Exercise 3

Ilana Pervoi 318271640, Omri Bar Oz 313325961

Question 1:

a) code

b) code

c)

Cluster number	Size	Common Label	Percentage
0	101	7	0.42
1	128	7	0.35
2	81	0	0.99
3	81	3	0.62
4	63	3	0.41
5	147	1	0.65
6	101	2	0.60
7	75	4	0.67
8	100	6	0.52
9	115	8	0.54

The algorithm classified correctly 568 out of 1000 samples → classification error of 43% on the sample.

We calculated it by searching how many samples were classified to the cluster with the label as their true label. We knew the true label of a sample by its index in the sample matrix, as we created the data by generating 100 samples from each digit.

d)

Cluster number	Size	Common Label	Percentage
0	291	1	0.1
1	1	0	1
2	1	2	1
3	1	0	1
4	1	4	1
5	1	5	1
6	1	5	1
7	1	5	1
8	1	6	1
9	1	6	1

The algorithm classified correctly 39 out of 300 samples → classification error of 87% on the sample.

The k-means clustering algorithm worked better for this problem than single linkage.

e)

K-means:

Cluster number	Size	Common Label	Percentage
0	192	3	0.39
1	182	7	0.38
2	95	0	0.80
3	148	6	0.36
4	211	1	0.47
5	172	2	0.19

The algorithm classified correctly 406 out of 1000 samples → classification error of 59% on the sample.

When moving from k=10 to k=6 we notice that the classification error grows. The reason behind it is that the number of the "real" clusters in the MNIST dataset is 10 (the groups of numbers from 0-9), and by running KMEANS on it with k=6, we force the algorithm to divide the dataset to only 6 clusters, resulting with at most 6 out of 10 possible labels coverage, which means more samples would be classified incorrectly.

Single Linkage:

Cluster number	Size	Common Label	Percentage
0	295	0	0.1
1	1	2	1
2	1	2	1
3	1	2	1
4	1	3	1
5	1	4	1

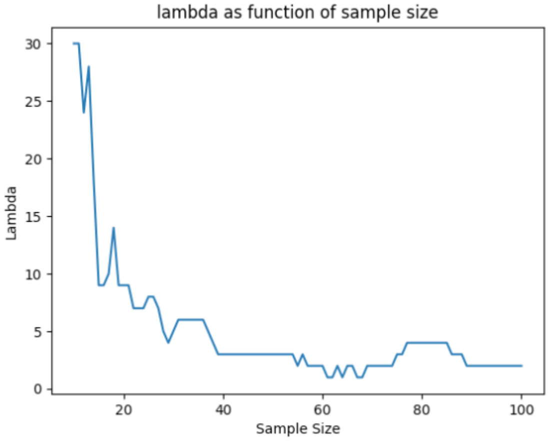
The algorithm classified correctly 35 out of 300 samples → classification error of 88% on the sample.

When moving from k=10 to k=6 we don't see a major difference in the results of the algorithm – most of the samples are assigned to one cluster, which results is about 87% error on the samples.
The reason behind it, is that the samples from MNIST dataset are quite close in respect to the Euclidean distance metric, hence the algorithm will merge clusters (wrongly) to one cluster, resulting in 1 big cluster and k-1 clusters with minimal

number of assigned samples (as we generate the initial clusters with examples from the sample set).

Question 2:

a)



b) We expect to see the value of optimal λ decreases as the sample size increases. This is because when the sample size is low, the sample does not represent the distribution properly, hence we will obtain large hypothesis class which will result in overfitting. To handle that, a higher λ will be required as a penalty to reduce the hypothesis class size. As the sample size increases, we expect to see a decrease in the optimal value of λ, until convergence. As the sample size increases, the hypothesis class size decreases. This results in less overfitting which means we don't want to penalize the norm of w as much.

c) Yes, this is we got what we expected as explained 2.b, in the plot submitted in 2.a.

