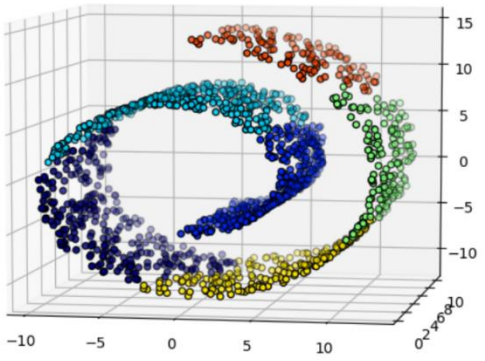


Unit 18: t-SNE Algorithm

Final Topic of the Course:

We will wrap up the new content in this course by discussing the algorithm and the theory behind the t-SNE algorithm.

Purpose of this Lecture:

We will introduce the algorithm and the theory behind the t-SNE algorithm.

In this lecture we will cover the following topics.

- Stochastic neighbor embedding.
- Drawbacks of stochastic neighbor embedding.
- What's different in SNE algorithm vs. t-SNE algorithm?
- T-SNE algorithm – general
- Full t-sne algorithm – with gradient descent algorithm.

Additional Resources

van der Maaten, L.J.P.; Hinton, G.E. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9:2579-2605, 2008.
https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

STOCHASTIC NEIGHBOR EMBEDDING (SNE)

Input:

- Dataset: $\mathbf{X_{m\times n}} = \begin{bmatrix} \boldsymbol{x_{1*}} \\ \boldsymbol{x_{2*}} \\ \vdots \\ \boldsymbol{x_{m*}} \end{bmatrix}$ comprised of *m* objects, where each object has a complete set of n attributes.
- Perplexity Values: $Perplexity(P_i)$ for *i=1,…,m*

Algorithm

- **Step 1: Define a projected/mapped 2-d coordinates matrix *of decision variables*:** $\mathbf{Y_{m\times 2}} = \begin{bmatrix} \boldsymbol{y_{1*}} \\ \boldsymbol{y_{2*}} \\ \vdots \\ \boldsymbol{y_{m*}} \end{bmatrix}$.

- **<u>Step 2</u>: Create a similarity matrix $P = [p_{j|i}]$ between each of the objects in $\mathbf{X_{m\times n}}$.**

<u>(i,j) entries represents</u>:

similarity between x_{i*} to x_{j*}

Read ahead for how we pick σ_i.

<u>(i,j) entries mathematically</u>:

$$p_{j|i} = \frac{\exp\left(-\frac{\left||x_{i*} - x_{j*}\right||^2}{2\sigma_i^2}\right)}{\sum_{k\neq i}\exp\left(-\frac{\left||x_{i*} - x_{k*}\right||^2}{2\sigma_i^2}\right)}$$

<u>(j,i) entries graphical interpretation:</u>

$$p_{j|i} = \frac{\exp\left(-\frac{\left||x_{i*} - x_{j*}\right||^2}{2\sigma_i^2}\right)}{\sum_{k\neq i}\exp\left(-\frac{\left||x_{i*} - x_{k*}\right||^2}{2\sigma_i^2}\right)} = \frac{\frac{1}{\sigma_i\sqrt{2\pi}}\exp\left(-\frac{(dist(x_{i*}, x_{j*}) - 0)^2}{2\sigma_i^2}\right)}{\sum_{k\neq i}\frac{1}{\sigma_i\sqrt{2\pi}}\exp\left(-\frac{(dist(x_{i*}, x_{k*}) - 0)^2}{2\sigma_i^2}\right)} = \frac{f(dist(x_{i*}, x_{j*}))}{\sum_{k\neq i} f(dist(x_{i*}, x_{k*}))}$$

