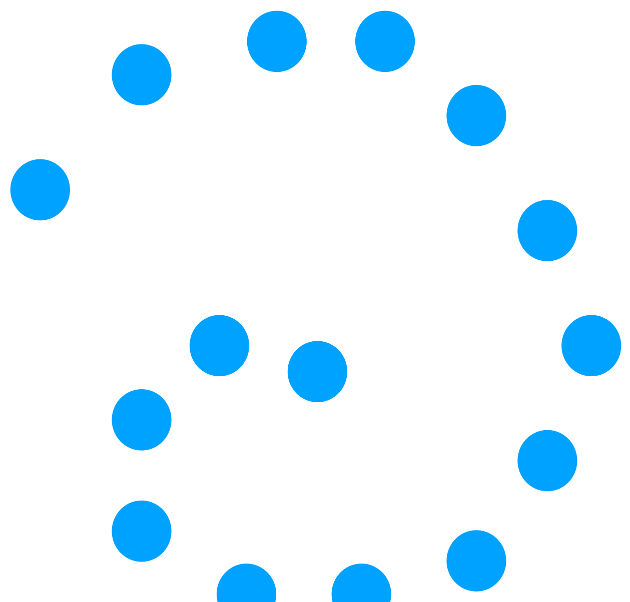
Machine Learning for Data Science (CS4786)
Lecture 9

Isomap + TSNE

Manifold Based Dimensionality Reduction

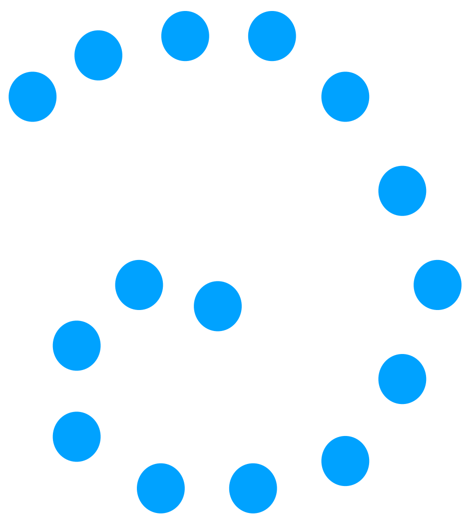
- Key Assumption: Points live on a low dimensional manifold
- Manifold: subspace that looks locally Euclidean
- Given data, can we uncover this manifold?



Can we unfold this?

Method I: Isomap

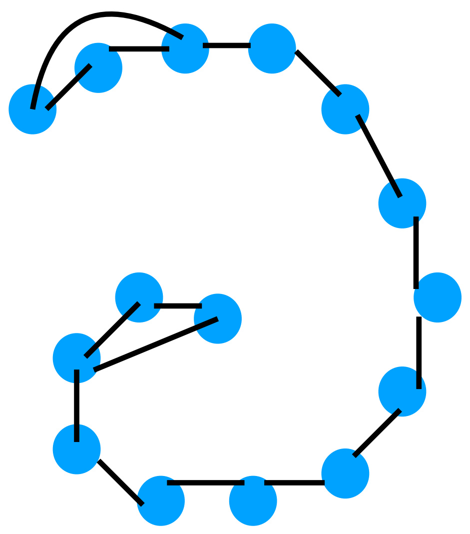
1. For every point, find its (k-) Nearest Neighbors



