An average-case Johnson-Lindenstrauss lemma

1 Overview

We derive an average-case version of the Johnson-Lindenstrauss lemma. We show that a n vectors in $\mathbf{R}^D$ can be mapped to $\mathbf{R}^d$ with average distortion of only $\Theta(1/d)$. Moreover, a simple random projection provides such a mapping.

Let $x_1, \ldots, x_n \in \mathbf{R}^D$. Let $R \in \mathbf{R}^{d \times D}$ be a random matrix of independent $N(0,1)$ entries. We consider the embedding provided by the map $f(x) = \frac{1}{\sqrt{d}} Rx$. The cost we measure is a weighted sum of square error:

$$g(R) = \frac{1}{\binom{n}{2}} \sum_{i<j} \left(\frac{\|f(x_i) - f(x_j)\|^2 - \|x_i - x_j\|^2}{\|x_i - x_j\|^2} \right)^2 \quad (1)$$

$$= \frac{1}{\binom{n}{2}} \sum_{i<j} \left(\frac{\frac{1}{d}\|R(x_i - x_j)\|^2 - \|x_i - x_j\|^2}{\|x_i - x_j\|^2} \right)^2. \quad (2)$$

We normalize each term in the sum by the original distance because we are using a projection technique which will distort distances at a rate conmeasurate with the scale. This is analagous to the standard Johnson-Lindenstrauss lemma, which uses as its cost function the worst-case ratio.

We will apply the following concentration of measure result to the cost function above to obtain a high-probability bound on the error.

Theorem 1 *Let f be a C-Lipschitz function of a gaussian random variable. Then,*

$$P(|f(x) - \mathbf{E}[f(x)]| > \epsilon) \leq 2 \exp\left(-\frac{\epsilon^2}{2C^2} \right).$$

We show in section 2 that $\mathbf{E}[g(R)]$ is $\Theta(\frac{1}{d})$. In section 3, we show that $g(R)$ is $\Theta(\frac{1}{d})$-Lipschitz. Section 4 details a couple of experiments verifying the bound.

2 Expectation calculation

We show that the $\mathbf{E}[g(R)]$ is $\Theta(1/d)$.

$$\mathbf{E}[g(R)] = \mathbf{E}\left[\frac{1}{\binom{n}{2}} \sum_{i<j} \left(\frac{\frac{1}{d}\|R(x_i - x_j)\|^2 - \|x_i - x_j\|^2}{\|x_i - x_j\|^2} \right)^2 \right] \quad (3)$$

$$= \frac{1}{\binom{n}{2}} \sum_{i<j} \mathbf{E}\left[\left(\frac{\frac{1}{d}\|R(x_i - x_j)\|^2 - \|x_i - x_j\|^2}{\|x_i - x_j\|^2} \right)^2 \right]. \quad (4)$$

Let us calculate the expectation of each term in the sum. To simplify notation, let $w = x_i - x_j$. We are interested in calculating the following expectation:

$$
\begin{aligned}
&\mathbf{E}\left[\left(\frac{\frac{1}{d}\|Rw\|^2 - \|w\|^2}{\|w\|^2}\right)^2\right] && (5)\\
&= \mathbf{E}\left[\frac{\frac{1}{d^2}\|Rw\|^4 - \frac{2}{d}\|w\|^2\|Rw\|^2 + \|w\|^4}{\|w\|^4}\right] && (6)\\
&= \frac{1}{d^2\|w\|^4}\mathbf{E}\left[\|Rw\|^4\right] - \frac{2}{d\|w\|^2}\mathbf{E}\left[\|Rw\|^2\right] + 1. && (7)
\end{aligned}
$$

2.1 Second moment

Let us calculate $\mathbf{E}\left[\|Rw\|^2\right]$ first. Recall that the entries of R are independent draws from $N(0,1)$.

$$
\begin{aligned}
\mathbf{E}\left[\|Rw\|^2\right] &= \mathbf{E}\left[\sum_{i=1}^{d}\langle r_i, w\rangle^2\right] && (8)\\
&= \sum_{i=1}^{d}\mathbf{E}\left[\left(\sum_{j=1}^{D} r_{ij}w_j\right)^2\right] && (9)\\
&= \sum_{i=1}^{d}\sum_{j=1}^{D}\mathbf{E}\left[r_{ij}^2 w_j^2\right] + \sum_{j\neq k}\mathbf{E}\left[r_{ij}w_j r_{ik}w_k\right] && (10)\\
&= \sum_{i=1}^{d}\sum_{j=1}^{D} w_j^2 = d\|w\|^2. && (11)
\end{aligned}
$$