<u>Interpretation of f(x):</u>

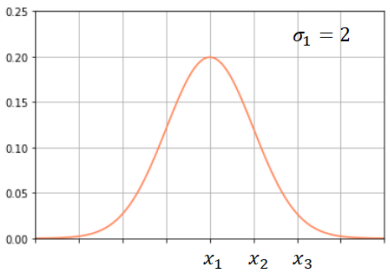
- $f(x)$ is the pdf of the _____ distribution with:
  - mean = _____
  - Standard deviation = _____

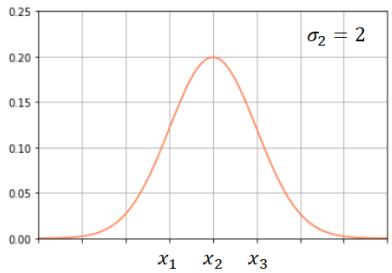
<u>Interpretation of $f(dist(x_{i*}, x_{j*}))$:</u>

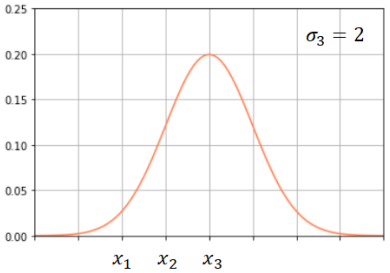
$f(dist(x_{i*}, x_{j*})) = g(dist(x_{i*}, x_{j*}))$

- $g(dist(x_{i*}, x_{j*}))$ is the _____ curve, centered at _____ with standard deviation _____

<u>Ex:</u> Generate the similarity matrix for the three 1-D data points shown below.







Properties of $\boldsymbol{P = [p_{j|i}]}$

- This matrix _____ to be symmetric.
- The rows in this matrix form a _____.

Step 3: Create a similarity matrix $\boldsymbol{Q = [q_{j|i}]}$ between each of the decision variable objects in $\mathbf{Y_{m\times 2}}$.

- **(j,i) entries represents**:
  similarity between $\boldsymbol{y_{j*}}$ to $\boldsymbol{y_{i*}}$

- **(j,i) entries mathematically:**

$$q_{j|i} = \frac{\exp\left(-\left||y_{i*} - y_{j*}\right|\right|^2\right)}{\sum_{k\neq i}\exp\left(-\left||y_{i*} - y_{k*}\right|\right|^2\right)}$$

- **(j,i) entries graphical interpretation:**

$$q_{j|i} = \frac{\exp\left(-\left\|\boldsymbol{y_{i*}} - \boldsymbol{y_{j*}}\right\|^2\right)}{\sum_{k\neq i}\exp\left(-\left\|\boldsymbol{y_{i*}} - \boldsymbol{y_{k*}}\right\|^2\right)} = \frac{\frac{1}{\left(\frac{1}{2}\right)\sqrt{2\pi}}\exp\left(-\frac{\left(dist\left(\boldsymbol{y_{i*}}, \boldsymbol{y_{j*}}\right) - 0\right)^2}{2\cdot\left(\frac{1}{2}\right)^2}\right)}{\sum_{k\neq i}\frac{1}{\left(\frac{1}{2}\right)\sqrt{2\pi}}\exp\left(-\frac{\left(dist\left(\boldsymbol{y_{i*}}, \boldsymbol{y_{k*}}\right) - 0\right)^2}{2\cdot\left(\frac{1}{2}\right)^2}\right)} = \frac{h\left(dist\left(\boldsymbol{y_{i*}}, \boldsymbol{y_{j*}}\right)\right)}{\sum_{k\neq i} h\left(dist\left(\boldsymbol{y_{i*}}, \boldsymbol{y_{k*}}\right)\right)}$$

Interpretation of h(x):

- $h(x)$ is the pdf of the _____ distribution with:
  - mean = _____
  - Standard deviation = _____

Interpretation of $h\left(dist\left(\boldsymbol{y}, \boldsymbol{y_{j*}}\right)\right)$:

$$h\left(dist\left(\boldsymbol{y_{i*}}, \boldsymbol{y_{j*}}\right)\right) = \bar{h}\left(dist\left(\boldsymbol{y_{i*}}, \boldsymbol{y_{j*}}\right)\right)$$

- $\bar{h}\left(dist\left(\boldsymbol{y_{i*}}, \boldsymbol{y_{j*}}\right)\right)$ is the _____ curve, centered at _____ with standard deviation _____

Properties of $\boldsymbol{Q = [q_{j|i}]}$

- This matrix _____ to be symmetric.
- The rows in this matrix form a _____.

