HW3 Handwritten Assignment

Lecturor: Pei-Yuan Wu
TAs: Po-Yang Hsieh, Le-Rong Hsu

October 2024

Problem 1 (Principle Component Analysis) (0.5%)

Given 10 samples in 3D:

$$(1,2,3),(4,8,5),(3,12,9),(1,8,5),(5,14,2),(7,4,1),(9,8,9),(3,8,1),(11,5,6),(10,11,7)$$

1. What are the principal axes? Please write down your derivation or provide your code of computation at HW3 math problems (NTU COOL) through "add another file".
2. Please compute the principal components for each sample and either write down your derivation or provide your code of computation in report.
3. What is the average reconstruction error if reduce dimension to 2D? Here the reconstruction error is defined as the squared loss.

```
Eigenvalues:
 [16.99715933 12.9228041   6.08003657]
Principal Axes (Eigenvectors):
 [[-0.6165947  -0.67817891  0.39985541]
 [-0.58881629  0.73439013  0.33758926]
 [-0.52259579 -0.02728563 -0.85214385]]
Principal Components:
 [[ 7.18658682 -1.37323947 -2.25104047]
 [ 0.75871342  0.94399334 -0.73022635]
 [-3.07034019  4.45059025 -3.1883001 ]
 [ 2.60849751  2.97853006 -1.92979259]
 [-1.82299166  4.75401212  4.25159619]
 [ 3.35457763 -3.91896138  2.52755823]
 [-4.41464321 -2.55604371 -2.13952468]
 [ 3.46569126  1.73131477  2.27849363]
 [-2.31359638 -6.03371503  0.2038499 ]
 [-5.75249521 -0.97648096  0.97738622]]
Average Reconstruction Error (2D): 5.472032912651864
```

Problem 2 (Reparameterization Trick)(0.5%)

By working on this problem, we hope that students will understand VAEs from the perspective of deep latent variable models, as a supplement to the course materials. [1]

Let's recall the problem setting of the VAE. Let $p_{\mathcal{D}}$ be the data distribution and for each $\mathbf{x} \in \mathcal{D}$, $p_{\mathcal{D}}(\mathbf{x})$ is the probability of $\mathbf{x}$. A deep latent variable model, which is a generative model, represents the joint distribution $p_\theta(\mathbf{x}, \mathbf{z})$ over data and latent variables, and the likelihood of data generated by θ is given by

$$\begin{aligned} p_\theta(\mathbf{x}) &= \sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) && (\mathbf{z} \text{ discrete}) \\ &= \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} && (\mathbf{z} \text{ continuous}) \end{aligned}$$

[1] The idea of this problem follows from "An Introduction to Variational Autoencoders" by Diederik P. Kingma and Max Welling.

For a better understanding of this relation, please refer to the Problem 2 of HW1 as an example. By the property of the joint distribution, one has

$$p_\theta(\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{z}|\mathbf{x})}.$$

Although $p_\theta(\mathbf{x}, \mathbf{z})$ is tractable, the marginal probability $p_\theta(\mathbf{x})$ is intractable because $p_\theta(\mathbf{z}|\mathbf{x})$, the probability of latent variables conditioning on the data distribution, is typically intractable.

To deal with the intractability, VAE is invented by considering an *inference model* $q_\phi(\mathbf{z}|\mathbf{x})$, or encoder, parametrized by a neural network, denoted ϕ, to approximate $p_\theta(\mathbf{z}|\mathbf{x})$, while the generative model $p_\theta(\mathbf{x}, \mathbf{z})$ can be decomposed as

$$\begin{aligned} p_\theta(\mathbf{x}, \mathbf{z}) &= p_\theta(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z}) \\ &= p(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z}) \end{aligned}$$

where $p_\theta(\mathbf{z}) = p(\mathbf{z})$ is called the *prior distribution*, meaning that it is not conditioned on any observation, and $p_\theta(\mathbf{x}|\mathbf{z})$ is called the decoder which reconstructs data from latent variables.

1. For $\mathbf{x} \in \mathcal{D}$, please decompose $\log p_\theta(\mathbf{x})$ into ELBO and KL divergence. You may refer to the lecture slides but don't just give the final answer.

2. Let $\mathcal{L}_{\phi,\theta}(\mathbf{x})$ denote the ELBO. To perform optimization on ϕ, we have to take gradient of $\mathcal{L}_{\phi,\theta}(\mathbf{x})$ with respect to ϕ:

$$\begin{aligned} \nabla_\phi \mathcal{L}_{\phi,\theta}(\mathbf{x}) &= \nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right] \\ &\neq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\nabla_\phi \left(\log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right)\right] \end{aligned}$$

   where the last equality is generally not true. What is an alternative way to perform optimization? Please refer to section 2.4 in the reference. Also, show that the gradient of ELBO with respect to ϕ becomes exchangeable and is unbiased.