2.2 Fourth moment

Next we calcuate $\mathbf{E}\left[\|Rw\|^4\right]$.

$$
\begin{aligned}
\mathbf{E}\left[\|Rw\|^4\right] &= \mathbf{E}\left[\|Rw\|^2\|Rw\|^2\right] \\
&= \mathbf{E}\left[\left(\sum_{i=1}^{d}\langle r_i, w\rangle^2\right)\left(\sum_{i=1}^{d}\langle r_i, w\rangle^2\right)\right] \\
&= \sum_{i=1}^{d}\mathbf{E}\left[\langle r_i, w\rangle^4\right] + \sum_{i\neq j}\mathbf{E}\left[\langle r_i, w\rangle^2\langle r_j, w\rangle^2\right] \\
&= \sum_{i=1}^{d}\mathbf{E}\left[\langle r_i, w\rangle^4\right] + \sum_{i\neq j}\mathbf{E}\left[\langle r_i, w\rangle^2\right]\mathbf{E}\left[\langle r_j, w\rangle^2\right] \\
&= \sum_{i=1}^{d}\mathbf{E}\left[\langle r_i, w\rangle^4\right] + \sum_{i\neq j}\|w\|^2\|w\|^2 \text{ (see section 2.1)} \\
&= \sum_{i=1}^{d}\mathbf{E}\left[\langle r_i, w\rangle^4\right] + \binom{d}{2}\|w\|^4. \qquad (12)
\end{aligned}
$$

It remains to calculate $\mathbf{E}\left[\langle r_i, w\rangle^4\right]$. To simplify notation further, let $s = r_i$.

$$
\begin{aligned}
\mathbf{E}\left[\langle s, w\rangle^4\right] &= \mathbf{E}\left[\left(\sum_{j=1}^{D} s_j w_j\right)^4\right] \\
&= \mathbf{E}\left[\left(\sum_{j=1}^{D} s_j^2 w_j^2 + \sum_{j\neq k} s_j w_j s_k w_k\right)^2\right] \\
&= \mathbf{E}\left[\left(\sum_{j=1}^{D} s_j^2 w_j^2\right)^2\right] + \mathbf{E}\left[\left(\sum_{j\neq k} s_j w_j s_k w_k\right)^2\right] \\
&\quad + 2\mathbf{E}\left[\left(\sum_{j=1}^{D} s_j^2 w_j^2\right)\left(\sum_{k\neq l} s_k w_k s_l w_l\right)\right]. \qquad (13)
\end{aligned}
$$

Let us look at the three parts of (13) Individually.

1.

$$
\begin{aligned}
\mathbf{E}\left[\left(\sum_{j=1}^{D} s_j^2 w_j^2\right)^2\right] &= \mathbf{E}\left[\sum_{j=1}^{D} s_j^4 w_j^4 + \sum_{j\neq k} s_j^2 w_j^2 s_k^2 w_k^2\right] \\
&= \sum_{j=1}^{D} 3w_j^4 + \sum_{j\neq k} w_j^2 w_k^2 \\
&= \sum_{j=1}^{D} 2w_j^4 + \left(\sum_{j=1}^{D} w_j^2\right)^2 \\
&= \|w\|^4 + 2\sum_{j=1}^{D} w_j^4. \qquad (14)
\end{aligned}
$$

2.

$$
\begin{aligned}
&\mathbf{E}\left[\left(\sum_{j\neq k} s_j w_j s_k w_k\right)^2\right] \\
&= \mathbf{E}\left[\sum_{j\neq k} s_j^2 w_j^2 s_k^2 w_k^2 + \sum_{\substack{(j,k)\neq(l,m),\\ j\neq k,\\ l\neq m}} s_j w_j s_k w_k s_l w_l s_m w_m\right] \\
&= \sum_{j\neq k} w_j^2 w_k^2 + 2\mathbf{E}\left[\sum_{\substack{j\neq l\\ j\neq k\\ l\neq m}} s_j w_j s_k w_k s_l w_l s_m w_m\right] \\
&= \sum_{j\neq k} w_j^2 w_k^2 + 2\mathbf{E}\left[\sum_{j\neq k} s_j^2 w_j^2 s_k^2 w_k^2\right] \\
&= 3\sum_{j\neq k} w_j^2 w_k^2. \qquad (15)
\end{aligned}
$$

3.

$$
\begin{aligned}
&2\mathbf{E}\left[\left(\sum_{j=1}^{D} s_j^2 w_j^2\right)\left(\sum_{k\neq l} s_k w_k s_l w_l\right)\right] \\
&= 2\mathbf{E}\left[\left(\sum_{j=1}^{D}\sum_{k\neq l} s_j^2 w_j^2 s_k w_k s_l w_l\right)\right] = 0. \qquad (16)
\end{aligned}
$$