Local Structure Preservation Idea:

- If the mapped points in $\mathbf{Y_{m\times 2}}$ correctly model the *similarity* between the high dimensional data points in $\mathbf{X_{m\times n}}$, then we would expect:

- **<u>Step 4</u>: Find the optimal values of the coordinates in $\mathbf{Y_{m\times 2}}$ that minimize the sum of the Kullback-Leibler (KL) divergence over all data points.**

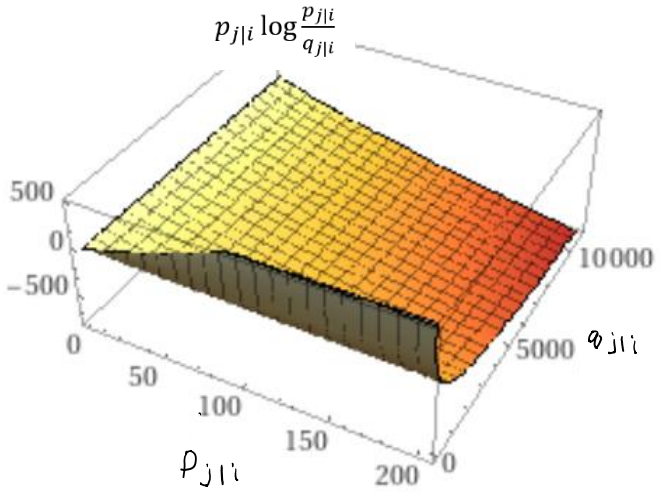
  <u>Optimization Problem:</u> $min_{\mathrm{Y_{m\times 2}}} C = min_{\mathrm{Y_{m\times 2}}} \sum_i KL(\boldsymbol{P_{i*}}||\boldsymbol{Q_{i*}}) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$

  <u>Common Algorithm</u>: Gradient descent methods

Downsides of this optimization problem:

- The optimization problem will work harder to avoid scenario _____ errors, at the expense of not caring about scenario _____ errors as much.
- The algorithm will focus harder on preserving the _____ structure of the original dataset, at the expense of preserving the _____ structure of the original dataset.
-

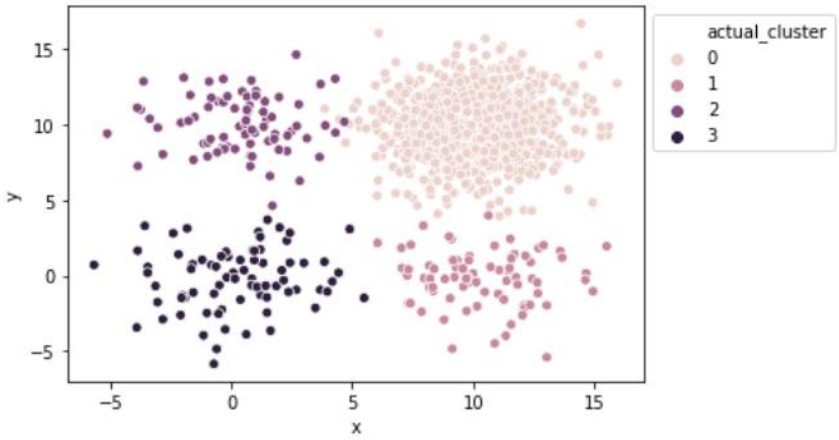
Ex: Below is a visualization of the (i,j) expression in the objective function for the SNE algorithm. With this in mind, which scenario will lead to a higher cost in the objective function?