Solution: 1.

$$\begin{aligned} \log p_\theta(\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x})\right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right] && \text{(joint distribution identity)} \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\right] \\ &= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]}_{\text{ELBO}} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\right]}_{\text{KL divergence}} \end{aligned}$$

2. Essentially, the reparametrization trick is to apply a change of variable to the noise $\mathbf{z}$ which originally follows the density $q_\phi(\mathbf{z}|\mathbf{x})$ so that the randomness of $\mathbf{z}$ does not come from ϕ. Fix $\mathbf{x}$. The "invertible" transformation is of the form

$$\mathbf{z} = g(\epsilon, \phi, \mathbf{x})$$

where the random vector ϵ is independent of ϕ and $\mathbf{x}$ and follows a known distribution, e.g. Gaussian. Denote this distribution function $p(\epsilon)$. That is, the reparametrization can be expressed as

$$\begin{aligned} \mathbf{z} &= g(\epsilon, \phi, \mathbf{x}) \\ \epsilon &\sim p(\epsilon). \end{aligned}$$

For a function $f(\cdot)$ differentiable with respect to ϕ,

$$\begin{aligned}
\frac{\partial}{\partial\phi}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})] &= \frac{\partial}{\partial\phi}\mathbb{E}_{p(\epsilon)}[f(\mathbf{z})] \\
&= \mathbb{E}_{p(\epsilon)}[\frac{\partial}{\partial\phi}f(\mathbf{z})] && (\because \epsilon \text{ is independent of } \phi) \\
&= \mathbb{E}_{p(\epsilon)}[\frac{\partial}{\partial\phi}f(\mathbf{g}(\epsilon,\phi,\mathbf{x}))] \\
&= \mathbb{E}_{p(\epsilon)}[\frac{\partial f}{\partial\mathbf{z}}(\mathbf{z})\frac{\partial\mathbf{z}}{\partial\phi}] && (\because \mathbf{z}=g(\epsilon,\phi,\mathbf{x}) \text{ and chain rule}) \\
&= \mathbb{E}_{p(\epsilon)}[\frac{\partial f}{\partial\mathbf{z}}(\mathbf{z})\frac{\partial g(\epsilon,\phi,\mathbf{x})}{\partial\phi}]
\end{aligned}$$

As a consequence, the derivative of ELBO w.r.t. ϕ can be rewritten as

$$\begin{aligned}
\frac{\partial}{\partial\phi}\text{ELBO} &= \frac{\partial}{\partial\phi}\left\{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x},\mathbf{z})\right] - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log q_\phi(\mathbf{z}|\mathbf{x})\right]\right\} \\
&= \frac{\partial}{\partial\phi}\left\{\mathbb{E}_{p(\epsilon)}\left[\log p_\theta(\mathbf{x},\mathbf{z})\right] - \mathbb{E}_{p(\epsilon)}\left[\log q_\phi(\mathbf{z}|\mathbf{x})\right]\right\} \\
&= \frac{\partial}{\partial\phi}\mathbb{E}_{p(\epsilon)}\left[\log p_\theta(\mathbf{x},\mathbf{z})\right] - \frac{\partial}{\partial\phi}\mathbb{E}_{p(\epsilon)}\left[\log q_\phi(\mathbf{z}|\mathbf{x})\right] \\
&= \text{Part 1} - \text{Part 2}
\end{aligned}$$

For Part 1,

$$\begin{aligned}
\frac{\partial}{\partial\phi}\mathbb{E}_{p(\epsilon)}\left[\log p_\theta(\mathbf{x},\mathbf{z})\right] &= \mathbb{E}_{p(\epsilon)}\left[\frac{\partial}{\partial\phi}\log p_\theta(\mathbf{x},\mathbf{z})\right] \\
&= \mathbb{E}_{p(\epsilon)}\left[\frac{\partial}{\partial\mathbf{z}}\log p_\theta(\mathbf{x},\mathbf{z})\frac{\partial g(\epsilon,\phi,\mathbf{x})}{\partial\phi}\right]
\end{aligned}$$

