t-SNE’s spectral regime

Noah Bergam

Advised by Nakul Verma

Outline of the Talk

1. Introduction to t-SNE
2. Introduction to Spectral Clustering
3. Cai and Ma (2022): the connection

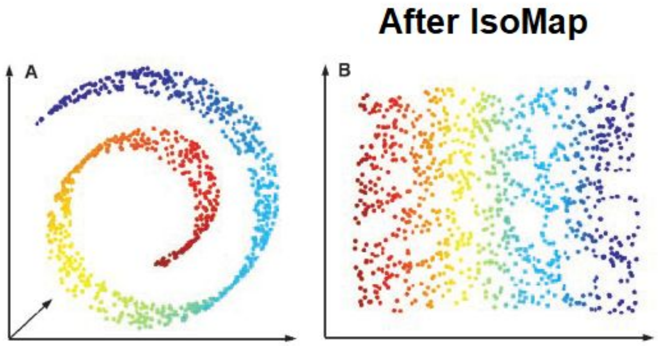
Dimensionality Reduction

High dimensional data is everywhere

- Images (#pixels)
- Language (#vocabulary)
- Single-cell transcriptomics (#genes)

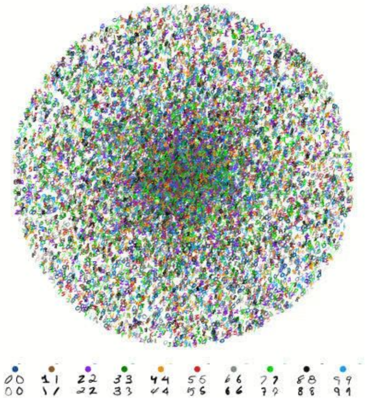
Oftentimes, it has low-dimensional **intrinsic structure** (e.g. a *manifold*).

Problem: Find a map into a lower-dimensional space, which preserves "information/structure"



The t-SNE approach (van der Maaten 2007)

1. Start with $\mathcal{X} = \{x_1, ..., x_n\} \subset \mathbb{R}^d$.
2. Randomly initialize the corresponding low-dimensional representations ("embeddings") $\mathcal{Y} = \{y_1, ..., y_n\} \subset \mathbb{R}^2$.
3. Iteratively update the $\mathcal{Y}$ embeddings, to match the local structure of $\mathcal{X}$.



How do we characterize “structure”?

Affinity matrix P associated with X.

For $i \neq j$, define

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)} \qquad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

Gaussian distribution

Affinity matrix Q associated with Y.

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

Cauchy (Student-T) distribution

Cost Function and Updates

P and Q are discrete probability distributions

We compute their “distance”

