**3.5 Lecture 17 (9/Nov): Johnson-Lindenstrauss embedding \( \ell_2 \rightarrow \ell_2 \)**

By a small sample we may judge of the whole piece.
Cervantes, Don Quixote de la Mancha §1-1-4

Today we’ll see a geometric application of the Chernoff bound. At first glance the question we solve, which originates in analysis, appears to have nothing to do with probability. But actually it illustrates a shared geometric core between analysis and probability.

**Definition 57** A metric space \((M, d_M)\) is a set \(M\) and a function \(d_M : M \times M \rightarrow (\mathbb{R} \cup \infty)\) that is symmetric; \(0\) on the diagonal; and obeys the triangle inequality, \(d_M(x, y) \leq d_M(x, z) + d_M(z, y)\).

Examples:

1. A Euclidean space is a vector space \(\mathbb{R}^n\) equipped with the metric \(d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}\).
2. The same vector space can be equipped with a different metric, for instance the \(\ell_{\infty}\) metric, \(\max_i |x_i - y_i|\), or the \(\ell_1\) metric, \(\sum_i |x_i - y_i|\).
   Actually in real vector spaces the metrics we use, like these, are usually derived from norms (see Sec. 3.5.1).
3. Sometimes we get important metrics as restrictions of another metric. For instance let \(\Delta_n\) denote the probability simplex, \(\Delta_n = \{x \in \mathbb{R}^n : \sum_i x_i = 1, x_i \geq 0\}\). In this space (half of) the \(\ell_1\) distance is referred to as “total variation distance”, \(d_{\text{TV}}\). It has another characterization, \(d_{\text{TV}}(p, q) = \max_{A \subseteq [n]} p(A) - q(A)\).
   **Exercise**: Usually a metric arises through a “min” definition (shortest path from one place to another), and in Example 3 we will see that \(d_{\text{TV}}\) does have that kind of definition. Why does it coincide with a “max” definition?
4. Many metric spaces have nothing to do with vector spaces. An important class of metrics are the shortest path metrics, derived from undirected graphs: If \(G = (V, E)\) is a graph and \(x, y \in V\), let \(d(x, y)\) denote the length (number of edges on) a shortest path between them.
5. If you start with a metric \(d\) on a measurable space \(M\) you can “lift” it to the transportation metric \(d_{\text{trans}}\). This is much bigger: the “points” of this new metric space are probability distributions on \(M\), and the transportation distance is how far you have to shift probability mass in order to transform one distribution to the other. Here is the formal definition for the case of a finite space \(M\). Let \(\mu, \nu\) be the two distributions. \(\pi\) will range over probability distributions on the direct product space \(M^2\).

\[
d_{\text{trans}}(\mu, \nu) = \min\{\sum_{x,y} d(x,y)\pi(x,y) | \forall x : \sum_y \pi(x,y) = \mu(x), \forall y : \sum_x \pi(x,y) = \nu(y), \forall x, y : \pi(x, y) \geq 0\}
\]

Sometimes this is called “earthmover distance” (imagine bulldozers pushing the probability mass around).

For example, if \(M\) is the graph metric on a clique of size \(k\) (as in Example 3) then \(d_{\text{trans}} = d_{\text{TV}}\) variation distance among probability distributions on the vertices (i.e., the metric space of Example 3).

**Definition 58** An embedding \(f : M \rightarrow M'\) is a mapping of a metric space \((M, d_M)\) into another metric space \((M', d_M')\). The distortion of the embedding is \(\sup_{a,b,c,d \in M} \frac{d_M'(f(a), f(b))}{d_M(a,b)} \cdot \frac{d_M'(f(c), f(d))}{d_M'(f(c'), f(d'))}\). The mapping is called isometric if it has distortion 1.
A finite metric space is one in which the underlying set is finite. A finite $\ell_2$ space is one that can be embedded isometrically into a Euclidean space of any dimension.

Exercise: The dimension need not be greater than $n - 1$. $(n$ points span only at most an $(n - 1)$-dimensional affine subspace.)

Exercise: Generically, the dimension must be $n - 1$. (Show the distances between points in Euclidean space determine their coordinates up to a rotation, reflection and translation. Then consider the volume of the simplex.)