Method I: Isomap

1. For every point, find its (k-) Nearest Neighbors

2. Form the Nearest Neighbor graph

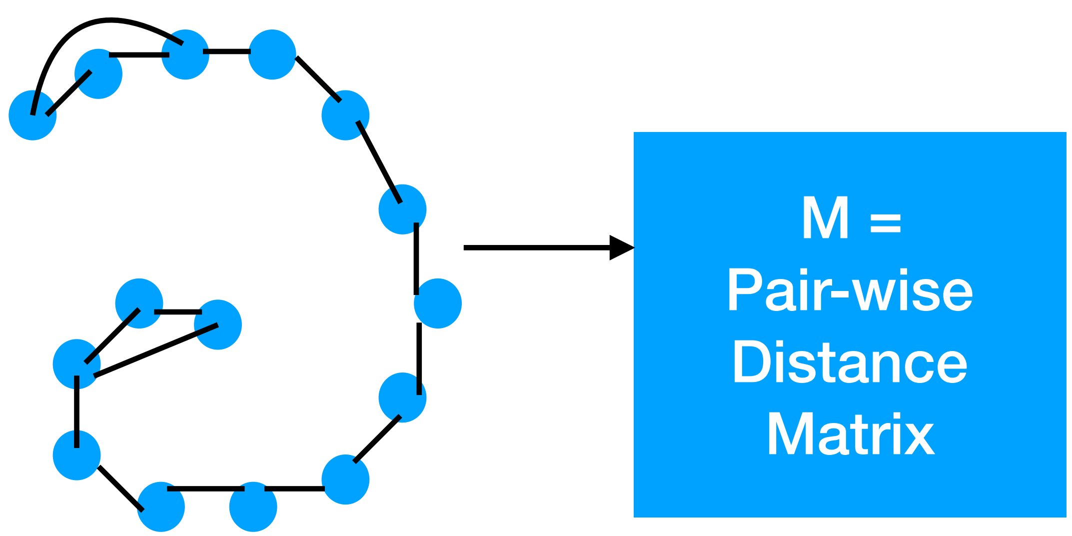


METHOD I: ISOMAP

1. For every point, find its (k-) Nearest Neighbors

2. Form the Nearest Neighbor graph

3. For every pair of points A and B, distance between point A to B is shortest distance between A and B on graph



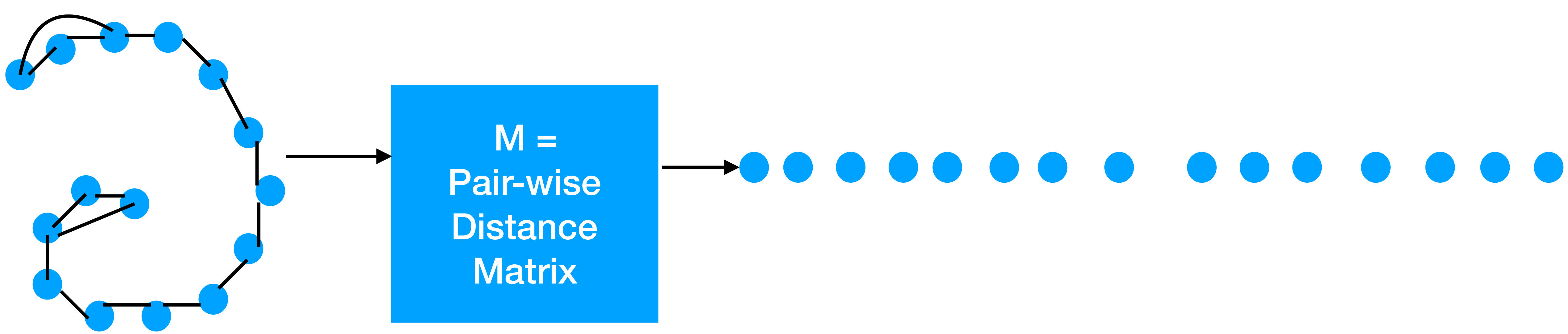
Method I: Isomap

1. For every point, find its (k-) Nearest Neighbors

2. Form the Nearest Neighbor graph

3. For every pair of points A and B, distance between point A to B is shortest distance between A and B on graph

4. Find points in low dimensional space such that distances between points in this space is equal to distance on graph.



Multidimensional Scaling

Question: Given $n \times n$ matrix M of pairwise distances, find $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^K$ such that $\forall i, j \in [n]$, $\|\mathbf{y}_i - \mathbf{y}_j\|_2 \approx M_{i,j}$

