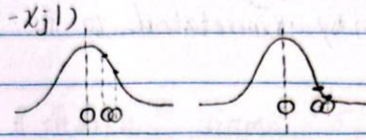
1. probability distribution

represents similarities between neighbors

$|x_i - x_j|$ Euclidean distance

$\downarrow$ proportional to probability density under a Gaussian centered at x_i

$g(|x_i - x_j|)$



You can distigui -sh between similar and non-similar points but absolute values of probability are much smaller

So, we fix that by dividing the current projection value by the sum of the projections

$$p_{j|i} = \frac{g(|x_i - x_j|)}{\sum_{k \neq i} g(|x_i - x_k|)}$$

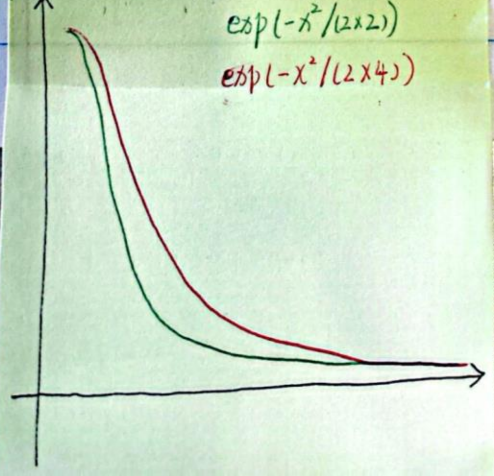
$\downarrow$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

N: number of data points.

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

this is what exactly in t-SNE paper.



2. perplexity, sigma and target number of neighbors

intuitive understanding:

A perplexity is more or less a target number of neighbors for our central point

Basically, ↑ perplexity → variance ↑

if we set perplexity to 4, it searches the right value of σ to "fit" 4 neighbors

Technical derivation.

"SNE performs a binary search for the value of σ that produces probability distribution with a fixed perplexity that is specified by the user

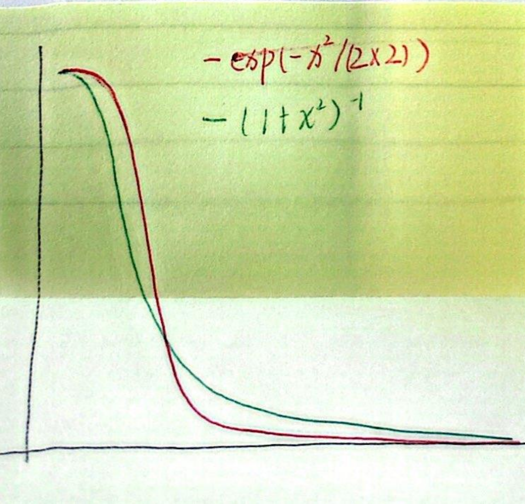
$$Perp(P_i) = 2^{-\sum_j P_{j|i} \log_2 P_{j|i}}$$

where $-\sum_j P_{j|i} \log_2 P_{j|i}$ is Shannon Entropy

Typical perplexity value ranges between 5 and 50

3. t-distribution & t-SNE.

for low-dimensional space, Gaussian creates crowding problem. To solve this, we are going to use student t-distribution with a single degree of freedom.



$$q_{ij} = \frac{(1+\|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1+\|y_k - y_l\|^2)^{-1}}$$

advantages of t :

1. It "falls" quickly and has a "long tail", so points won't get squashed into a single point.
2. we don't have to bother with σ^2 because we don't have one in q_{ij}

4. finding embedding in low dimensinal space.

~~$C = KL\ D_{KL}(P\|Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$~~.

$$C = KL(P\|Q) \qquad \text{loss function}$$

$$= \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

$$= \sum_i \sum_j (p_{ij} \log p_{ij} - p_{ij} \log q_{ij})$$

$\downarrow$ Gradient descent

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}$$

5. Tricks done in t-SNE to perform better.

early compression (not to focus on local groups)

early exaggeration.