Dimension Reduction: Johnson–Lindenstrauss Lemma (Lecture 10 and Part of Lecture 11)

DS-563 / CD-543 @ Boston University
Instructor: Krzysztof Onak

Spring 2024

1 The Johnson–Lindenstrauss Lemma

The main result that we are discussing now is the Johnson–Lindenstrauss lemma, which was originally published by Johnson and Lindenstrauss in 1984.

Theorem 1. *For any set S of n points in $\mathbb{R}^k$ and $\epsilon \in \left(0, \frac{1}{2}\right)$, there is a linear embedding $f : \mathbb{R}^k \to \mathbb{R}^d$, where $d = \left\lceil \frac{24 \ln n}{\epsilon^2} \right\rceil$, such that for any pair of points $u, v \in S$,*

$$(1-\epsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\epsilon)\|u-v\|^2.$$

Informally, the Johnson–Lindenstrauss lemma states that it is possible to map a set of size n in the Euclidean space of arbitrary dimension to the Euclidean space of dimension at most $O(\log n)$, while roughly preserving all pairwise distances. The Johnson–Lindenstrauss lemma allows for reducing the dimensionality of a data set. This can be used for optimizing storage and more efficient processing, because some algorithmic techniques heavily depend on the dimension.

2 Our Construction

We will not only prove that such a linear mapping is possible, but also that a random mapping has the desired properties with non-zero probability. As noted later, by increasing the dimension in our construction, it is possible to make it work with probability very close to 1. This allows for performing this kind of dimension reduction in practice.

Our transformation has the following form:

$$f(x) = \frac{1}{\sqrt{d}} Ax$$

in which we interpret $x \in \mathbb{R}^k$ as a "vertical" vector that has one column and k rows, A is a matrix with k columns and d rows, and $\frac{1}{\sqrt{d}}$ is a normalizing factor. It is clear that $f(x)$ is a "vertical" vector in $\mathbb{R}^d$ that has one column and d rows. To fully describe the transformation, we have to describe how we select A. We do this by independently drawing each entry in A from $\mathcal{N}(0,1)$. Depicting it slightly differently, our

transformation looks as follows:

$$f(x) = \frac{1}{\sqrt{d}} \begin{bmatrix} & d \times k \text{ matrix } A, & \\ & \text{entries from } \mathcal{N}(0,1) & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_k \end{bmatrix}.$$

In the next section, we show the following lemma.

Lemma 2. *For any $x \in \mathbb{R}^k$,*

$$(1-\epsilon)\|x\|^2 \leq \|f(x)\|^2 \leq (1+\epsilon)\|x\|^2$$

with probability at least $1 - 2e^{-\frac{\epsilon^2 d}{12}}$.

It is relatively easy to prove the Johnson–Lindenstrauss lemma, using this lemma.

Proof of Theorem 1. For any pair u and v of points in S, let $x = u - v$. Since our transformation f is linear, we have $f(u) - f(v) = f(u - v) = f(x)$. By Lemma 2, we have that $\|f(x)\|^2 = \|f(u) - f(v)\|^2$ is at least $(1-\epsilon)\|x\|^2 = (1-\epsilon)\|u - v\|^2$ and at most $(1+\epsilon)\|x\|^2 = (1+\epsilon)\|u - v\|^2$ with probability at least $1 - 2e^{-\frac{\epsilon^2 d}{12}}$. By the union bound over all pairs of points, this property has to hold for all pairs with probability at least $1 - \binom{n}{2} \cdot 2e^{-\frac{\epsilon^2 d}{12}} > 1 - n^2 e^{-\frac{\epsilon^2 d}{12}}$. By setting $d = \frac{24 \ln n}{\epsilon^2}$, we make the last term 0. Since it strictly lower bounds the probability that a randomly selected f has the desired properties, this proves that with positive probability f accurately preserves all pairwise distances. Hence the desired linear embedding f exists. □

3 Review of Basic Properties of the Gaussian Distribution

For any distribution $\mathcal{D}$, we write $X \sim \mathcal{D}$ to say that X is a random variable with distribution $\mathcal{D}$. We write $\mathcal{N}(\alpha, \sigma^2)$ to denote the Gaussian distribution with mean α and variance σ^2.

We use the following properties, which we state mostly without a proof:

- Let $X \sim \mathcal{N}(\alpha, \sigma_1^2)$ and $Y \sim \mathcal{N}(\beta, \sigma_2^2)$ be two independent Gaussian variables. The distribution of their sum, $X + Y$, is $\mathcal{N}(\alpha + \beta, \sigma_1^2 + \sigma_2^2)$.
- Let $X \sim \mathcal{N}(0, \sigma^2)$ and $\alpha \in \mathbb{R}$. Then $\alpha X \sim \mathcal{N}(0, \alpha^2\sigma^2)$.
- Let $X \sim \mathcal{N}(0, \sigma^2)$. Then $\mathrm{E}[X^2] = \sigma^2$. Why? $E[X^2] = (E[X^2] - (E[X])^2) + (E[X])^2 = \mathrm{Var}(X) + 0 = \sigma^2$.
- Let $X \sim \mathcal{N}(0, 1)$ and $\lambda < \frac{1}{2}$. Then $\mathrm{E}[e^{\lambda X^2}] = (1 - 2\lambda)^{-1/2}$. This is known as the moment generating function of the chi-squared distribution. It is not too difficult to prove this via integration, using standard calculus tools.

