Question 1

1) We note that:

$$\begin{aligned}
p(y=k|\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\sigma}) &= \frac{p(y=k,\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\sigma})}{p(x,\boldsymbol{\mu},\boldsymbol{\sigma})} \\
&= \frac{p(\boldsymbol{x}|y=k,\boldsymbol{\mu},\boldsymbol{\sigma})p(y=k|\boldsymbol{\mu},\boldsymbol{\sigma})p(\boldsymbol{\mu},\boldsymbol{\sigma})}{p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\sigma})p(\boldsymbol{\mu},\boldsymbol{\sigma})} \\
&= \frac{p(\boldsymbol{x}|y=k,\boldsymbol{\mu},\boldsymbol{\sigma})p(y=k|\boldsymbol{\mu},\boldsymbol{\sigma})}{p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\sigma})} \\
&= \frac{p(\boldsymbol{x}|y=k,\boldsymbol{\mu},\boldsymbol{\sigma})p(y=k|\boldsymbol{\mu},\boldsymbol{\sigma})}{\sum_{t=1}^{k} p(\boldsymbol{x}|y=t,\boldsymbol{\mu},\boldsymbol{\sigma})p(y=t|\boldsymbol{\mu},\boldsymbol{\sigma})} \\
&= \frac{p(\boldsymbol{x}|y=k,\boldsymbol{\mu},\boldsymbol{\sigma})p(y=k)}{\sum_{t=1}^{k} p(\boldsymbol{x}|y=t,\boldsymbol{\mu},\boldsymbol{\sigma})p(y=t)}
\end{aligned}$$

2) Observe that:

$$\begin{aligned}
\ell(\boldsymbol{\theta};D) &= -\log p(y^{(1)}=t_1,\boldsymbol{x}^{(1)},\ldots,y^N=t_N,\boldsymbol{x}^{(N)}|\boldsymbol{\theta}) \\
&= -\log(\prod_{i=1}^{N} p(y^{(i)}=t_i,\boldsymbol{x}^{(i)}|\boldsymbol{\theta})) \\
&= -\sum_{i=1}^{N} \log(p(\boldsymbol{x}^{(i)}|y^{(i)}=t_i,\boldsymbol{\theta})\cdot p(y^{(i)}=t_i|\boldsymbol{\theta})) \\
&= -\sum_{i=1}^{N} \log(p(\boldsymbol{x}^{(i)}|y^{(i)}=t_i,\boldsymbol{\mu},\boldsymbol{\sigma})\cdot \alpha_{t_i}) \\
&= -\sum_{i=1}^{N} \log(p(\boldsymbol{x}^{(i)}|y^{(i)}=t_i,\boldsymbol{\mu},\boldsymbol{\sigma})) + \log(\alpha_{t_i}) \\
&= -\sum_{i=1}^{N} \Big(-\frac{1}{2}\log(\prod_{p=1}^{D} 2\pi\sigma_p^2) - \sum_{p=1}^{D} \frac{1}{2\sigma_p^2}(x_p-\mu_{t_ip})^2 + \log(\alpha_{t_i})\Big) \\
&= -\sum_{i=1}^{N} \Big(-\frac{1}{2}\sum_{p=1}^{D}\log(2\pi\sigma_p^2) - \sum_{p=1}^{D} \frac{1}{2\sigma_p^2}(x_p-\mu_{t_ip})^2 + \log(\alpha_{t_i})\Big)
\end{aligned}$$

3) We first compute the partial derivative of the likelihood with the respect to

μ_{rj} for some $1 \leq r \leq k$ and $1 \leq j \leq D$:

$$\begin{aligned}
\frac{\partial}{\partial \mu_{rj}} \ell(\boldsymbol{\theta}; D) &= -\sum_{i=1}^{N} \Big(\cancelto{0}{-\frac{1}{2} \sum_{p=1}^{D} \frac{\partial}{\partial \mu_{rj}} \log(2\pi\sigma_p^2)} - \sum_{p=1}^{D} \frac{\partial}{\partial \mu_{rj}} \frac{1}{2\sigma_p^2} (x_p^{(i)} - \mu_{t_i p})^2 + \cancelto{0}{\frac{\partial}{\partial \mu_{rj}} \log(\alpha_{t_i})} \Big) \\
&= \sum_{i=1}^{N} \sum_{p=1}^{D} \frac{\partial}{\partial \mu_{rj}} \frac{1}{2\sigma_p^2} (x_p^{(i)} - \mu_{t_i p})^2 \\
&= \sum_{i=1}^{N} \sum_{p=1}^{D} \frac{\partial}{\partial \mu_{rj}} \frac{1}{2\sigma_p^2} (x_p^{(i)} - \mu_{t_i p})^2 \mathbb{I}(t_i = r) + \cancelto{0}{\sum_{i=1}^{N} \sum_{p=1}^{D} \frac{\partial}{\partial \mu_{rj}} \frac{1}{2\sigma_p^2} (x_p^{(i)} - \mu_{t_i p})^2 \mathbb{I}(t_i \neq r)} \\
&= \sum_{i=1}^{N} \frac{1}{\sigma_j^2} (x_j^{(i)} - \mu_{t_i j}) \mathbb{I}(t_i = r) \\
&= \frac{1}{\sigma_j^2} \sum_{i=1}^{N} (x_j^{(i)} - \mu_{t_i j}) \mathbb{I}(t_i = r) \\
&= \frac{1}{\sigma_j^2} \Big(\sum_{i=1}^{N} x_j^{(i)} \mathbb{I}(t_i = r) - \sum_{i=1}^{N} \mu_{t_i j} \mathbb{I}(t_i = r) \Big) \\
&= \frac{1}{\sigma_j^2} \Big(\sum_{i=1}^{N} x_j^{(i)} \mathbb{I}(t_i = r) - \mu_{rj} \sum_{i=1}^{N} \mathbb{I}(t_i = r) \Big)
\end{aligned}$$

Now we computer the partial derivative of the likelihood with respect to σ_j^2 for some $1 \leq j \leq D$:

$$\begin{aligned}
\frac{\partial}{\partial \sigma_j^2} \ell(\boldsymbol{\theta}; D) &= -\sum_{i=1}^{N} \Big(-\frac{1}{2} \sum_{p=1}^{D} \frac{\partial}{\partial \sigma_j^2} \log(2\pi\sigma_p^2) - \sum_{p=1}^{D} \frac{\partial}{\partial \sigma_j^2} \frac{1}{2\sigma_p^2} (x_p^{(i)} - \mu_{t_i p})^2 + \cancelto{0}{\frac{\partial}{\partial \sigma_j^2} \log(\alpha_{t_i})} \Big) \\
&= \sum_{i=1}^{N} \frac{1}{2\sigma_j^2} - \sum_{i=1}^{N} \frac{1}{2(\sigma_j^2)^2} (x_j^{(i)} - \mu_{t_i j})^2 \\
&= \frac{N}{2\sigma_j^2} - \frac{1}{2(\sigma_j^2)^2} \sum_{i=1}^{N} (x_j^{(i)} - \mu_{t_i j})^2 \\
&= \frac{N\sigma_j^2 - \sum_{i=1}^{N} (x_j^{(i)} - \mu_{t_i j})^2}{2(\sigma_j^2)^2}
\end{aligned}$$

4) We first find the MLE of $\boldsymbol{\mu}$. First we set $\frac{\partial}{\partial \mu_{rj}} \ell(\boldsymbol{\theta}; D) = 0$, arriving at:

$$\frac{1}{\sigma_j^2} \Big(\sum_{i=1}^{N} x_j^{(i)} \mathbb{I}(t_i = r) - \mu_{rj} \sum_{i=1}^{N} \mathbb{I}(t_i = r) \Big) = 0$$

from which we get:

$$\hat{\mu}_{rj} = \frac{\sum_{i=1}^{N} x_j^{(i)} I(t_i = r)}{\sum_{i=1}^{N} \mathbb{I}(t_i = r)}$$

and so the MLE $\hat{\boldsymbol{\mu}}$ is given by the matrix $(\boldsymbol{a_{rj}})$ where $a_{rj} = \hat{\mu}_{rj}$ for all $1 \leq r \leq k$ and $1 \leq j \leq D$.

