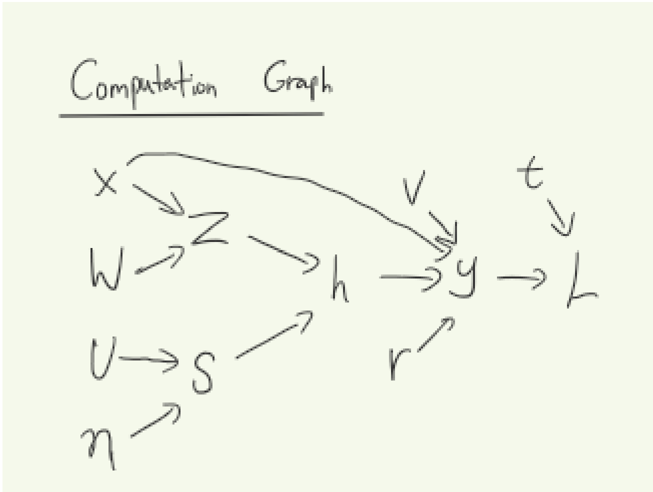
CSC311, Fall 2022, Homework 3

Huipeng Lei, 1005908154

Due Monday, November 7, 2022

Problem 1

a) 

b) Given $\sigma(x) = \frac{1}{1+e^{-x}}$. We can calculate

$$\begin{aligned} 1 - \sigma(x) &= 1 - \frac{1}{1+e^{-x}} \\ &= \frac{1+e^{-x}-1}{1+e^{-x}} \\ &= \frac{e^{-x}}{1+e^{-x}} \end{aligned}$$

Then

$$\begin{aligned} \frac{d\sigma(x)}{dx} &= \frac{-(-e^{-x})}{(1+e^{-x})^2} \\ &= \frac{1}{1+e^{-x}} \frac{e^{-x}}{1+e^{-x}} \\ &= \sigma(x)(1-\sigma(x)) \end{aligned}$$

c)

$$
\begin{aligned}
&\overline{\mathcal{L}} = 1 \\
&\overline{y} = \overline{\mathcal{L}}\frac{\partial \mathcal{L}}{\partial y} = \overline{\mathcal{L}}(\frac{t}{y} + \frac{t-1}{1-y}) = \overline{\mathcal{L}}\frac{t-y}{y(1-y)} \\
&\overline{\mathbf{v}} = \overline{y}\frac{\partial y}{\partial \mathbf{v}} = \overline{y}\sigma(\mathbf{v}^T\mathbf{h} + \mathbf{r}^T\mathbf{x})(1 - \sigma(\mathbf{v}^T\mathbf{h} + \mathbf{r}^T\mathbf{x}))\mathbf{h}^T \\
&\overline{\mathbf{r}} = \overline{y}\frac{\partial y}{\partial \mathbf{r}} = \overline{y}\sigma(\mathbf{v}^T\mathbf{h} + \mathbf{r}^T\mathbf{x})(1 - \sigma(\mathbf{v}^T\mathbf{h} + \mathbf{r}^T\mathbf{x}))\mathbf{x}^T \\
&\overline{\mathbf{h}} = \overline{y}\frac{\partial y}{\partial \mathbf{h}} = \overline{y}\sigma(\mathbf{v}^T\mathbf{h} + \mathbf{r}^T\mathbf{x})(1 - \sigma(\mathbf{v}^T\mathbf{h} + \mathbf{r}^T\mathbf{x}))\mathbf{v}^T \\
&\overline{\mathbf{z}} = \overline{\mathbf{h}}\frac{\partial \mathbf{h}}{\partial \mathbf{z}} = \overline{\mathbf{h}}J_{\mathbf{z}}(\mathbf{h}) = \overline{h}\begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_n \end{bmatrix} \\
&\qquad\qquad (J_{\mathbf{z}} \text{ means to apply the Jacobian Matrix wrt. } \mathbf{z}) \\
&\overline{\mathrm{s}} = \overline{h}\frac{\partial \mathbf{h}}{\partial \mathrm{s}} = \overline{h}J_{\mathrm{s}}(\mathbf{h}) = \overline{h}\begin{bmatrix} z_1 & 0 & \cdots & 0 \\ 0 & z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_n \end{bmatrix} \\
&\qquad\qquad (J_{\mathrm{s}} \text{ means to apply the Jacobian Matrix wrt. s}) \\
&\overline{\mathbf{W}} = \overline{\mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{W}} = \overline{\mathbf{z}}(\mathbf{x}^T \otimes \mathbf{I}) \\
&\qquad\qquad (\mathbf{x}^T \otimes \mathbf{I} \text{ a tensor; } \otimes \text{ the Kronecker product; } \mathbf{I} \text{ identity matrix}) \\
&\overline{\mathbf{x}} = \overline{\mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{x}} + \overline{y}\frac{\partial y}{\partial \mathbf{x}} = \overline{\mathbf{z}}\mathbf{W} + \overline{y}\sigma(\mathbf{v}^T\mathbf{h} + \mathbf{r}^T\mathbf{x})(1 - \sigma(\mathbf{v}^T\mathbf{h} + \mathbf{r}^T\mathbf{x}))\mathbf{r}^T \\
&\overline{\mathbf{U}} = \overline{\mathrm{s}}\frac{\partial \mathrm{s}}{\partial \mathbf{U}} = \overline{\mathrm{s}}(\eta^T \otimes \mathbf{I}) \\
&\qquad\qquad (\eta^T \otimes \mathbf{I} \text{ a tensor; } \otimes \text{ the Kronecker product; } \mathbf{I} \text{ identity matrix}) \\
&\overline{\eta} = \overline{\mathrm{s}}\frac{\partial \mathrm{s}}{\partial \eta} = \overline{\mathrm{s}}\mathbf{U}
\end{aligned}
$$

