Johnson-Lindenstrauss Dimensionality Reduction
A Simple Proof

Allan Grønlund

April 30, 2018

1 Intro

The goal of dimensionality reduction is simply to represent a high dimensional data set in a lower dimensional space while preserving much of the important structure. Principal Component Analysis is a well known dimensionality reduction technique that finds (a few) directions of high variance in the data and projects the data onto these vectors of maximal variance. The idea is to maintain as much variance in the data as possible (the structure) while reducing the dimensionality. The mapping of data from a high dimensional space to a low dimensional space is called an embedding, i.e. we embed the higher dimensional data points in the lower dimensional space.

In this note we describe a dimensionality reduction technique that approximately preserves all pairwise distances between the data points. A fundamental result of Johnson and Lindenstrauss [1] says that any m point subset of Euclidean space can be linearly embedded in $k = O(\lg m/\varepsilon^2)$ dimensions without distorting the distances between any pair of points by more than a factor of $(1 \pm \varepsilon)$, for any $0 < \varepsilon < \frac{1}{2}$. For Principal Component Analysis to be relevant the original data points $x_1, \ldots, x_m$ must be inherently low dimensional. The Johnson-Lindenstraus theorem requires no assumption on the original data and the target dimension is even independent of the input dimension. Another fundamental result og Larsen and Nelson [4], shows that that the Johnson-Lindenstrauss Lemma is tight even for non-linear embeddings!

Besides the obvious application of using dimensionality for compression, the Johnson–Lindenstrauss theorem has found numerous applications in algorithms and machine learning for instance for Approximate k-Means, Linear Regression, and Approximate Nearest Neighbours.

2 Simple JL Lemma

The Johnson-Lindenstrauss Lemma comes in many forms. In these notes we will prove the following version of the Johson-Lindenstrauss Lemma.

Theorem 1 ([1]). *For any $0 < \varepsilon < \frac{1}{2}$ and any integer m, then for integer*

$$k = O\left(\frac{1}{\varepsilon^2} \lg m\right),$$

large enough and any m points $x_1, \ldots, x_m \subset \mathbb{R}^d$ there exists a linear map (matrix) $L : \mathbb{R}^d \to \mathbb{R}^k$ such that for any $1 \leq i, j \leq m$

$$(1-\varepsilon)\|x_i - x_j\|_2^2 \leq \|Lx_i - Lx_j\|_2^2 \leq (1+\varepsilon)\|x_i - x_j\|_2^2$$

The really cool thing is: the linear transformation L in Theorem 1 is simply multiplication by a matrix whose entries are sampled independently from a standard Gaussian. To be precise, define a random variable A as a $k \times d$ matrix, where each entry $A_{i,j} \sim \mathcal{N}(0,1)$, i.e. is a zero-mean Normal Distribution (Gaussian) with standard deviation 1. The final embedding matrix is sample of the random variable $L = \frac{1}{\sqrt{k}}A$.

First we prove a seemingly simpler result for such random variables.

Lemma 1. *Fix any vector unit vector x. For any $0 < \varepsilon, \delta < \frac{1}{2}$. For $k = O(\varepsilon^{-2} \log \frac{1}{\delta})$ large enough let $L = \frac{1}{\sqrt{k}} A$ be a random variable, where A is a $k \times d$ random matrix whose entries are independent zero mean Gaussians ($\sim \mathcal{N}(0,1)$) Then:*

$$\Pr_L(|\|Lx\|^2 - 1| > \varepsilon) \leq \delta$$

Stated differently, for any unit vector x and values of $0 < \varepsilon, \delta < \frac{1}{2}$. If we pick a random matrix as described, then with probability at least $1 - \delta$ the norm of x is distorted by a factor at most $(1 \pm \varepsilon)$ by L.

Note that this generalizes to any non-unit vector since a vector v can be written as $v = \|v\| \frac{v}{\|v\|}$. Since the embedding is linear, $Lv = \|v\| L \frac{v}{\|v\|}$, thus with probability at least $1 - \delta$,

$$\|Lv\|^2 = \left\| \|v\| L \frac{v}{\|v\|} \right\|^2 = \|v\|^2 \left\| L \frac{v}{\|v\|} \right\|^2 \in \|v\|^2 (1 \pm \varepsilon)$$

and the same bound applies for any vector v.

