



UNIVERSITY OF  
ARKANSAS

# tSNE

Jiahui Chen

Department of Mathematical Sciences  
University of Arkansas

# Why tSNE and UMAP?

- Aims to solve the problems of PCA
  - Non-linear scaling to represent changes at different levels
  - Optimal separation in 2-dimensions
- T-Distributed Stochastic Neighbour Embedding (tSNE)
- Uniform Manifold Approximation and Projection (UMAP)

# Dimensionality Reduction

- The Dimensionality Reduction focus on preserving distances. A cost function should be developed to measure this reduction.
- Ideas for Dimensionality Reduction
  - Distance preservation
  - Topology preservation
  - Information preservation

tSNE is a distance-based method but tends to preserve topology

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^h\} \rightarrow \mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n \in \mathbb{R}^l\}$$

$$\min_{\mathbf{Y}} C(\mathbf{X}, \mathbf{Y})$$

# Stochastic Neighbour Embedding

SNE converts Euclidean distances to similarities, that can be interpreted as probabilities. It computes pair-wise similarities.

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

$$q_{j|i} = \frac{\exp(-\| \mathbf{y}_i - \mathbf{y}_j \|^2)}{\sum_{k \neq i} \exp(-\| \mathbf{y}_i - \mathbf{y}_k \|^2)}$$

$$p_{j|i} = 0, q_{j|i} = 0$$

$\sigma_i$  is either set by hand or found by a binary search for the value of  $\sigma_i$  that makes the entropy of the distribution over neighbors equal to  $\log k$

# Kullback-Leiber Divergence

SNE converts Euclidean distances to similarities, that can be interpreted as probabilities. It computes pair-wise similarities.

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)} \text{ and } q_{j|i} = \frac{\exp(-\|\mathbf{y}_i - \mathbf{y}_j\|^2)}{\sum_{k \neq i} \exp(-\|\mathbf{y}_i - \mathbf{y}_k\|^2)}$$

$P_i = \{p_{1|i}, p_{2|i}, \dots, p_{n|i}\}$  and  $Q_i = \{q_{1|i}, q_{2|i}, \dots, q_{n|i}\}$  are the distributions on the neighbors of datapoint  $i$ .

Kullback-Leiber Divergence (KL) compares two distributions

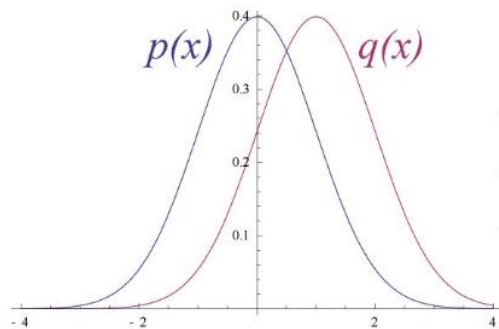
$$C = \sum_i KL(P_i \parallel Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

**Minimization**

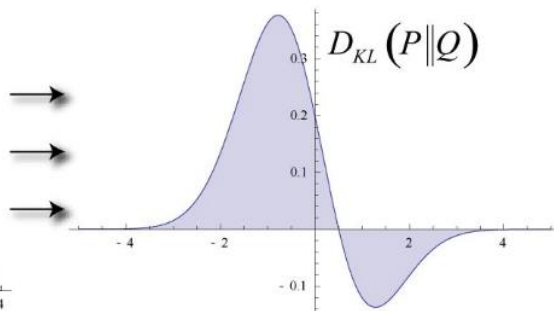
KL divergence is asymmetric and always positive

$$\frac{dC}{d\mathbf{y}_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(\mathbf{y}_i - \mathbf{y}_j)$$

# Kullback-Leiber Divergence

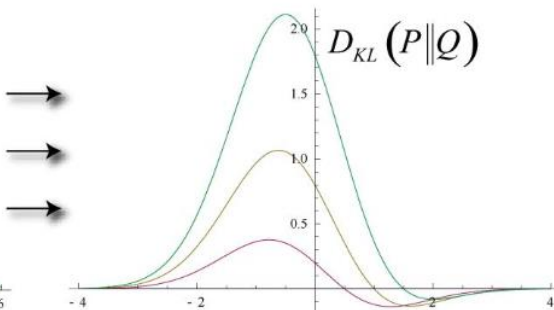
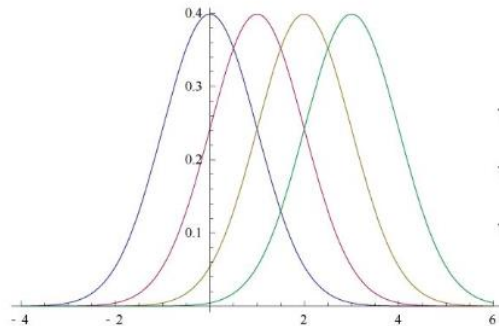