Multidimensional Scaling

Define for all i, j, $D_{i,j} = M_{i,j}^2$

Multidimensional Scaling

Define for all i, j, $D_{i,j} = M_{i,j}^2$

$$D_{i,j} = \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 = \|\mathbf{y}_i\|_2^2 + \|\mathbf{y}_j\|_2^2 - 2\mathbf{y}_i^\top \mathbf{y}_j$$

Multidimensional Scaling

Define for all i, j, $D_{i,j} = M_{i,j}^2$

$$D_{i,j} = \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 = \|\mathbf{y}_i\|_2^2 + \|\mathbf{y}_j\|_2^2 - 2\mathbf{y}_i^\top \mathbf{y}_j$$

$$\mathbf{y}_i^\top \mathbf{y}_j = \frac{1}{2}\left(\|\mathbf{y}_i\|_2^2 + \|\mathbf{y}_j\|_2^2 - D_{i,j}\right)$$

Multidimensional Scaling

Define for all i, j, $D_{i,j} = M_{i,j}^2$

$$D_{i,j} = \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 = \|\mathbf{y}_i\|_2^2 + \|\mathbf{y}_j\|_2^2 - 2\mathbf{y}_i^\top \mathbf{y}_j$$

$$\mathbf{y}_i^\top \mathbf{y}_j = \frac{1}{2}\left(\|\mathbf{y}_i\|_2^2 + \|\mathbf{y}_j\|_2^2 - D_{i,j}\right)$$

w.l.o.g. assume $\sum_i \mathbf{y}_i = 0$

Multidimensional Scaling

Define for all i, j, $D_{i,j} = M_{i,j}^2$

$$D_{i,j} = \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 = \|\mathbf{y}_i\|_2^2 + \|\mathbf{y}_j\|_2^2 - 2\mathbf{y}_i^\top \mathbf{y}_j$$

$$\mathbf{y}_i^\top \mathbf{y}_j = \frac{1}{2}\left(\|\mathbf{y}_i\|_2^2 + \|\mathbf{y}_j\|_2^2 - D_{i,j}\right)$$

w.l.o.g. assume $\sum_i \mathbf{y}_i = 0$

Show that:

$$\mathbf{y}_i^\top \mathbf{y}_j = \frac{1}{n}\sum_i D_{i,j} + \frac{1}{n}\sum_j D_{i,j} - \frac{1}{n^2}\sum_i \sum_j D_{i,j} - D_{i,j}$$

Method:

Multidimensional Scaling

Method:

1. Compute matrix D so that $D_{i,j} = M_{i,j}^2$

Method:

1. Compute matrix D so that $D_{i,j} = M_{i,j}^2$
2. Compute matrix $\tilde{K} = -\frac{1}{2}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)D(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)$

Multidimensional Scaling

Method:

1. Compute matrix D so that $D_{i,j} = M_{i,j}^2$
2. Compute matrix $\tilde{K} = -\frac{1}{2}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)D(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)$

$$3\left[\underset{\text{K}}{{}^{\text{n}}P}, \gamma\right] = \text{eigs}\left(\tilde{K}, K\right)$$

Method:

1. Compute matrix D so that $D_{i,j} = M_{i,j}^2$
2. Compute matrix $\tilde{K} = -\frac{1}{2}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)D(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)$

3 $$\Big[\underbrace{P}_{n \times K}, \gamma\Big] = \mathrm{eigs}\big(\tilde{K}, K\big)$$

4. $$\underbrace{Y}_{n \times K} = \Big[\, P_1\sqrt{\gamma_1} \;\cdots\; P_K\sqrt{\gamma_K} \,\Big]$$

K

Isomap: Pitfalls

1. If we don't take enough nearest neighbors, then graph may not be connected
2. If we connect points too far away, points that should not be connected can get connected
3. There may not be a right number of nearest neighbors we should consider!

Stochastic Neighborhood Embedding

- Use a probabilistic notion of which points are neighbors.

Stochastic Neighborhood Embedding

- Use a probabilistic notion of which points are neighbors.

  Stochastic neighborhood distribution P

Stochastic Neighborhood Embedding

- Use a probabilistic notion of which points are neighbors.

