**Blockcourse on High-Dimensional Computational Geometry** Spring 2013

# Lecture 9: Embeddings in Lower Dimensional Spaces

*Helmut Alt* *Scribe: Paul Seiferth & Michael Dobbins*

## 1 Background

### 1.1 Motivation

Nearest neighbour algorithms can be used to identify objects by relating a new object to a collection of known reference objects. For example, given an image of a letter, identify the letter. Or, given a proposed trademark symbol, check if there are established trademark symbols with similar appearance.

To do this, we first define a function that associates a feature vector to any applicable object, then use a nearest neighbour algorithm (or some approximation version) to find a reference object that is close to the new object in an appropriate metric. If the feature vector has very high dimension $d$, the known nearest neighbour algorithms may not perform well, i.e. the running time depends exponentially on the $d$. A way to deal with the "curse of dimensionality" is to first project the feature vectors into a lower dimensional space with low distortion and then apply nearest neighbour search in this subspace.

There is an extensive theory of low distortion embeddings going back to the 1930s, and due to its applications it has seen renewed interest over the last 15 years, especially with a lot of results by Motwani, Indyk, and Matoušek.

### 1.2 Metric Spaces

**Definition 1.** *A metric space $(X, \rho)$ is a set $X$ with a real non-negative function $\rho : X \times X \to [0, \infty)$ such that for all $x, y, z \in X$,*

- $\rho(x, y) = 0$ *if and only if* $x = y$
- $\rho(x, y) = \rho(y, x)$
- $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$

**Example 1:** $\ell_p^d$, the $\ell_p$-metric on $\mathbb{R}^d$ for $p \in \mathbb{N} \cup \infty$

This is defined by the $\ell_p$-norm, $\ell_p(x, y) = \|x - y\|_p$ where for $p < \infty$

$$\|x\|_p = \left( \sum_{i=1}^{d} x_i^p \right)^{1/p}$$

and $\|x\|_\infty = \max_i x_i$. Note that $\lim_{p\to\infty} \|x\|_p = \|x\|_\infty$ when the limit on the left side exists, and $\|\cdot\|_2$ is simply denoted by $\|\cdot\|$.

More generally, every normed vector space is a metric space.

**Definition 2.** *A norm $\nu$ for a vector space $X$ is a real non-negative function $\nu : X \to [0, \infty)$ such that*

- $\nu(x) = 0$ *if and only if* $x = 0$
- $\nu(sx) = s\nu(x)$
- $\nu(x + y) \leq \nu(x) + \nu(y)$

*for all $x, y \in X$ and $s \in \mathbb{R}$.*

**Example 2:** The discrete metric

Let $X$ can be an arbitrary set.The discrete distance between $x, y \in X$ is

$$\delta(x, y) = \begin{cases} 0 & x = y \\ 1 & x \neq y \end{cases}$$

**Example 3:** Hausdorff distance

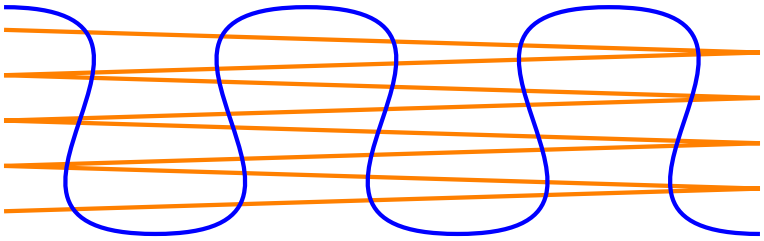
Hausdorff distance measures similarity of closed and bounded subsets of a given metric space $(X, \rho)$. Let $\mathcal{H}$ be the set of all closed and bounded subsets of $X$. For $A, B \in \mathcal{H}$ the Hausdorff distance with respect to $\rho$ is

$$\rho_H(A, B) = \max(\rho'_H(A, B), \rho'_H(B, A)) \quad \text{where}$$

$$\rho'_H(A, B) = \max_{a \in A} \min_{b \in B} \rho(a, b)$$

For $\rho = \ell_p$ and $X = \mathbb{R}^d$ the set $\mathcal{H}$ consists of all sets that are closed and bounded in the sense of Euclidean space. For more general metrics, the set $\mathcal{H}$ consists of sets that are closed and bounded in the topology defined by the metric. A set is open in the topology defined by a metric when it is a union of open metric balls (possibly uncountably many balls).