Original Gaussian PDF's



KL Area to be Integrated

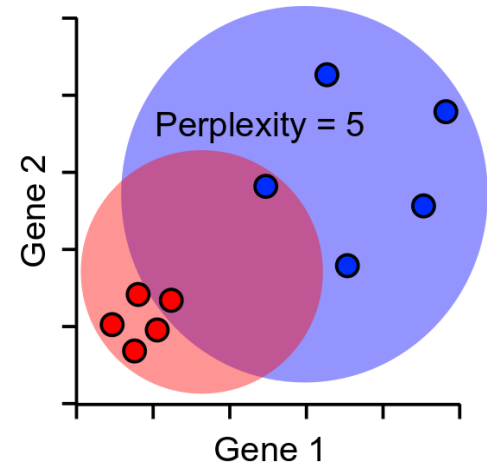
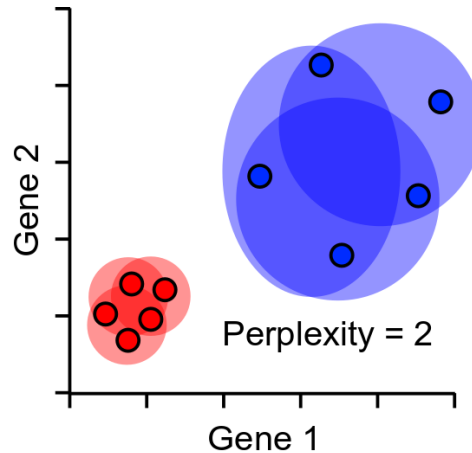
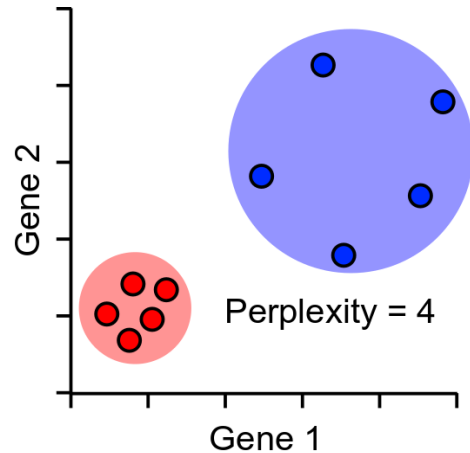
Measures the similarity between two probability distributions & it is asymmetric



$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

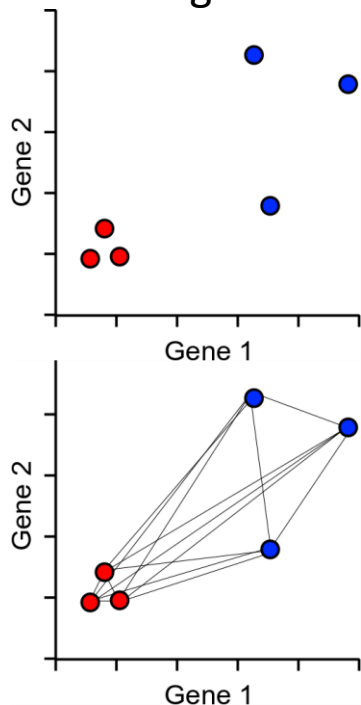
# Perplexity

- Perplexity = expected number of neighbours within a cluster. It can be thought of as a guess about the number of close neighbours each point has.
- Distances scaled relative to perplexity neighbours



# Perplexity

For the distance  $d_{ij}^2 = -\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2$  in  $p_{j|i}$ ,  $\sigma_i$  is either set by hand or found by a binary search for the value of  $\sigma_i$  that makes the entropy of the distribution over neighbors equal to  $\log k$ .

