B1

(*) SNE

Define : $q_{j|i} = \dfrac{e^{-\|y_i - y_j\|^2}}{\sum_{k \neq i} e^{-\|y_i - y_k\|^2}} = \dfrac{E_{ij}}{\sum_{k\neq i} E_{ik}} = \dfrac{E_{ij}}{z_i}$

See that $E_{ij} = E_{ji}$, the loss function is defined as:

$$C = \sum_{k,l\neq k} p_{l|k} \log \frac{p_{l|k}}{q_{l|k}} = \sum_{k,l\neq k} p_{l|k} \log p_{l|k} - p_{l|k} \log q_{l|k}$$

$$= \sum_{k,l\neq k} p_{l|k} \log p_{l|k} - p_{l|k} \log E_{kl} + p_{l|k} \log z_k$$

Derive with respect to y_i. To make the derivation less cluttered, I will omit the ∂y_i term at the denominator.

$$\frac{\partial C}{\partial y_i} = \sum_{k,l\neq k} - p_{l|k}\, \partial \log E_{kl} + \sum_{k,l\neq k} p_{l|k}\, \partial \log z_k$$

Start with the first term, noting that the derivation is non-zero when $\forall j \neq i$, $k = i$ or $l = i$

$$\sum_{k,l\neq k} - p_{l|k}\, \partial \log E_{k,l} = \sum_{j\neq i} - p_{j|i}\, \partial \log E_{ij} - p_{i|j}\, \partial \log E_{ji}$$

Since $\partial E_{ij} = E_{ij}(-2(y_i - y_j))$, we have:

$$\sum_{j\neq i} - p_{j|i} \frac{E_{ij}}{E_{ij}}(-2(y_i - y_j)) - p_{i|j} \frac{E_{ji}}{E_{ji}}(2(y_j - y_i))$$

$$= 2\sum_{j\neq i} (p_{j|i} + p_{i|j})(y_i - y_j) \quad (1)$$

We conclude with the second term, since $\sum_{l\neq j} p_{l|j} = 1$ and z_j does not depend on k, we can write

$$\sum_{j,k\neq j} p_{k|j}\, \partial \log z_j = \sum_j \partial \log z_j$$

The derivation is non zero when $k = i$ or $j = i$

$$= \sum_j \frac{1}{z_i} \sum_{k\neq j} \partial E_{jk}$$

$$= \sum_{j \neq i} \frac{E_{ji}}{Z_j} (2(y_j - y_i)) + \sum_{j \neq i} \frac{E_{ij}}{Z_i} (-2(y_j - y_i))$$

$$= 2 \sum_{j \neq i} (-q_{j|i} - q_{j|i}) (y_i - y_j) \quad (2)$$

From (1) + (2), we have the final result:

$$\frac{\partial C}{\partial y_i} = 2 \sum_{j \neq i} (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j}) (y_i - y_j)$$

(*) t - SNE

+ Define:

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

$$= \frac{E_{ij}^{-1}}{\sum_{k, l \neq k} E_{kl}^{-1}} = \frac{E_{ij}^{-1}}{Z}$$

Notice that $E_{ij} = E_{ji}$, the loss function is defined:

$$C = \sum p_{lk} \log \frac{p_{lk}}{q_{lk}} = \sum_{k, l \neq k} p_{lk} \log p_{lk} - p_{lk} \log q_{lk}$$

$$= \sum_{k, l \neq k} p_{lk} \log p_{lk} - p_{lk} \log E_{kl}^{-1} + p_{lk} \log Z$$

We derive respect to y_i. To make the derivation less cluttered, I will omit the ∂y_i term at the denominator

$$\frac{\partial C}{\partial y_i} = \sum_{k, l \neq k} - p_{lk} \, \partial \log E_{kl}^{-1} + \sum_{k, l \neq k} p_{lk} \, \partial \log Z$$

With the first term, noting that the derivative is non zero when $\forall j$, $k = i$ or $l = i$ that $p_{ji} = p_{ij}$ and $E_{ji} = E_{ij}$

$$\sum_{k, l \neq k} - p_{lk} \, \partial \log E_{kl}^{-1} = -2 \sum_{j \neq i} p_{ji} \, \partial \log E_{ij}^{-1}$$

Thứ ngày . .

Since $\partial E_{ij}^{-1} = E_{ij}^{-2}(-2(y_i - y_j))$ we have:

$$-2\sum_{j\neq i} p_{ji}\frac{E_{ji}^{-2}}{E_{ij}^{-1}}(-2(y_i - y_j)) = 4\sum_{j\neq i} p_{ij}E_{ij}^{-1}(y_i - y_j) \quad (3)$$

We conclude with the second term, using the fact that $\sum_{k,l\neq k} p_{kl} = 1$ and that Z does not depend on k or l:

$$\sum_{k,l\neq k} p_{lk}\,\partial\log Z = \frac{1}{Z}\sum_{k',l'\neq k'}\partial E_{k'l'}^{-1}$$

$$= 2\sum_{j\neq i}\frac{E_{ji}^{-2}}{Z}(-2(y_j - y_i))$$

$$= -4\sum_{j\neq i} q_{ij}E_{ji}^{-1}(y_i - y_j) \quad (4)$$

From (3) + (4), we have final result:

$$\frac{\partial C}{\partial y_i} = 4\sum_{j\neq i}(p_{ji} - q_{ji})E_{ji}^{-1}(y_i - y_j)$$

$$\frac{\partial C}{\partial y_i} = 4\sum_{j\neq i}(p_{ji} - q_{ji})(1 + \|y_i - y_j\|^2)^{-1}(y_i - y_j)$$

B4 Different between PCA and t-SNE.

	PCA	t-SNE
1)	+ It is a linear dimensionallity reduction technique	+ It is non-linear dimensionality reduction technique.
2)	It tries to preserve the global structure of the data	+ It tries to preserve the local structure cluster of data
3)	It does not work well as compared to t-SNE	+ It involves Hyperparameters such as perplexity, learning rate and no of steps
4)	It does not involve Hyperparameters	+ It is one of the best dimension reduction technique.

5)	It gets highly affected by outliers	It can handle outliers
6)	PCA is a deterministic algorithm	It is non-deterministic or randomised algorithm
7)	It works by rotating the vectors for preserving variance	It works by minimising the distance between the point in a gaussian.
8)	We can find decide on how much variance to preserve using eigen values	We can not preserve var instead we can preserve distance using hyperparameters