d) In class we proved that the Bayes-optimal predicator with respect to the **squared loss** is

$$h_{bayes}(x) = E[Y|X = x]$$

$$\begin{aligned}
\rightarrow h_{bayes}(x) &= E[Y|X = x] = E[< w, x > +N|X = x] \\
&\underbrace{=}_{linearity\ of\ expectation} E[< w, x > |X = x] + E[N|X = x] \\
&= E[< w, x > |X = x] + E[N] \\
&=< w, x > +0 \\
&= < w, x >
\end{aligned}$$

In class we proved that the Bayes-optimal predicator with respect to the **absolute loss** is

$$h_{bayes}(x) = MEDIAN_{(X,Y)\sim D}[Y|X = x] =$$

$$b\ s.t.\ P(Y \geq b|X = x) \geq \frac{1}{2}\ \ and\ \ (Y \leq b|X = x) \geq \frac{1}{2}$$

Given

$$y =< w, x > +\eta$$
$$\eta \sim N(0, \sigma)\ drawn\ independently\ for\ each\ example$$
$$\sigma > 0$$

Let $N \sim N(0, \sigma)$ be the random variable that outputs η. We obtain:

$$h_{bayes}(x) =$$
$$b\ s.t.\ P(< w, x > +N \geq b|X = x) \geq \frac{1}{2}\ \ and\ \ P(< w, x > +N \leq b|X = x) \geq \frac{1}{2}$$
$$= b\ s.t.\ P(N \geq b - < w, x > |X = x) \geq \frac{1}{2}\ \ and\ \ P(N \leq b - < w, x > |X = x) \geq \frac{1}{2}$$
$$= b\ s.t.\ 1 - P(N \leq b - < w, x > |X = x) \geq \frac{1}{2}\ \ and\ \ P(N \leq b - < w, x > |X = x) \geq \frac{1}{2}$$
$$= b\ s.t.\ P(N \leq b - < w, x > |X = x) \leq \frac{1}{2}\ \ and\ \ F_{N|X=x}(b - < w, x >) \geq \frac{1}{2}$$
$$= b\ s.t.\ \ F_{N|X=x}(b - < w, x >) \leq \frac{1}{2}\ \ and\ \ F_{N|X=x}(b - < w, x >) \geq \frac{1}{2}$$

$N \sim N(0, \sigma)$ so:

$$\rightarrow b - < w, x > \geq 0\ \ and\ \ b - < w, x > \leq 0$$
$$\rightarrow b - < w, x > = 0$$
$$\rightarrow b =< w, x >$$

$$\rightarrow h_{bayes}(x) =< w, x >$$

Question 3:

a)

$$w^{(t+1)} := w^{(t)} - \eta \nabla f(w^{(t)})$$
$$f(w) = \lambda \|w\| + \sum_{i=1}^{m} (< w, x_i > -y_i)^2$$
$$\frac{\partial f}{\partial w} = \lambda \cdot \frac{1}{2} \cdot \frac{1}{\|w\|} \cdot 2w + 2\sum_{i=1}^{m} x_i(< w, x_i > -y_i) = \frac{\lambda w}{\|w\|} + 2X(X^T w - Y)$$
$$\rightarrow\ w^{(t+1)} = w^{(t)} - \eta(\frac{\lambda w^{(t)}}{\|w^{(t)}\|} + 2X(X^T w^{(t)} - Y))$$

b)

$$w^{(t+1)} := w^{(t)} - \eta_t\left(\nabla R\left(w^{(t)}\right) +\ \nabla \ell(w^{(t)}, (x_t, y_t))\right)$$