Plugging (14), (15), and (16) back into (13), we get

$$\mathbf{E}\left[\langle s, w\rangle^4\right] = \|w\|^4 + 2\sum_{j=1}^{D} w_j^4 + 3\sum_{j\neq k} w_j^2 w_k^2 = 3\|w\|^4 + \sum_{j\neq k} w_j^2 w_k^2.$$

Plugging this back into 12, we finally come to

$$\begin{aligned} \mathbf{E}\left[\|Rw\|^4\right] &= 3d\|w\|^4 + d\sum_{j\neq k} w_j^2 w_k^2 + \binom{d}{2}\|w\|^4 && (17) \\ &\leq \left(\frac{d^2+7d}{2}\right)\|w\|^4. && (18) \end{aligned}$$

2.3 Putting it all together

We now plug our second and fourth moment estimates into (7), yielding

$$\begin{aligned} \mathbf{E}\left[\left(\frac{\frac{1}{d}\|Rw\|^2 - \|w\|^2}{\|w\|^2}\right)^2\right] &\leq \frac{1}{d^2\|w\|^4}\left(\frac{d^2+7d}{2}\right)\|w\|^4 - \frac{2}{d\|w\|^2}d\|w\|^2 + 1 \\ &= 1/2 + \frac{7}{2d} - 1 \leq \frac{4}{d}. && (19) \end{aligned}$$

Finally, since w does not appear in the estimate, we see that

$$\mathbf{E}\left[g(R)\right] \leq \frac{4}{d} = \Theta\left(\frac{1}{d}\right).$$

3 Lipschitz constant calculation

To be completed.

4 Experiments

We illustrate the hypothesized error on two datasets. First, we ran the random projection procedure on a subset of the 'ones' digits in the mnist dataset. We reduced the dimension from 784 to every value in the range $[1, 200]$. Figure 4 shows the error as a function of d; $\frac{5}{d}$ is plotted for comparison.

The second experiment was identical, except that the dataset consisted of 1000 random points in 1000-dimensions. Results are shown in figure 4.

In both examples, the error is bounded above by $\frac{5}{d}$, which validates the theoretical results discussed.

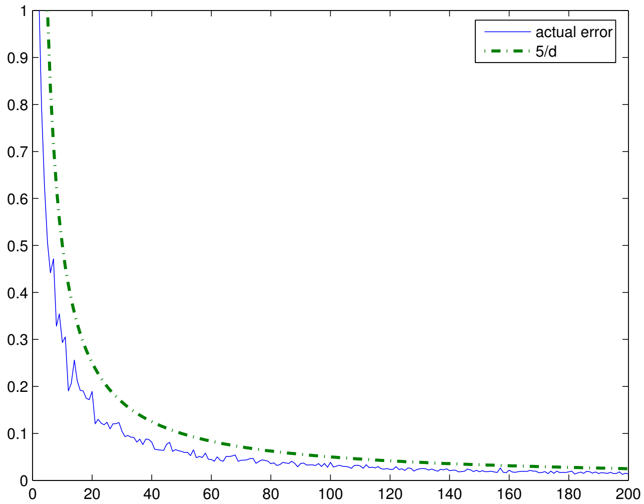

Figure 1: Error as a function of dimensionality for the mnist data. $\frac{5}{d}$ is also plotted for comparison.

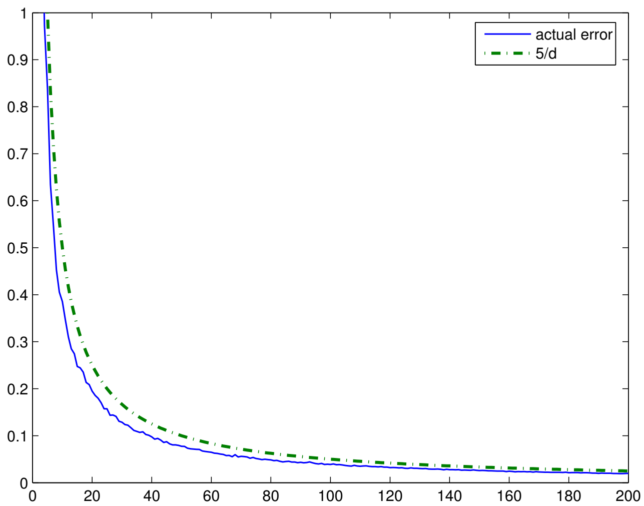

Figure 2: Error as a function of dimensionality for the synthetic data. $\frac{5}{d}$ is also plotted for comparison.