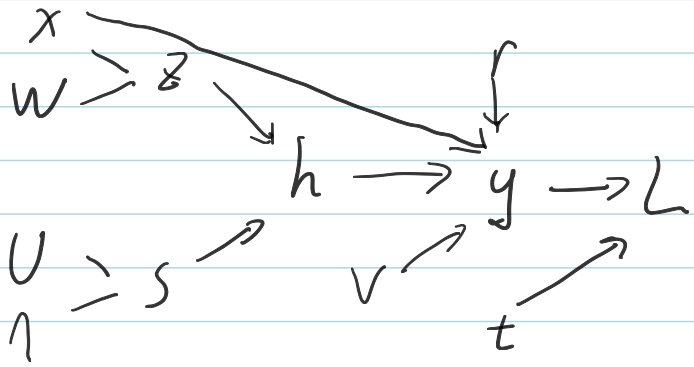
Question 1:

a)

computation graph:



b)

backprop: get the element wise forward prop.

$z_i = \sum_j w_{ij} \cdot x_j$

$s_i = \sum_j U_{ij} \cdot \eta_j$

$h_i = z_i \cdot \sigma(s_i)$

$y = \sum_i v_i \cdot h_i + \sum_i r_i \cdot x_i$

$L = \frac{1}{2}(y-t)^2$

get element wise back prop:

$\bar{L} = 1$

$\bar{y} = \bar{L} \cdot (y-t)$

$\bar{v_i} = \bar{y} h_i$

$\bar{r_i} = \bar{y} \cdot x_i$

$\bar{h_i} = \bar{y} \cdot v_i$

$\bar{s_i} = \bar{h_i} \cdot z_i \cdot \sigma'(s_i)$

$\bar{z_i} = \bar{h_i} \cdot \sigma(s_i)$

$\bar{w_{ij}} = \bar{z_i} \cdot x_j$

$\bar{U_{ij}} = \bar{s_i} \cdot \eta_j$

$\bar{x_i} = \bar{y} \cdot r_i + \sum_j \bar{z_j} w_{ji}$

$\bar{\eta_i} = \sum_j \bar{s_j} U_{ji}$

In vectorized form.

$\bar{v} = \bar{y} h$ $\quad \bar{h} = \bar{y} V$

$\bar{r} = \bar{y} x$ $\quad \bar{\delta} = \bar{h} \circ z \circ G'(s)$

$\bar{z} = \bar{h} \circ G(s)$

$\bar{w} = \bar{z} x^T$ $\quad \bar{U} = \bar{\delta} \eta^T$

$\bar{x} = \bar{y} r + w^T \bar{z}$

$\bar{\eta} = U^T \bar{\delta}$

we get $\bar{v}, \bar{r}, \bar{w}, \bar{U}, \bar{x}, \bar{\eta}$ as desired

Question 2:

Using Naive Bayes.

a) Given independence.

Know $p(x, c|\theta, \pi) = P(c|\pi) \cdot \prod_{j=1}^{784} P(x_j|c, \theta_{jc})$

from lecture, we know: loglikelihood:

$$l(\theta, \pi) = \sum_{i=1}^{N} \log(p(c^{(i)}|\pi) + \sum_{i=1}^{N} \sum_{j=1}^{784} P(x_j^{(i)}|c, \theta_{jc})$$

first maximize the Bernoulin log-likelihood of labels term

Note $P(c^{(i)}) = \prod_{j=0}^{9} \pi_{c=j}^{t_j^{(i)}}$ $(t_j^{(i)})$ here represent the the value of j^{th} index of $t^{(i)}$

$$\frac{d}{d\pi_c} l(\theta, \pi) = \frac{d}{d\pi_c} \sum_{i=1}^{N} \log(p(c^{(i)}|\pi))$$

$$= \frac{d}{d\pi_c} \sum_{i=1}^{N} \log(\prod_{j=0}^{9} \pi_{c=j}^{t_j^{(i)}})$$