For the second part, we need to deal with $\log q_\phi(\mathbf{z}|\mathbf{x})$ where $\mathbf{z}$ is a function of ϵ. We hope that $\log q_\phi(\mathbf{z}|\mathbf{x})$ can be written as a function of ϵ so that we can take expectation. By the change of variable theorem,

$$\begin{aligned}
q_\phi(\mathbf{z}|\mathbf{x}) &= p(\epsilon)\left|\det\frac{\partial\epsilon}{\partial\mathbf{z}}\right| \\
&= p(\epsilon)\left|\left(\det\frac{\partial\mathbf{z}}{\partial\epsilon}\right)^{-1}\right| \\
&= p(\epsilon)\left|\det\frac{\partial\mathbf{z}}{\partial\epsilon}\right|^{-1}
\end{aligned}$$

where $\mathbf{z}=g(\epsilon,\phi,\mathbf{x})$ and $\epsilon=g^{-1}(\mathbf{z})$ (this is the reason that we require g to be invertible). $\frac{\partial\mathbf{z}}{\partial\epsilon}$ is the Jacobian matrix

$$\frac{\partial\mathbf{z}}{\partial\epsilon} = \begin{bmatrix} \frac{\partial z_1}{\partial\epsilon_1} & \cdots & \frac{\partial z_1}{\partial\epsilon_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_k}{\partial\epsilon_1} & \cdots & \frac{\partial z_k}{\partial\epsilon_k} \end{bmatrix}.$$

Here, we also use the fact that *the inverse of the Jacobian matrix of an invertible function is the Jacobian matrix of the inverse of the function*, which comes from the inverse function theorem. We thus have

$$\log q_\phi(\mathbf{z}|\mathbf{x}) = \log p(\epsilon) - \log\left|\det\frac{\partial\mathbf{z}}{\partial\epsilon}\right| =: \log q(\epsilon).$$

Then,

$$\begin{aligned}\text{Part 2} &= \frac{\partial}{\partial \phi}\mathbb{E}_{p(\epsilon)}\left[\log q_\phi(\mathbf{z}|\mathbf{x})\right] \\ &= \frac{\partial}{\partial \phi}\mathbb{E}_{p(\epsilon)}\left[\log q(\epsilon)\right] \\ &= \mathbb{E}_{p(\epsilon)}\left[\frac{\partial}{\partial \phi}\log q(\epsilon)\right]\end{aligned}$$

and the remaining calculation is omitted.

Remark. As a TA, I didn't want to be too strict about these details. As long as you understand that the reparametrization trick is essentially a change of variables ensuring that the model parameter ϕ does not contribute to the randomness of $\mathbf{z}$, you will receive full marks.

Problem 3 (Laplacian Eigenmaps)(2%)

Consider an undirected connected graph G, which is shown below. We want to utilize Laplacian Eigenmaps method to reduce these 10 points to 3-dimensional space. Here, undirected graph means that edges in the graph do not have a direction, and connected graph means that there is a path from any node to any other node in the graph.

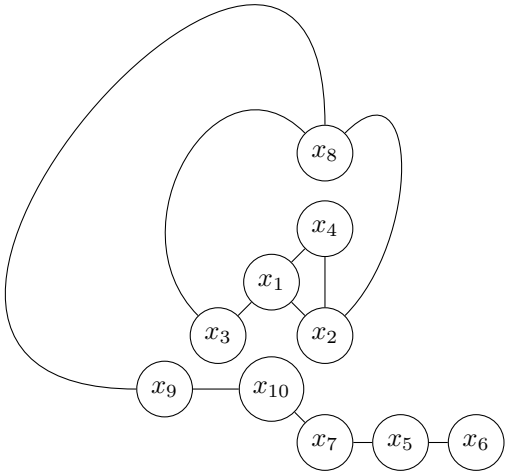

Figure 1: Problem 2 undirected connected graph G

1. Write down the adjacency matrix $\boldsymbol{W}$

2. Write down the diagonal matrix $\boldsymbol{D} = diag(d_1, ..., d_{10})$, where $d_i = \sum_{j=1}^{10} \frac{\boldsymbol{W_{ij}} + \boldsymbol{W_{ji}}}{2}$ and the Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$.

Solution:

```
W=
[[0 1 1 1 0 0 0 0 0 0]
 [1 0 0 1 0 0 0 1 0 0]
 [1 0 0 0 0 0 0 1 0 0]
 [1 1 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 1 1 0 0 0]
 [0 0 0 0 1 0 0 0 0 0]
 [0 0 0 0 1 0 0 0 0 1]
 [0 1 1 0 0 0 0 0 1 0]
 [0 0 0 0 0 0 0 1 0 1]
 [0 0 0 0 0 0 1 0 1 0]]
```

Solution:

```
D=
[[3 0 0 0 0 0 0 0 0 0]
 [0 3 0 0 0 0 0 0 0 0]
 [0 0 2 0 0 0 0 0 0 0]
 [0 0 0 2 0 0 0 0 0 0]
 [0 0 0 0 2 0 0 0 0 0]
 [0 0 0 0 0 1 0 0 0 0]
 [0 0 0 0 0 0 2 0 0 0]
 [0 0 0 0 0 0 0 3 0 0]
 [0 0 0 0 0 0 0 0 2 0]
 [0 0 0 0 0 0 0 0 0 2]]

L=
 [[ 3 -1 -1 -1  0  0  0  0  0  0]
 [-1  3  0 -1  0  0  0 -1  0  0]
 [-1  0  2  0  0  0  0 -1  0  0]
 [-1 -1  0  2  0  0  0  0  0  0]
 [ 0  0  0  0  2 -1 -1  0  0  0]
 [ 0  0  0  0 -1  1  0  0  0  0]
 [ 0  0  0  0 -1  0  2  0  0 -1]
 [ 0 -1 -1  0  0  0  0  3 -1  0]
 [ 0  0  0  0  0  0  0 -1  2 -1]
 [ 0  0  0  0  0  0 -1  0 -1  2]]
```