What we’ll see today is a method of embedding an $n$-point $\ell_2$ metric into a very low-dimensional Euclidean space with only slight distortion. This is useful in the theory of computation because many algorithms for geometric problems have complexity that scales exponentially in the dimension of the input space. We’ll have to skip giving example applications, but there are quite a few by now, and because of these, a variety of improvements and extensions of the embedding method have also been developed.

Our goal is to prove the following claim:

**Theorem 59 (Johnson and Lindenstrauss [56])** Given a set $A$ of $n$ points in a Euclidean space, there exists a map $f : A \rightarrow (\mathbb{R}^k, \ell_2)$ with $k = O(\epsilon^{-2} \log n)$ that is of distortion $\epsilon$ on the metric restricted to $A$. Moreover, the map $f$ can be taken to be linear and can be found with a simple randomized algorithm in expected time polynomial in $n$.

Although the points of $A$ generically span an $(n - 1)$-dimensional affine space, and the map is linear, nonetheless observe that we are not embedding all of $\mathbb{R}^{n-1}$ with low distortion—that is impossible, as the map is many-one—we care only about the distances among our $n$ input points.

### 3.5.1 Normed spaces

A real normed space is a vector space $V$ equipped with a nonnegative real-valued “norm” $\| \cdot \|$ satisfying $\|cv\| = c\|v\|$ for $c \geq 0$, $\|v\| \neq 0$ for $v \neq 0$, and $\|v + w\| \leq \|v\| + \|w\|$. Norms automatically define metrics, as in examples 1, 2, by taking the distance between $v$ and $w$ to be $\|v - w\|$.

Let $S = (S, \mu)$ be any measure space. For $p \geq 1$, the $L_p$ normed space w.r.t. the measure $\mu$, $L_p(S)$, is defined to be the vector space of functions

$$f : S \rightarrow \mathbb{R}$$

of finite “$L_p$ norm,” defined by

$$\|f\|_p = \left(\int_S \|f(x)\|^p d\mu(x)\right)^{1/p}$$

Exercise:

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p$$

So (like any normed space), $L_p(S)$ is also automatically a metric space.

This framework allows us to discuss the collection of all $L_2$ (Euclidean) spaces, all $L_1$ spaces, etc. The most commonly encountered cases are indeed $L_1, L_2$ and $L_\infty$, which is defined to be the sup norm (so $\mu$ doesn’t matter). Today we discuss embeddings $L_2 \rightarrow L_2$. Time permitting we may also discuss embeddings of general metrics into $L_1$.

We will use the shorthand $L^k_p$ to refer to an $L_p$ space on a set $S$ of cardinality $k$, with the counting measure.
3.5.2 JL: the original method

Returning to the statement of the Johnson-Lindenstrauss Theorem ([59]), how do we find such a map \( f \)? Here is the original construction: pick an orthogonal projection, \( \tilde{W} \), onto \( \mathbb{R}^k \) uniformly at random, and let \( f(x) = \tilde{W}x \) for \( x \in A \).

For \( k \) as specified, this is satisfactory with high (constant) probability (which depends on the constant in \( k = O(\varepsilon^{-2} \log n) \)).

An equivalent description of picking a projection \( \tilde{W} \) at random is as follows: choose \( U \) uniformly (i.e., using the Haar measure) from \( O_n \) (the orthogonal group). Let \( \tilde{Q} \) be the \( n \times n \) matrix which is the projection map onto the first \( k \) basis vectors:

\[
\tilde{Q} = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 \\
& & & \ddots & & \\
0 & 0 & \cdots & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.
\]

Then set \( W = U^{-1}\tilde{Q}U \). I.e., a point \( x \in A \) is mapped to \( U^{-1}\tilde{Q}Ux \).

Let’s start simplifying this. The final multiplication by \( U^{-1} \) doesn’t change the length of any vector so it is equivalent to use the mapping

\[ x \rightarrow \tilde{Q}Ux \]

and ask what this does to the lengths of vectors between points of \( A \).

Having simplified the mapping in this way, we can now discard the all-0 rows of \( \tilde{Q} \), and use just \( Q \):

\[
Q = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 \\
& & & \ddots & & \\
0 & 0 & \cdots & 1 & 0 & 0
\end{pmatrix}.
\]

So JL’s final mapping is

\[ f(x) = QUx. \]

In order to analyze this map, we will consider a vector \( v \), the difference between two points in \( A \), i.e. \( v = x - y \) for some \( x, y \in A \).

