On Johnson-Lindenstrauss Lemma

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be   Homepage: angms.science

First draft : July 22, 2017
Last update : October 31, 2018

Overview

What is Johnson-Lindenstrauss Lemma

What is Johnson-Lindenstrauss Lemma (JL Lemma) : a theorem related to dimensional reduction.

What is dimensional reduction (DR) : a process of reducing the number of variables (dimension) of the data vector with respect to some conditions. A common method of DR is projection.

What JL Lemma says : in Euclidean space, there always exists a low-distortion projection of points from high dimension to lower dimension.

What are the applications of JL Lemma :

- Dimension reduction
- Random projection
- Compressive sensing
- Graph embedding

What is low-distortion

The distance between any two points after the projection is distorted within a factor of $(1 \pm \epsilon)$, for any $0 < \epsilon < 1$.
i.e. the transform nearly-preserves the pairwise distances between points.

Given two points x, y, let $u = f(x)$ and $v = f(y)$ be the corresponding points after the projection $f(\,.\,)$. With a distortion factor ϵ, mathematically the word "low-distortion" means :

$$(1-\epsilon)\|x-y\|_2^2 \leq \|u-v\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2$$

or equivalently

$$1-\epsilon \leq \frac{\|u-v\|_2^2}{\|x-y\|_2^2} \leq 1+\epsilon$$

Thus, in more accurate sense, the projection f does not preserve the absolute difference but the *relative difference* (the ratio of distances before and after the transform)

The Johnson-Lindenstrauss Lemma

Theorem (JL Lemma). Given a tolerance value $0 < \epsilon < 1$, a number of data pints n, a number $k \geq \dfrac{4 \ln n}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}}$, then there exists a linear mapping $f : \mathbb{R}^m \rightarrow \mathbb{R}^k$ such that for a m-dimensional data matrix with n points : $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{m \times n}$, we have

$$(1 - \epsilon)\|x_i - x_j\|_2^2 \leq \|f(x_i) - f(x_j)\|_2^2 \leq (1 + \epsilon)\|x_i - x_j\|_2^2$$

for all $1 \leq i, j \leq n$

Notation:

- m : original dimension of the data
- n : number of data points
- ϵ : tolerance value, $0 \leq \epsilon \leq 1$
- k : dimension of projected data, we want $k < m$
- $f(\,.\,)$: the projection
- X : data matrix with n columns in $\mathbb{R}^m$
- $\|x_i - x_j\|_2^2$: pairwise l_2 distance between columns of X

The philosophy behind the proof of JL Lemma

- The proof is *constructive*.
  What we need to proof is the existence of a mathematical object (the projection f). A constructive proof is to demonstrates the existence of such mathematical object by creating or providing a method to create the object.
  We will see such f is $f(x) = \frac{1}{\sqrt{k}} Ax$
- The proof is *probabilistic*.
  - It uses many probabilistic techniques
  - It proves the existence of such mathematical object : we construct an probability space and show that a randomly chosen element in this space has desired properties with positive probability (probability > 0 implies such object exists).

i.e. We construct a random matrix $\frac{1}{\sqrt{k}} A$, show that such matrix satisfies the low-distortion property when it is applied to a vector x

The outline of the proof of JL Lemma

1. Construct a random projection over k dimensional subspaces.
   We will see that this projection is the matrix $\frac{1}{\sqrt{k}}A$
2. Prove the expected value of the l_2 distance of the random projection is equal to that of the original subspace:

$$\mathbb{E}\|\frac{1}{\sqrt{k}}Au\|_2^2 = \mathbb{E}\|u\|_2^2$$

3. Prove the variance of the l_2 distance is greater than the specified error bound only with a probability $\frac{2}{n^2}$:

$$\mathbb{P}\Big[\|\frac{1}{\sqrt{k}}Au\|_2^2 \notin \big[(1-\epsilon)\|u\|_2^2,\ (1+\epsilon)\|u\|_2^2\big]\Big] \leq \frac{2}{n^2}$$

4. Prove the union bound of this probability across all pairs of points is less than $1 - \frac{1}{n}$

The proof of JL Lemma - construct a random projection

Consider a matrix $A \in \mathbb{R}^{k \times m}$ that each entry are sampled independently from $N(0,1)$ (i.e. Independent and identically distributed)

$$A_{ij} \overset{iid}{\sim} N(0,1)$$

So $\mathbb{E}A_{ij}^2 = 1$ and $\mathbb{E}A_{ij}A_{ik} = 0$ for $j \neq k$.