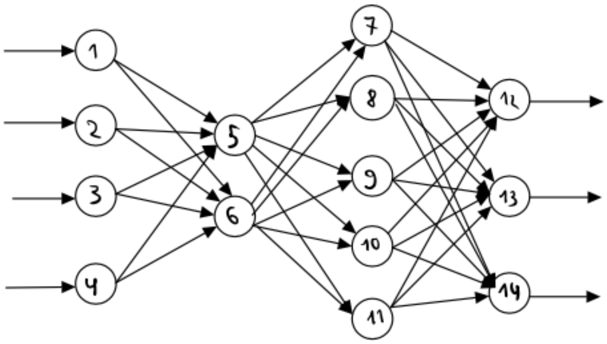
$$R(w) = \lambda\|w\| \rightarrow \nabla R(w) = \frac{\lambda w}{\|w\|}$$

$$\ell(w, (x_i, y_i) = (< w, x_i > -y_i)^2 \rightarrow \nabla\ell(w, (x_i, y_i) = 2x_i(< w, x_i > -y_i)$$

$$\rightarrow w^{(t)} - \eta_t \left(\frac{\lambda w^{(t)}}{\|w^{(t)}\|} + 2x_t(< w^{(t)}, x_t > -y_t) \right)$$

Question 4:

a) The graph $G = (V, E)$ that describes the neural network architecture:



b) $\mathcal{X} = \mathbb{R}^4$
c) $\mathcal{Y} = \mathbb{R}^3$
d) $h((x_1, x_2, x_3, x_4)) = \underset{12 \le i \le 14}{\text{argmax}}\left(\sum_{j=7}^{11} w_{j,i} \cdot \sigma\left(\sum_{k=5}^{6} w_{k,j} \cdot \sigma\left(\sum_{l=1}^{4} w_{l,k} \cdot x_l\right)\right)\right)$

Question 5:

a) Let $\mathcal{X} = [0,1]^d$ and $\mathcal{Y} = \{0,1\}$. Let $\overline{\mathcal{H}_n} \subseteq \{0,1\}^{\mathcal{X}}$ be the hypothesis class consisting of decision trees with depth at most n and binary attribute tests of the form $x(i) \ge \theta$ for $\theta \in \left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$.

For each tree in $\overline{\mathcal{H}_n}$ it has at most 2^{n+1} nodes. This is because the longest path is n. Then the tree with the largest number of nodes is a perfect binary tree with height n. Then the number of nodes is: $\sum_{i=0}^{n} 2^i = 2^{n+1} - 1 \le 2^{n+1}$.
For each node in the tree, we can select an attribute i, and then choose one of the possible values of θ to check if it is larger then. Then there are $3 \cdot d$ such options. A node can also be a leaf with label 0 or 1. Then every node has $3 \cdot d + 2$ options. Then:

$$|\overline{\mathcal{H}_n} | \le (3 \cdot d + 2)^{2^{n+1}} \quad \blacksquare$$

b) Danny is trying to use PAC boundaries equations we learned in class. The problem in this case is that ID3 is not an ERM algorithm.
This means that Danny is wrong using this equation.

Question 6:

a) No. We will show a contradiction to the Naïve-Bayes assumption:

$$\mathbb{P}[X = (-1,1)|Y = -1] = 0$$
$$\mathbb{P}(X(1) = -1|Y = -1) = 0.25$$
$$\mathbb{P}(X(2) = 1|Y = -1) = 0.2$$

$$\mathbb{P}(X(1) = -1|Y = -1) \cdot \mathbb{P}(X(2) = 1|Y = -1) = 0.25 \cdot 0.2$$
$$\neq 0 = \mathbb{P}[X = (-1,1)|Y = -1]$$

b) For each $x \in \mathcal{X}$ we will run:

$$h(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathbb{P}[Y = y] \cdot \mathbb{P}[X[1] = x[1]|Y = y] \cdot \mathbb{P}[X[2] = x[2]|Y = y]$$