Hausdorff distance can be problematic. Consider the following two sets:



These two curves have low Hausdorff distance, but they do not look very similar according to a common intuitive notion of similarity. This motivates the following example of a metric.

**Example 4:** "Dog leash" Fréchet distance

Consider two distinct curves. The Fréchet distance measures the shortest leash that you would need if you and your dog each walk along one of the curves without going backwards. This metric dates back to 1906. Formally, let $\mathfrak{P}$ be the space of monotone reparameterizations of $[0, 1]$. That is, $\sigma \in \mathfrak{P}$ is a continuous, surjective, and non-decreasing function $\sigma : [0, 1] \to [0, 1]$. For two curves $f, g : [0, 1] \to \mathbb{R}^d$, their Fréchet distance with respect to $\rho$ is

$$\rho_F(f, g) = \inf_{\sigma, \tau \in \mathfrak{P}} \max_{t \in [0,1]} \rho(f \circ \sigma(t), f \circ \tau(t))$$

**Example 5:** Hamming distance

Hamming distance counts the letter that differ between strings of the same length $d$. For some alphabet $\Sigma$, let $X = \Sigma^d$. For $x, y \in X$, the Hamming distance is

$$\rho(x, y) = |\{i : x_i \neq y_i\}|$$

This is metric is used in error correcting codes.

**Example 5a:** Edit distance

Edit distance counts the number of edits needed to transform one string into another. For some alphabet $\Sigma$, let $X = \Sigma^*$ be the set of strings of arbitrary finite length. For $x, y \in X$, their edit distance is the minimum number of deletions, insertions, and replacements of characters needed to change $x$ into $y$. This metric is used to study mutations in DNA.

**Example 6:** $\ell_p$-space

This is the infinite dimensional version of the $\ell_p$-metric on $\mathbb{R}^d$. Here $X$ consist of infinite sequences such that the $\ell_p$-norm converges. That is for $x : \mathbb{N} \to \mathbb{R}$, $x \in X$ we have

$$\|x\|_p = \left( \sum_{i \in \mathbb{N}} x_i^p \right)^{1/p} < \infty$$

### 1.3 Distortion

A map between to metric spaces that preserves distances is called an *isometric embedding*. It is not always possible to isometrically embed a point set coming from one metric space into another one. For example, the vertices of regular $d$-simplex cannot be isometrically embedded in $\mathbb{R}^{d-1}$. Any connected graph is a metric space with distance between vertices given by the length of the shortest path between them. The *claw* is the graph with vertices $\{0, 1, 2, 3\}$ and edges $\{(0, 1), (0, 2), (0, 3)\}$, and it cannot be embedded in $\mathbb{R}^d$ for any $d$. Any metric space $(X, \rho)$ has an isometric embedding in $\ell_p$ if and only if there is some $d$ such that $(X, \rho)$ has an isometric embedding in $\ell_p^d$.

Let $(X, \rho)$ and $(Y, \sigma)$ be metric spaces and $D \geq 1$. A map $f : X \to Y$ is called a *$D$-embedding*, if and only if there is an $r > 0$, such that for all $x, y \in X$ the following holds:

$$r \cdot \rho(x, y) \leq \sigma(f(x), f(y)) > rD \cdot \rho(x, y).$$

The infimum over all $D$ with this property is called the *distortion* of $D$.

There are very few possibilities for isometrically embedding a metric space, but surprisingly many if small distortion is allowed.

## 2 Johnson and Lindenstrauss Flattening Lemma

**Theorem 3.** *Let $X \subseteq l_2$, $|X| = n$ and $\varepsilon \in (0, 1]$. Then there exists a $(1 + \varepsilon)$-embedding of $X$ into $l_2^k$ where $k$ is $O(\varepsilon^{-1} \log n)$.*

The proof of Theorem 3 is based on the following Lemma which we will prove afterwards.

**Lemma 4.** *For a random vector $x \in_u S^{n-1}$ where $S^{n-1} = \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\}$ let $f(x) = \sqrt{x_1^2 + x_2^2 + \cdots + x_k^2}$ be the length of the projection onto the subspace spanned by the first $k$ coordinates. Then there exists a number $a = a(n,k)$, s.t. for all $t$, we get*

$$Pr[f(x) > a + t] \leq 2e^{-t^2 n/2} \text{ and}$$
$$Pr[f(x) < a - t] \leq 2e^{-t^2 n/2}.$$

*Furthermore, there exists a constant $c$ s.t. if $n \geq c$ and $k \geq 10 \ln n$ then $a \geq \frac{1}{2}\sqrt{\frac{k}{n}}$.*

*Proof.* (of Theorem 3) Suppose $n$ is sufficiently large (exact value will be determined later) and we have $X \subseteq l_2^n$ with $|X| = n$. Let $k = 200\varepsilon^{-}2 \ln n$. Note that, if $k > n$ we are done, so assume $k < n$.