Let the projection f be : $f(x) = \frac{1}{\sqrt{k}}Ax$.

Consider a fixed vector $u \in \mathbb{R}^m$.

Let $v = f(u) = \frac{1}{\sqrt{k}}Au \in \mathbb{R}^k$.

Note that $v \in \mathbb{R}^k$. i.e., for a vector u from the original m-dimensional space, v is the projected vector in the k-dimensional subspace.

The proof of JL Lemma - showing $\mathbb{E}\|v\|_2^2 = \|u\|_2^2$

Showing $\mathbb{E}\|v\|_2^2 = \|u\|_2^2$ means statistically the mapping f does not change the squared l_2-length of a vector

$$\mathbb{E}\|v\|_2^2 = \mathbb{E}\sum_{i=1}^{k} v_i^2 = \sum_{i=1}^{k} \mathbb{E}v_i^2$$

As $v = f(u) = \frac{1}{\sqrt{k}} Au$ so $v_i = \frac{1}{\sqrt{k}} \sum_j A_{ij} u_j$ and $v_i^2 = \frac{1}{k}(\sum_j A_{ij} u_j)^2$

$$\begin{aligned}\mathbb{E}\|v\|_2^2 &= \sum_{i=1}^{k} \mathbb{E}\frac{1}{k}(\sum_j A_{ij} u_j)^2 = \sum_{i=1}^{k} \frac{1}{k}\mathbb{E}(\sum_j A_{ij} u_j)^2 \\ &= \sum_{i=1}^{k} \frac{1}{k}\mathbb{E}(A_{i1}u_1 + A_{i2}u_2 + \ldots)(A_{i1}u_1 + A_{i2}u_2 + \ldots)\end{aligned}$$

As $\mathbb{E}A_{ij}A_{ik} = 0$ for $j \neq k$

$$\mathbb{E}\|v\|_2^2 = \sum_{i=1}^{k} \frac{1}{k}\mathbb{E}(A_{i1}^2 u_1^2 + A_{i2}^2 u_2^2 + \ldots + A_{ik}^2 u_k^2)$$

u is fixed vector so it can be taken out from $\mathbb{E}$ and $\mathbb{E}A_{ij}^2 = 1$.

$$\begin{aligned}\mathbb{E}\|v\|_2^2 &= \sum_{i=1}^{k} \frac{1}{k}(u_1^2 \underbrace{\mathbb{E}A_{i1}^2}_{1} + u_2^2 \underbrace{\mathbb{E}A_{i2}^2}_{1} + \ldots) \\ &= \sum_{i=1}^{k} \frac{1}{k}(u_1^2 + u_2^2 + \ldots) \\ &= \|u\|_2^2 \quad \square\end{aligned}$$

The proof of JL Lemma - show $\mathbb{P}\Big[\|v\|_2^2 \notin \big[(1-\epsilon)\|u\|_2^2,\ (1+\epsilon)\|u\|_2^2\big]\Big] \leq \frac{2}{n^2}$

Let x be the product between matrix A and unit vector $\hat{u} = \dfrac{u}{\|u\|_2^2}$.

$$x = A\frac{u}{\|u\|_2}$$

So the element of x will be $x_i = \frac{A_{i:}^T u}{\|u\|_2}$ where $A_{i:}$ is the i^{th} row of A. As u is fix and A_{ij} is iid, x_i are also iid.

We already have $v = \frac{1}{\sqrt{k}}Au$ so $Au = \sqrt{k}v$ and thus

$$x = \frac{\sqrt{k}v}{\|u\|_2} \quad \text{and} \quad \|x\|_2^2 = k\frac{\|v\|_2^2}{\|u\|_2^2}$$

In words: a vector x is defined that its norm-squared is the ratio of the norm-squared of v and norm-squared of u (times a constant k)

The proof of JL Lemma - show $\mathbb{P}\Big[\|v\|_2^2 \notin \big[(1-\epsilon)\|u\|_2^2, (1+\epsilon)\|u\|_2^2\big]\Big] \leq \frac{2}{n^2}$ ... 2