Problem 2

a) $\mathbf{L}(\boldsymbol{\theta}) = \prod_{i=1}^{N}(P(t^{(i)}|\pi)\prod_{j=1}^{784} P(x_j^{(i)}|c,\theta_{jc}))$. Then

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{N} log(P(t^{(i)}|\pi)) + \sum_{i=1}^{N}\sum_{j=1}^{784} log(P(x_j^{(i)}|c^{(i)},\theta_{jc}) = \sum_{i=1}^{N} log(\prod_{j=0}^{9}\pi_j^{t_j^{(i)}}) + \sum_{i=1}^{N}\sum_{j=1}^{784} log(P(x_j^{(i)}|c^{(i)},\theta_{jc})$$

<u>**Find the MLE for $\boldsymbol{\pi}$**</u>

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\pi}} &= \begin{bmatrix} \frac{\partial l(\boldsymbol{\theta})}{\partial \pi_0} & \frac{\partial l(\boldsymbol{\theta})}{\partial \pi_1} & \cdots & \frac{\partial l(\boldsymbol{\theta})}{\partial \pi_9} \end{bmatrix} \\
\frac{\partial l(\boldsymbol{\theta})}{\partial \pi_j} &= \frac{\partial}{\partial \pi_j}\sum_{i=1}^{N} log(\prod_{j=0}^{9}\pi_j^{t_j^{(i)}}) \\
&= \frac{\partial}{\partial \pi_j}\sum_{i=1}^{N}\sum_{j=0}^{9} log(\pi_j^{t_j^{(i)}}) \\
&= \frac{\partial}{\partial \pi_j}\sum_{i=1}^{N}(t_0^{(i)}log(\pi_0) + \cdots + t_j^{(i)}log(\pi_j) + \cdots + t_8^{(i)}log(\pi_8) + t_9^{(i)}log(1-\sum_{k=0}^{8}\pi_k)) \\
&= \sum_{i=1}^{N}[\frac{t_j^{(i)}}{\pi_j} - \frac{t_9^{(i)}}{1-\sum_{k=0}^{8}\pi_k}] \\
0 &= \frac{1}{\pi_j}\sum_{i=1}^{N} t_j^{(i)} - \frac{1}{\pi_9}\sum_{i=1}^{N} t_9^{(i)} \\
\implies \hat{\pi}_j &= \hat{\pi}_9\sum_{i=1}^{N}\frac{t_j^{(i)}}{t_9^{(i)}} = \frac{\sum_{i=1}^{N} t_j^{(i)}}{N}, \forall j \in \{0,\ldots,8\}, \qquad (\pi_9 = \frac{\sum_{i=1}^{N} t_9^{(i)}}{N}) \\
&= \frac{\#\text{ class } t_j \text{ in dataset}}{\text{Total } \# \text{ of samples}}
\end{aligned}$$

