SEEM5020 Algorithms for Big Data
Dimension Reduction with the Johnson-Lindenstrauss Lemma

Sibo WANG

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Prevalence of Vector Data

Recent advances in deep learning encode various unstructured data, like text, video, and image, into high-dimensional vectors. Many real-world applications now directly deal with these vector data.

- YouTube video to vector data for video recommendation
- Airbnb maps their property descriptions to vectors for search ranking
- Alibaba represents products as vectors for product recommendation
- ...
Definition 1 (Dimension reduction)

Given a set of points $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ in $\mathbb{R}^d$ (where $d$ is the original high dimension), the goal of dimension reduction is to find a function $f : \mathbb{R}^d \to \mathbb{R}^k$ such that $k < d$, and each point $\mathbf{x}_i$ is mapped to a point in $\mathbb{R}^k$ with preservation of certain properties (e.g., pairwise distances) in the lower-dimensional space.

Why do we need dimension reduction?

- **Efficiency**: Reducing the dimensionality can speed up algorithms without significantly compromising the quality of results.
- **Curse of dimensionality**: High-dimensional spaces can be counterintuitive; algorithms can behave poorly as dimension increases.
- **Storage**: Reducing the storage space.
- ...
Preserving the Pairwise Euclidean Distance

Problem 1

Given a set of points $X = \{x_1, x_2, \ldots, x_n\}$ in $\mathbb{R}^d$, we want to find a function $f : \mathbb{R}^d \to \mathbb{R}^k$ such that $k < d$, and each point $x_i$ is mapped to a point in $\mathbb{R}^k$, such that for any $i, j$ we have:

$$(1 - \varepsilon)\|x_i - x_j\|_2 \leq \|f(x_i) - f(x_j)\|_2 \leq (1 + \varepsilon)\|x_i - x_j\|_2 \quad (1)$$

Applications:

- Approximate all-pair distances.
- Distance-based clustering like $k$-means, DBSCAN, Hierarchical Clustering.
- Linear-regression
- ...
The Johnson-Lindenstrauss Lemma [1]

Lemma 1 (JL Lemma)

There is a linear mapping function $f$ that maps $x_i$ ($1 \leq i \leq n$) with $d$ dimensions to $x'_i$ with $k = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensions that satisfies Equation (1).

Some tricks that we might start with based on our previous lectures:

- Random Gaussian projection:
- Random Sign projection;
- Random coordinate selection
- ...
The (ε, δ)-JL Property

Theorem 1 ((ε, δ)-JL Property)

Let \( \Pi \) be a \( k \times d \) random matrix with each entry being a normalized random Gaussian variable, i.e., \( \Pi_{i,j} \sim \frac{1}{\sqrt{k}} \mathcal{N}(0,1) \). Then by setting \( f(x) \) as \( f(x) = \Pi \cdot x \), the following holds for an arbitrary vector \( x \):

\[
(1 - \varepsilon) \|x\|_2^2 \leq \|\Pi x\|_2^2 \leq (1 + \varepsilon) \|x\|_2^2.
\]

with \( 1 - \delta \) probability when \( k = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)\).

Remark 1: The above theorem actually provides a tighter bound as required by Equation (1). It is simple to obtain the bound following that of Equation (1) by taking a square root on each side.

Remark 2: Given the above theorem, by setting \( \delta' = \frac{\delta}{\binom{n}{2}} \) in Theorem 1, we can guarantee that Equation (1) holds for an arbitrary pair.
Analysis of the \((\varepsilon, \delta)\)-JL Property

Define \(G = \sqrt{k} \cdot \Pi\), i.e., each entry in \(G\) is a Gaussian random variable following \(\mathcal{N}(0, 1)\). Let

\[
G = \begin{pmatrix}
g_1^T \\
g_2^T \\
\vdots \\
g_k^T
\end{pmatrix}
\]

Then, for an arbitrary vector \(x\), let \(w = G \cdot x\) and \(w_i\) be the \(i\)-th entry in \(w\). Then, we have

\[w_i = g_i^T \cdot x.\]

Accordingly,

\[
\|\Pi \cdot x\|_2^2 = \left\| \frac{1}{\sqrt{k}} G \cdot x \right\|_2^2 = \frac{1}{k} \left\| (g_1^T x, g_2^T x, \cdots, g_k^T x) \right\|_2^2 = \frac{1}{k} \sum_{i=1}^k w_i^2.
\]
Let $x = (x_1, x_2, \cdots, x_d)^T$ and $g_i = (g_{i,1}, g_{i,2}, \cdots, g_{i,d})^T$. For $w_i$, then it is

$$w_i = x_1 \cdot g_{i,1} + x_2 \cdot g_{i,2} + \cdots + x_d \cdot g_{i,d}.$$ 

Notice that $g_{i,j}$ is a Gaussian random variable from $\mathcal{N}(0,1)$.