Consider $\mathbb{P}\Big[\|v\|_2^2 \geq (1+\epsilon)\|u\|_2^2\Big]$. As $\|x\|_2^2 = k\frac{\|v\|_2^2}{\|u\|_2^2}$ so $\|v\|_2^2 = \frac{\|x\|_2^2\|u\|_2^2}{k}$ and

$$\begin{aligned}
\mathbb{P}\Big[\|v\|_2^2 \geq (1+\epsilon)\|u\|_2^2\Big] &= \mathbb{P}\Big[\frac{\|x\|_2^2\|u\|_2^2}{k} \geq (1+\epsilon)\|u\|_2^2\Big] \\
&= \mathbb{P}\Big[\|x\|_2^2 \geq (1+\epsilon)k\Big] \\
\text{(with } \lambda \geq 0) &= \mathbb{P}\Big[\lambda\|x\|_2^2 \geq \lambda(1+\epsilon)k\Big] \\
&= \mathbb{P}\Big[\exp(\lambda\|x\|_2^2) \geq \exp(\lambda(1+\epsilon)k)\Big]
\end{aligned}$$

Markov's inequality $\mathbb{P}\Big[x \geq a\Big] \leq \dfrac{\mathbb{E}[x]}{a}$

$$\begin{aligned}
\mathbb{P}\Big[\|v\|_2^2 \geq (1+\epsilon)\|u\|_2^2\Big] &\leq \frac{\mathbb{E}[\exp(\lambda\|x\|_2^2)]}{\exp(\lambda(1+\epsilon)k)} \\
&= \frac{\mathbb{E}[e^{\lambda x_1^2 + \lambda x_2^2 + \ldots}]}{e^{\lambda(1+\epsilon)k}}
\end{aligned}$$

The proof of JL Lemma - show $\mathbb{P}\Big[\|v\|_2^2 \notin \big[(1-\epsilon)\|u\|_2^2,\ (1+\epsilon)\|u\|_2^2\big]\Big] \leq \frac{2}{n^2}$ ... 3

As x_i iid, $e^{\lambda x_1^2} = e^{\lambda x_2^2} = \ldots$

$$\begin{aligned}
\mathbb{P}\Big[\|v\|_2^2 \geq (1+\epsilon)\|u\|_2^2\Big] &\leq \frac{\prod_{i=1}^{k} \mathbb{E}[e^{\lambda x_i^2}]}{e^{\lambda(1+\epsilon)k}} \\
&= \frac{\Big(\mathbb{E}[e^{\lambda x_i^2}]\Big)^k}{e^{\lambda(1+\epsilon)k}} \\
&= \Big(\frac{\mathbb{E}[e^{\lambda x_i^2}]}{e^{\lambda(1+\epsilon)}}\Big)^k
\end{aligned}$$

For random variable $\theta \sim N(0,1)$

$$\mathbb{E}(e^{s\theta^2}) = \frac{1}{\sqrt{2\pi}} \int e^{s\theta^2} e^{-\theta^2/2} d\theta = \frac{1}{\sqrt{2\pi}} \int e^{-(1-2s)\theta^2/2} d\theta = \frac{1}{\sqrt{1-2s}}$$

$$\mathbb{P}\Big[\|v\|_2^2 \geq (1+\epsilon)\|u\|_2^2\Big] \leq \Big(\frac{1}{\sqrt{1-2\lambda} e^{\lambda(1+\epsilon)}}\Big)^k$$

The proof of JL Lemma - show $\mathbb{P}\Big[\|v\|_2^2 \notin \big[(1-\epsilon)\|u\|_2^2,\, (1+\epsilon)\|u\|_2^2\big]\Big] \leq \frac{2}{n^2}$ ... 4

Let $\lambda = \frac{\epsilon}{2(1+\epsilon)}$

$$\mathbb{P}\Big[\|v\|_2^2 \geq (1+\epsilon)\|u\|_2^2\Big] = \Big((1+\epsilon)e^{-\epsilon}\Big)^{k/2} = \Big(e^{\ln(1+\epsilon)}e^{-\epsilon}\Big)^{k/2}$$

Inequality trick: for $x \in (0,\, 1)$, $\ln(1+x)$ is bounded above by it's odd orders of Taylor series approximation:

$$\begin{aligned}
\ln(1+x) &\leq x \\
\ln(1+x) &\leq x - \frac{x^2}{2} + \frac{x^3}{3} \\
\ln(1+x) &\leq x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5}
\end{aligned}$$

So $e^{\ln(1+\epsilon)} \leq e^{\epsilon - \epsilon^2/2 + \epsilon^3/3}$ and $e^{\ln(1+\epsilon)}e^{-\epsilon} \leq e^{-\epsilon^2/2+\epsilon^3/3}$ and

$$\mathbb{P}\Big[\|v\|_2^2 \geq (1+\epsilon)\|u\|_2^2\Big] \leq \Big(e^{-\epsilon^2/2+\epsilon^3/3}\Big)^{k/2}$$