$$= \frac{d}{d\pi_c} \sum_{i=1}^{N} \sum_{j=0}^{9} t_j^{(i)} \log(\pi_{c=j})$$

replace $\pi_9 = 1 - \sum_{j=0}^{8} \pi_j$

$$= \frac{d}{d\pi_c} \sum_{i=1}^{N} t_9^{(i)} \log(1 - \sum_{j=0}^{8} \pi_j) + \frac{d}{d\pi_c} \sum_{i=1}^{N} \sum_{j=0}^{8} t_j^{(i)} \log(\pi_j)$$

$$= \sum_{i=1}^{N} \frac{t_c^{(i)}}{\pi_c} - \sum_{i=1}^{N} \frac{t_9^{(i)}}{1 - \sum_{j=0}^{8} \pi_j}$$

$$= \sum_{i=1}^{N} \frac{t_c^{(i)}}{\pi_c} - \sum_{i=1}^{N} \frac{t_9^{(i)}}{\pi_9}$$

Set derivative to 0:

$$\sum_{i=1}^{N} \frac{t_c^{(i)}}{\pi_c} - \sum_{i=1}^{N} \frac{t_9^{(i)}}{\pi_9} = 0$$

$$\sum_{i=1}^{N} \frac{t_c^{(i)}}{\pi_c} = \sum_{i=1}^{N} \frac{t_q^{(i)}}{\pi_q}$$

$$\frac{\sum_{i=1}^{N} t_c^{(i)}}{\pi_c} = \frac{\sum_{i=1}^{N} t_q^{(i)}}{\pi_q}$$

$$\hat{\pi}_c = \hat{\pi}_q \cdot \frac{\#\text{ of sample of class } c}{\#\text{ of sample of class } q}$$

since $\sum_{i=0}^{q} \hat{\pi}_i = 1$

$$\hat{\pi}_q + \sum_{j=0}^{8} \hat{\pi}_q \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_q^{(i)}} = 1$$

$$\hat{\pi}_q + \hat{\pi}_q \frac{\sum_{j=0}^{8} \sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_q^{(i)}} = 1$$

$$\hat{\pi}_q = (1 - \hat{\pi}_q) \cdot \frac{\sum_{i=1}^{N} t_q^{(i)}}{\sum_{j=0}^{8} \sum_{i=1}^{N} t_j^{(i)}}$$

$$\hat{\pi}_q = \frac{\sum_{i=1}^{N} t_q^{(i)}}{\sum_{j=0}^{8} \sum_{i=1}^{N} t_j^{(i)}} - \frac{\sum_{i=1}^{N} t_q^{(i)}}{\sum_{j=0}^{8} \sum_{i=1}^{N} t_j^{(i)}} \hat{\pi}_q$$

$$\hat{\pi}_q \left(1 + \frac{\sum_{i=1}^{N} t_q^{(i)}}{\sum_{j=0}^{8} \sum_{i=1}^{N} t_j^{(i)}}\right) = \frac{\sum_{i=1}^{N} t_q^{(i)}}{\sum_{j=0}^{8} \sum_{i=1}^{N} t_j^{(i)}}$$

$$\hat{\pi}_q = \frac{\sum_{i=1}^{N} t_q^{(i)}}{\sum_{i=1}^{N} \left(\sum_{j=0}^{8} t_j^{(i)} + t_q^{(i)}\right)}$$

$$= \frac{\sum_{i=1}^{N} t_q^{(i)}}{N}$$ (only one of index of t is 1

$$\hat{\pi}_q = \frac{\#\text{ of smaple of class } q}{\#\text{ of samples}}$$

Hence for c in 1~8

$$\hat{\pi}_c = \hat{\pi}_q \frac{\#\text{ of sample of class } c}{\#\text{ of sample of class } q}$$