4 Proof of Lemma 2

Let us start by introducing auxiliary notation. We write A_i to refer to the i-th row of A and $A_{i,j}$ to refer to the j-th entry in the A_i.

4.1 Useful properties

We first make an observation that the entries of Ax have a distribution that, as we will see later, is very convenient.

Lemma 3. *For any $i \in [d]$ and any $x \in \mathbb{R}^k$, the distribution of $A_i x$ is $\mathcal{N}(0, \|x\|^2)$.*

Proof. We have $A_{i,j} \sim \mathcal{N}(0,1)$, for each $j \in [k]$, and hence $A_{i,j}x_j \sim \mathcal{N}(0, x_j^2)$. Since $A_i x = \sum_{j=1}^k A_{i,j}x_j$, where each $A_{i,j}x_j$ is an independent random variable, we can write $A_i x \sim \mathcal{N}(0, \sum_{j=1}^k x_j^2)$, using basic properties of Gaussian variables. Since $\|x\|^2 = \sum_{j=1}^k x_k^2$, this finishes the proof of the lemma. □

We rely on the following concentration result about a sum of squares of random Gaussian variables. (This type of distribution is known as the chi-squared distribution.) We prove it later but for now we just state it here.

Lemma 4. *Let $X_1, \ldots, X_d$ be independent variables distributed according to $\mathcal{N}(0,1)$. For each $i \in [d]$, let $Z_i = X_i^2$. For any $\epsilon \in (0,1)$,*

$$\Pr\left(\left|d - \sum_{i=1}^{d} Z_i\right| > \epsilon d\right) \leq 2e^{-\frac{\epsilon^2 d}{12}}.$$

4.2 The proof

Proof of Lemma 2. We want to show that

$$(1-\epsilon)\|x\|^2 \leq \left\|\frac{1}{\sqrt{d}}Ax\right\|^2 \leq (1+\epsilon)\|x\|^2$$

holds with probability close to 1. This expression is equivalent to

$$(1-\epsilon)d \leq \left\|\frac{Ax}{\|x\|}\right\|^2 \leq (1+\epsilon)d,$$

which we obtain by multiplying the previous inequality by $\frac{d}{\|x\|^2}$. This can be restated as

$$\left|d - \left\|\frac{Ax}{\|x\|}\right\|^2\right| \leq \epsilon d$$

By Lemma 3, we know that the distribution of each entry in Ax is $\mathcal{N}(0, \|x\|^2)$. Therefore, the distribution of each entry in $\frac{Ax}{\|x\|}$ is $\mathcal{N}(0,1)$. Since the entries of $\frac{Ax}{\|x\|}$ are independent for a fixed vector x, $\left\|\frac{Ax}{\|x\|}\right\|^2$ is really a sum of squares of d independent random variables distributed according to $\mathcal{N}(0,1)$. The proof of the lemma then follows directly from the concentration result stated as Lemma 4. □

4.3 Proof of the concentration result

Proof of Lemma 4. In order to prove the desired bound, we prove separately that $\sum_{i=1}^{d} Z_i$ is bounded from above and from below with probability almost 1. (These proofs are almost identical.) Let us define p_{high} to be the probability that $\sum_{i=1}^{d} Z_i$ is too high. We introduce an additional variable $\lambda \in (0, \frac{1}{2})$, which we set later. We have

$$p_{\text{high}} = \Pr\left(\sum_{i=1}^{d} Z_i > d(1+\epsilon)\right) = \Pr\left(e^{\lambda \sum_{i=1}^{d} Z_i} > e^{d\lambda(1+\epsilon)}\right).$$

By applying Markov's inequality, we get

$$p_{\text{high}} \leq \frac{\mathrm{E}\left[e^{\lambda \sum_{i=1}^{d} Z_i}\right]}{e^{d\lambda(1+\epsilon)}} = \frac{\prod_{i=1}^{d} \mathrm{E}\left[e^{\lambda Z_i}\right]}{e^{d\lambda(1+\epsilon)}} = \frac{\left(\mathrm{E}\left[e^{\lambda Z_1}\right]\right)^d}{e^{d\lambda(1+\epsilon)}} = \frac{\left((1-2\lambda)^{-1/2}\right)^d}{e^{d\lambda(1+\epsilon)}} = \left((1-2\lambda)e^{2\lambda(1+\epsilon)}\right)^{-d/2},$$

where the first equality follows from the independence of Z_i's, the second follows from the fact that all of them have the same distribution, and the third one uses an identity mentioned in Section 3. Now set $\lambda = \frac{\epsilon}{2(1+\epsilon)}$. This value clearly lies in $(0, 1/2)$, since $\epsilon \in (0,1)$, and therefore our entire analysis so far holds. We obtain

$$p_{\text{high}} \leq \left((1+\epsilon)e^{-\epsilon}\right)^{d/2}.$$

From the Taylor series for $\ln(1+t)$, we have $\ln(1+t) \leq t - \frac{t^2}{2} + \frac{t^3}{3}$ for $t \in [0,1]$. Therefore,

$$p_{\text{high}} \leq \left(e^{\epsilon - \epsilon^2/2 + \epsilon^3/3 - \epsilon}\right)^{d/2} \leq \left(e^{-\epsilon^2/6}\right)^{d/2} = e^{-\frac{\epsilon^2 d}{12}}.$$

