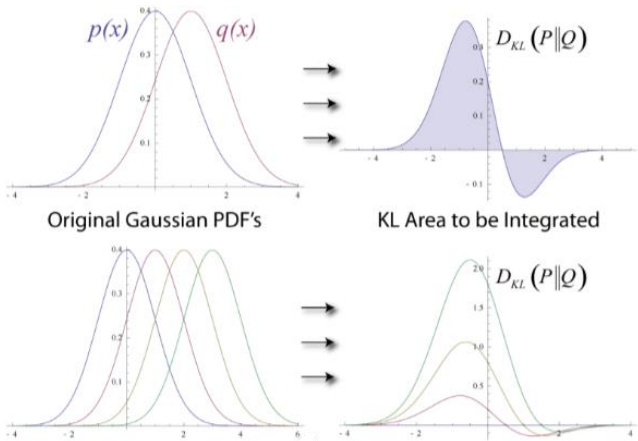
Dimensionality Reduction

❖ Stochastic Neighbor Embedding (SNE)

- 가까운 이웃 객체들과 거리 정보를 잘 보존하는 것이 먼 객체들과의 거리 정보를 보존하는 것보다 더 중요함
- SNE는 LLE와 다르게 이웃을 확정적으로 정하는 것이 아니라 모든 이웃에 대해서 확률적으로 정의를 함
- 저차원 공간 상에서 임베딩이 잘 되었는지 KL divergence를 비용 함수로 사용하여 평가함

고차원에서 객체 i가 j를 이웃으로 선택할 확률 $p_{j|i} = \frac{e^{-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}}}{\sum_{k \neq i} e^{-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}}}$, 저차원에서 객체 i가 j를 이웃으로 선택할 확률 $q_{j|i} = \frac{e^{-\|y_i - y_j\|^2}}{\sum_{k \neq i} e^{-\|y_i - y_k\|^2}}$

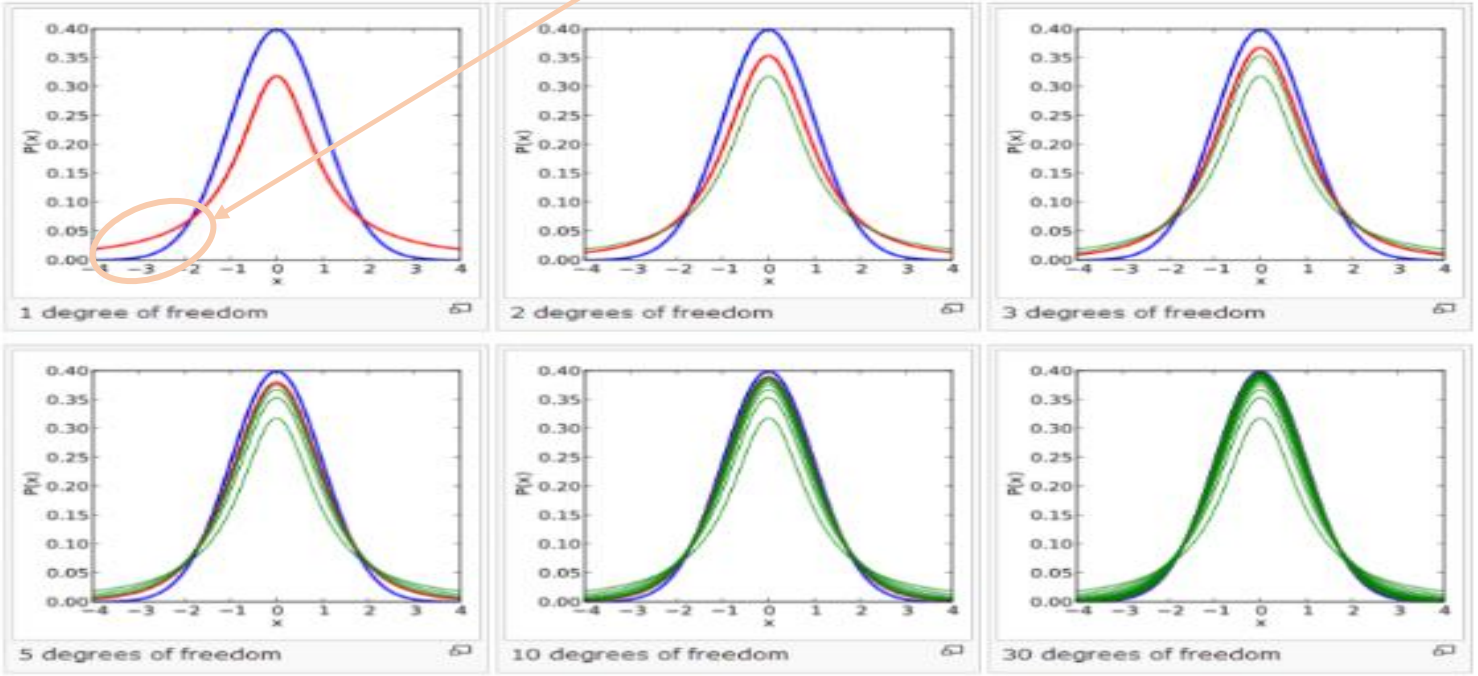
$$Cost = \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$



Dimensionality Reduction

❖ t-distributed Stochastic Neighbor Embedding (t-SNE)

- 기존의 SNE(symmetric SNE)는 가우시안 분포를 사용하기 때문에 crowding problem이 존재함
- 이를 해결하고자 저차원 공간에서 가우시안 분포보다 덜 급격하게 감소하는 분포함수인 t-분포 함수를 사용(자유도 1)
- t-분포도 표본 평균, 표본 분산으로 정의되는 확률변수이므로 표본 수가 많아질수록 중심 극한정리에 의해 가우시안 분포로 수렴



Dimensionality Reduction

❖ t-SNE example: Hand digit recognition

- sklearn 패키지에서 제공하는 digits 데이터 셋이며 각 이미지는 64차원으로 이루어진 손글씨 데이터셋
- 기존 방법론들과 t-SNE를 적용했을 때 embedding 비교

A selection from the 64-dimensional digits dataset