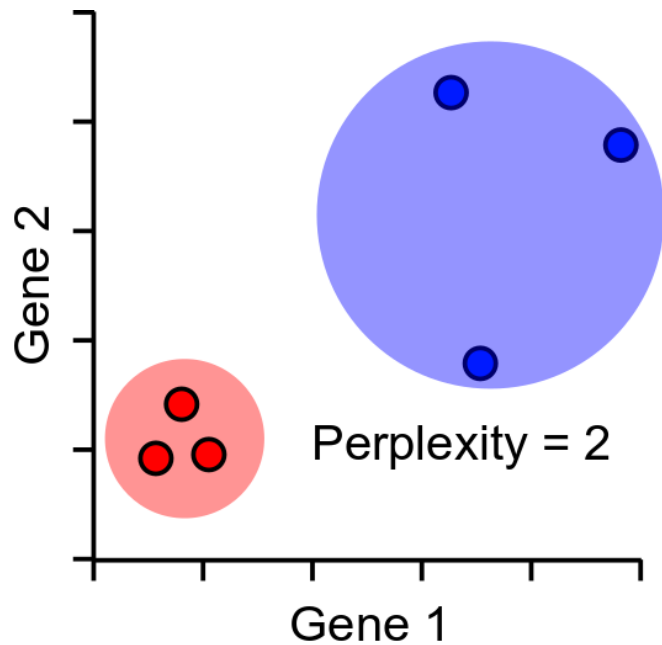


	0	10	10	295	158	153
	9	0	1	217	227	213
	1	8	0	154	225	238
	205	189	260	0	23	45
	248	227	246	44	0	54
	233	176	184	41	36	0



# Perplexity

For the distance  $d_{ij}^2 = -\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2$  in  $p_{j|i}$ ,  $\sigma_i$  is either set by hand or found by a binary search for the value of  $\sigma_i$  that makes the entropy of the distribution over neighbors equal to  $\log k$ .



	0	4	6	586	657	836
	4	0	4	815	527	776
	9	3	0	752	656	732
	31	28	29	0	4	7
	31	24	25	4	0	7
	40	37	32	8	8	0

# Shannon entropy

For the distance  $d_{ij}^2 = -\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2$  in  $p_{j|i}$ ,  $\sigma_i$  is either set by hand or found by a binary search for the value of  $\sigma_i$  that makes the entropy of the distribution over neighbors equal to  $\log k$ .

Here,  $k$  - is the effective number of local neighbors or “perplexity” and is chosen by hand.

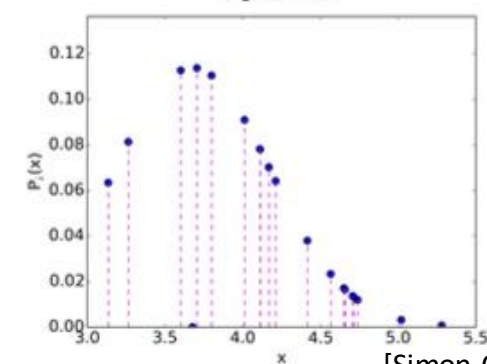
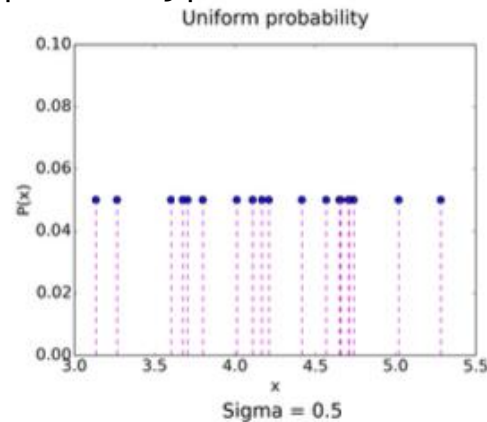
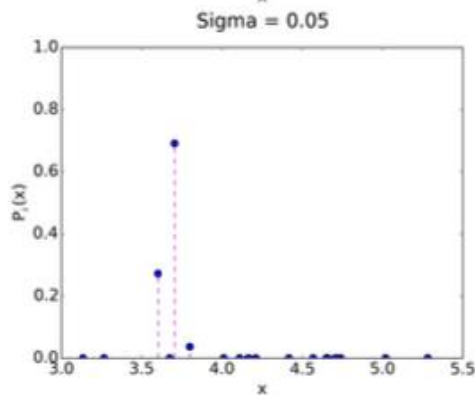
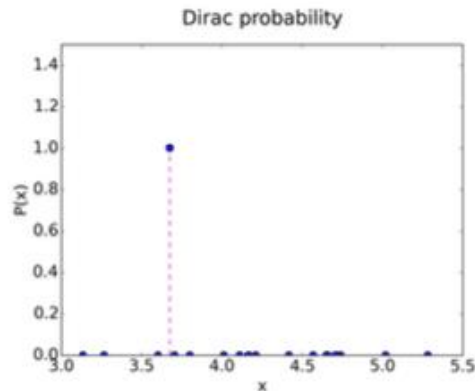
$$\text{Perplexity}(P_i) = 2^{H(P_i)}$$

where  $H(P_i)$  is the Shannon entropy of the conditional probability distribution  $P_i$

$$H(P_i) = - \sum_j p_{j|i} \log_2 p_{j|i}$$

# Shannon entropy

$$\text{Perplexity}(P_i) = 2^{H(P_i)}, H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$



# Shannon entropy

For the distance  $d_{ij}^2 = -\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2$  in  $p_{j|i}$ ,  $\sigma_i$  is either set by hand or found by a binary search for the value of  $\sigma_i$  that makes the entropy of the distribution over neighbors equal to  $\log k$ .