**Lemma 2 (Gaussian Distribution Properties (i))**

$X \sim \mathcal{N}(0, \sigma^2)$. Then, given a positive $a$, $a \cdot X \sim \mathcal{N}(0, a^2 \cdot \sigma^2)$.

**Lemma 3 (Gaussian Distribution Properties (ii))**

$X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2) \rightarrow Z = X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Therefore, $x_j \cdot g_{i,j}$ is a Gaussian random variable following the distribution $\mathcal{N}(0, x_j^2)$ and $w_j$ is a Gaussian random variable following the distribution $\mathcal{N}(0, x_1^2 + x_2^2 + \cdots + x_d^2) = \mathcal{N}(0, \|x\|_2^2)$. 
Define $z_i = \frac{w_i}{\|x\|_2}$. Then, we have that

- $z_i \sim \mathcal{N}(0, 1)$;
- $z_i^2 = \frac{w_i^2}{\|x\|_2^2}$;
- $\mathbb{E}[w_i^2] = \|x\|_2^2$.

Besides, summing all $k$ terms of $w_i^2$ together, we have:

$$\mathbb{E}\left[\frac{1}{k} \sum_{i=1}^{k} w_i^2 \right] = \|x\|_2^2.$$

Hence, we only need to care if the random variable $Y = \frac{1}{k} \sum_{i=1}^{k} w_i^2$ concentrates on its expectation or not. Rewrite $Y$ as follows:

$$Y = \frac{1}{k} \sum_{i=1}^{k} w_i^2 = \|x\|_2^2 \cdot \frac{\sum_{i=1}^{k} z_i^2}{k}.$$
Definition 2 ($\chi^2$ random variable)

A $\chi^2$-squared random variable $\chi_k^2$ with $k$ degrees of freedom can be represented as the sum $\sum_{i=1}^{k} z_i^2$ where $z_i$ are i.i.d. from $\sim \mathcal{N}(0,1)$.

Theorem 2 (Concentration bound of $\chi_k^2$ random variables [2])

For a $\chi_k^2$ random variable $\sum_{i=1}^{k} z_i^2$, the following bound holds:

$$\mathbb{P}[|\frac{\sum_{i=1}^{k} z_i^2}{k} - 1| \geq \varepsilon] \leq 2e^{-k \cdot \varepsilon^2 / 8}$$

Then, we can get that when $k = \frac{8 \log(2/\delta)}{\varepsilon^2}$,

$$\mathbb{P}[|\frac{\sum_{i=1}^{k} z_i^2}{k} - 1| \geq \varepsilon] \leq \delta \iff \mathbb{P}[|Y - \|x\|_2^2| \geq \varepsilon \|x\|_2^2] \leq \delta.$$

This finishes the proof.
Lemma 4

For a $\chi^2_k$ random variable $Z = \sum_{i=1}^{k} Z_i^2$, the following bound holds:

$$\mathbb{P}[Z > (1 + \varepsilon)^2 \cdot k] \leq e^{-\frac{3k \cdot \varepsilon^2}{4}}$$

Firstly, let $t \in (0, 0.5)$, we have that:

$$\mathbb{P}[Z > (1 + \varepsilon)^2 k] = \mathbb{P}[e^{tZ} > e^{(1+\varepsilon)k \cdot t}] \leq \frac{\mathbb{E}[e^{tZ}]}{e^{(1+\varepsilon)^2 k \cdot t}} \quad (2)$$

Also note that:

$$\mathbb{E}[e^{tZ}] = \mathbb{E}[e^{t\sum_{i=1}^{k} Z_i^2}] = \prod_{i=1}^{k} \mathbb{E}[e^{t \cdot Z_i^2}].$$
We consider $\mathbb{E}[e^{t\cdot z^2_i}]$ separately.

\[
\mathbb{E}[e^{t\cdot z^2_i}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{ty^2} \cdot e^{-\frac{y^2}{2}} \, dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(1-2t)y^2}{2}} \, dy
\]

\[
= \frac{1}{\sqrt{2\pi \sqrt{1-2t}}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} \, dz = \frac{1}{\sqrt{1-2t}}.
\]