  Stochastic neighborhood distribution P

- Close by points are neighbors with high probability, ...

Stochastic Neighborhood Embedding

- Use a probabilistic notion of which points are neighbors.

  Stochastic neighborhood distribution P

- Close by points are neighbors with high probability, ...

  Eg: For point $\mathbf{x}_t$, point $\mathbf{x}_s$ is picked as neighbor with probability

$$p_{t \to s} = \frac{\exp\left(-\frac{\|\mathbf{x}_s - \mathbf{x}_t\|^2}{2\sigma^2}\right)}{\sum_{u \neq t} \exp\left(-\frac{\|\mathbf{x}_u - \mathbf{x}_t\|^2}{2\sigma^2}\right)}$$

  Probability that points s and t are connected $P_{s,t} = P_{t,s} = \frac{p_{t \to s} + p_{s \to t}}{2n}$

Stochastic Neighborhood Embedding

- Use a probabilistic notion of which points are neighbors.

  Stochastic neighborhood distribution P

- Close by points are neighbors with high probability, ...
  Eg: For point $\mathbf{x}_t$, point $\mathbf{x}_s$ is picked as neighbor with probability

$$p_{t \to s} = \frac{\exp\left(-\frac{\|\mathbf{x}_s - \mathbf{x}_t\|^2}{2\sigma^2}\right)}{\sum_{u \neq t} \exp\left(-\frac{\|\mathbf{x}_u - \mathbf{x}_t\|^2}{2\sigma^2}\right)}$$

  Probability that points s and t are connected $P_{s,t} = P_{t,s} = \frac{p_{t \to s} + p_{s \to t}}{2n}$

- Goal: Find $\mathbf{y}_1, \dots, \mathbf{y}_n$ with stochastic neighborhood distribution Q such that "P and Q are similar"

Stochastic Neighborhood Embedding

- Use a probabilistic notion of which points are neighbors.

  Stochastic neighborhood distribution P

- Close by points are neighbors with high probability, ...
  Eg: For point $\mathbf{x}_t$, point $\mathbf{x}_s$ is picked as neighbor with probability

$$p_{t \to s} = \frac{\exp\left(-\frac{\|\mathbf{x}_s - \mathbf{x}_t\|^2}{2\sigma^2}\right)}{\sum_{u \neq t} \exp\left(-\frac{\|\mathbf{x}_u - \mathbf{x}_t\|^2}{2\sigma^2}\right)}$$

  Probability that points s and t are connected $P_{s,t} = P_{t,s} = \frac{p_{t \to s} + p_{s \to t}}{2n}$

- Goal: Find $\mathbf{y}_1, \ldots, \mathbf{y}_n$ with stochastic neighborhood distribution Q such that "P and Q are similar"

  i.e. minimize:

$$\mathrm{KL}(P\|Q) = \sum_{s,t} P_{s,t} \log\left(\frac{P_{s,t}}{Q_{s,t}}\right) = \sum_{s,t} P_{s,t} \log\left(P_{s,t}\right) - \sum_{s,t} P_{s,t} \log\left(Q_{s,t}\right)$$

Choice for Q

- Just like we defined P, we can define Q for a given $\mathbf{y}_1, \ldots, \mathbf{y}_n$ by

$$q_{t \to s} = \frac{\exp(-\frac{\|\mathbf{y}_s - \mathbf{y}_t\|^2}{2\sigma^2})}{\sum_{u \neq t} \exp(-\frac{\|\mathbf{y}_u - \mathbf{y}_t\|^2}{2\sigma^2})}$$

and then set $Q_{s,t} = \frac{q_{t \to s} + q_{s \to t}}{2n}$

Choice for Q

- Just like we defined P, we can define Q for a given $\mathbf{y}_1, \ldots, \mathbf{y}_n$ by

$$q_{t \to s} = \frac{\exp(-\frac{\|\mathbf{y}_s - \mathbf{y}_t\|^2}{2\sigma^2})}{\sum_{u \neq t} \exp(-\frac{\|\mathbf{y}_u - \mathbf{y}_t\|^2}{2\sigma^2})}$$

  and then set $Q_{s,t} = \frac{q_{t \to s} + q_{s \to t}}{2n}$

- However we are faced with the crowding problem:

CHOICE FOR Q

- Just like we defined P, we can define Q for a given $\mathbf{y}_1, \ldots, \mathbf{y}_n$ by

$$q_{t \to s} = \frac{\exp\left(-\frac{\|\mathbf{y}_s - \mathbf{y}_t\|^2}{2\sigma^2}\right)}{\sum_{u \neq t} \exp\left(-\frac{\|\mathbf{y}_u - \mathbf{y}_t\|^2}{2\sigma^2}\right)}$$

  and then set $Q_{s,t} = \frac{q_{t \to s} + q_{s \to t}}{2n}$

- However we are faced with the crowding problem:
  - In high dimension we have a lot of space, Eg. in d dimension we have $d+1$ equidistant point

Choice for Q

- Just like we defined P, we can define Q for a given $\mathbf{y}_1, \ldots, \mathbf{y}_n$ by

$$q_{t \to s} = \frac{\exp\left(-\frac{\|\mathbf{y}_s - \mathbf{y}_t\|^2}{2\sigma^2}\right)}{\sum_{u \neq t} \exp\left(-\frac{\|\mathbf{y}_u - \mathbf{y}_t\|^2}{2\sigma^2}\right)}$$

  and then set $Q_{s,t} = \frac{q_{t \to s} + q_{s \to t}}{2n}$

- However we are faced with the crowding problem:
  - In high dimension we have a lot of space, Eg. in d dimension we have $d + 1$ equidistant point
  - For d dimensional gaussians, most points are found at distance $\sqrt{d}$ from mean!

CHOICE FOR Q

- Just like we defined P, we can define Q for a given $\mathbf{y}_1, \ldots, \mathbf{y}_n$ by

$$q_{t \to s} = \frac{\exp(-\frac{\|\mathbf{y}_s - \mathbf{y}_t\|^2}{2\sigma^2})}{\sum_{u \neq t} \exp(-\frac{\|\mathbf{y}_u - \mathbf{y}_t\|^2}{2\sigma^2})}$$

  and then set $Q_{s,t} = \frac{q_{t \to s} + q_{s \to t}}{2n}$

- However we are faced with the crowding problem:
  - In high dimension we have a lot of space, Eg. in d dimension we have $d + 1$ equidistant point
  - For d dimensional gaussians, most points are found at distance $\sqrt{d}$ from mean!
  - If we use gaussians in both high and low dimensional space, all the points are squished in to a small space
  - Too many points crowd the center!

METHOD II: T-SNE

- Instead for Q we use, student t distribution which is heavy tailed:

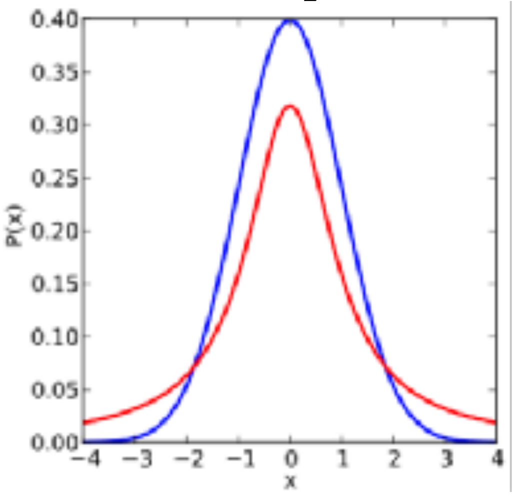
$$q_{t \to s} = \frac{\left(1 + \|\mathbf{y}_s - \mathbf{y}_t\|^2\right)^{-1}}{\sum_{u \neq t} \left(1 + \|\mathbf{y}_u - \mathbf{y}_t\|^2\right)^{-1}}$$

and then set $Q_{s,t} = \frac{q_{t \to s} + q_{s \to t}}{2n}$

Method II: t-SNE

- Instead for Q we use, student t distribution which is heavy tailed:

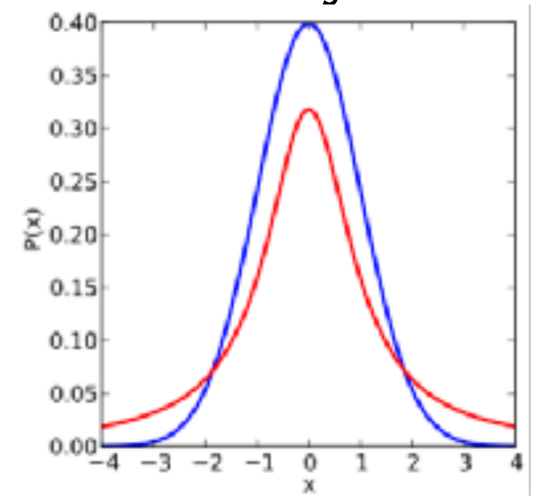
$$q_{t \to s} = \frac{\left(1 + \|\mathbf{y}_s - \mathbf{y}_t\|^2\right)^{-1}}{\sum_{u \neq t} \left(1 + \|\mathbf{y}_u - \mathbf{y}_t\|^2\right)^{-1}}$$



and then set $Q_{s,t} = \frac{q_{t \to s} + q_{s \to t}}{2n}$

METHOD II: T-SNE

- Instead for Q we use, student t distribution which is heavy tailed:



$$q_{t \to s} = \frac{\left(1 + \|\mathbf{y}_s - \mathbf{y}_t\|^2\right)^{-1}}{\sum_{u \neq t} \left(1 + \|\mathbf{y}_u - \mathbf{y}_t\|^2\right)^{-1}}$$

and then set $Q_{s,t} = \frac{q_{t \to s} + q_{s \to t}}{2n}$

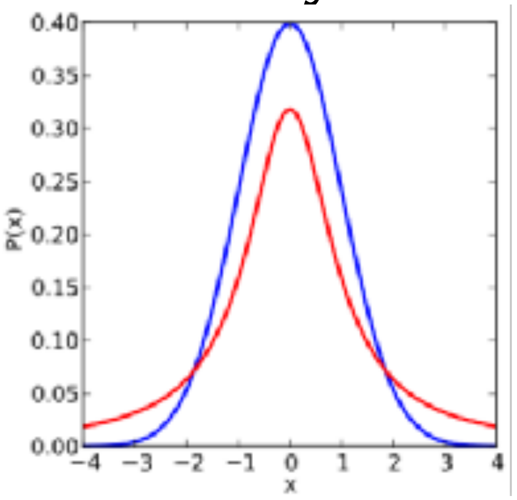
- It can be verified that

$$\nabla_{\mathbf{y}_t} \mathrm{KL}(P \| Q) = 4 \sum_{s=1}^{n} (P_{s,t} - Q_{s,t})(\mathbf{y}_t - \mathbf{y}_s) \left(1 + \|\mathbf{y}_s - \mathbf{y}_t\|^2\right)^{-1}$$

Method II: t-SNE

- Instead for Q we use, student t distribution which is heavy tailed:

$$q_{t \to s} = \frac{\left(1 + \|\mathbf{y}_s - \mathbf{y}_t\|^2\right)^{-1}}{\sum_{u \neq t} \left(1 + \|\mathbf{y}_u - \mathbf{y}_t\|^2\right)^{-1}}$$



and then set $Q_{s,t} = \frac{q_{t \to s} + q_{s \to t}}{2n}$

- It can be verified that

$$\nabla_{\mathbf{y}_t} \mathrm{KL}(P \| Q) = 4 \sum_{s=1}^{n} (P_{s,t} - Q_{s,t})(\mathbf{y}_t - \mathbf{y}_s) \left(1 + \|\mathbf{y}_s - \mathbf{y}_t\|^2\right)^{-1}$$

- Algorithm: Find $\mathbf{y}_1, \ldots, \mathbf{y}_n$ by performing gradient descent

Demo