10 Lecture 10, November 3, 2014

Returning to the statement of the Johnson-Lindenstrauss Theorem (45), 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 $\mathcal{O}^n$ (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

$$x \to \tilde{Q}Ux$$

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,

$$\Pr_U(QUv = w) = \Pr_U(QU(Av) = w).$$

So we may as well consider that v is the vector $v = (1, 0, 0, \ldots, 0)^\dagger$.

In that case, the length^2 of $\|QUv\| = \|(QU)_1\|$ where $(QU)_1$ is the first column of QU. But $(QU)_1 = (U_{1,1}, U_{2,1}, \ldots, U_{k,1})^\dagger$.

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.

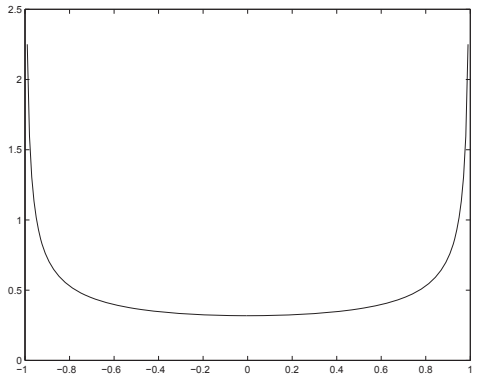

Figure 2: Density of projection of a unit vector in 2D onto a random unit vector

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

However, in higher dimensions, this density looks more like Figure (3). The phenomenon we are encountering is truly a feature of high dimension.

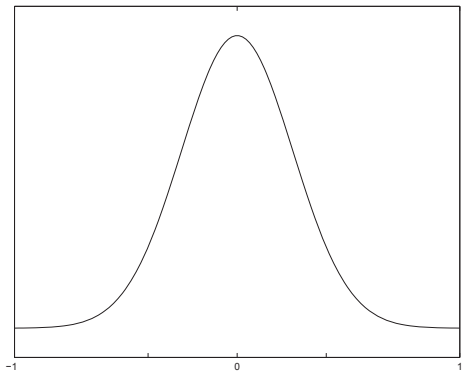

Figure 3: Density of projection of a unit vector in high dimension onto a random unit vector

Remarks:

1. In the density some constant fraction of the probability is contained in the interval $\left[\frac{-1}{\sqrt{dim}}, \frac{1}{\sqrt{dim}}\right]$.
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.

10.1 A related 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.)

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

$$W = \begin{pmatrix} \dots & \dots & w_1 & \dots & \dots \\ \dots & \dots & w_2 & \dots & \dots \\ & & \vdots & & \\ \dots & \dots & w_k & \dots & \dots \end{pmatrix}.$$

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

By the first note and third notes, 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 $\|w^1\|^2 = \sum_1^k w_{i1}^2$ 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.

We now have independent random variables $w_{11}, \dots, w_{k1}$, each normally distributed with $E(w_{i1}^2) = 1$. So $E(\sum w_{i1}^2) = k$. We want a deviation bound on $\sum w_{i1}^2$.

There is a name for these rvs: each w_{i1}^2 is a χ^2 rv with parameter 1, and their sum is a χ^2 rv with parameter k.

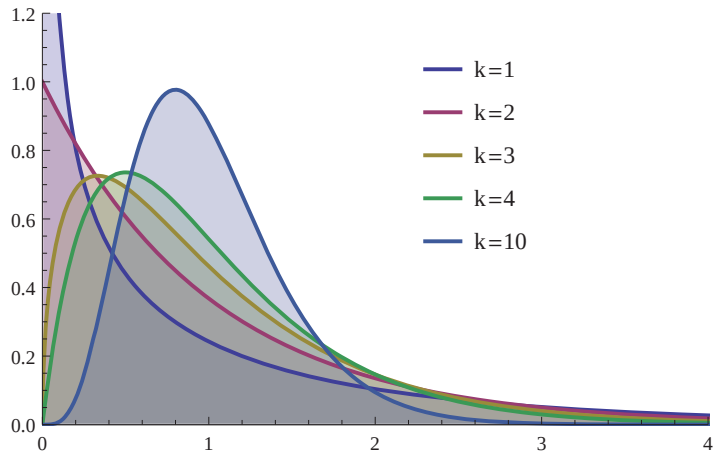

Figure 4: Probability density of $\frac{1}{k}\sum w_{i1}^2$

Set random variables $y_i = w_{i1}^2 - 1$ so that $E(y_i) = 0$. With this change of variables we now want a bound on the deviation from 0 of the rv $\overline{y} = \frac{1}{k}\sum_{i=1}^k y_i$.

To get a Chernoff bound, we need the mgf, $g(\beta)$, for y_i, in order to use Eqn. 22 to write:

$$P(\overline{y}/\varepsilon > 1) < [\inf_{\beta>0} e^{-\varepsilon\beta} g(\beta)]^k \quad \text{for } \varepsilon \neq 0. \tag{27}$$

So what is $g(\beta)$?

$$\begin{aligned} g(\beta) = E(e^{\beta y}) &= E(e^{\beta(w^2-1)}) \\ &= e^{-\beta} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{w^2(\beta-1/2)} dw \\ &= \frac{e^{-\beta}}{\sqrt{1-2\beta}} \int_{-\infty}^{\infty} \frac{\sqrt{1-2\beta}}{\sqrt{2\pi}} e^{-\frac{1}{2}w^2(1-2\beta)} dw \\ &= \frac{e^{-\beta}}{\sqrt{1-2\beta}} \end{aligned}$$

The last equality follows as the integrand is the density of a normal random variable with standard deviation $\frac{1}{\sqrt{1-2\beta}}$.