$$\mathrm{KL}(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

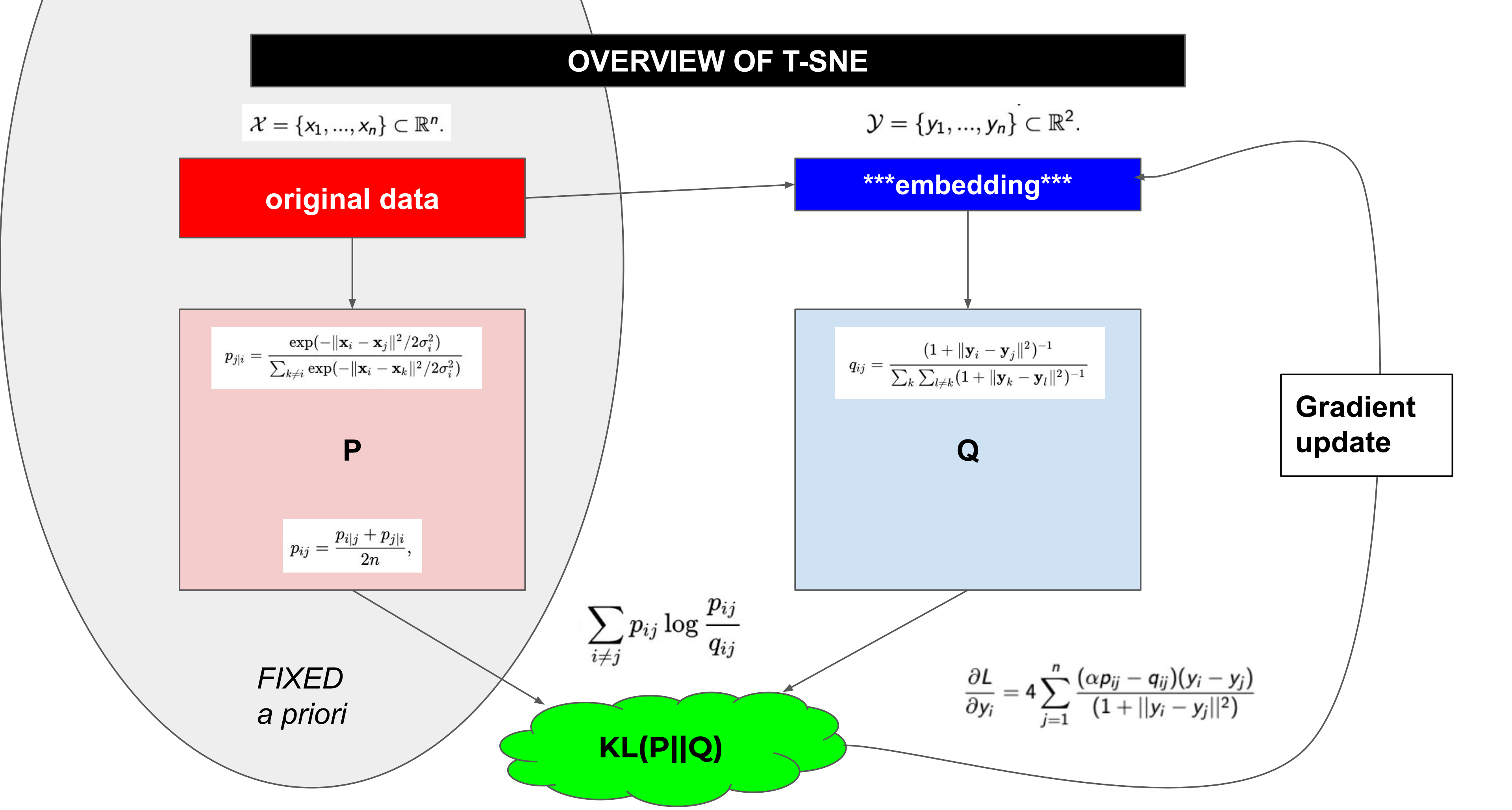
We **update the embeddings** according to gradient descent.

$$\frac{\partial KL(P||Q)}{\partial y_i} = 4 \sum_{j \neq i}^{n} \frac{(\alpha p_{ij} - q_{ij})(y_i - y_j)}{(1 + ||y_i - y_j||^2)}$$

(alpha is the “early exaggeration” parameter. Helps experimentally.)

“Dynamical Systems Interpretation”

$$
\begin{aligned}
\frac{dC}{dy_i} &= 4 \sum_{j=1, j\neq i}^{n} (p_{ij} - q_{ij})(1 + ||y_i - y_j||^2)^{-1}(y_i - y_j) \\
&= 4 \sum_{j=1, j\neq i}^{n} (p_{ij} - q_{ij}) q_{ij} Z (y_i - y_j) \\
&= 4\Big(\sum_{j \neq i} p_{ij} q_{ij} Z (y_i - y_j) - \sum_{j \neq i} q_{ij}^2 Z (y_i - y_j) \Big) \\
&= 4(F_{attraction} + F_{repulsion})
\end{aligned}
$$


OVERVIEW OF T-SNE
$\mathcal{X} = \{x_1, ..., x_n\} \subset \mathbb{R}^n.$
$\mathcal{Y} = \{y_1, ..., y_n\} \subset \mathbb{R}^2.$
original data
embedding
$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)}$
P
$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n},$
$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$
Q
Gradient update
$\sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$
FIXED
a priori
KL(P||Q)
$\frac{\partial L}{\partial y_i} = 4 \sum_{j=1}^{n} \frac{(\alpha p_{ij} - q_{ij})(y_i - y_j)}{(1 + \|y_i - y_j\|^2)}$

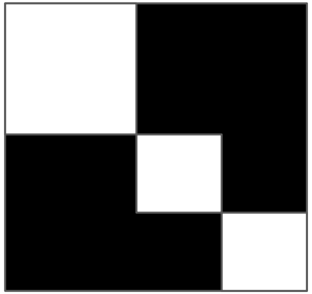
Spectral Dimensionality Reduction

1. Start with $\mathcal{X} = \{x_1, ..., x_n\} \in \mathcal{M}_{d\times n}$.
2. Construct an adjacency matrix $A_{\mathcal{X}}$ corresponding to some kind "similarity graph" on $\mathcal{X}$ (like k-nearest neighbors, or affinity matrix)
3. Compute the eigenvectors of $L(A_{\mathcal{X}})$, the graph Laplacian.
4. Construct $\mathcal{Y} = \{y_1, ..., y_n\} \in \mathcal{M}_{k\times n}$, where the rows are the k lowest eigenvectors.