3. By HW2 Problem 3, Neighbor Embedding Slide p.7-p.10 and programming tools(MATLAB, Python...), solve the optimization problem
   minimize $Trace(\boldsymbol{\Psi}^T \boldsymbol{L} \boldsymbol{\Psi})$
   subject to $\boldsymbol{\Psi}^T \boldsymbol{D} \boldsymbol{\Psi} = \boldsymbol{I}_3$
   variables $\boldsymbol{\Psi} \in \mathbb{R}^{10 \times 3}$
   Also, please plot the reduced points $\boldsymbol{z_1}, ..., \boldsymbol{z_{10}}$ in 3-D scatter plot.

Solution: The scatter plot is shown below

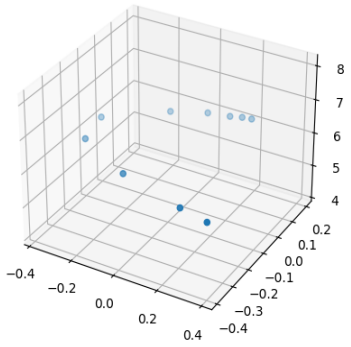

Figure 2: Scatter plot

and the Ψ matrix is given by the eigenvectors of the smallest three eigenvalues

```
[[ 0.21320072  0.18831754  0.14189576]
 [ 0.21320072  0.17464619  0.09357859]
 [ 0.21320072  0.15764162 -0.00046639]
 [ 0.21320072  0.19459486  0.18210673]
 [ 0.21320072 -0.3627844   0.25247422]
 [ 0.21320072 -0.38899742  0.39050752]
 [ 0.21320072 -0.28767812 -0.06404401]
 [ 0.21320072  0.10571999 -0.14249883]
 [ 0.21320072 -0.03650001 -0.36950084]
 [ 0.21320072 -0.17380083 -0.33528676]]
```

4. You may find that the minimal eigenvalue of $\boldsymbol{L}$ is 0, and the corresponding eigenvector is

$$\begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix} \tag{1}$$

where c is a constant. Since all the points fall into a plane, the span of these points is $\mathbb{R}^2$. In order to construct $\boldsymbol{z_1}, ..., \boldsymbol{z_{10}}$ such that $\text{span}\{\boldsymbol{z_1}, ..., \boldsymbol{z_{10}}\} = \mathbb{R}^3$, we need choose the second, third, fourth smallest eigenvalue and the corresponding eigenvectors. Please plot the reduced points by the updated $\boldsymbol{z_1}, ..., \boldsymbol{z_{10}}$ in 3-D scatter plot and verify that whether $Trace(\boldsymbol{\Psi}^T \boldsymbol{L} \boldsymbol{\Psi}) = 1.098$ and $\boldsymbol{\Psi}^T \boldsymbol{D} \boldsymbol{\Psi} = \boldsymbol{I}_3$.

Solution: We choose the second, third, and fourth smallest eigenvectors, so Ψ becomes

```
[[ 0.18831754  0.14189576 -0.09069099]
 [ 0.17464619  0.09357859 -0.09972695]
 [ 0.15764162 -0.00046639  0.30654733]
 [ 0.19459486  0.18210673 -0.29471481]
 [-0.3627844   0.25247422  0.09353918]
 [-0.38899742  0.39050752  0.28954605]
 [-0.28767812 -0.06404401 -0.22910952]
 [ 0.10571999 -0.14249883  0.28875405]
 [-0.03650001 -0.36950084  0.0730296 ]
 [-0.17380083 -0.33528676 -0.24156895]]
```

and the scatter plot is given by

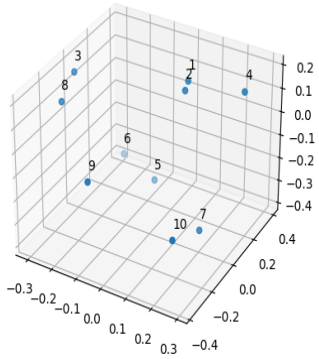

Figure 3: Scatter plot

5. Show that for no matter the graph is, there is an eigenvector of $\boldsymbol{L}$

$$\begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix} \tag{2}$$

where c is a constant, and the corresponding eigenvalue is 0.