And after that we will tell explicitly what h returns.
We will now calculate the probabilities
$\mathbb{P}[Y = y],\ \mathbb{P}[X[1] = x[1]|Y = y],\ \mathbb{P}[X[2] = x[2]|Y = y]$:

$$\mathbb{P}[Y = 1] = \frac{6 + 24 + 2 + 8}{60} = \frac{2}{3}$$
$$\mathbb{P}[Y = -1] = 1 - \mathbb{P}[Y = 1] = \frac{1}{3}$$
$$\mathbb{P}[X[1] = 1|Y = 1] = \frac{\frac{2 + 8}{60}}{\frac{2}{3}} = \frac{1}{4}$$
$$\mathbb{P}[X[2] = 1|Y = 1] = \frac{\frac{24 + 8}{60}}{\frac{2}{3}} = \frac{4}{5}$$
$$\mathbb{P}[X[1] = -1|Y = 1] = 1 - \mathbb{P}[X[1] = 1|Y = 1] = \frac{3}{4}$$
$$\mathbb{P}[X[2] = -1|Y = 1] = 1 - \mathbb{P}[X[2] = 1|Y = 1] = \frac{1}{5}$$
$$\mathbb{P}[X[1] = 1|Y = -1] = \frac{\frac{11 + 4}{60}}{\frac{1}{3}} = \frac{3}{4}$$
$$\mathbb{P}[X[2] = 1|Y = -1] = \frac{\frac{0 + 4}{60}}{\frac{1}{3}} = \frac{1}{5}$$
$$\mathbb{P}[X[1] = -1|Y = -1] = 1 - \mathbb{P}[X[1] = 1|Y = -1] = \frac{1}{4}$$
$$\mathbb{P}[X[2] = -1|Y = -1] = 1 - \mathbb{P}[X[2] = 1|Y = -1] = \frac{4}{5}$$

Now lets calculate $h\big((1,1)\big), h\big((1,-1)\big), h\big((-1,1)\big), h\big((-1,-1)\big)$:

$$h\big((1,1)\big) = \underset{y\in\mathcal{Y}}{\operatorname{argmax}} \begin{cases} \mathbb{P}[Y=1]\cdot\mathbb{P}[X[1]=1|Y=1]\cdot\mathbb{P}[X[2]=1|Y=-1] = \frac{2}{3}\cdot\frac{1}{4}\cdot\frac{4}{5} = \frac{2}{15} \\ \mathbb{P}[Y=-1]\cdot\mathbb{P}[X[1]=1|Y=-1]\cdot\mathbb{P}[X[2]=1|Y=-1] = \frac{1}{3}\cdot\frac{3}{4}\cdot\frac{1}{5} = \frac{1}{20} \end{cases} = 1$$

$$h\big((1,-1)\big) = \underset{y\in\mathcal{Y}}{\operatorname{argmax}} \begin{cases} \mathbb{P}[Y=1]\cdot\mathbb{P}[X[1]=1|Y=1]\cdot\mathbb{P}[X[2]=-1|Y=-1] = \frac{2}{3}\cdot\frac{1}{4}\cdot\frac{1}{5} = \frac{1}{30} \\ \mathbb{P}[Y=-1]\cdot\mathbb{P}[X[1]=1|Y=-1]\cdot\mathbb{P}[X[2]=-1|Y=-1] = \frac{1}{3}\cdot\frac{3}{4}\cdot\frac{4}{5} = \frac{1}{5} \end{cases} = -1$$

$$h\big((-1,1)\big) = \underset{y\in\mathcal{Y}}{\operatorname{argmax}} \begin{cases} \mathbb{P}[Y=1]\cdot\mathbb{P}[X[1]=-1|Y=1]\cdot\mathbb{P}[X[2]=1|Y=-1] = \frac{2}{3}\cdot\frac{3}{4}\cdot\frac{4}{5} = \frac{2}{5} \\ \mathbb{P}[Y=-1]\cdot\mathbb{P}[X[1]=-1|Y=-1]\cdot\mathbb{P}[X[2]=1|Y=-1] = \frac{1}{3}\cdot\frac{1}{4}\cdot\frac{1}{5} = \frac{1}{60} \end{cases} = 1$$