Example

- e.g. 2-nearest neighbors
- $\mathcal{X} = \{(1,3),(1,1),(2,0),(-2,-2),(-3,-3),(-5,0)\}$:

$$\left(\begin{array}{ccc|ccc} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{array}\right)$$



"Block matrix," indicative of cluster structure

We want to use spectral decomposition to detect clusters of points.

The Graph Laplacian

The heart of **spectral graph theory**; many nice properties

- Analogous to the Laplace operator in calcu∇^2s:

Operates on a graph G.

- The adjacency matrix records whether
- The degree matrix (diagonal) records how many edges on a given node

Formula:

$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

Cai and Ma (2022): the spectral regime

1. **Rewrite** the t-SNE gradient update in matrix form.
2. Find conditions for when the update matrix is roughly constant. This is a **power iteration.**
3. Show that the power iteration converges.

1. $y_k = A_k y_{k-1}$
2. $A_k = A$. Therefore $y_k = A^k y_0$.
3. $\lim_{k\to\infty} A^k y_0$?

Cai and Ma (2022)

"Adjacency Matrix"

$$S_{ij}^{(k)}(\alpha) = \frac{\alpha p_{ij} - q_{ij}^{(k)}}{1 + ||y_i^{(k)} - y_j^{(k)}||^2}$$

Rewrite the t-SNE update.

$$y_i^{(k+1)} = y_i^{(k)} + h \sum_{1 \le j \le n, j \ne i} (y_j^{(k)} - y_i^{(k)}) S_{ij}^{(k)}(\alpha), \quad i = 1, ..., n,$$

Look at the row space of the embedding.

$$\boldsymbol{y}_\ell^{(k+1)} = [\mathbf{I}_n - h\mathbf{L}(\mathbf{S}_\alpha^{(k)})]\boldsymbol{y}_\ell^{(k)}, \quad \ell = 1, 2,$$

Graph
Laplacian!

The path to POWER ITERATIONS

$$\boldsymbol{y}_\ell^{(k+1)} = [\mathbf{I}_n - h\mathbf{L}(\mathbf{S}_\alpha^{(k)})]\boldsymbol{y}_\ell^{(k)}, \quad \ell = 1, 2,$$

1) Original.

$$\boldsymbol{y}_\ell^{(k+1)} \approx [\mathbf{I}_n - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)]\boldsymbol{y}_\ell^{(k)}, \quad \ell = 1, 2,$$

2) Roughly constant adjacency matrix

$$\boldsymbol{y}_\ell^{(k+1)} \approx [\mathbf{I}_n - h\mathbf{L}(\alpha\mathbf{P} - \mathbf{H}_n)]^k\boldsymbol{y}_\ell^{(0)}.$$

3) Power iterations

$$\mathbf{H}_n = \tfrac{1}{n(n-1)}(\mathbf{1}_n\mathbf{1}_n^\top - \mathbf{I}_n),$$

Question: Where do these power iterations lead?

Answer: Power iterations lead to the null space of L(P)!

Let R be the dimension of the null space of L(P)

Let U be a n by R matrix, whose columns are the orthogonal basis for the null space of L(P).

$$\boldsymbol{y}_\ell^{(k)} \approx \mathbf{U}\mathbf{U}^\top \boldsymbol{y}_\ell^{(0)}, \quad \ell \in [2].$$

The Laplacian null-space records clusters…

Consider well-clustered data (P effectively a block matrix!)

Proposition 6 (Laplacian null space) *Suppose* $\mathbf{A} \in \mathbb{R}^{n \times n}$ *is symmetric and well conditioned. Then the smallest eigenvalue of the Laplacian* $\mathbf{L}(\mathbf{A})$ *is* 0 *and has multiplicity* R*, and the associated eigen subspace is spanned by* $\{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_R\}$ *where for each* $r \in \{1, ..., R\}$*,*

$$[\boldsymbol{\theta}_r]_j = \begin{cases} 1/\sqrt{n_r} & \text{if the } j\text{-th node belongs to the } r\text{-th component} \\ 0 & \text{otherwise} \end{cases},$$

and n_r *is the number of nodes in the* r*-th connected component. In particular, up to possible permutation of coordinates, any vector* $\mathbf{u}$ *in the null space of* $\mathbf{L}(\mathbf{A})$ *can be expressed as*