We now turn to the proof that $\sum_{i=1}^{d} Z_i$ is bounded from below with high probability. Let p_{low} be the probability that $\sum_{i=1}^{d} Z_i$ is too low. Following almost the same reasoning as before we get for any $\lambda > 0$:

$$p_{\text{low}} = \Pr\left(\sum_{i=1}^{d} Z_i < d(1-\epsilon)\right) = \Pr\left(e^{-\lambda \sum_{i=1}^{d} Z_i} > e^{-d\lambda(1-\epsilon)}\right) \leq \frac{\mathrm{E}\left[e^{-\lambda \sum_{i=1}^{d} Z_i}\right]}{e^{-d\lambda(1-\epsilon)}}$$
$$= \frac{\prod_{i=1}^{d} \mathrm{E}\left[e^{-\lambda Z_i}\right]}{e^{-d\lambda(1-\epsilon)}} = \frac{\left(\mathrm{E}\left[e^{-\lambda Z_1}\right]\right)^d}{e^{-d\lambda(1-\epsilon)}} = \frac{\left((1+2\lambda)^{-1/2}\right)^d}{e^{-d\lambda(1-\epsilon)}} = \left((1+2\lambda)e^{-2\lambda(1-\epsilon)}\right)^{-d/2}.$$

We now set $\lambda = \frac{\epsilon}{2(1-\epsilon)} > 0$. We get

$$p_{\text{low}} \leq \left((1-\epsilon)e^{\epsilon}\right)^{d/2}.$$

Using the Taylor series for $\ln(1-t)$, we have $\ln(1-t) \leq -t - \frac{t^2}{2}$ for $t \in [0,1]$. Hence

$$p_{\text{low}} \leq \left(e^{-\epsilon - \epsilon^2/2 + \epsilon}\right)^{d/2} \leq e^{-\frac{\epsilon^2 d}{4}} \leq e^{-\frac{\epsilon^2 d}{12}}.$$

By the union bound, the probability that $\sum_{i=1}^{d} Z_i$ is out of the desired range is therefore bounded by $p_{\text{high}} + p_{\text{low}} \leq 2e^{-\frac{\epsilon^2 d}{12}}$. □

5 Additional notes

Probability of success. The way our proof is written, it is focused on proving that the random linear transformation has the desired properties with non-zero probability. In practice it is likely that checking whether a given transform preserves all pairwise distances is not feasible. However, it is easy to show that the probability of finding a good transformation can be made $1 - n^{-t}$ for an arbitrarily large t by increasing the constant in the target dimension.

Oblivious embedding. A nice property of our linear transformation is that it is oblivious, i.e., we do not have to know the set of point to decided on what the transformation is. This could be useful in a setting in which data points arrive over time, and we have to process them as they appear. Contrast this with multidimensional scaling, which is a popular technique for visualizing data sets, but requires knowing the entire data set.

Simpler matrices and faster embeddings. Our construction used the Gaussian distribution for each entry of the transformation matrix. It is, in fact, possible to get the same guarantees by independently selecting each entry from the uniform distribution on $\{-1, 1\}$ (known as the Rademacher distribution). This is much simpler to implement then drawing entries from the Gaussian distribution. It also gives clear bounds on the range of any number that may appear when we compute the transformation.

One downside of the transformation we presented is that it requires $O(dk)$ time. It is possible to improve on this by using sparse matrices. In particular, it is possible to use matrices that have a fixed number of randomly selected non-zero entries in every column. Each entry is selected from the Rademacher distribution (with appropriate scaling). This leads to a significantly more complicated proof, but allows for a transformation that can be executed in time close to linear in the sum of the dimensions.

Optimality of the Johnson–Lindenstrauss lemma. It turns out that it is optimal (up to constant factors). It is possible to construct an arbitrarily large set of n points that cannot be embedded into $o(\epsilon^{-2} \log n)$ dimensions, while preserving all the pairwise distances up to a factor of roughly $1 + \epsilon$.

Dimensionality reduction for k-means and k-median. k-means and k-median are popular clustering methods. Given a set of n points in a high dimensional space, one could map it to $O(\epsilon^{-2} \log n)$ dimensions, using the transformation provided by the Johnson–Lindenstrauss lemma. This would preserve all pairwise distances up to a factor of $1 \pm \epsilon$ and would therefore preserve the cost of k-means and k-median up to the same factor. It turns out, however, that $O(\epsilon^{-2} \log(k/\epsilon))$ dimensions suffice to achieve this goal, despite the fact that this is in general not enough to preserve all pairwise distances.