Then, we have $\mathbb{E}[e^{tZ}] = (1-2t)^{-k/2}$. Putting it into Equation (2), we have:

\[
\mathbb{P}[Z > (1+\varepsilon)^2 \cdot k] \leq e^{-(1+\varepsilon)^2 \cdot k \cdot t \cdot (1-2t)^{-k/2}}. \tag{3}
\]

Choose the minimum of $f(t) = e^{-(1+\varepsilon)^2 \cdot k \cdot t \cdot (1-2t)^{-k/2}}$. To do this, we may first take $f_1(t) = \ln f(t)$ and then take the derivative on $f_1(t)$. We can derive that: $t = \left(1 - \frac{1}{(1+\varepsilon)^2}\right)/2 \in (0, 0.5)$. 
Another Proof via Moment Generating Function (Cont.)

Putting \( t = \left( 1 - \frac{1}{(1+\varepsilon)^2} \right)/2 \) into Equation (3), we have:

\[
\mathbb{P}[Z > (1 + \varepsilon)^2 k] \leq e^{-(1+\varepsilon)^2 \cdot k \cdot t \cdot (1 - 2t)^{-k/2}} = e^{-k(\varepsilon + \varepsilon^2/2 - \ln(1+\varepsilon))}. \tag{4}
\]

Using the fact that \( \ln(1 + x) \leq x - \frac{x^2}{4} \), we have that

\[
\mathbb{P}[Z > (1 + \varepsilon)^2 k] \leq e^{-\frac{3k\varepsilon^2}{4}}.
\]

This finishes the proof of Lemma 4
We also have the following theorem for the lower tail.

**Lemma 5**

*For a \( \chi_k^2 \) random variable \( Z = \sum_{i=1}^{k} Z_i^2 \), the following bound holds:*

\[
\mathbb{P}[Z < (1 - \varepsilon)^2 \cdot k] \leq e^{-\frac{k\varepsilon^2}{2}}
\]
The previous solution uses a dense matrix for projection. Do we really need such a dense matrix to preserve the pairwise distance? The next theorem provides an alternative solution with almost $\frac{2}{3}$ entries of the projection matrix to be zero.

**Theorem 3 (Achlioptas [3])**

Let $\Pi$ be a $k \times d$ matrix where each entry is i.i.d. draw from the following distribution:

$$
\Pi_{ij} = \begin{cases} 
\frac{\sqrt{3}}{\sqrt{k}} & \text{with probability } \frac{1}{6}, \\
0 & \text{with probability } \frac{2}{3}, \\
-\frac{\sqrt{3}}{\sqrt{k}} & \text{with probability } \frac{1}{6}.
\end{cases}
$$

Then, $\Pi$ has the $(\varepsilon, \delta)$-JL property.
Theorem 4 (Kane and Nelson[4])

There exist distributions on $\Pi \in \mathbb{R}^{k \times d}$ such that, when $k = O(\log n/\epsilon^2)$ and each row of dimension $d$ includes $O\left(\frac{\log n}{\epsilon}\right)$ non-zero entries, the $(\epsilon, \delta)$-JL property is satisfied.
Approximate near neighbor search (ANNS) under $L_2$-norm. First, reduce the dimension $d$ to $k$ and then apply the ANNS solution, e.g., tree structure, LSH, Production Quantization, and graph-based indices, on the newly mapped space.

Approximate linear regression. Given a set of data points $X = \{x_1, x_2, \ldots, x_n\}$ in $\mathbb{R}^d$ and corresponding target values $y = \{y_1, y_2, \ldots, y_n\}$. The goal is to find $a \in \mathbb{R}^d$ such that:

$$\min_{a \in \mathbb{R}^d} \sum_{i=1}^{n} (x_i \cdot a - y_i)^2 = \min_{a \in \mathbb{R}^d} \|Xa - y\|_2^2$$

By JL Lemma, we can solve the approximate version:

$$\min_{a \in \mathbb{R}^d} \|\Pi Xa - \Pi y\|_2.$$
References

William B. Johnson and Joram Lindenstrauss.
Extensions of lipschitz maps into a hilbert space.

Lecture notes on concentration bounds.

Dimitris Achlioptas.
Database-friendly random projections: Johnson-lindenstrauss with binary coins.

Daniel M. Kane and Jelani Nelson.
Sparser johnson-lindenstrauss transforms.