The proof of JL Lemma - show $\mathbb{P}\Big[\|v\|_2^2 \notin \big[(1-\epsilon)\|u\|_2^2,\,(1+\epsilon)\|u\|_2^2\big]\Big] \le \frac{2}{n^2}$ ... 5

$$\mathbb{P}\Big[\|v\|_2^2 \ge (1+\epsilon)\|u\|_2^2\Big] \quad \le \quad \left(e^{-\epsilon^2/2+\epsilon^3/3}\right)^{k/2}$$

In the beginning we assumed $k \ge \dfrac{4\ln n}{\frac{\epsilon^2}{2}-\frac{\epsilon^3}{3}}$, then we have

$$-\frac{\epsilon^2}{2}+\frac{\epsilon^3}{3} \le -\frac{4\ln n}{k}$$

So

$$\mathbb{P}\Big[\|v\|_2^2 \ge (1+\epsilon)\|u\|_2^2\Big] \quad \le \quad \left(e^{-\frac{4\ln n}{k}}\right)^{k/2} = e^{-2\ln n} = n^{-2}$$

Next we consider the other side $\mathbb{P}\Big[\|v\|_2^2 \le (1-\epsilon)\|u\|_2^2\Big]$

The proof of JL Lemma - show $\mathbb{P}\Big[\|v\|_2^2 \notin \big[(1-\epsilon)\|u\|_2^2,\ (1+\epsilon)\|u\|_2^2\big]\Big] \le \frac{2}{n^2}$ ... 6

$$
\begin{aligned}
\mathbb{P}\Big[\|v\|_2^2 \le (1-\epsilon)\|u\|_2^2\Big] &= \mathbb{P}\Big[\frac{\|x\|_2^2\|u\|_2^2}{k} \le (1-\epsilon)\|u\|_2^2\Big] \\
\text{(with } \lambda \ge 0) &= \mathbb{P}\Big[\lambda\|x\|_2^2 \le \lambda(1-\epsilon)k\Big] \\
&= \mathbb{P}\Big[\exp(-\lambda\|x\|_2^2) \ge \exp(-\lambda(1-\epsilon)k)\Big] \\
\text{Markov's inequality} &\le \frac{\mathbb{E}[\exp(-\lambda\|x\|_2^2)]}{\exp(-\lambda(1-\epsilon)k)} \\
\text{(as } x_i \text{ iid)} &= \frac{\prod_{i=1}^{k}\mathbb{E}[\exp(-\lambda x_i^2)]}{\exp(-\lambda(1-\epsilon)k)} \\
&= \Big(\frac{\mathbb{E}[\exp(-\lambda x_i^2)]}{\exp(-\lambda(1-\epsilon))}\Big)^k \\
\because \mathbb{E}(\exp(s\theta^2)) = \frac{1}{\sqrt{1-2s}} &\le \Big(\frac{1}{\sqrt{1+2\lambda}\exp(-\lambda(1-\epsilon))}\Big)^k \\
\text{let } \lambda = \frac{\epsilon}{2(1-\epsilon)} &\le \Big((1-\epsilon)e^{+\epsilon}\Big)^{k/2} = \Big(e^{\ln(1-\epsilon)}e^{\epsilon}\Big)^{k/2}
\end{aligned}
$$

The proof of JL Lemma - show $\mathbb{P}\Big[\|v\|_2^2 \notin \big[(1-\epsilon)\|u\|_2^2,\ (1+\epsilon)\|u\|_2^2\big]\Big] \leq \frac{2}{n^2}$ ... 7

Inequality trick: for $x \in (0,\, 1)$, $\ln(1-x) < -x - \frac{x^2}{2} + \frac{x^3}{3}$

$$\mathbb{P}\Big[\|v\|_2^2 \leq (1-\epsilon)\|u\|_2^2\Big] \quad \leq \quad \left(e^{-\epsilon^2/2+\epsilon^3/3}\right)^{k/2}$$

$$k \geq \frac{4\ln n}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}} \implies -\frac{\epsilon^2}{2} + \frac{\epsilon^3}{3} \leq -\frac{4\ln n}{k}$$

$$\mathbb{P}\Big[\|v\|_2^2 \leq (1-\epsilon)\|u\|_2^2\Big] \quad \leq \quad \left(e^{-\frac{4\ln n}{k}}\right)^{k/2} = n^{-2}$$