We can solve for

$$\begin{aligned}
\hat{\pi}_9 &= 1 - \sum_{j=0}^{8}\hat{\pi}_j \\
&= 1 - \frac{1}{N}\sum_{i=1}^{N}\left(t_0^{(i)} + t_1^{(i)} + \cdots + t_8^{(i)}\right) \\
&= \frac{N - \sum_{i=1}^{N}\left(t_0^{(i)} + t_1^{(i)} + \cdots + t_8^{(i)}\right)}{N} \\
&= \frac{\sum_{i=1}^{N} t_9^{(i)}}{N}
\end{aligned}$$

Thus,

$$\begin{aligned}
\hat{\pi} &= \begin{pmatrix} \hat{\pi}_0 & \hat{\pi}_1 & \cdots & \hat{\pi}_8 & \hat{\pi}_9 \end{pmatrix} \\
&= \begin{pmatrix} \frac{\sum_{i=1}^N t_0^{(i)}}{N} & \frac{\sum_{i=1}^N t_1^{(i)}}{N} & \dots & \frac{\sum_{i=1}^N t_8^{(i)}}{N} & \frac{\sum_{i=1}^N t_9^{(i)}}{N} \end{pmatrix}
\end{aligned}$$

<u>Find the MLE for $\boldsymbol{\theta}$</u>

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_{jc}} &= \frac{\partial}{\partial \theta_{jc}} \sum_{i=1}^{N} \sum_{j=1}^{784} log(P(x_j^{(i)}|c^{(i)}, \theta_{jc}) \\
&= \frac{\partial}{\partial \theta_{jc}} \sum_{i=1}^{N} \sum_{j=1}^{784} \Big[x_j^{(i)} log\Big(\theta_{jc}\Big) + (1 - x_j^{(i)}) log\Big(1 - \theta_{jc}\Big)\Big] \\
& \qquad\qquad\qquad\qquad (\text{Since } P(x_j|c, \theta_{jc}) = \theta_{jc}^{x_j}(1-\theta_{jc})^{(1-x_j)}) \\
0 &= \sum_{i=1}^{N} \mathbb{1}(c^{(i)} = c)\Big(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1 - x_j^{(i)}}{1 - \theta_{jc}}\Big) \\
0 &= \frac{1}{\theta_{jc}(1-\theta_{jc})} \sum_{i=1}^{N} \mathbb{1}(c^{(i)} = c)\Big((1-\theta_{jc})x_j^{(i)} - \theta_{jc}(1 - x_j^{(i)})\Big) \\
0 &= \sum_{i=1}^{N} \mathbb{1}(c^{(i)} = c)(x_j^{(i)} - \theta_{jc}x_j^{(i)} - \theta_{jc} + \theta_{jc}x_j^{(i)}) \\
\sum_{i=1}^{N} \mathbb{1}(c^{(i)} = c)\theta_{jc} &= \sum_{i=1}^{N} \mathbb{1}(c^{(i)} = c)x_j^{(i)} \\
\implies \widehat{\theta_{jc}} &= \frac{\sum_{i=1}^N \mathbb{1}(c^{(i)} = c)x_j^{(i)}}{\sum_{i=1}^N \mathbb{1}(c^{(i)} = c)} \forall j = 1, 2, \dots, 784 \\
\implies \widehat{\theta_{jc}} &= \frac{N_{jc}}{N_c}
\end{aligned}$$

(N_{jc} denotes # of class c images with jth pixel 1; N_c denotes # class c images)

Therefore,

$$\boldsymbol{\theta} = \begin{bmatrix} \frac{N_{10}}{N_0} & \frac{N_{11}}{N_1} & \cdots & \frac{N_{19}}{N_9} \\ \frac{N_{20}}{N_0} & \frac{N_{21}}{N_1} & \cdots & \frac{N_{29}}{N_9} \\ \cdots & \cdots & \ddots & \cdots \\ \frac{N_{7840}}{N_0} & \frac{N_{7841}}{N_1} & \cdots & \frac{N_{7849}}{N_9} \end{bmatrix}$$