Now we find the MLE of $\boldsymbol{\sigma}$. By setting $\frac{\partial}{\partial \sigma_j^2} \ell(\boldsymbol{\theta}; D) = 0$, we get that:

$$\frac{N\sigma_j^2 - \sum_{i=1}^{N} (x_j^{(i)} - \mu_{t_i j})^2}{2(\sigma_j^2)^2} = 0$$

and so we get that:

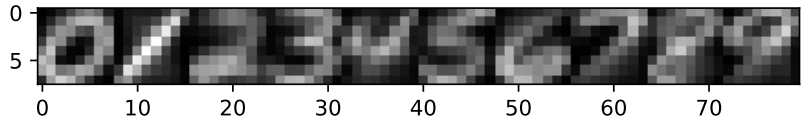
$$\hat{\sigma}_j^2 = \frac{1}{N} \sum_{i=1}^{N} (x_j^{(i)} - \mu_{t_i j})^2$$

Thus we get that the MLE for $\boldsymbol{\sigma}$ is:

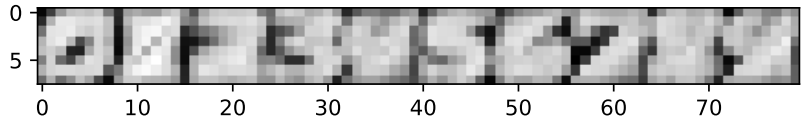
$$\hat{\boldsymbol{\sigma}} = \left(\frac{1}{N} \sum_{i=1}^{N} (x_1^{(i)} - \mu_{t_i 1})^2 \quad \frac{1}{N} \sum_{i=1}^{N} (x_2^{(i)} - \mu_{t_i 2})^2 \quad \dots \quad \frac{1}{N} \sum_{i=1}^{N} (x_D^{(i)} - \mu_{t_i D})^2 \right)$$

Question 2

2.0. Plot of the means for each digit classes in the training data:



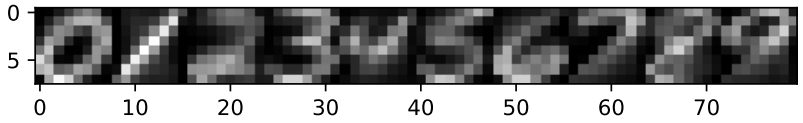
2.1.1. 8×8 plot of the log of the diagonal elements of each covariance matrix $\sum_k$:



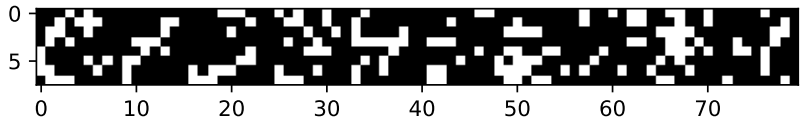
2.1.2. The average conditional log-likelihood for training set: -0.125
The average conditional log-likelihood for testing set: -0.197

2.1.3. Test Accuracy: 0.97275
Train Accuracy: 0.98143

2.2.3. Plot of η_k vectors as an 8×8 image for $k = 0, \ldots, 9$:



2.2.4. Plot of new data point using generative model using 8×8 grey-scale:



2.2.5. The average condition log-likelihood for training set: -0.944
Average Condition log-likelihood for Testing Set: -0.987

2.2.6. Test Accuracy: 0.76425
Train Accuracy: 0.77414

2.3. The Condition Gaussian Classifier performed very well on classifying hand-written digits with test accuracy of 97% and train accuracy 98%. This provides evidence that the classifier is generalizing well as the test accuracy is close to perfect and there is not much discrepancy between train and test accuracy. Furthermore, the average condition log-likelihood for the train and test set is relatively close to 0. This tells us that the classifier, more times than not, predicts the right classifier with high probability, which gives us more confidence that the classifier is generalizing well to the handwritten digits.

On the other hand, the Naive Bayes Classifier is much less impressive. The classifier has test accuracy of 76% and train accuracy of 77&, which is relatively much less accurate than the above classifier. Furthermore, the average condition log-likelihood for the train and test relatively is much further away from 0 than the Gaussian classifier. This tells us that, on average, the predictor puts relatively much less probability on the right class relative to the Gaussian classifier. Moreover, there is no discernible difference in the empirical run time of the two classifiers.

Hence, it is save to say that the Condition Gaussian Classifier performed the best, whereas, the Naive Bayes Classifier performed the worst. This matches my expectation because a lot of data is lost from the hand-written digit when the pixels are binarized, hence it makes it a lot harder to decide the right class based on lesser data. Therefore, it makes sense why the Naive Bayes Classifier had a hard time to classify the right digits, relative to the Gaussian.