of dimensions ensures that, if the centers of the Gaussians were "well-separated" to begin with, a random projection would maintain the distances between them. There is a worry that the other points from the Gaussian would become scattered all over, but one can use another fact about random projections that projecting points down causes distances to contract more readily than they expand, and the Gaussians tend to stay concentrated around their centers. Details in a later class.

3.3 Random Tree Embeddings

Since many algorithmic problems are simple on trees, one would want to embed a given metric into a tree with low distortion. However, this is not possible, and one can show that the n-cycle requires distortion $\Omega(n)$ to embed into any tree. However, much better results can be obtained if one wants to embed the given metric (X, d) into a *distribution of trees*, instead of one tree, so that the *expected distances* are preserved.

Formally, we say that the metric (X, d) α-probabilistically embeds into a distribution $\mathcal{D}$ of trees if

- Each tree $T = (V_T, E_T)$ in the support of the distribution $\mathcal{D}$ contains the points of the metric; i.e., $X \subseteq V_T$. Furthermore, the distances in T *dominate* those in d; i.e., $d_T(x, y) \geq d(x, y)$ for all $x, y \in X$.

- Given a pair of vertices $x, y \in X$, the expected distance is not too much larger than $d(x, y)$, i.e.,

$$\mathbf{E}[\, d_T(x, y)\,] \leq \alpha\, d(x, y).$$

The added power of randomization can be seen by taking the n-cycle with unit length edges: the distribution we want is obtained by picking one of the edges uniformly at random and deleting it. It can be verified that this gives us a $2(1 - \frac{1}{n})$-probabilistic embedding of the cycle into its subtrees.

This concept has been widely studied in many papers, and culminated in the following result due to Fakcharoenphol, Rao and Talwar (2003):

Theorem 3.5 *Any n-point metric $O(\log n)$-probabilistically embeds into a distribution of trees; furthermore, samples from this distribution can be generated in polynomial time.*

Embeddings into random trees (as we shall refer to them) have enjoyed very wide applicability, and we will see examples later in the course. Essentially, any intractable problem on a metric space (that has a cost function linear in the distances) can be now solved on a tree instead, and Theorem 3.5 guarantees that we lose only $O(\log n)$ in the cost (in expectation).

References

Bourgain, J. (1985). On Lipschitz embeddings of finite metric spaces in Hilbert space. *Israel Journal of Mathematics*, 52(1-2):46–52.

Dasgupta, S. (1999). Learning mixtures of gaussians. In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science*, pages 634–644.

Fakcharoenphol, J., Rao, S., and Talwar, K. (2003c). A tight bound on approximating arbitrary metrics by tree metrics. In *Proceedings of the thirty-fifth ACM symposium on Theory of computing*, pages 448–455. ACM Press.

Linial, N., London, E., and Rabinovich, Y. (1995). The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245. (Preliminary version in *35th FOCS*, 1994).

Leighton, F. T. and Rao, S. B. (1999). Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM*, 46(6):787–832. (Preliminary version in *29th FOCS*, pages 422–431, 1988).

Matoušek, J. (1997). On embedding expanders into l_p spaces. *Israel J. Math.*, 102:189–197.

Matoušek, J. (2002). *Lectures on discrete geometry*, volume 212 of *Graduate Texts in Mathematics*. Springer-Verlag, New York.