Solution: Let $\boldsymbol{v} =$

$$\begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix} \tag{3}$$

.
$\boldsymbol{L}\boldsymbol{v}_i = \sum_{k=1}^{N} \boldsymbol{L}_{ik} c = c \sum_{k=1}^{N} \boldsymbol{L}_{ik} = c(\sum_{k=1}^{N} \boldsymbol{D}_{ik} - \sum_{k=1}^{N} \boldsymbol{D}_{ik}) = c(\text{deg of node i} - \text{deg of node i}) = c \times 0 = 0.$
Hence, $\boldsymbol{L}\boldsymbol{v} = \boldsymbol{0} = 0\boldsymbol{v}$, which indicates that $\boldsymbol{v}$ is an eigenvector of $\boldsymbol{L}$ that corresponds to eigenvalue 0.

6. By Neighbor Embedding Slide p.9, please show that

$$\forall \boldsymbol{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix} \in \mathbb{R}^N, \boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} = \frac{1}{2} \sum_{1 \le i,j \le N} w_{ij}(f_i - f_j)^2.$$

Solution: Under the knowledge on Neighbor Embedding Slide p.9 that

$$\text{Trace}(\boldsymbol{\Psi}^T \boldsymbol{L} \boldsymbol{\Psi}) = \frac{1}{2} \sum_{i,j=1,\ldots,N} w_{ij} ||z_i - z_j||_2^2$$

. Let $\boldsymbol{f} = \boldsymbol{\Psi} \in \mathbb{R}^{N \times 1}$, $\text{Trace}(\boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f}) = \boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f}$, $\sum_{i,j=1,\ldots,N} w_{ij} ||z_i - z_j||_2^2 = \sum_{i,j=1,\ldots,N} w_{ij(f_i - f_j)^2}$.
Hence we get $\boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} = \frac{1}{2} \sum_{1 \le i,j \le N} w_{ij}(f_i - f_j)^2$.

7. Show that if $\boldsymbol{f}$ is an eigenvector of $\boldsymbol{L}$ which corresponds to eigenvalue 0, then $\boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} = 0$.

Solution: $\boldsymbol{L}\boldsymbol{f} = 0\boldsymbol{f} = \boldsymbol{0}$, hence $\boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} = 0$.

8. Show that if the graph is connected, the second smallest eigenvalue of $\boldsymbol{L}$ will be nonzero.

Solution: By 6, we know that $\boldsymbol{L}$ is positive semidefinite matrix, which indicates that all the eigenvalues of $\boldsymbol{L}$ is non-negative. From the above result and 5. we deduced that 0 is the smallest eigenvalue, which corresponds to eigenvector $\boldsymbol{v_1}$ =

$$\begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix} \tag{4}$$

Also, because $\boldsymbol{L}$ is a symmetric matrix, it its diagonalizable, which means that $\boldsymbol{L}$ has N independent eigenvectors. Suppose $\boldsymbol{v_2}$ is an eigenvector corresponding to the second smallest eigenvalue λ_2. From the above facts, $\boldsymbol{v_2}$ can not be the multiply of $\boldsymbol{v_1}$, which means $\boldsymbol{v_2}$ can not be the form

$$\begin{bmatrix} d \\ d \\ \vdots \\ d \end{bmatrix} \tag{5}$$

On the other hand, because G is a connected graph, we can find a path from node i to node j. We define the nodes on path be $n_1 = i, n_2, ..., n_{k-1}, n_k = j$. $\sum_{1 \le i,j \le N} w_{ij}(v_{2_i} - v_{2_j})^2 \ge \sum_{1 \le r \le k-1} (v_{2_{n_r}} - v_{2_{n_{r+1}}})^2 > 0$ due to the fact that v_2 is not the multiply of all one vector. Also by 6., we get $\boldsymbol{v_2}^T \boldsymbol{L} \boldsymbol{v_2} = \boldsymbol{v_2}^T \boldsymbol{\lambda_2} \boldsymbol{v_2} = \boldsymbol{\lambda_2} ||\boldsymbol{v_2}||_2 = \sum_{1 \le i,j \le N} w_{ij}(v_{2_i} - v_{2_i})^2 > 0$, where $\boldsymbol{\lambda_2}$ is the second smallest eigenvalue. Hence by the above inequality, we show that the second smallest eigenvalue of $\boldsymbol{L}$ will be nonzero.