With

$$\mathbb{P}\Big[\|v\|_2^2 \geq (1+\epsilon)\|u\|_2^2\Big] \quad \leq \quad n^{-2}$$

We have

$$\mathbb{P}\Big[\|v\|_2^2 \notin \big[(1-\epsilon)\|u\|_2^2,\ (1+\epsilon)\|u\|_2^2\big]\Big] \quad \leq \quad \frac{2}{n^2}$$

The proof of JL Lemma - prove of union bound

We have

$$\mathbb{P}\Big[\|v\|_2^2 \notin \big[(1-\epsilon)\|u\|_2^2,\ (1+\epsilon)\|u\|_2^2\big]\Big] \quad \leq \quad \frac{2}{n^2}$$

Now if we set $u = x_i - x_j$, and $f(x_i) - f(x_j) = \dfrac{A(x_i - x_j)}{\sqrt{k}}$, we have

$$\mathbb{P}\Big[\|\frac{A(x_i - x_j)}{\sqrt{k}}\|_2^2 \notin \big[(1-\epsilon)\|x_i - x_j\|_2^2,\ (1+\epsilon)\|x_i - x_j\|_2^2\big]\Big] \quad \leq \quad \frac{2}{n^2}$$

let E_i be such event for one pair of points x_i, x_j. So for all $\binom{n}{2}$ pairs of points the union bound for the probability is

$$\mathbb{P}\Big[\bigcup_i^n E_i\Big] \quad \leq \quad \sum_{i=1}^n \mathbb{P}(E_i) = \frac{n(n-1)}{2}\frac{2}{n^2} = 1 - \frac{1}{n} \quad \square$$

Hence the probability of the existence of the projection that fulfil the property of low-distortion for all pairs of points is $1 - \frac{1}{n} > 0$, so such projection exists.

$k(n, \epsilon)$, m are independent of each other

k is a function of n and ϵ : $k(n, \epsilon) \geq \dfrac{4 \ln n}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}}$.

In application we can use $k = \left\lceil \dfrac{4 \ln n}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}} \right\rceil$.

For example, $k(n = 10, \epsilon = 0.1) = 1974$: there is a f projects 10 data into a 1974-dimensional space that all the pairwise relative distances between the data are bounded within $[0.9,\ 1.1]$.

It is very important to note that : $k(n, \epsilon)$ and m are independent of each other ! That means we need $m > 1974$ to achieve "dimension reduction" $k < m$ for such n and ϵ.

But m can also be smaller than k: if original data points are 3-dimensional vectors, such low-distortion transform is actually a dimension expansion.

$k(n, \epsilon)$, m are independent of each other ... 2

The reason is that, JL Lemma is a **worst-case guarantee**. It consider **all** the pairwise distances between data points: it also consider the two data points that are **furthest away from each other**.

In this case the low-distortion guarantee is built on top of these two points that are furthest away from each other, and hence k growth very large for a small ϵ.

As k is independent of m and actual data input, **JL Lemma works for any data** X (and any m) as long as they have the same (n, ϵ).

The idea behind: once the error tolerance ϵ and the number of data point n are fixed, there is a fixed probability distribution on maps which will work (with high probability) for any vectors.

Comparing JL Lemma VS SVD

Both JL Lemma and Singular Value Decomposition (SVD) are dimension reduction technique, but they are different:

- inputs of JL Lemma are the number of data point and tolerance ϵ
- inputs of SVD are data matrix and the dimension of reduction
- JL Lemma is independent of dimension m
- JL Lemma is **minimax** or **worst-case**
- SVD is **on average** : find the approximation that has lowest rank, where rank is the collective behaviour of all the data points
- As SVD is on average, so it works better with *future data*: if the training data and future data are coming from the same source, then SVD should work well for the future data. But JL only works for the current data, increasing n will only increase k

Last page - summary

- JL Lemma: given a tolerance $0 < \epsilon < 1$, number of data pints n, then with $k \geq \frac{4 \ln n}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}}$, there exists a linear mapping f such that for any data matrix $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{m \times n}$, we have $(1-\epsilon)\|x_i - x_j\|_2^2 \leq \|f(x_i) - f(x_j)\|_2^2 \leq (1+\epsilon)\|x_i - x_j\|_2^2$ for all i, j
- The way to construct the f is surprisingly simple, $f = \frac{1}{\sqrt{k}} A$, where A is random matrix : $A_{ij} \overset{iid}{\sim} N(0,1)$
- The proof of JL Lemma
- $k(n, \epsilon)$ and m are independent of each other in the JL Lemma

End of document