$$\hat{\pi}_c = \frac{\#\text{ of sample of class } c}{\#\text{ of samples}}.$$

notice this form works for c in $1 \sim 9$

maximize Bernoulli log-likelihood for feature x_j

notice $p(x_j \mid c, \theta_{jc}) = \theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j}$ since $x_j \in \{0,1\}$

$$\frac{d}{d\theta_{jc}} l(\theta, \pi_c) = \frac{d}{d\theta_{jc}} \sum_{i=1}^{N} \sum_{j'=1}^{784} \log\left(\theta_{j'c}^{x_{j'}^{(i)}} (1-\theta_{j'c})^{(1-x_{j'}^{(i)})}\right)$$

$$= \frac{d}{d\theta_{jc}} \sum_{i=1}^{N} \sum_{j'=1}^{784} x_{j'}^{(i)} \log \theta_{j'c^{(i)}}$$

$$+ \frac{d}{d\theta_{jc}} \sum_{i=1}^{N} \sum_{j'=1}^{784} (1-x_{j'}^{(i)}) \log(1-\theta_{j'c^{(i)}})$$

$$= \sum_{i=1}^{N} \sum_{j'=1}^{784} \frac{d}{d\theta_{jc}} x_{j'}^{(i)} \log \theta_{j'c^{(i)}} + \sum_{i=1}^{N} \sum_{j'=1}^{784} \frac{d}{d\theta_{jc}} (1-x_{j'}^{(i)}) \log(1-\theta_{j'c^{(i)}})$$

$$= \sum_{i=1}^{N} \frac{d}{d\theta_{jc}} x_j^{(i)} \log \theta_{jc^{(i)}} + \sum_{i=1}^{N} \frac{d}{d\theta_{jc}} (1-x_j^{(i)}) \log (1-\theta_{jc^{(i)}})$$

$$= \sum_{i \in \{1 \le i \le N \mid c^{(i)}=c\}} \frac{x_j^{(i)}}{\theta_{jc}} - \sum_{i \in \{1 \le i \le N \mid c^{(i)}=c\}} \frac{(1-x_j^{(i)})}{1-\theta_{jc}}$$

setting it to zero,

$$\sum_{i \in \{1 \le i \le N \mid c^{(i)}=c\}} \frac{x_j^{(i)}}{\hat{\theta}_{jc}} - \sum_{i \in \{1 \le i \le N \mid c^{(i)}=c\}} \frac{(1-x_j^{(i)})}{1-\hat{\theta}_{jc}} = 0$$

$$\frac{\sum_{i \in \{1 \le i \le N \mid c^{(i)}=c\}} x_j^{(i)}}{\hat{\theta}_{jc}} = \frac{\sum_{i \in \{1 \le i \le N \mid c^{(i)}=c\}} (1-x_j^{(i)})}{1-\hat{\theta}_{jc}}$$

$$(1-\hat{\theta}_{jc}) \sum_{i \in \{1 \le i \le n \mid C^{(i)}=c\}} x_j^{(i)}$$

$$= \hat{\theta}_{jc} \sum_{i \in \{1 \le i \le n \mid C^{(i)}=c\}} (1 - x_j^{(i)})$$

$$\hat{\theta}_{jc} \sum_{i \in \{1 \le i \le n \mid C^{(i)}=c\}} 1 = \sum_{i \in \{1 \le i \le n \mid C^{(i)}=c\}} x_j^{(i)}$$

$$\hat{\theta}_{jc} = \frac{\sum_{i \in \{1 \le i \le n \mid C^{(i)}=c\}} x_j^{(i)}}{\sum_{i \in \{1 \le i \le n \mid C^{(i)}=c\}} 1}$$

$$= \frac{\#\text{ samples of class } c \text{ with } x_j = 1}{\#\text{ samples of class } c}$$