Problem 4 (Expectation Maximization Interpretation behind Semi-Supervised Learning)(1.5%)

Given N samples $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$ as well as their labels $y_1, \ldots, y_N \in \{0, 1, \ldots, K\}$. Consider the generative model where each sample $\mathbf{x}_i$ is generated independently according to Gaussian mixture model that depends on the label y_i, as represented by random variable

$$X_i \sim \begin{cases} \sum_{j=1}^{K} \pi_j \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) & , \text{ if } y_i = 0 \\ \mathcal{N}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) & , \text{ if } y_i = k \neq 0 \end{cases}$$

where $\pi_1 + \ldots + \pi_K = 1$, and $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with probability density function

$$\mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{\sqrt{(2\pi)^m \left|\boldsymbol{\Sigma}\right|}} \exp\left(-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}\right)\right)$$

We would like to apply Expectation Maximization algorithm to find the maximum likelihood estimation of parameters $\theta = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$.

1. Please write down the E-step and M-step and show that the parameters are updated from $\theta^{(t)} = \left\{\left(\pi_k^{(t)}, \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)\right\}_{k=1}^K$ to $\theta^{(t+1)} = \left\{\left(\pi_k^{(t+1)}, \boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t+1)}\right)\right\}_{k=1}^K$ in the following form:

$$\pi_k^{(t+1)} = \frac{\sum_{i:y_i=0} \delta_{ik}^{(t)}}{\sum_{i:y_i=0} 1}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i:y_i=k} \mathbf{x}_i + \sum_{i:y_i=0} \delta_{ik}^{(t)} \mathbf{x}_i}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i:y_i=k} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T + \sum_{i:y_i=0} \delta_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

where $N_k = \sum_{i:y_i=k} 1$ is the number of samples in class k. Please show your derivations.

2. What is the closed form expression of $\delta_{ik}^{(t)}$? Please show your derivations.

Solution: Given data set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, the likelihood and log-likelihood functions are given by

$$p(S;\theta) = \left(\prod_{i:y_i=0} \sum_{j=1}^K \pi_j \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \right) \cdot \prod_{k=1}^K \prod_{i:y_i=k} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right),$$

$$\log p(\mathcal{X};\theta) = \sum_{i:y_i=0} \log \left(\sum_{j=1}^K \pi_j \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \right) + \sum_{k=1}^K \sum_{i:y_i=k} \log \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

Denote latent variable $z_i \in \{1, \ldots, K\}$ indicating which Gaussian distribution x_i is drawn from.

- Expectation Step (E-step): Compute

$$Q\left(\theta \mid \theta^{(t)}\right) = \sum_{i=1}^N \mathbb{E}_{z_i|\mathbf{x}_i;y_i;\theta^{(t)}} \left[\log p\left(\mathbf{x}_i, y_i, z_i; \theta\right)\right]$$

Posterior prob. dist. of latent variables z_i based on current parameters $\theta^{(t)}$: If $y_i = 0$, then

$$\delta_{ik}^{(t)} = \mathbb{P}\left[z_i = k \mid \mathbf{x}_i; y_i = 0; \theta^{(t)}\right] = \frac{p\left(\mathbf{x}_i, y_i = 0, z_i = k; \theta^{(t)}\right)}{\sum_{j=1}^K p\left(\mathbf{x}_i, y_i = 0, z_i = j; \theta^{(t)}\right)} = \frac{\pi_k^{(t)} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)}{\sum_{j=1}^K \pi_j^{(t)} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right)}$$

If $y_i = l \neq 0$, then

$$\delta_{ik}^{(t)} = \mathbb{P}\left[z_i = k \mid \mathbf{x}_i; y_i = l; \theta^{(t)}\right] = \begin{cases} 1 & , \text{ if } l = k \\ 0 & , \text{ if } l \neq k \end{cases}$$

- Log-likelihood of parameter θ given data (x_i, y_i) and latent variable z_i :

If $y_i = 0$, then

$$\log p\left(\mathbf{x}_i, y_i = 0, z_i = k; \theta\right) = \log \left(\frac{\pi_k}{\sqrt{(2\pi)^m \left|\boldsymbol{\Sigma}_k\right|}} \right) - \frac{1}{2} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)$$

If $y_i = l \neq 0$, then

$$\log\left(\mathbf{x}_i, y_i = k, z_i = l; \theta\right) = \begin{cases} \log\left(\frac{1}{\sqrt{(2\pi)^m \left|\Sigma_k\right|}} \right) - \frac{1}{2} \left(\mathbf{x}_1 - \mu_k\right)^T \Sigma_k^{-1} \left(\mathbf{x}_1 - \mu_k\right) & l = k \\ \log(0) & l \neq k \end{cases}$$

Hence

$$Q\left(\theta \mid \theta^{(t)}\right) = \sum_{i=1}^N \sum_{k=1}^K \delta_{ik}^{(t)} \left\{ \log \left(\frac{\pi_k^{\mathbb{1}(y_i=0)}}{\sqrt{(2\pi)^m \left|\boldsymbol{\Sigma}_k\right|}} \right) - \frac{1}{2} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right) \right\}$$

