Problem 1
Reconstructing the t-SNE problem. Calculating the loss derivative with parameters (y) in the t-SNE problem

Solution
The similarity of data point x_j to data point x_i is the conditional probability, $p_{(j|i)}$.
For nearby data points, $p_{(j|i)}$ is relatively high, whereas for widely separated data points, $p_{(j|i)}$ will be almost infinitesimal
The conditional probability $p_{(j|i)}$ is given by:

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

where σ_i is the variance of the Gaussian that is centered on data point x_i For the low-dimensional counterparts y_i and y_j of the high-dimensional data points x_i and x_j, it is possible to compute a similar conditional probability, which we denote by $q_{(j|i)}$:

$$q_{(j|i)} = \frac{e^{-\|y_i - y_j\|^2}}{\sum_{k \neq i} e^{-\|y_i - y_k\|^2}}$$

Since we are only interested in modeling pairwise similarities, we set $q_{(i|i)} = 0$.
If the map points y_i and y_j correctly model the similarity between the high-dimensional data points x_i and x_j, the conditional probabilities $p_{j|i}$ and $q_{j|i}$ will be equal
SNE minimizes the sum of Kullback-Leibler divergences over all data points using a gradient descent method. The cost function C is given:

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

in which P_i represents the conditional probability distribution over all other data points given data point x_i, and Q_i represents the conditional probability distribution over all other map points given map point y_i

Because the Kullback-Leibler divergence is not symmetric, different types of error in the pairwise distances in the low-dimensional map are not weighted equal

SNE performs a binary search for the value of σ_i that produces a P_i with a fixed perplexity that is specified by the user. The perplexity is defined as:

$$Perp(P_i) = 2^{H(P_i)}$$

where $H(P_i)$ is the Shannon entropy of P_i measured in bits

$$H(P_i) = -\sum_j p_{(j|i)} log_2 p_{(j|i)}$$

The perplexity can be interpreted as a smooth measure of the effective number of neighbors. The performance of SNE is fairly robust to changes in the perplexity, and typical values are between 5 and 50.
The minimization of the cost function is performed using a gradient descent method. The gradient has a surprisingly simple form:

$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

Mathematically, the gradient update with a momentum term is given by:

$$\gamma^{(t)} = \gamma^{(t-1)} + \eta \frac{\delta C}{\delta y} + \alpha(t)(\gamma^{(t-1)} - \gamma^{(t-2)})$$

where $\gamma^{(t)}$ indicates the solution at iteration t, η indicates the learning rate, and $\alpha(t)$ represents the momentum at iteration t.

As an alternative to minimizing the sum of the Kullback-Leibler divergences between the conditional probabilities $p_{j|i}$ and $q_{j|i}$ it is also possible to minimize a single Kullback-Leibler divergence between a joint probability distribution, P, in the high-dimensional space and a joint probability distribution, Q, in the low-dimensional space:

$$C = \sum_i KL(P||Q) = \sum_i \sum_j p_{(j|i)} log \frac{p_{ij}}{q_{ij}}$$

Set p_{ii} and q_{ii} to zero. We refer to this type of SNE as symmetric SNE, because it has the property that $p_{ij} = p_{ji}$ and $q_{ij} = q_{ji} \forall i, j$
In symmetric SNE, the pairwise similarities in the low-dimensional map q_{ij} are given by:

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

The obvious way to define the pairwise similarities in the high-dimensional space $p_{(ij)}$ is:

$$p_{ij} = \frac{e^{-\|x_i - x_j\|^2/2\sigma^2}}{\sum_{k \neq l} e^{-\|x_k - x_l\|^2/2\sigma^2}}$$

The gradient of symmetric SNE is fairly similar to that of asymmetric SNE, and is given:

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)$$