Here we derived form of $\hat{\pi}_c$ and $\hat{\theta}_{jc}$

b)

using Bayes' Rule:

$$p(t|x,\theta,\pi) = \frac{p(x|t,\theta)\cdot p(t|\pi)}{\sum_{c=0}^{9} p(c|\pi)\cdot p(x|t_c=1,\theta)}$$

we know $p(t|\pi) = \prod_{c=0}^{9} \pi_c^{t_c}$

$$p(x|t,\theta) = \prod_{j=1}^{784} p(x_j|t,\theta)$$

$$= \prod_{j=1}^{784} \prod_{c=0}^{9} (\theta_{jc}^{x_j}(1-\theta_{jc})^{(1-x_j)})^{t_c}$$

$$p(t|x,\theta,\pi) = \frac{\prod_{j=1}^{784} \prod_{c=0}^{9} (\theta_{jc}^{x_j}(1-\theta_{jc})^{(1-x_j)})^{t_c} \prod_{c=0}^{9} \pi_c^{t_c}}{\sum_{c=0}^{9} \pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j}\cdot(1-\theta_{jc})^{1-x_j}}$$

for log-likelihood:

$$\log(p(t|x,\theta,\pi)) = \log \frac{\prod_{j=1}^{784} \prod_{c=0}^{9} (\theta_{jc}^{x_j}(1-\theta_{jc})^{(1-x_j)})^{t_c} \prod_{c=0}^{9} \pi_c^{t_c}}{\sum_{c=0}^{9} \pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j}\cdot(1-\theta_{jc})^{1-x_j}}$$

$$= \log\left(\prod_{j=1}^{784} \prod_{c=0}^{9} (\theta_{jc}^{x_j}(1-\theta_{jc})^{(1-x_j)})^{t_c} \prod_{c=0}^{9} \pi_c^{t_c}\right) - \log\left(\sum_{c=0}^{9} \pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j}\cdot(1-\theta_{jc})^{1-x_j}\right)$$

Separate it into 2 parts: Simplify each

$$\log\left(\prod_{j=1}^{784} \prod_{c=0}^{9} (\theta_{jc}^{x_j}(1-\theta_{jc})^{(1-x_j)})^{t_c} \prod_{c=0}^{9} \pi_c^{t_c}\right)$$

$$= \sum_{j=1}^{784} \sum_{c=0}^{9} t_c(x_j \log\theta_{jc} + (1-x_j)\log(1-\theta_{jc})) + \sum_{c=0}^{9} t_c \log\pi_c$$

$$\log\left(\sum_{c=0}^{9} \pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j} \cdot (1-\theta_{jc})^{x_j}\right)$$

$$= \log \sum_{c=0}^{9} \pi_c \cdot e^{\log \prod_{j=1}^{784} \theta_{jc}^{x_j} \cdot (1-\theta_{jc})^{1-x_j}}$$

$$= \log \sum_{c=0}^{9} \pi_c \cdot e^{\sum_{j=1}^{784} (x_j \log \theta_{jc} + (1-x_j) \cdot \log(1-\theta_{jc}))}$$

Hence the log-likelihood $\log p(t \mid x, \theta, \pi_c)$

$$= \sum_{j=1}^{784} \sum_{c=0}^{9} t_c (x_j \log \theta_{jc} + (1-x_j) \log(1-\theta_{jc})) + \sum_{c=0} t_c \log \pi_c$$

$$- \log \sum_{c=0}^{9} \pi_c \cdot e^{\sum_{j=1}^{784} (x_j \log \theta_{jc} + (1-x_j) \cdot \log(1-\theta_{jc}))}$$

c) in order to calculate the average log likelihood with MLE need to calculate the individual log likelihood.

from part b, we know need to calculate $\log \theta_{jc}$

from part a, we know $\hat{\theta}_{jc} = \frac{\#\text{ samples of class } c \text{ with } X_j = 1}{\#\text{ samples of class } c}$

we can see that around corner. for each image. $X_j = 0$ as white. Hence $\#\text{ samples of class } c \text{ with } X_j = 1$ is 0 in some cases