- Maximization Step (M-step): Choose

$$\theta^{(t+1)} = \arg\max_{\theta \in \Theta} Q\left(\theta \mid \theta^{(t)}\right)$$

Note that $Q\left(\theta \mid \theta^{(t)}\right)$ takes exactly the same form as in the unsupervised GMM scenario, hence the optimal

solution $\theta^{(t+1)}$ is specified by

$$\pi_k^{(t+1)} = \frac{\sum_{i:y_i=0} \delta_{ik}^{(t)}}{\sum_{i:y_i=0} 1}$$

$$\boldsymbol{\mu}^{(t+1)} = \frac{\sum_{i=1}^N \delta_{ik}^{(t)} \mathbf{x}_i}{\sum_{i=1}^N \delta_{ik}^{(t)}} = \frac{\sum_{i:y_i=k} \mathbf{x}_i + \sum_{i:y_i=0} \delta_{ik}^{(t)} \mathbf{x}_i}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

$$= \sum_k^{(t+1)} = \frac{\sum_{i=1}^N \delta_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{\sum_{i=1}^N \delta_{ik}^{(t)}}$$

$$= \frac{\sum_{i:y_i=k} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T + \sum_{i:y_i=0} \delta_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

Problem 5 (EM for Mixture of Linear Models)(1.5%)

Consider the generative model parameterized by $\theta = ((\pi_k, \mathbf{w}_k, \rho_k))_{k=1}^K$, where $\mathbf{w}_k \in \mathbb{R}^m, \rho_k > 0, \pi_k > 0$ for each k, and that $\sum_{k=1}^K \pi_k = 1$, so that the probabilistic density function (PDF) of generating a real-valued number y given $\mathbf{x} \in \mathbb{R}^m$ is

$$p(y \mid \mathbf{x}; \theta) = \sum_{k=1}^K \pi_k \mathcal{N}\left(y; f_k(\mathbf{x}; \theta), \rho_k\right)$$

where $f_k(\mathbf{x}; \theta) = \mathbf{w}_k^\top \mathbf{x}$ and

$$\mathcal{N}(y; \mu, \rho) = \frac{1}{\sqrt{2\pi\rho}} \exp\left(-\frac{(y-\mu)^2}{2\rho}\right)$$

denotes the PDF of Gaussian distribution with mean μ and variance ρ. That is, $p(\cdot \mid \mathbf{x}; \theta)$ describes how the output is generated based on a mixture of linear models $f_k(\cdot; \theta)$ with mixing coefficient π_k and uncertainty ρ_k. Suppose we observe N data inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$ and their corresponding outputs $y_1, \ldots, y_N$. The maximum likelihood estimation is given by

$$\theta^{\text{opt}} = \operatorname*{argmax}_\theta \sum_{i=1}^N \log p\left(y_i \mid \mathbf{x}_i; \theta\right)$$

Derive the E-step and M-step equations of the EM algorithm for optimizing the linear model coefficients $\mathbf{w}_k$, the mixing coefficients π_k, as well as the variances ρ_k.

Solution: Denote $\mathcal{X} = (x_i)_{i=1}^N, \mathcal{Y} = (y_i)_{i=1}^N$, and latent variables $\mathcal{Z} = (z_1, \ldots, z_N) \in [1, K]^N$, and consider the joint probability distribution of $(\mathcal{Y}, \mathcal{Z})$ given $\mathcal{X}$ as

$$p(\mathcal{Y}, \mathcal{Z} \mid \mathcal{X}; \theta) = \prod_{i=1}^N p\left(y_i, z_i \mid x_i; \theta\right) \quad \text{where } p(y, z = k \mid \mathbf{x}; \theta) = \pi_k \mathcal{N}\left(y; f_k(\mathbf{x}; \theta), \rho_k\right)$$

At iteration t, the EM algorithm estimates $\theta^{(t+1)} = \left(\pi_k^{(t+1)}, \mathbf{w}_k^{(t+1)}, \rho_k^{(t+1)}\right)_{k=1}^K$ from $\theta^{(t)} = \left(\pi_k^{(t)}, \mathbf{w}_k^{(t)}, \rho_k^{(t)}\right)_{k=1}^K$ as follows: - Expectation Step: Compute

$$Q\left(\theta \mid \theta^{(t)}\right) = \mathbb{E}_{\mathcal{Z}|\mathcal{Y},\mathcal{X};\theta^{(t)}}[\log p(\mathcal{Y}, \mathcal{Z} \mid \mathcal{X}; \theta)] = \sum_{i=1}^N \mathbb{E}_{z_i|\mathbf{x}_i,y_i;\theta^{(t)}}\left[\log p\left(y_i, z_i \mid \mathbf{x}_i; \theta\right)\right]$$