Thus, $g(\beta)$ is well defined and differentiable in $(-\infty, \frac{1}{2})$, with (necessarily) $g(0) = 1$ and $g'(0) = 0$.

For a given ε what β should be used in the Chernoff bound (Eqn. 27)? After some calculus, we find that $\beta = \frac{\varepsilon}{2(1+\varepsilon)}$ is the best value (for both signs of ε). Figure (5) shows the dependence of β on ε.

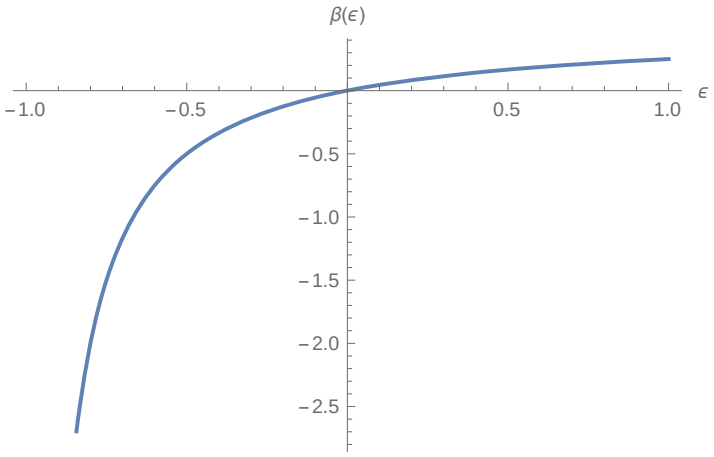

Figure 5: Best choice of β as a function of ε for the χ^2 distribution: $\beta = \frac{\varepsilon}{2(1+\varepsilon)}$

Plugging this value of β above into the bound, we get

$$P(\overline{y}/\varepsilon > 1) < ((1+\varepsilon)^{\frac{1}{2}} e^{-\frac{\varepsilon}{2}})^k \tag{28}$$

which we incidentally note is $(1 - \frac{1}{2}\varepsilon^2 + O(\varepsilon^3))^k$. The function $(1+\varepsilon)^{\frac{1}{2}} e^{-\frac{\varepsilon}{2}}$ is shown in Fig. 6.

Now let's apply this bound to the modified JL construction. We will ensure distortion (Defn. 44) e^{ε} (with positive probability) by showing that for each of our $\binom{n}{2}$ vectors v, with probability $> 1 - 1/\binom{n}{2}$,

$$\|v\| e^{-\varepsilon/2} \leq \|Wv\|/\sqrt{k} \leq \|v\| e^{\varepsilon/2}.$$

This event has the same probability, for a specific v, as the event

$$e^{-\varepsilon} \leq \frac{1}{k} \sum w_{i1}^2 \leq e^{\varepsilon}$$

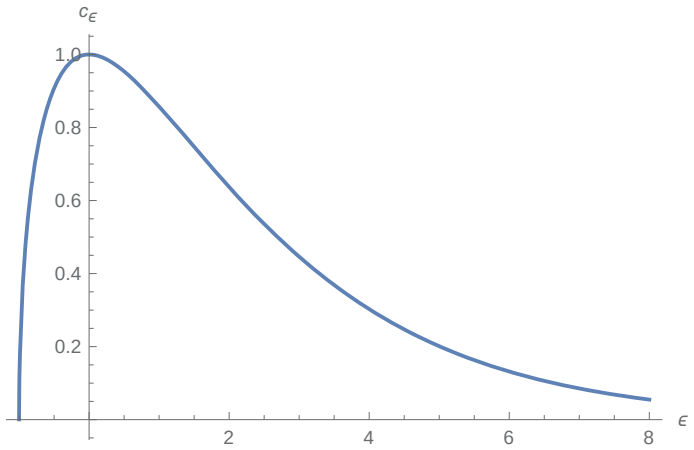

Figure 6: Base of the Chernoff bound for the χ^2 distribution: $c_\varepsilon = (1+\varepsilon)^{\frac{1}{2}} e^{-\frac{\varepsilon}{2}}$

or equivalently

$$e^{-\varepsilon} - 1 \leq \bar{y} \leq e^{\varepsilon} - 1. \tag{29}$$

Applying Inequality (28) first on the right of (29), we have

$$\Pr(\bar{y} > e^{\varepsilon} - 1) < e^{k(\varepsilon/2 - (e^{\varepsilon}-1)/2)} = e^{(k/2)(1+\varepsilon-e^{\varepsilon})} < e^{-k\varepsilon^2/4}$$

Next applying Inequality (28) on the left of (29), we have

$$\Pr(\bar{y} < e^{-\varepsilon} - 1) < e^{k(-\varepsilon/2 - (e^{-\varepsilon}-1)/2)} = e^{(k/2)(1-\varepsilon-e^{-\varepsilon})} < e^{-k(\varepsilon^2/4 + O(\varepsilon^3))}$$

In all, taking $k = 8(1+O(\varepsilon))\varepsilon^{-2}\log n$ suffices so that $\Pr(\frac{1}{k}\sum w_{i1}^2 \notin [e^{-\varepsilon}, e^{\varepsilon}]) < 1/n^2$ and therefore so the mapping with probability at least $1/2$ has distortion bounded by e^{ε}.

Finally, for the computational aspect: to get a randomized "Las Vegas" algorithm simply try matrices W at random and test whether their distortion is satisfactory.

Note: About another embedding question: Finite l_2 metric spaces can be embedded in l_1 isometrically. There's also an algorithm—deterministic, in fact—to find such an embedding, but it takes exponential time in the number of points in the space.

Comment: There are deterministic poly-time algorithms producing an embedding up to the standards of the Johnson-Lindenstrauss theorem, see [18, 62].