- **Scenario a**: Widely separated data points $\boldsymbol{x_{i*}}$ and $\boldsymbol{x_{j*}}$ and nearby map points $\boldsymbol{y_{i*}}$ and $\boldsymbol{y_{j*}}$**, or**

- **Scenario b**: nearby data points $\boldsymbol{x_{i*}}$ and $\boldsymbol{x_{j*}}$ and widely separated map points $\boldsymbol{y_{i*}}$ and $\boldsymbol{y_{j*}}$**?**

Additional Consideration: **How to pick σ_i?**



If all objects in the SNE algorithm used the same standard deviation $\sigma = \sigma_i$ in the calculation of $p_{j|i}$, then the objects in cluster 0 would have _____ close neighbors than the objects in the other 3 clusters.

Goal: Pick σ_i to Adapt to Different Cluster Sparsities

Clusters of different sparsities need different values of σ_i to effectively preserved the local pairwise distances of all the objects in the cluster.

- If we have a cluster that is more dense, then we want the objects in this cluster to have values of σ_i that are _____ when calculating $p_{j|i}$.
- If we have a cluster that is more sparse, then we want the objects in this cluster to have values of σ_i that are _____ when calculating $p_{j|i}$.

Solution: Perform a binary search for σ_i that satisfies:

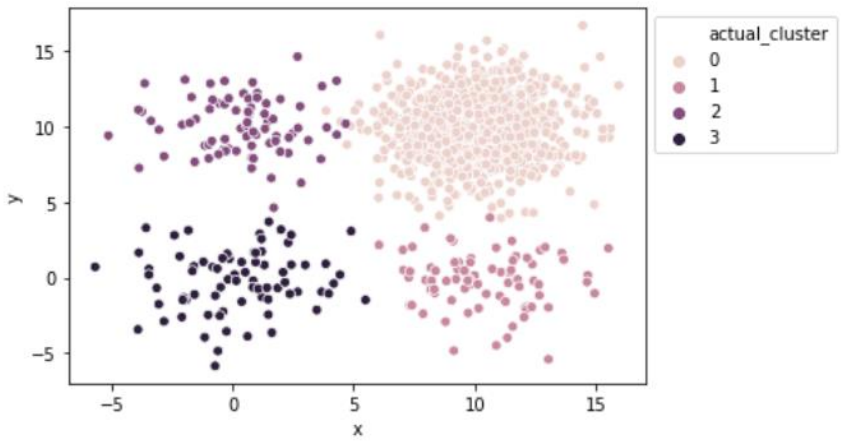
$$\_\_\_\_\_ = Perplexity(P_i) = 2^{-\sum_j p_{j|i} \log_2 p_{j|i}}$$

What $Perplexity(P_i)$represents:

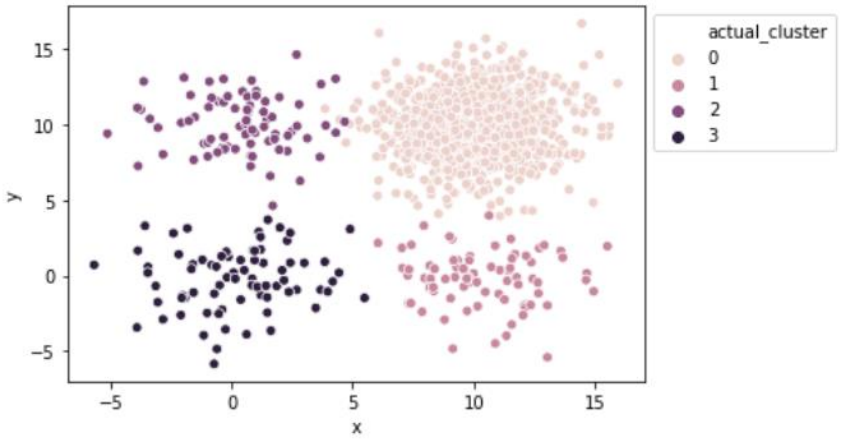
What $-\sum_j p_{j|i} \log_2 p_{j|i}$represents:

$$\_\_\_\_\_ \leq -\sum_j p_{j|i} \log_2 p_{j|i} \leq \_\_\_\_\_$$

Ex: Let's assume for now, that we have chosen a perplexity of 1. Use the method described above to estimate a value for σ_i corresponding to an observation in cluster 1 below.