Here,  $k$  - is the effective number of local neighbors or “perplexity” and is chosen by hand.

The binary search adjusts  $\sigma_i$  until the perplexity of the conditional distribution  $P_i$  is approximately equal to the user-defined perplexity.

If the entropy is too high (the distribution is too spread out),  $\sigma_i$  needs to be decreased, leading to a less dispersed distribution.

Conversely, if the entropy is too low,  $\sigma_i$  needs to be increased.

# Symmetric SNE

- Kullback-Leiber Divergence of SNE

$$C = \sum_i KL(P_i \parallel Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

where,  $p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)}$  and  $q_{j|i} = \frac{\exp(-\|\mathbf{y}_i - \mathbf{y}_j\|^2)}{\sum_{k \neq i} \exp(-\|\mathbf{y}_i - \mathbf{y}_k\|^2)}$

- Asymmetric  $\rightarrow$  Symmetric, HOW?
- $p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)} \rightarrow p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$  and  $q_{j|i} \rightarrow q_{ji}$
- $C = KL(P \parallel Q)$ , and  $\frac{dC}{dy_i} = 4 \sum_j (p_{ij} - q_{ij})(\mathbf{y}_i - \mathbf{y}_j)$
- Faster Computation

# tSNE

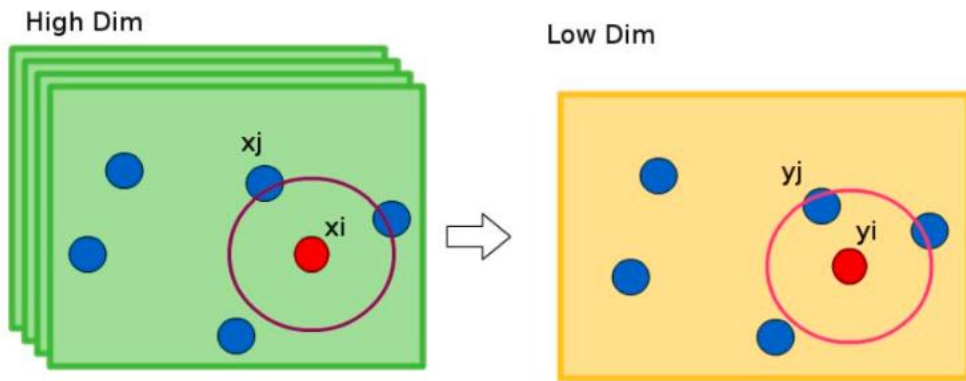
- Use heavier tail distribution than Gaussian in low-dim space (t-Distribution)

$$q_{ji} \propto (1 + \| \mathbf{y}_i - \mathbf{y}_j \|^2)^{-1}$$

- Why Student-t Distribution?

# Student-t Distribution

- The crowding problem: when embedding neighbors from a high-dim space into a low-dim space, there is too little space near a point for all of its close-by neighbors.
- Suppose data is intrinsically high dimensional
- We try to model the local structure of this data in the map
- Result: Dissimilar points have to be modeled as too far apart in the map!



# tSNE

- Use heavier tail distribution than Gaussian in low-dim space (t-Distribution)

$$q_{ij} \propto (1 + \| \mathbf{y}_i - \mathbf{y}_j \|^2)^{-1}$$

- $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$  and  $q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|\mathbf{y}_i - \mathbf{y}_k\|^2)^{-1}}$
- The gradient:  $\frac{dC}{d\mathbf{y}_i} = 4 \sum_j (p_{ij} - q_{ij})(\mathbf{y}_i - \mathbf{y}_j) (1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}$
- Even Faster Computation
- Better Behavior



# Discussion

- tSNE aims to maintain the local structure of the data by preserving the local neighbor relationships.
- PCA is a linear algorithm that projects the data onto a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate.

# References

- Simon Andrews, Babraham Bioinformatics
- Simon Carbonnelle, Université Catholique de Louvain, ICTEAM