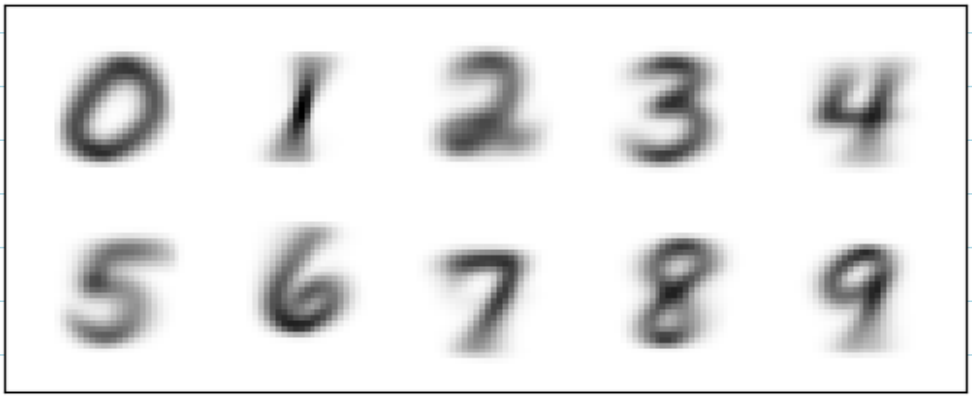
Hence $\hat{\theta}_{jc} = 0$ and $\log(0)$ is undefined.

since we involve undefined term when calculate average.

the average is also undefined.

d)

MLE-estimator $\hat{\theta}$ for each class



e) from lecture, we know,

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \log p(\theta) + \log p(D|\theta)$$

let $f = \log p(\theta) + \log p(D|\theta)$

Since $P(\theta)$ follows Beta(3, 3),

$$P(\theta) = \text{constant} \cdot \prod_{j=1}^{784} \prod_{c=0}^{9} \theta_{jc}^{2} (1-\theta_{jc})^{2}$$

$$p(D|\theta) = \prod_{i=1}^{N} \prod_{j=1}^{784} P(x_j^{(i)} | t^{(i)}, \theta)$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{784} \prod_{c=0}^{9} \theta_{jc}^{t_c^{(i)} x_j^{(i)}} (1-\theta_{jc})^{t_c^{(i)}(1-x_j^{(i)})}$$

$$\hat{\theta}_{map} = \arg\max_{\theta} \sum_{j=1}^{784} \sum_{c=0}^{9} (2\log\theta_{jc} + 2\log(1-\theta_{jc})) + \sum_{i=1}^{N} \sum_{j=1}^{784} \sum_{c=0}^{9} [t_c^{(i)} x_j^{(i)} \cdot \log(\theta_{jc}) + t_c^{(i)}(1-x_j^{(i)}) \log(1-\theta_{jc})]$$

$$= \arg\max_{\theta} \left(\sum_{j=1}^{784} \sum_{c=0}^{9} \left(\left(2 + \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}\right) \log(\theta_{jc}) + \left(2 + \sum_{i=1}^{N} t_c^{(i)}(1-x_j^{(i)})\right) \log(1-\theta_{jc}) \right) \right)$$

differentiate by θ_{jc} and set it to 0

$$\frac{df}{d\theta_{jc}} = \frac{d}{d\theta_{jc}} \sum_{j=1}^{784} \sum_{c=0}^{9} \left(\left(2 + \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}\right) \log(\theta_{jc}) + \left(2 + \sum_{i=1}^{N} t_c^{(i)}(1-x_j^{(i)})\right) \log(1-\theta_{jc}) \right)$$

$$= 0$$

$$= \frac{2 + \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{\hat{\theta}_{jc}} - \frac{2 + \sum_{i=1}^{N} t_c^{(i)} (1 - x_j^{(i)})}{1 - \hat{\theta}_{jc}}$$

we get $\hat{\theta}_{jc} = \frac{2 + \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{4 + \sum_{i=1}^{N} t_c^{(i)}}$

$$\hat{\theta}_{jc} = \frac{2 + \#\text{ of samples that } x_j = 1 \text{ and class } C}{4 + \#\text{ of samples and class } C}$$