Ex: Let's assume for now, that we have chosen a perplexity that is equal to the total number of observation m in the dataset above. Use the method described above to estimate a value for σ_i corresponding to an observation in cluster 2 below.



Good guesses for $Perplexity(P_i)$**:**

DRAWBACKS OF STOCHASTIC NEIGHBOR EMBEDDING (SNE)

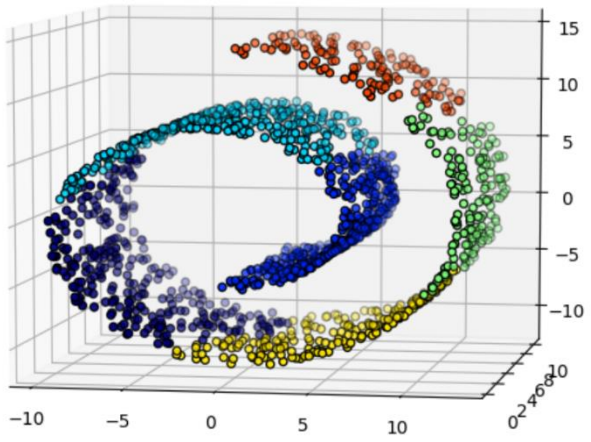
1. <u>Because of non-symmetric KL-Divergence Function:</u>
    a. Focuses more on preserving the local structure.

2. <u>Because of non-symmetric distances matrices:</u>

    a. Cost function of SNE is computationally inefficient. The gradient descent algorithm needs to be run several times to select the right initial parameters such that the algorithm is less likely to get stuck in poor local minimum.

3. <u>Because of Gaussian probability in mapped distance matrix Q:</u>
    a. "Crowding problem": Some types of datasets ***X*** have the property that the available area (in 2-dimensions) to map _____ distant points isn't large enough to as the amount of available area that accurately maps _____ points. So _____ distances tend to be mapped much _____ than in the original dataset.

    b. The exponential function in the cost function of the gradient descent algorithm makes for slower run time.



https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_swiss_roll.html

WHAT'S DIFFERENT IN T-SNE ALGORITHM VS. SNE ALGORITHM?

1. T-SNE uses a _____ version of P and Q (as used in SNE).

2. Uses a _____ (as opposed to Gaussian distribution as used in SNE) to measure distance between two points *in the low dimensional space.*

T-SNE ALGORITHM – GENERAL

Input:

- Dataset: $\mathbf{X_{m\times n}} = \begin{bmatrix} \boldsymbol{x_{1*}} \\ \boldsymbol{x_{2*}} \\ \vdots \\ \boldsymbol{x_{m*}} \end{bmatrix}$ comprised of *m* objects, where each object has a complete set of n attributes.
- Parameters Used to Generate Similarity Matrix P:
  - $Perplexity$

Algorithm

- **Step 1: Define projected/mapped 2-d coordinates matrix of decision variables:** $\mathbf{Y_{m\times 2}} = \begin{bmatrix} \boldsymbol{y_{1*}} \\ \boldsymbol{y_{2*}} \\ \vdots \\ \boldsymbol{y_{m*}} \end{bmatrix}$.