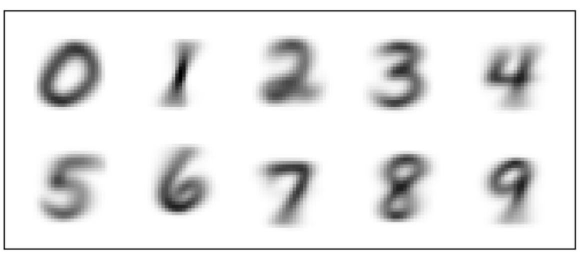
b) By Bayes' Rule, we can obtain

$$P(\mathrm{t}|\mathrm{x}, \boldsymbol{\theta}, \boldsymbol{\pi}) = P(c|\mathrm{x}, \boldsymbol{\theta}, \boldsymbol{\pi}) = \frac{P(\boldsymbol{x}, c|\boldsymbol{\theta}, \boldsymbol{\pi})}{\sum_{c'} P(\boldsymbol{x}, c'|\boldsymbol{\theta}, \boldsymbol{\pi})}$$

The denominator sums the conditional probability over all classes.
Then,

$$
\begin{aligned}
logP(\mathrm{t}|\mathrm{x},\boldsymbol{\theta},\boldsymbol{\pi}) &= log\frac{P(\boldsymbol{x},c|\boldsymbol{\theta},\boldsymbol{\pi})}{\sum_{c'}P(\boldsymbol{x},c'|\boldsymbol{\theta},\boldsymbol{\pi})}\\
&= logP(\boldsymbol{x},c|\boldsymbol{\theta},\boldsymbol{\pi}) - log\sum_{c'}P(\boldsymbol{x},c'|\boldsymbol{\theta},\boldsymbol{\pi})\\
&= log(P(c|\boldsymbol{\pi})\prod_{j=1}^{784}P(x_j|c,\theta_{jc}) - log(\sum_{c}P(c|\boldsymbol{\pi})\prod_{j=1}^{784}P(x_j|c,\theta_{jc}))\\
&= log(P(c|\boldsymbol{\pi})) + \sum_{j=1}^{784}log(P(x_j|c,\theta_{jc})) - log(\sum_{c}P(c|\boldsymbol{\pi})\prod_{j=1}^{784}P(x_j|c,\theta_{jc}))\\
&= log(\pi_c) + \sum_{j=1}^{784}\Big(x_j log\theta_{jc} + (1-x_j)log(1-\theta_{jc})\Big)\\
&\qquad - log\Big(\sum_{i=0}^{9}\pi_i\prod_{j=1}^{784}\theta_{jc^{(i)}}^{x_j}(1-\theta_{jc^{(i)}})^{1-x_j}\Big)
\end{aligned}
$$

c) Average log-likelihood for MLE is Nan due to numerical errors like division by zero or log 0 error. From these errors, there exists some division by zero or log 0 error due to some θ_{jc} being 0. This is due to data sparsity of input.



d)

e) Using a $Beta(\alpha,\beta)$ prior on each θ_{jc}, we have $P(\theta_{jc}) = \theta_{jc}^{\alpha-1}(1-\theta_{jc})^{\beta-1}$.

Find MAP estimator for $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}_{MAP}} = \underset{\boldsymbol{\theta}}{argmax} P(\boldsymbol{\theta}|\boldsymbol{x}, c, \boldsymbol{\pi}) = \underset{\boldsymbol{\theta}}{argmax} P(\boldsymbol{\theta}|c)P(\boldsymbol{x}|c, \boldsymbol{\theta})$

$$P(\boldsymbol{\theta}|c)P(\boldsymbol{x}|c,\boldsymbol{\theta}) = \theta_{jc}^{\alpha-1}(1-\theta_{jc})^{\beta-1}\prod_{i=1}^{N}\prod_{j=1}^{784}\theta_{jc}^{x_j^{(i)}}(1-\theta_{jc})^{1-x_j^{(i)}}$$

$$\begin{aligned} log(P(\boldsymbol{\theta}|c)P(\boldsymbol{x}|c,\boldsymbol{\theta})) &= (\alpha-1)log\theta_{jc} + (\beta-1)log(1-\theta_{jc})) \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{784}\left[x_j^{(i)}log\left(\theta_{jc}\right) + (1-x_j^{(i)})log\left(1-\theta_{jc}\right)\right] \end{aligned}$$