Since the question of distortion of the length of \( v \) is scale invariant, we can simplify by supposing that \( \|v\| = 1 \).

Moreover, the process described above has the same distribution for all rotations of \( v \). That is to say, for any \( v, w \in \mathbb{R}^n \) and any orthogonal matrix \( A \),

\[
\Pr_U(QUv = w) = \Pr_U(QU(Av) = w). \quad \text{(prob. densities)}
\]

So we may as well consider that \( v \) is the vector \( v = (1,0,0,\ldots,0)^\ast \). (Where \( ^\ast \) denotes transpose.)

In that case, \( \|QUv\| \) equals \( \|(QU)_{\ast 1}\| \) where \( (QU)_{\ast 1} \) is the first column of \( QU \). But \( (QU)_{\ast 1} = (U_{1,1}, U_{2,1}, \ldots, U_{k,1})^\ast \), i.e., the top few entries of the first column of \( U \).

Since \( U \) is a random orthogonal matrix, the distribution of its first column (or indeed of any other single column) is simply that of a random unit vector in \( \mathbb{R}^n \).
So the whole question boils down to showing concentration for the length of the projection of a random unit vector onto the subspace spanned by the first \( k \) standard basis vectors.

This distribution is somewhat deceptive in low dimensions. For \( n = 2, k = 1 \) the density looks like Figure (3.4).

![Figure 3.4](image1.png)

**Figure 3.4: Density of projection of a unit vector in 2D onto a random unit vector**

However, in higher dimensions, this density looks more like Figure (3.5). The phenomenon we are encountering is truly a feature of high dimension.

![Figure 3.5](image2.png)

**Figure 3.5: Density of projection of a unit vector in high dimension onto a random unit vector**

**Remarks:**

1. In the one-dimensional projection density (Fig. 3.5) some constant fraction of the probability is contained in the interval \( \left[ -\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}} \right] \).

2. The squares of the projection-lengths onto each of the \( k \) dimensions are “nearly independent” random variables, so long as \( k \) is small relative to \( n \).

Johnson and Lindenstrauss pushed this argument through but there is an easier way to get there, by just slightly changing the construction.
3.5.3 JL: a similar, and easier to analyze, method

Pick \( k \) vectors \( w_1, w_2, \ldots, w_k \) independently from the spherically symmetric Gaussian density with standard deviation 1, i.e., from the probability density

\[
\eta(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} \exp \left( -\frac{1}{2} \sum_{i=1}^{n} x_i^2 \right)
\]

**Note 1:** the projection of this density on any line through the origin is the 1D Gaussian with standard deviation 1, i.e., the density

\[
\frac{1}{\sqrt{2\pi}} \exp \left( -\frac{x^2}{2} \right)
\]

(Follows immediately from the formula, by factoring out the one dimension against an entire “conditioned” Gaussian on the remaining \( n - 1 \) dimensions.)

**Note 2:** The distribution is invariant under the orthogonal group. (Follows immediately from the formula.)

**Note 3:** The coordinates \( x_1, x_2 \) etc. are independent rvs. (Follows immediately from the formula.)

Set

\[
W = \begin{pmatrix}
\cdots & \cdots & w_1 & \cdots & \cdots \\
\cdots & \cdots & w_2 & \cdots & \cdots \\
& \ddots & & & \\
\cdots & \cdots & w_k & \cdots & \cdots
\end{pmatrix}
\]

(The rows of \( W \) are the vectors \( w_i \).) Then, for \( v \in \mathbb{R}^n \) set \( f(v) = Wv \).

By Notes 1 & 3, each entry of \( W \) is an i.i.d. random variable with density \( \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \).

Informally, this process is very similar to that of JL, although it is certainly not identical. Individual entries of \( W \) can (rarely) be very large, and rows are not going to be exactly orthogonal, although they will usually be quite close to orthogonal.

Because of Note 2, analysis of this method boils down, just as for the original JL construction, to showing a concentration result for the length of the first column of \( W \), which we denote \( w^1 \).

Because of Note 3, the expression \( \|w^1\|^2 = \sum_{i=1}^{k} w_i^2 \) gives the LHS as the sum of independent, and by Note 1 iid, rvs. This will enable us to show concentration through a Chernoff bound.