Before we prove Lemma 1, lets see how we can use if to prove Theorem 1.

Proof of Theorem 1. Fix a point set $x_1, \ldots, x_m$. Set $\delta = \frac{1}{m^3}$ and $k = O(\varepsilon^{-2} \log \frac{1}{\delta})$ large enough and consider the random variable matrix $L = \frac{1}{\sqrt{k}} A$ where $A_{i,j} \sim \mathcal{N}(0,1)$.

First observe that from linearity of L, it follows that $\|Lx_i - Lx_j\| = \|L(x_i - x_j)\|$. This means that the distance between x_i and x_j is distorted by more than $(1 \pm \varepsilon)$ if and only if the norm of the vector $(x_i - x_j)$ is distorted by more than $(1 \pm \varepsilon)$. Assuming Lemma 1, for any pair of indices $1 \leq i, j \leq m$, the probability that the matrix L distorts the norm of the vector $x_i - x_j$ by more that $(1 \pm \varepsilon)$ is at most δ. Since we need all pairwise distances to be approximately maintained, we need to bound the probability that there will be at least one index pair i, j such that $x_i - x_j$ is distorted by at least $(1 \pm \varepsilon)$. We call an index pair i, j bad if L does not maintain the norm of $x_i - x_j$.

Applying the Union Bound, the probability (over L) that there is a bad index pair is at most the sum over all i, j the probability that L distorts the norm of $x_i - x_j$ by more than $(1 \pm \varepsilon)$. We get:

$$\Pr_L(\exists (i,j) \text{ that is bad }) \leq \sum_{i,j=1, i<j}^{m} \Pr_L(i,j \text{ bad}) = \binom{m}{2} \frac{1}{m^3} \leq \frac{1}{m}$$

Thus with probability at least $1 - \frac{1}{m} > 0$ ($m > 1$ otw. trivial) the random matrix L maintains all the pairwise norms in the dataset within a factor $(1 \pm \varepsilon)$. In particular, this means that there exists a matrix that can embed the points and maintain all pairwise distances (otherwise the probability would be zero), and to find one just sample a matrix and it works with probability at least $1 - \frac{1}{m}$. □

What remains is to prove Lemma 1. To do that we first need a few results about the Normal Distribution and the χ^2 distribution that we will not prove.

Fact 1. *For any constants a, b if $X \sim \mathcal{N}(0,1)$, and $Y \sim \mathcal{N}(0,1)$ then $aX + bY \sim \mathcal{N}(0, a^2\sigma_x^2 + b^2\sigma_y^2)$.*

Definition 1. *If $Z_1, ..., Z_k$ are independent, standard normal random variables ($Z_i \sim \mathcal{N}(0,1)$), then the sum of their squares,*

$$Q = \sum_{i=1}^{k} Z_i^2$$

is distributed according to the chi-squared distribution with k degrees of freedom. This is usually denoted as $Q \sim \chi_k^2$.

Lemma 2 ([2])**.** *Let $Y \sim \chi_k^2$. Then*

$$\Pr_Y \left(\left| \frac{Y}{k} - 1 \right| \geq x \right) \leq e^{-\frac{3}{16} k x^2}, x \in [0, \frac{1}{2})$$

With these results in place we are ready to prove Lemma 1. First we prove that the random variable defined as the norm of $\|Ax\|^2$ is χ^2_k distributed. Then we apply Lemma 2 and Lemma 1 follows.