The posterior probability of latent variables based on current parameters $\theta^{(t)}$ is given by

$$\delta_{i,k}^{(t)} = \mathbb{P}\left[z_i = k \mid \mathbf{x}_i, y_i; \theta^{(t)}\right] = \frac{p\left(y_i, z_i = k \mid \mathbf{x}_i; \theta^{(t)}\right)}{\sum_{l=1}^K p\left(y_i, z_i = l \mid \mathbf{x}_i; \theta^{(t)}\right)} = \frac{\pi_k \mathcal{N}\left(y_i; f_k\left(\mathbf{x}_i, \theta^{(t)}\right), \rho_k^{(t)}\right)}{\sum_{l=1}^K \pi l \mathcal{N}\left(y_i; f_l\left(\mathbf{x}_i, \theta^{(t)}\right), \rho_l^{(t)}\right)}.$$

The log-likelihood of parameter θ for jointly generating y_i, z_i given $\mathbf{x}_i$ is

$$\log p\left(y_i, z_i = k \mid \mathbf{x}_i; \theta\right) = \log \frac{\pi_k}{\sqrt{2\pi\rho_k}} - \frac{1}{2\rho_k}\left(y_i - f_k\left(\mathbf{x}_i; \theta\right)\right)^2$$

We can now write $Q\left(\theta \mid \theta^{(t)}\right)$ in explicit form as

$$\begin{aligned} Q\left(\theta \mid \theta^{(t)}\right) &= \sum_{i=1}^{N}\sum_{k=1}^{K} \delta_{i,k}^{(t)} \log p\left(y_i, z_i = k \mid \mathbf{x}_i; \theta\right) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} \delta_{i,k}^{(t)} \left(\log \frac{\pi_k}{\sqrt{2\pi\rho_k}} - \frac{\left(y_i - \mathbf{w}_k^\top \mathbf{x}_i\right)^2}{2\rho_k}\right) \end{aligned}$$

- Maximization Step: Choose $\theta^{(t+1)} = \text{argmax}_\theta\, Q\left(\theta \mid \theta^{(t)}\right)$. Taking partial derivatives of $Q\left(\theta \mid \theta^{(t)}\right)$ w.r.t. $\mathbf{w}_k$ and ρ_k yields

$$\nabla_{\mathbf{w}_k} Q\left(\theta \mid \theta^{(t)}\right) = \sum_{i=1}^{N} \delta_{i,k}^{(t)} \frac{\left(y_i - \mathbf{w}_k^\top \mathbf{x}_i\right)}{\rho_k} \mathbf{x}_i = \left(\sum_{i=1}^{N} \frac{\delta_{i,k}^{(t)}}{\rho_k} y_i \mathbf{x}_i\right) - \left(\sum_{i=1}^{N} \frac{\delta_{i,k}^{(t)}}{\rho_k} \mathbf{x}_i \mathbf{x}_i^\top\right) \mathbf{w}_k$$

$$\frac{\partial}{\partial \rho_k} Q\left(\theta \mid \theta^{(t)}\right) = \frac{1}{2}\sum_{i=1}^{N} \delta_{i,k}^{(t)} \left(\frac{\left(y_i - \mathbf{w}_k^\top \mathbf{x}_i\right)^2}{\rho_k^2} - \frac{1}{\rho_k}\right)$$

$$= \frac{1}{2}\left(\frac{1}{\rho_k^2}\sum_{i=1}^{N} \delta_{i,k}^{(t)} \left(y_i - \mathbf{w}_k^\top \mathbf{x}_i\right)^2 - \frac{1}{\rho_k}\sum_{i=1}^{N} \delta_{i,k}^{(t)}\right)$$

It follows that the maximal solution for $\mathbf{w}_k$ and ρ_k are

$$\mathbf{w}_k^{(t+1)} = \left(\sum_{i=1}^{N} \frac{\delta_{i,k}^{(t)}}{\rho_k} \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \left(\sum_{i=1}^{N} \frac{\delta_{i,k}^{(t)}}{\rho_k} y_i \mathbf{x}_i\right)$$

$$\rho_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{i,k}^{(t)} \left(y_i - \mathbf{w}_k^\top \mathbf{x}_i\right)^2}{\sum_{i=1}^{N} \delta_{i,k}^{(t)}}$$

As for π_k, we introduce Lagrange multiplier λ to deal with the constraint $\sum_{k=1}^{K} \pi_k = 1$, and note that

$$\frac{\partial}{\partial \pi_k}\left(Q\left(\theta \mid \theta^{(t)}\right) - \lambda \sum_{l=1}^{K} \pi_l\right) = \frac{1}{\pi_k}\sum_{i=1}^{N} \delta_{i,k}^{(t)} - \lambda$$

By setting the above quantity identially zero for all k, we solve $\lambda = N$, and the maximal solution for π_k is

$$\pi_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} \delta_{i,k}^{(t)}$$