$$h\big((-1,-1)\big) = \underset{y\in\mathcal{Y}}{\operatorname{argmax}} \begin{cases} \mathbb{P}[Y=1]\cdot\mathbb{P}[X[1]=-1|Y=1]\cdot\mathbb{P}[X[2]=-1|Y=-1] = \frac{2}{3}\cdot\frac{3}{4}\cdot\frac{1}{5} = \frac{1}{10} \\ \mathbb{P}[Y=-1]\cdot\mathbb{P}[X[1]=-1|Y=-1]\cdot\mathbb{P}[X[2]=-1|Y=-1] = \frac{1}{3}\cdot\frac{1}{4}\cdot\frac{4}{5} = \frac{1}{15} \end{cases} = 1$$

In total, the predicator h we got:

$$\begin{gathered} h\big((1,1)\big) = 1 \\ h\big((1,-1)\big) = -1 \\ h\big((-1,1)\big) = 1 \\ h\big((-1,-1)\big) = 1 \end{gathered}$$

It is also the bayes optimal predicator, as it chooses for each of the examples from $\mathcal{X}$ the label with the highest probability for it, and as we saw in class, $h_{bayes}(x) \in \underset{y\in\mathcal{Y}}{\operatorname{argmax}}\, \eta_y(x)$ where $\eta_y(x) \coloneqq \mathbb{P}[Y=y\,|X=x]$. Which means that for each example x, the h_{bayes} predicts the label with the highest probability given x.

Question 7:

a) Since we want to reduce the dimensionality from 4 to 2, the distortion would be the sum of the lowest 2 eigenvalues of X^TX.
Notice that in the experiment the 3rd dimension is linearly dependent of the 1st and 2nd dimensions, and the 4th dimension linearly dependent of the 2nd and 3rd dimensions → linearly dependent of the 1st and 2nd dimensions.
Hence, the rank of X^TX would be 2 and it would have 2 eigenvalues that are 0, which are the 2 lowest ((*) as X^TX is positive semi-definite matrix, its eigenvalues are ≥ 0).
In conclusion, the distortion is $0+0=0$.

b) $X = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & -1 & 0 & 1 \end{bmatrix}$ satisfies the equations

$$x_t(3) = x_t(1)^2 + x_t(2)^3$$
$$x_t(4) = \left(x_t(3) - x_t(1)\right)^2$$

And since $A = X^T X = \begin{bmatrix} 2 & -1 & 1 & 1 \\ -1 & 2 & 1 & 0 \\ 1 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{bmatrix}$ has only one vector $v = \begin{bmatrix} 1 \\ 1 \\ -1 \\ 0 \end{bmatrix}$ such that $Av = 0v$, $X^T X$ has **only one** eigenvalue that is 0.

Combining with the fact in (*) we receive that the distortion is $\lambda_3 + 0 = \lambda_3 > 0$ ■

<u>Question 8:</u>

Let $\mathcal{X} = \{0,1,2\}$. Let $\Theta \subseteq [0,1]^3$ such that $\forall \theta \in \Theta: \theta(0) + \theta(1) + \theta(2) = 1$. Define Trinomial distribution $\mathcal{D}_\theta$ for $\theta \in \Theta$ as follows: $\mathbb{P}_{X\sim\mathcal{D}_\theta}[X = i] = \theta(i)$. Assume that we have a sample $S = x_1, \dots, x_m \sim \mathcal{D}_\theta$.

a) $\Theta' = \{\theta \in \Theta | \theta(0) = 3 \cdot \theta(1)\}$

Also assume that $\theta \in \Theta'$.

$$\hat{\theta} = \underset{\theta\in\Theta'}{\operatorname{argmax}} L(S, \theta)$$

$$L(S,\theta) = \sum_{i=1}^{m} \log\left(\mathbb{P}_{X\sim\mathcal{D}_\theta}[x_i = i]\right) =$$
$$= \sum_{i=1}^{m} \log\left(\theta(x_i)\right) =$$
$$= \sum_{i:x_i=0} \log(\theta(0)\,) + \sum_{i:x_i=1} \log(\theta(1)\,) + \sum_{i:x_i=2} \log(\theta(2)\,) =^*$$

Denote $\theta(1) = a$. Then we have $\theta(0) = 3a$.