Take the derivative wrt. θ_{jc}

$$\begin{aligned} 0 &= \frac{\alpha-1-\alpha\theta_{jc}+\theta_{jc}-\beta\theta_{jc}+\theta_{jc}}{\theta_{jc}(1-\theta_{jc})} + \frac{1}{\theta_{jc}(1-\theta_{jc})}\sum_{i=1}^{N}\mathbb{1}(c^{(i)}=c)(x_j^{(i)}-\theta_{jc}) \\ &= \alpha-1-\theta_{jc}(\alpha+\beta-2) + \sum_{i=1}^{N}\mathbb{1}(c^{(i)}=c)x_j^{(i)} - \sum_{i=1}^{N}\mathbb{1}(c^{(i)}=c)\theta_{jc} \end{aligned}$$

$$\begin{aligned} MAP(\widehat{\theta_{jc}}) &= \frac{\sum_{i=1}^{N}\mathbb{1}(c^{(i)}=c)x_j^{(i)}+\alpha-1}{\sum_{i=1}^{N}\mathbb{1}(c^{(i)}=c)+\alpha+\beta-2} \\ &= \frac{N_{jc}+\alpha-1}{N_c+\alpha+\beta-2} \end{aligned}$$

(N_{jc} denotes # of class c images with jth pixel 1; N_c denotes # class c images)

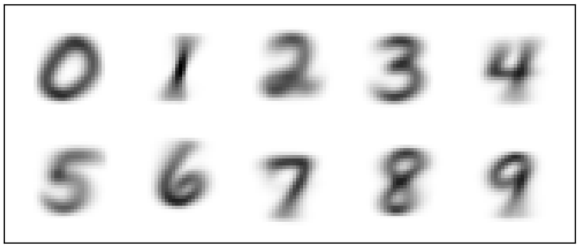
When $\alpha = 3, \beta = 3$,

$$MAP(\hat{\theta_{jc}}) = \frac{N_{jc}+2}{N_c+4}$$

In contrast to $\hat{\theta}_{jcMLE} = \frac{N_{jc}}{N_c}$, the numerator and denominator of the MAP estimator has more counts than the MLE, where α and β are can be think as pseudo-counts.

f)

```
Average log-likelihood for MLE is  nan
Average log-likelihood for MAP is  -3.3570625208614815
Training accuracy for MAP is  0.8352166666666667
Test accuracy for MAP is  0.816
```



g)

h) The Naive Bayes assumption may be reasonable in this problem where images are independent.The Naive Bayes assumption may not be reasonable since in reality , it is almost impossible to have a set of completely independent predictors, which in this problem, the pixels may be dependent on other pixels.

Problem 3

a) From the question, we know

$$P(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{x_k}$$

$$P(\boldsymbol{\theta}) \propto \theta_1^{a_1-1} \cdots \theta_K^{a_k-1}$$

Then the posterior distribution can be calculated with

$$\begin{aligned}
P(\boldsymbol{\theta}|\mathcal{D}) &\propto P(\boldsymbol{\theta})P(\mathcal{D}|\boldsymbol{\theta}) \\
&\propto \theta_1^{a_1-1}\theta_2^{a_2-1} \cdots \theta_K^{a_k-1} \prod_{i=1}^{N}\prod_{k=1}^{K} \theta_k^{x_k^{(i)}} \\
&\propto \theta_1^{a_1-1}\theta_2^{a_2-1} \cdots \theta_K^{a_k-1} \prod_{k=1}^{K} \theta_k^{\sum_{i=1}^{N} x_k^{(i)}} \\
&\propto \theta_1^{a_1-1}\theta_2^{a_2-1} \cdots \theta_K^{a_k-1} \prod_{k=1}^{K} \theta_k^{N_k} \\
&\propto \theta_1^{N_1+a_1-1}\theta_2^{N_2+a_2-1} \cdots \theta_K^{N_k+a_k-1}
\end{aligned}$$

We have computed the posterior distribution to be a Dirichlet distribution with

$$\boldsymbol{\theta} \sim \text{Dirichlet}(N_1 + a_1, N_2 + a_2, \ldots, N_k + a_k)$$

Hence, the Dirichlet distribution is a conjugate prior for the categorical distribution.

b) The MAP estimate of $\boldsymbol{\theta}$ is

$$\hat{\theta}_{MAP} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ P(\boldsymbol{\theta}|\mathcal{D}) = \underset{\boldsymbol{\theta}}{\text{argmax}}\ log(P(\boldsymbol{\theta}|\mathcal{D}))$$