- **Step 2: Create a similarity matrix $\boldsymbol{P} = [\boldsymbol{p_{ij}}]$ between each of the objects in $\mathbf{X_{m\times n}}$.**

  - (j,i) entries represents: similarity between x_{j*} to x_{i*}
  - (j,i) entries mathematically:
    - $p_{j|i} = \frac{\exp\left(-\frac{\left||x_{i*}-x_{j*}|\right|^2}{2\sigma_i^2}\right)}{\sum_{k\neq i}\exp\left(-\frac{\left||x_{i*}-x_{k*}|\right|^2}{2\sigma_i^2}\right)}$
    - $p_{i|j} = \frac{\exp\left(-\frac{\left||x_{i*}-x_{j*}|\right|^2}{2\sigma_j^2}\right)}{\sum_{k\neq j}\exp\left(-\frac{\left||x_{j*}-x_{k*}|\right|^2}{2\sigma_j^2}\right)}$
    - $p_{ij} = \frac{1}{2m}\left(p_{j|i} + p_{i|j}\right)$
  - Is $\boldsymbol{P} = [\boldsymbol{p_{ij}}]$ symmetric?

  Finding σ_i:

  Perform a binary search for σ_i that satisfies: $Perplexity = 2^{-\sum_j p_{j|i}\log_2 p_{j|i}}$

- **Step 3: Create a similarity matrix $\boldsymbol{Q} = [\boldsymbol{q_{ij}}]$ between each of the objects in $\mathbf{Y_{m\times 2}}$.**

  - (j,i) entries represents: similarity between y_{j*} to y_{i*}
  - (j,i) entries mathematically: $q_{ij} = \frac{\left(1+\left||y_{i*}-y_{j*}|\right|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\left||y_{k*}-y_{l*}|\right|^2\right)^{-1}}$
  - Is $\boldsymbol{Q} = [\boldsymbol{q_{ij}}]$ symmetric?

  Why student t-distribution with df=1?

  - _____ tails than Gaussian.
  - Pairs of mapped points with moderate distances will have higher similarity (than if you had used Gaussian.) Helps with the crowding problem.

- (j,i) entries graphical interpretation:

$$q_{ij} = \frac{\left(1 + \left||y_{i*} - y_{j*}\right|\right|^2\right)^{-1}}{\sum_{k \neq l}\left(1 + \left||y_{k*} - y_{l*}\right|\right|^2\right)^{-1}} = \frac{\frac{\Gamma(\frac{1+1}{2})}{\sqrt{1\pi}\Gamma(\frac{1}{2})}\left(1 + \frac{\left||y_{i*} - y_{j*}\right|\right|^2}{1}\right)^{-\frac{1+1}{2}}}{\sum_{k \neq l}\frac{\Gamma(\frac{1+1}{2})}{\sqrt{1\pi}\Gamma(\frac{1}{2})}\left(1 + \frac{\left||y_{i*} - y_{j*}\right|\right|^2}{1}\right)^{-\frac{1+1}{2}}} = \frac{f(dist(\boldsymbol{y_{i*}}, \boldsymbol{y_{j*}}))}{\sum_{k \neq l} f(dist(\boldsymbol{y_{k*}}, \boldsymbol{y_{l*}}))}$$

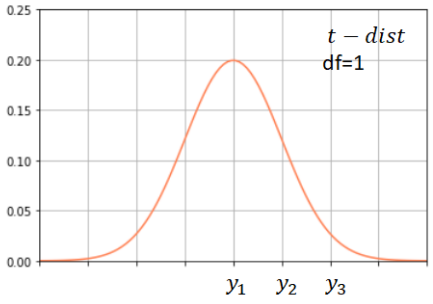
One Interpretation of f(x):

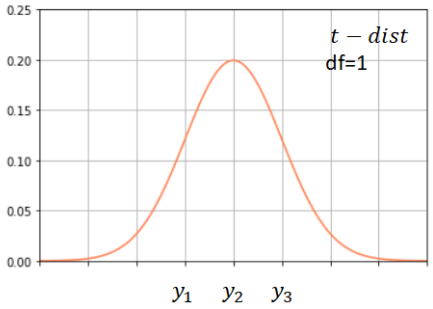
  - $f(x)$ is the pdf of the _____ distribution with df = _____.