$$\mathbf{u} = \frac{a_1}{\sqrt{n_1}} \begin{bmatrix} \mathbf{1}_{n_1} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} + \frac{a_2}{\sqrt{n_2}} \begin{bmatrix} \mathbf{0} \\ \mathbf{1}_{n_2} \\ \vdots \\ \mathbf{0} \end{bmatrix} + ... + \frac{a_R}{\sqrt{n_R}} \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{1}_{n_R} \end{bmatrix}, \tag{17}$$

for some $a_1, ..., a_R \in \mathbb{R}$*.*

Hence, under certain conditions, we know exactly where the embeddings are going…

Theorem 7 (Implicit clustering and early stopping) *Suppose the similarity* $\mathbf{P}$ *and the tuning parameters* (α, h, k) *satisfy (T1.D) and (T2.D), and the initialization satisfies (I1) and (I2). Then there exists some permutation matrix* $O \in \mathbb{R}^{n \times n}$ *such that, for* $\ell \in [2]$,

$$\lim_{(k,n)\to\infty} \frac{\|\boldsymbol{y}_\ell^{(k)} - O\mathbf{z}_\ell\|_2}{\|\boldsymbol{y}_\ell^{(0)}\|_2} = 0, \tag{18}$$

where

$$\mathbf{z}_\ell = (\underbrace{z_{\ell 1}, ..., z_{\ell 1}}_{n_1}, \underbrace{z_{\ell 2}, ..., z_{\ell 2}}_{n_2}, ..., \underbrace{z_{\ell R}, ..., z_{\ell R}}_{n_R})^\top \in \mathbb{R}^n, \tag{19}$$

and $z_{\ell r} = \boldsymbol{\theta}_r^\top \boldsymbol{y}_\ell^{(0)}/\sqrt{n_r}$ *for* $r \in [R]$.

Conclusion

t-SNE is powerful but not very well-understood

Spectral clustering is well-understood

Cai and Ma show a deep connection between t-SNE and spectral clustering.

Question (Linderman): Is t-SNE just spectral clustering is disguise? It seems to perform better, so there should be more to this story…

Works Cited

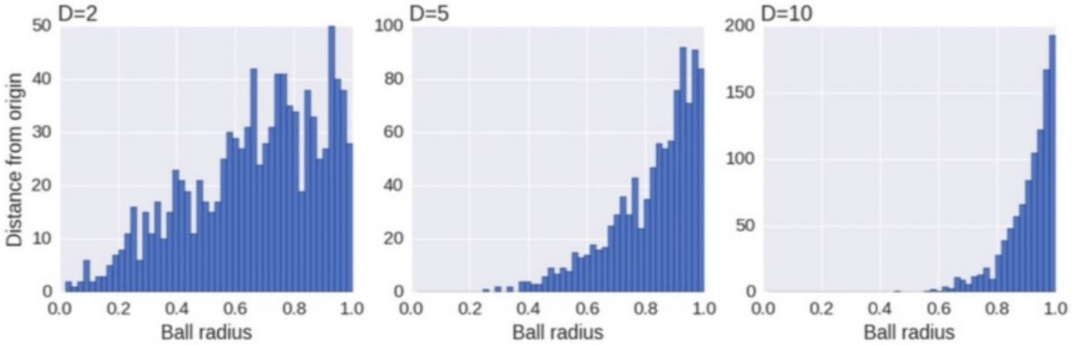
Cai and Ma, *Theoretical Foundations of t-SNE for Visualizing High-Dimensional Clustered Data* (2022)

Van der Maaten and Hinton, *Visualizing Data using t-SNE* (2008)

Ulrike von Luxburg, *A Tutorial on Spectral Graph Theory* (2007)

Problem with SNE: "crowding problem"

SNE suffers from the "crowding problem": The area of the 2D map that is available to accommodate moderately distant data points will not be large enough compared with the area available to accommodate nearby data points.



Unified Framework of Linear Dimensionality Reduction

Discussion

We can put most linear dimensionality reduction algorithms in a unified framework. Essentially, they are all special cases of Kernel-PCA.

- PCA: $K = X^T X$(Linear Kernel).
- Classical-MDS: $K = \frac{-1}{2} H D^{Euclidean} H$ where H is the centering matrix.
- Isomap: $K = \frac{-1}{2} H D^{Geodesic} H$.
- LLE: once W is learned, $K = M^{-1}$ or $K = (\lambda_{max} I - M)$, where $M = (I - W)(I - W)^T$. (Difference is in the scale of coordinate of the embedding. $K = \wedge^{1/2} V$).
- LE: $K = L^{-1}$ or $K = (\lambda_{max} I - L)$ and the result is also off in the scale of coordinate of the embedding as LLE.