Let $L$ be a random, uniformly distributed $k$-dimensional linear subspace of $l_2^n$ and let $p : l_2^n \to L$ be the corresponding orthogonal projection. According to Lemma 4 we get a number $a$ where $\|p(x)\|_2$ is concentrated. (Note that here we use the lemma with a randomly chosen subspace and a fixed point $x \in S^{n-1}$ instead of doing it the other way around.)

Then the following is true:

**Claim 5.** *For any 2 points $x, y \in l_2^n$ with $x \neq y$ the inequalities*

$$\left(1 - \frac{\varepsilon}{3}\right) a\|x - y\|_2 \leq \|p(x) - p(y)\|_2 \leq \left(1 + \frac{\varepsilon}{3}\right) a\|x - y\|_2 \tag{1}$$

*are violated with probability $\leq 1/n^2$.*

*Proof.* Let $x, y \in X$ with $x \neq y$ be fixed and let $u = x - 1$. Then, by linearity of the map $p$, we get $p(u) = p(x) - p(y)$. We need to show that

$$\left(1 - \frac{\varepsilon}{3}\right) a\|u\|_2 \leq \|p(u)\|_2 \leq \left(1 + \frac{\varepsilon}{3}\right) a\|u\|_2$$

is false with low probability. By scaling we can assume that $\|u\|_2 = 1$, since again, $p$ is linear. Thus, we can fix a $u \in S^{n-1}$. Now, the inequalities are violated if and only if

$$\|u\|_2 < \left(1 - \frac{\varepsilon}{3}\right) a = a - \frac{\varepsilon}{3}a \text{ or,}$$
$$\|u\|_2 < \left(1 + \frac{\varepsilon}{3}\right) a = a + \frac{\varepsilon}{3}a.$$

By Lemma 4 the probability for this is at most $2 \cdot 2e^{-t^2 n/2}$, and in our case for $t = \frac{\varepsilon}{3}a$ we get $2 \cdot 2e^{-\varepsilon^2 a^2 n/18}$. Taking a detailed glance at the exponent reveals

$$\frac{1}{18}\varepsilon^2 a^2 n \geq \frac{1}{18}\varepsilon^2 \frac{1}{4}\frac{k}{n}n = \frac{1}{72}\varepsilon^2 k \geq \frac{200}{72}\ln n,$$

where we used $a \geq \frac{1}{4}\sqrt{\frac{k}{n}}$ (by Lemma 4). Therefore, the probability of failure for a fixed pair $x, y \in X$ with $x \neq y$ is less than

$$< 4e^{-200 \ln n/72} = 4n^{-200/72} \leq n^2,$$

for sufficiently large $n$. □

But there are less than $n^2$ such pairs $x, y \in X$ with $x \neq y$ and thus, the probability that there is at least one pair that violates Equation 1 is strictly less than 1. On the other hand, we have with probability $> 0$ that for all pairs $x, y$ Equation 1 holds and we follow by the probabilistic method the existence of our desired subspace $L$.

It remains to verify that the distortion factor is appropriate. For this, we check that

$$\frac{1+\frac{\varepsilon}{3}}{1-\frac{\varepsilon}{3}} \leq 1+\varepsilon,$$

which is true for $\varepsilon < 1$. □

**Remark:** We can increase $k$ by constant in order to make the probability that a projection into an randomly, uniformly distributed chosen subspace satisfies the Johnson-Lindenstrauss Lemma arbitrarily close to 1, i.e. $1 - 1/n^r$ for every constant $r$ (see exercises).

Furthermore, Theorem 3 gives (somehow implicitly) the following polynomial time algorithm for computing such a projection.

1. Create a $k \times n$ matrix, where every entry is chosen according to the normal distribution, i.e. $\sim N(0, 1)$.
2. Normalize each row to get a matrix $P$.
3. Compute $y_1 = Px_1, y_2 = Px_2, \ldots, y_n = Px_n \in R^k$.