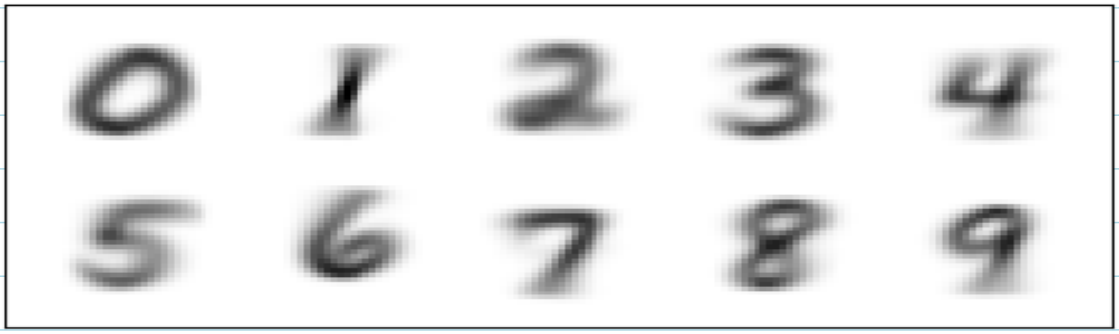
Hence $\hat{\theta}_{jc} = \frac{2 + \#\text{ of samples that } x_j = 1 \text{ and class } C}{4 + \#\text{ of samples and class } C}$

for each ij entry for $\hat{\theta}$ map

f)

```
Average log-likelihood for MAP is  -3.357063137860287
Training accuracy for MAP is  0.8352166666666667
Test accuracy for MAP is  0.816
```

g) MAP estimator ~~θ~~



Question 3:

a) based on given,

$$p(D|\theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{x_k^{(i)}}, \qquad p(\theta) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

since $p(\theta|D) \propto p(\theta) \cdot p(D|\theta)$

$$p(\theta|D) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{x_k^{(i)}}$$

$$\propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \prod_{k=1}^{K} \theta_k^{\sum_{i=1}^{N} x_k^{(i)}}$$

$$\propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \prod_{k=1}^{K} \theta_k^{N_k}$$

$$\propto \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}$$

Hence $p(\theta|D) \propto \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}$,

which is a conjugate prior for the categorical distribution

b)

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D)$$

$$= \arg\max_{\theta}$$

$$= \arg\max_{\theta} \log\left(\prod_{k=1}^{K} \theta_{k'}^{N_{k'}+a_{k'}-1}\right)$$

$$= \arg\max \sum_{k'=1}^{K} (N_{k'}+a_{k'}-1) \log \theta_{k'}$$

let $f = \sum_{k'=1}^{K} (N_{k'}+a_{k'}-1) \log \theta_{k'}$,

differentiate on θ_j $(j \neq K)$

$$\frac{d}{d\theta_j} f = \frac{d}{d\theta_j} \sum_{k'=1}^{K-1} (a_{k'} - 1 + N_{k'}) \log(\theta_{k'}) + \frac{d}{d\theta_j} (a_K - 1 + N_K) \log(\theta_K).$$

$$= \frac{d}{d\theta_j} \sum_{k'=1}^{K-1} (a_{k'} - 1 + N_{k'}) \log(\theta_{k'}) + \frac{d}{d\theta_j} (a_K - 1 + N_K) \log\left(1 - \sum_{k'=1}^{K-1} \theta_{k'}\right)$$