We also have $\theta(0) + \theta(1) + \theta(2) = 1$ then $3a + a + \theta(2) = 1$ which means that $\theta(2) = 1 - 4a$. Then: $\theta = (3a, a, 1 - 4a)$.

Also, denote the amount of x_i in S with $x_i = j$ to be z_j. More formally:

$$\forall j \in \{0,1,2\}: z_j \coloneqq |\{x_i \in S | x_i = j\}$$

$$=^* \sum_{i:x_i=0} \log(3a\,) + \sum_{i:x_i=1} \log(a\,) + \sum_{i:x_i=2} \log(1 - 4a\,) =$$
$$= z_0 \cdot \log(3a) + z_1 \cdot \log(a) + z + 2 \cdot \log(1 - 4a).$$

we changed the likelihood function to be $L(S, a)$.

the maximum value of a for this problem:

$$\frac{\partial L}{\partial a} = \frac{3\cdot z_0}{3a} + \frac{z_1}{a} + \frac{(-4)\cdot z_2}{1-4a} = \frac{z_0}{a} + \frac{z_1}{a} + \frac{(-4)\cdot z_2}{1-4a} \overset{to\ find\ maximum}{\hat{=}} 0$$

$$\frac{z_0 + z_1}{a} = \frac{4\cdot z_2}{1-4a}$$

and because $z_0 + z_1 + z_2 = m$:

$$\frac{m - z_2}{a} = \frac{4\cdot z_2}{1-4a}$$
$$(m - z_2)\cdot(1-4a) = 4a\cdot z_2$$
$$m - 4ma - z_2 + 4a\cdot z_2 = 4a\cdot z_2$$
$$m - 4ma - z_2 = 0$$
$$a = \frac{m - z_2}{4m}$$

Hence the estimator is:

$$\hat{\theta}(S) = \left(\frac{3m - 3z_2}{4m}, \frac{m - z_2}{4m}, \frac{z_2}{m}\right)$$

b) $X \sim \mathcal{D}(p_1, \dots, p_k, \sigma_1, \dots, \sigma_k)$ where $\forall 1 \le i \le k: p_i \in [0,1], \sigma_i > 0 \wedge \Sigma_{i=1}^{k} p_i = 1$ and $N_i \sim N(1, \sigma_i^2)$ with $f_{N_i} = f_{\sigma_i}$ then:

$$X = \Sigma_{i=1}^{k} p_i \cdot N_i$$

X has the following density:

$$f_X(x) = f_{p_1,\dots,p_k,\sigma_1,\dots,\sigma_k}(x) = \Sigma_{i=1}^{k} p_i \cdot f_{N_i}(x) = \Sigma_{i=1}^{k} p_i \cdot f_{\sigma_i}(x)$$

We define Θ as follows:

$$\Theta = \left\{(p_1, \dots, p_k, \sigma_1, \dots, \sigma_k) \middle|\ (p_1, \dots, p_k) \in [0,1]^k \wedge \Sigma_{\mathrm{i}=1}^{k} p_i = 1 \wedge \forall 1 \le i \le k: \sigma_i > 0 \right\}$$

Let $S \sim D_\theta^m$ and let $Z \sim Multinomial(p_1, \dots, p_k, 1)^m$.
Then every $Z_i \sim Multinomial(p_1, \dots, p_k, 1)$ which means that $Z_i = j$ if x_i is from the $j'th$ gaussian distribution.