Lemma 3. *Given any unit vector $x \in \mathbb{R}^d$. Let A be the random variable matrix with each $A_{i,j} \sim \mathcal{N}(0,1)$ independently of the other entries. Then the random variable that is the squared norm of Ax is $\chi_{k,d}$ distributed. To be precise:*

$$\|Ax\|^2 \sim \chi^2_k$$

Proof. The Lemma follows from just writing out the computation of the squared norm and applying Fact 1. Let a_i be the i'th row of A, then by definition:

$$||Ax||^2 = \sum_{i=1}^{k} \langle a_i, x\rangle^2 = \sum_{i=1}^{k} \left(\sum_{j=1}^{d} a_{i,j} x_j \right)^2$$

Now apply Fact 1. Since $a_{i,j} \sim \mathcal{N}(0,1)$, we know $a_{i,j}x_j \sim \mathcal{N}(0, x_j^2)$ and we are computing a sum of these. This means that $\sum_{j=1}^{d} a_{i,j}x_j \sim \mathcal{N}(0, \sum_{j=1}^{d} x_j^2) = \mathcal{N}(0,1)$, since $\|x\|^2 = 1$. In total we get that $||Ax||^2$ is distributed as a sum of k standard Gaussians i.e.

$$||Ax||^2 \sim \sum_{i=1}^{k} \mathcal{N}(0,1)^2 \sim \chi^2_k$$

□

With this final piece in place we are ready to prove Lemma 1.

Proof of Lemma 1. Fix a unit vector x and consider the random variable matrix $L = \frac{1}{\sqrt{k}}A$ defined as usual.

First note that the expected value $\mathbb{E}[||Ax||^2] = k$, which means that $\mathbb{E}[\|Lx\|^2] = 1$, i.e. in expectation the norm of x is preserved by L. This is a very relevant property indeed. But, it is not enough. Even if we have the right expectation we may still get a large distortion of the norm, and we need to prove that we do not with probabability at least $1-\delta$. This is usually proved with what is called *concentration* or *tail* bounds that says that the probability that a random variable deviates much from its expected value is small and the concentration bound we need is exactly Lemma 2

To be precise, we need to argue that

$$\Pr_L(|\|Lx\|^2 - 1| > \varepsilon) \leq \delta$$

Luckily this is exactly captured in Lemma 2. Since $\|Ax\|^2 \sim \chi^2_k$, then

$$\Pr_A\left(\left| \frac{\|Ax\|^2}{k} - 1 \right| > \varepsilon \right) \leq e^{-\frac{3}{16}k\varepsilon^2}$$

since $\|Ax\|^2/k = \|\frac{1}{\sqrt{k}}\ Ax\|^2 = \|Lx\|^2$, this means that

$$\Pr_L(|\|Lx\|^2 - 1| > \varepsilon) \leq e^{-\frac{3}{16}k\varepsilon^2}$$

To finalize the proof, we need to find a k such that $e^{-\frac{3}{16}k\varepsilon^2} \leq \delta$ which is easily achieved by setting $k \geq \frac{16}{3}\varepsilon^{-2} \lg \frac{1}{\delta}$. □

3 Final Remarks

In practice the fact that you have to fill a matrix with random gaussian samples may be a implementationwise expensive. Luckily, the JL Lemma generalizes to the case where the entries in A are instead uniform random signs (± 1). Also, it can be shown that the matrix can be made sparse and the amount of randomnes required reduced, all steps to make the embedding more efficient, see [3] for more.

In this note we showed that we could reduce the dimensionality of any data set by using a simple randomized linear transformation (multiplication with a matrix). But what if we allow non-linear embeddings i.e. what if we allow an algorithm to try all possible embeddings can we then do better. The amazing thing is that we cannot. In [4], Larsen and Nelson shows that the JL embedding Lemma is tight even for non-linear embeddings by showing that there exists (high dimensional) data sets such that there is no way to embed into a space with lower dimension than what is promised in the JL Lemma without violating some of the pairwise distances. Of course, in practice, the data set may be easy and better embeddings are possible and several resarch papers has investigated such an approach.

References

[1] William Johnson and Joram Lindenstrauss. Extensions of lipschitz maps into a hilbert space. *Contemporary Mathematics*, 26:189–206, 01 1984.

[2] I.M. Johnstone. Chi-square oracle inequalities. *Lecture Notes-Monograph Series*, 36:399–418, 2001.

[3] Daniel M. Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *J. ACM*, 61(1):4:1–4:23, January 2014.

[4] Kasper Green Larsen and Jelani Nelson. Optimality of the johnson-lindenstrauss lemma. *58th Annual IEEE Symposium on Foundations of Computer Science*, 2017.