Johnson-Lindenstrauss Lemma Beyond Euclidean Geometry

Speaker: Kevin Lu

Chengyuan Deng, Jie Gao, Kevin Lu, Feng Luo, Cheng Xin

Rutgers University

October 2025

Classical Multi-Dimensional Scaling (cMDS)

Given an input matrix of Euclidean distances between n points in $\mathbb{R}^d$, recover the coordinates of the points. [Torgerson 1958]

- Euclidean distance matrix (EDM) $D = \{d_{ij}^2\}$
- Centering: $B = -\frac{1}{2}CDC$, where $C = I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$.
- B is the Gram matrix of D. We find its orthogonal diagonalization $U^T \Lambda U$, $\Lambda = diag(\lambda_1, ...\lambda_n)$
- Since D is Euclidean, B is positive semi-definite.
- $X = \sqrt{Diag(\Lambda)}U$ is a $n \times n$ dimensional matrix of rank d (or less), specifying the coordinates of n points in $\mathbb{R}^n$.

Pseudo-Euclidean MDS

- We still start from Gram Matrix $B = U^T \Lambda U$. Now, let $X = \sqrt{Diag(|\Lambda|)}U$.

Pseudo-Euclidean MDS

- We still start from Gram Matrix $B = U^T \Lambda U$. Now, let $X = \sqrt{Diag(|\Lambda|)}U$.
- We redefine Gram Matrix to be $X^T A X$ where $A = diag(sgn(\lambda_1), ...sgn(\lambda_n))$. In this way, B is still the Gram matrix of X.

Pseudo-Euclidean MDS

- We still start from Gram Matrix $B = U^T \Lambda U$. Now, let $X = \sqrt{Diag(|\Lambda|)}U$.
- We redefine Gram Matrix to be $X^T A X$ where $A = diag(sgn(\lambda_1), ...sgn(\lambda_n))$. In this way, B is still the Gram matrix of X.
- In doing this, we have changed the inner product to :

$$\Phi(u, v) = \sum_{i=1}^{p} u_i v_i - \sum_{i=p+1}^{p+q} u_i v_i.$$

with the addition and subtraction corresponding to the positive and negative eigenvalues.

Pseudo-Euclidean Johnson Lindenstrauss

- Now with $X = (x_1, x_2, ...x_n)$ in Pseudo-Euclidean space, we can find $P^p X^p$ and $P^q X^q$ where P^p and P^q are random Gaussian matrices of dimensions $p' \times p$ and $q' \times q$
- By the Johnson Lindenstrauss lemma:

$$(1-\epsilon)\|x_i^p - x_j^p\|_E^2 \leq \|P^p x_i^p - P^p x_j^p\|_E^2 \leq (1+\epsilon)\|x_i^p - x_j^p\|_E^2$$

$$(1-\epsilon)\|x_i^q - x_j^q\|_E^2 \leq \|P^q x_i^q - P^q x_j^q\|_E^2 \leq (1+\epsilon)\|x_i^q - x_j^q\|_E^2$$

- Combined, we get:

$$\|Px_i - Px_j\|_{p',q'}^2 - \epsilon\|Px_i - Px_j\|_E^2 \leq \|Px_i - Px_j\|_{p',q'}^2 \leq$$

$$\|Px_i - Px_j\|_{p',q'}^2 + \epsilon\|Px_i - Px_j\|_E^2$$

Comparing Euclidean and Pseudo-Euclidean distances

- Consider vectors $v \in \mathbb{R}^{p,q}$ chosen randomly where each coordinate is chosen from the same distribution whose square has constant variance and mean.
- If $q < \frac{C-1}{C+1}p$, then with high probability:

$$\|v\|_E \leq C\|v\|_{p,q}$$

Power Distance MDS

- Given any $n \times n$ symmetric hollow dissimilarity matrix, D, we can rewrite $D = E + 4r^2(I - J)$ where $r \in \mathbb{R}^+$, E is a Euclidean distance matrix, I is an $n \times n$ identity matrix, $J = \mathbf{1}_n \mathbf{1}_n^T$.
- More specifically E is a Euclidean distance matrix if and only if $2r^2 \geq |e_n|$ where e_n is the least eigenvalue of (D).
- Using cMDS on E, we then get a set of points X. If we redefine the square distances to be:

$$\|x_i - x_j\|^2 - 4r^2$$

we recover D as the square distance matrix for X.

Power Distance JL

- Then, we can find random Gaussian projection PX
- From Johnson Lindenstruass, we have the error bound:

$$(1-\varepsilon)\|x_i - x_j\|^2 - 4\epsilon r^2 \leq \|Px_i - Px_j\|^2 \leq (1+\varepsilon)\|x_i - x_j\|^2 + 4\epsilon r^2$$

Experiments

Dataset	Simplex	Ball	Brain	Breast	Renal	MNIST	CIFAR10	Email	Facebook	Mooc
Size	1000	1000	130	151	143	1000	1000	986	4039	7047
$\#\{\lambda < 0\}$	900	887	53	59	57	454	399	465	1566	268
Metric	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓

Table 1: Non-Euclidean/Non-metric Datasets used in experiments

Limitations and Future Work

- A limitation is that both of these methods require finding eigenvalues of a matrix which brings the runtime to $O(n^3)$
- Future work could find meaningful ways to apply these methods such as reducing runtime or using the embeddings in downstream tasks
- Further geometric characterization of power distances and Pseudo-Euclidean spaces can also be useful to apply these methods