$$\left(\text{by } \textstyle\sum_k \theta_k = 1\right)$$

$$= \frac{a_j - 1 + N_j}{\theta_j} - \frac{a_K - 1 + N_K}{1 - \sum_{k'=1}^{K-1} \theta_{k'}}$$

$$= \frac{a_j - 1 + N_j}{\theta_j} - \frac{a_K - 1 + N_K}{\theta_K}$$

Set derivative to 0

$$\frac{a_j - 1 + lV_j}{\hat{\theta}_j} = \frac{a_k - 1 + lV_k}{\hat{\theta}_k}$$

$$\hat{\theta}_j = \frac{\hat{\theta}_k (a_j - 1 + lV_j)}{a_k - 1 + lV_k}$$

since $\sum_k \theta_k = 1$

$$\hat{\theta}_k + \sum_{j=1}^{k-1} \hat{\theta}_j = 1$$

$$\hat{\theta}_k + \sum_{j=1}^{k-1} \frac{\hat{\theta}_k (a_j - 1 + lV_j)}{a_k - 1 + lV_k} = 1$$

$$\sum_{j=1}^{k} \frac{\hat{\theta}_k (a_j - 1 + lV_j)}{a_k - 1 + lV_k} = 1$$

$$\hat{\theta}_k = \frac{a_k - 1 + lV_k}{\sum_{j=1}^{k} a_j - 1 + lV_j}$$

Hence $\hat{\theta}_j = \frac{a_k - 1 + lV_k}{\sum_{j=1}^{k} a_j - 1 + lV_j} \cdot \frac{a_j - 1 + lV_j}{a_k - 1 + lV_k}$

$$\hat{\theta}_j = \frac{a_j - 1 + lV_j}{\sum_{j=1}^{k} a_j - 1 + lV_j}$$

Hence for all $j \in \{j \in \mathbb{Z} \mid 1 \le j \le k\}$

$$\hat{\theta}_j = \frac{a_j - 1 + lV_j}{\sum_{j=1}^{k} a_j - 1 + lV_j}$$

c) want to know

$P(x_k^{D+1} = 1 | D)$

by definition,

$$P(x_k^{D+1} = 1 | D) = \int P(x_k^{D+1} = 1 | \theta) \, P(\theta | D)$$

$$= \int \theta_k \prod_{j=1}^{k} \theta_j^{a_j - 1 + N_j}$$

In addition, if $\theta \sim \text{Dirichlet}(a_1, \ldots, a_k)$

$E(\theta_k) = \int \theta_k \cdot P(\theta)$ since θ is continuous,

$$= \int \theta_k \prod_{j=1}^{k} \theta_j^{a_j - 1}$$

$$= \frac{a_k}{\sum_{k'} a_{k'}}$$

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

$E(\theta_k) = \int \theta_k \cdot P(\theta)$

$$= \int \theta_k \prod_{j=1}^{k} \theta_j^{a_j + N_j - 1} = P(x_k^{D+1} = 1 | D)$$

(has same form)

since if $\theta \sim \text{Dirichlet}(a_1, \ldots, a_k)$

$$E(\theta_k) = \frac{a_k}{\sum_{k'} a_{k'}}$$

then if $\theta \sim \text{Dirichlet}(\alpha_1 + N_1, \ldots, \alpha_K + N_K)$

$$E(\theta_k) = \frac{\alpha_k + N_k}{\sum_{k'} \alpha_{k'} + N_{k'}} = p(x_k^{n+1} = 1 \mid D)$$

Hence probability of the $N+1$ outcome was k,

$$p(x_k^{n+1} = 1 \mid D) = \frac{\alpha_k + N_k}{\sum_{k'} \alpha_{k'} + N_{k'}}$$

Question 4

a)

```
average log-likelihood for training set is -0.12462443666863016
average log-likelihood for training set is -0.19667320325525606
```

b)

```
training set accuracy is 0.9814285714285714
test set accuracy is 0.97275
```

c)