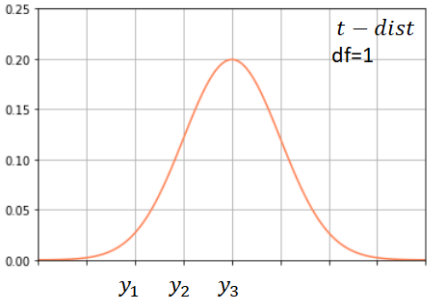
Another Interpretation of f(x):

  - $f(x)$ is the curve of the _____ distribution with df = _____, shifted to be centered at _____.

- **Ex:** Generate the similarity matrix for the three 1-d MAPPED data points shown below.







- **<u>Step 4</u>: Find the optimal values of the coordinates in $\mathbf{Y_{m\times 2}}$ that minimize the sum of the Kullback-Leibler (KL) divergence over all data points.**

  <u>Optimization Problem:</u> $min_{Y_{m\times 2}} C = min_{Y_{m\times 2}} KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$

  <u>Common Algorithm</u>: Gradient descent methods

FULL T-SNE ALGORITHM – WITH GRADIENT DESCENT ALGORITHM

Input:

- Dataset: $\mathbf{X_{m\times n}} = \begin{bmatrix} \boldsymbol{x_{1*}} \\ \boldsymbol{x_{2*}} \\ \vdots \\ \boldsymbol{x_{m*}} \end{bmatrix}$ comprised of *m* objects, where each object has a complete set of n attributes.
- Parameters Used to Generate Similarity Matrix P:
  - $Perplexity$
- Parameters Used to Generate Similarity Matrix P:
  - Number of iterations T
  - Learning rate η
  - Momentum $\alpha(t)$

Algorithm

- **Step 1: Create a similarity matrix $\boldsymbol{P} = [\boldsymbol{p_{ij}}]$ between each of the objects in $\mathbf{X_{m\times n}}$.**
  - $p_{ij} = \frac{1}{2}(p_{j|i} + p_{i|j})$
  - $p_{j|i} = \frac{\exp\left(-\frac{\left\|x_{i*} - x_{j*}\right\|^2}{2\sigma_i^2}\right)}{\sum_{k\neq i}\exp\left(-\frac{\left\|x_{i*} - x_{k*}\right\|^2}{2\sigma_i^2}\right)}$
  - Perform a binary search for σ_i that satisfies: $Perplexity = 2^{-\sum_j p_{j|i}\log_2 p_{j|i}}$
- **Step 2: Randomly sample projected/mapped 2-d coordinates matrix: $\mathbf{Y_{m\times 2}} = \begin{bmatrix} \boldsymbol{y_{1*}} \\ \boldsymbol{y_{2*}} \\ \vdots \\ \boldsymbol{y_{m*}} \end{bmatrix}$ from $\boldsymbol{N}(\mathbf{0}, 0.0001 \cdot \boldsymbol{I_{2\times 2}})$.**
- **For *t=1 to T* *do:***
  - Create a similarity matrix $Q = [q_{ij}]$ between each of the objects in $\mathrm{Y_{m\times 2}}$.
    - $q_{ij} = \frac{\left(1+\left||y_{i*} - y_{j*}\right|\right|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\left||y_{k*} - y_{l*}\right|\right|^2\right)^{-1}}$
    - Compute cost function gradient as follows (for each i=1,…,m)
      - $\frac{\delta C}{\delta \boldsymbol{y_{i*}}} = 4\sum_j (p_{ij} - q_{ij})(\boldsymbol{y_{i*}} - \boldsymbol{y_{j*}})\left(1 + \left||\boldsymbol{y_{i*}} - \boldsymbol{y_{j*}}\right|\right|^2\right)^{-1}$
    - Set $\mathbf{Y_{m\times 2}^{(t)}} := \mathbf{Y_{m\times 2}^{(t-1)}} + \gamma\frac{\delta C}{\delta \boldsymbol{Y}} + \alpha(t)(\mathbf{Y_{m\times 2}^{(t-1)}} - \mathbf{Y_{m\times 2}^{(t-2)}})$

  End do