Find $\hat{\theta}_{j_{MAP}}, \forall j = 1, 2, \ldots, k-1$:

$$\begin{aligned}
log(P(\boldsymbol{\theta}|\mathcal{D})) &= log\Big(\theta_1^{N_1+a_1-1}\theta_2^{N_2+a_2-1} \cdots \theta_K^{N_k+a_k-1}\Big) \\
&= (N_1 + a_1 - 1)log(\theta_1) + (N_2 + a_2 - 1)log(\theta_2) + \cdots + (N_k + a_k - 1)log(\theta_k) \\
&= (N_1 + a_1 - 1)log(\theta_1) + \cdots + (N_k + a_k - 1)log(1 - \sum_{k=1}^{K-1} \theta_k) \\
&\qquad\qquad\qquad\qquad\qquad\qquad (\text{Substitute } \theta_k = 1 - \textstyle\sum_{k=1}^{K-1} \theta_k)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} log(P(\boldsymbol{\theta}|\mathcal{D})) &= \frac{N_j + a_j - 1}{\theta_j} - \frac{N_k + a_k - 1}{1 - \sum_{k=1}^{K-1} \theta_k} \\
0 &= \frac{N_j + a_j - 1}{\theta_j} - \frac{N_k + a_k - 1}{\theta_k} \\
\frac{N_j + a_j - 1}{\theta_j} &= \frac{N_k + a_k - 1}{\theta_k} \\
\implies \hat{\theta}_{j_{MAP}} &= \hat{\theta}_{k_{MAP}} \frac{N_j + a_j - 1}{N_k + a_k - 1} \qquad \forall j = 1, 2, \ldots, k-1
\end{aligned}$$

We can now solve for $\hat{\theta}_{k_{MAP}}$

$$\begin{aligned}
\hat{\theta}_{k_{MAP}} &= 1 - \sum_{j=1}^{K-1} \hat{\theta}_{j_{MAP}} \\
&= 1 - \hat{\theta}_{k_{MAP}} \sum_{j=1}^{K-1} \frac{N_j + a_j - 1}{N_k + a_k - 1} \\
&= \frac{1}{1 + \sum_{j=1}^{K-1} \frac{N_j + a_j - 1}{N_k + a_k - 1}} \\
&= \frac{N_k + a_k - 1}{\sum_{j=1}^{K} (N_j + a_j - 1)}
\end{aligned}$$

Hence $\hat{\theta}_{k_{MAP}} = \frac{N_k + a_k - 1}{\sum_{j=1}^{K}(N_j + a_j - 1)}, \forall k = 1, 2, \ldots, K$.

c) To find the probability of $\boldsymbol{x}^{N+1}$ being some class smaller than K. We first find the probability of $\boldsymbol{x}^{N+1}$ being class k, i.e., $P(x_k^{(N+1)}|\mathcal{D}) = \int_\theta P(x_k^{(N+1)}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$.

Since $\boldsymbol{x}^{(N+1)}$ is a 1-of-K encoding vector, thus we have $P(x_k^{(N+1)}|\boldsymbol{\theta}) = \theta_k$. Hence, the probability becomes $P(x_k^{(N+1)}|\mathcal{D}) = \int_\theta \theta_k P(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$.
Computation for the probability of $\boldsymbol{x}^{N+1}$ being some class smaller than K:

$$\begin{aligned}
P(x^{(N+1)} < K) &= \sum_{k=1}^{K-1} P(x_k^{(N+1)} = 1) \\
&= \sum_{k=1}^{K-1} P(x_k^{(N+1)}|\mathcal{D}) \\
&= \sum_{k=1}^{K-1} \int_\theta \theta_k P(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} \\
&= \sum_{k=1}^{K-1} \mathbb{E}[\theta_k] \qquad \text{(since } \boldsymbol{\theta} \sim \text{Dirichlet}(N_1 + a_1, \ldots, N_K + a_K)) \\
&= \sum_{k=1}^{K-1} \frac{N_k + a_k}{\sum_{k'}(N_{k'} + a_{k'})} \qquad (\boldsymbol{\theta} \sim \text{Dirichlet}(N_1 + a_1, \ldots, N_K + a_K))
\end{aligned}$$

Problem 4

a)
```
The average conditional log-likelihood on training set is -0.12462443666863014
The average conditional log-likelihood on test set is -0.19667320325525525
```

b)
```
The accuracy on training set is 0.9814285714285714
The accuracy on test set is 0.97275
```

c) The performance is worse compared with full-covariance matrix (lower likelihood and accuracy). Diagonal covariance matrix cannot model dependence between pixels