The sample is $\big((x_1, z_1), \dots, (x_m, z_m)\big)$ and the joint distribution is:

$$g_\theta(x, z) = p_z \cdot f_{\sigma_z}(x)$$

And the augmented likelihood is:

$$L(S, Z; \theta) = \Sigma_{i=1}^{m} \log\big(g(x_i, z_i)\big) = \Sigma_{i=1}^{m} \log\left(p_{z_i} \cdot f_{\sigma_{z_i}}(x_i)\right)$$
$$= \Sigma_{i=1}^{m} \log\big(p_{z_i}\big) + \Sigma_{i=1}^{m} \log\left(f_{\sigma_{z_i}}(x_i)\right)$$

Now the partial derivatives $\frac{\partial L}{\partial p_j}$ and $\frac{\partial L}{\partial \sigma_j}$ for every $1 \le j \le k$:

$$\frac{\partial L}{\partial p_j} = \Sigma_{i=1}^{m} \frac{1}{p_j} \cdot \mathbb{I}[z_i = j] = \frac{1}{p_j} \cdot \Sigma_{i=1}^{m} \mathbb{I}[z_i = j]$$

So the partial derivatives for every p_j is never 0, then every legal choose of $(p_1, \dots, p_k)$ won't change the maximization of the likelihood.

$$\frac{\partial L}{\partial \sigma_j} = \Sigma_{i=1}^{m} \frac{f_{\sigma_j}(x_i)'}{f_{\sigma_j}(x_i)} \cdot \mathbb{I}[z_i = j]$$

$$= \Sigma_{i=1}^{m} \frac{-\frac{1}{\sigma_j^2\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\frac{(x_i-1)^2}{\sigma_j^2}} + \frac{1}{\sigma_j\sqrt{2\pi}} \cdot \left(\frac{(x_i-1)^2}{\sigma_j^3}\right) \cdot e^{-\frac{1}{2}\cdot\frac{(x_i-1)^2}{\sigma_j^2}}}{\frac{1}{\sigma_j\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\frac{(x_i-1)^2}{\sigma_j^2}}}$$

$$\cdot \mathbb{I}[z_i = j] = \Sigma_{i=1}^{m} \frac{\frac{1}{\sigma_j^2\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\frac{(x_i-1)^2}{\sigma_j^2}} \left(-1 + \left(\frac{(x_i-1)^2}{\sigma_j^2}\right)\right)}{\frac{1}{\sigma_j\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\frac{(x_i-1)^2}{\sigma_j^2}}} \cdot \mathbb{I}[z_i = j]$$

$$= \Sigma_{i=1}^{m} \frac{1}{\sigma_j} \cdot \left(-1 + \left(\frac{(x_i-1)^2}{\sigma_j^2}\right)\right) \cdot \mathbb{I}[z_i = j]$$

$$= \frac{1}{\sigma_j} \cdot \left(\Sigma_{i=1}^{m} \frac{(x_i-1)^2}{\sigma_j^2} \cdot \mathbb{I}[z_i = j] - \Sigma_{i=1}^{m} \mathbb{I}[z_i = j]\right) = 0$$

$$\rightarrow \Sigma_{i=1}^{m} \frac{(x_i-1)^2}{\sigma_j^2} \cdot \mathbb{I}[z_i = j] = \Sigma_{i=1}^{m} \mathbb{I}[z_i = j]$$

$$\rightarrow \frac{1}{\sigma_j^2} \cdot \Sigma_{i=1}^{m} (x_i - 1)^2 \cdot \mathbb{I}[z_i = j] = \Sigma_{i=1}^{m} \mathbb{I}[z_i = j]$$

$$\rightarrow \sigma_j = \sqrt{\frac{\Sigma_{i=1}^{m} (x_i - 1)^2 \cdot \mathbb{I}[z_i = j]}{\Sigma_{i=1}^{m} \mathbb{I}[z_i = j]}}$$

In conclusion, for $\sigma_1, \dots \sigma_j$ with values as above, and for any $p_1, \dots, p_k$ where $p_i \in [0,1]$ for every $1 \le i \le k$ and $\Sigma_{i=1}^{k} p_i = 1$ the estimator $\theta = (\sigma_1, \dots, \sigma_k, p_1, \dots, p_k)$ is the maximum likelihood estimator.