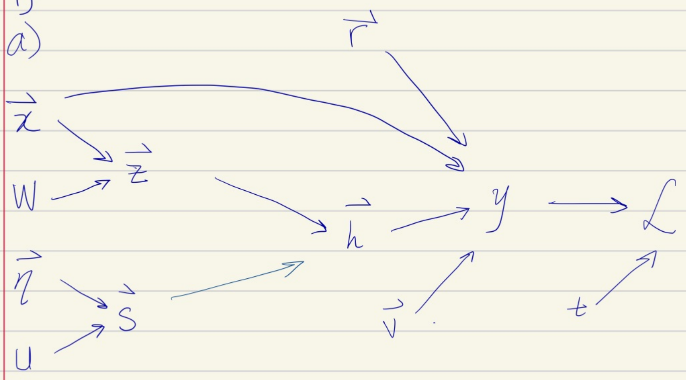
1)

a)



b) $\bar{\mathcal{L}} = 1$

$\bar{y} = \bar{\mathcal{L}} \frac{d\mathcal{L}}{dy} = \bar{\mathcal{L}}(y-t)$

$\bar{\vec v} = \bar y \frac{dy}{d\vec v} = \bar y \vec h$

$\bar{\vec h} = \bar y \frac{dy}{d\vec h} = \bar y \vec v$

$\bar{\vec r} = \bar y \frac{dy}{d\vec r} = \bar y \vec x$

$\bar{\vec z} = \bar{\vec h} \frac{d\vec h}{d\vec z} = \bar{\vec h} \odot \sigma(\vec s)$

$\bar{\vec s} = \bar{\vec h} \frac{d\vec h}{d\vec s} = \frac{d\mathcal{L}}{dy} \cdot \frac{dy}{d\vec s} = \bar y \left(\vec v \odot \vec z \odot \sigma'(\vec s)\right) = \bar{\vec h} \odot \vec z \odot \sigma'(\vec s)$

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

$\bar W = \bar{\vec z} \vec x^T$

$\bar{\vec \eta} = U^T \bar{\vec s}$

$\bar U = \bar{\vec s} \eta^T$

2)

a)

$$L(\vec{\theta},\vec{\pi}) = \prod_{i=1}^{N} p(\vec{x}^{(i)}, c^{(i)} \mid \vec{\theta},\vec{\pi}) = \prod_{i=1}^{N} p(c^{(i)} \mid \vec{\theta},\vec{\pi}) \cdot p(\vec{x}^{(i)} \mid c^{(i)},\vec{\theta},\vec{\pi}) = \prod_{i=1}^{N}\left(p(c^{(i)}\mid\vec{\pi}) \prod_{j=1}^{784} p(x_j^{(i)} \mid c^{(i)}, \theta_{jc^{(i)}})\right)$$

$$\ell(\vec{\theta},\vec{\pi}) = \sum_{i=1}^{N} \log\left(p(c^{(i)}\mid\vec{\pi}) \prod_{j=1}^{784} p(x_j^{(i)} \mid c^{(i)}, \theta_{jc^{(i)}})\right)$$

$$= \sum_{i=1}^{N} \log\left[p(c^{(i)}\mid\vec{\pi}) + \sum_{j=1}^{784} \log p(x_j^{(i)} \mid c^{(i)}, \theta_{jc^{(i)}})\right]$$

$$= \boxed{\sum_{i=1}^{N} \log p(c^{(i)}\mid\vec{\pi})} + \sum_{j=1}^{784} \boxed{\sum_{i=1}^{N} \log p(x_j^{(i)} \mid c^{(i)}, \theta_{jc^{(i)}})}$$

To optimize for $\vec{\pi}$, we maximize $\sum_{i=1}^{N} \log p(c^{(i)}\mid\vec{\pi})$

$$\ell = \sum_{i=1}^{N} \log p(c^{(i)}\mid\vec{\pi})$$

$$= \sum_{i=1}^{N} \log p(\vec{t}_{c^{(i)}} = 1 \mid \vec{\pi})$$

$$= \sum_{i=1}^{N} \log \pi_{c^{(i)}}$$

$$= \sum_{i=1}^{N} \log\left(\prod_{j=0}^{9} \pi_j^{t_j^{(i)}}\right)$$

$$= \sum_{i=1}^{N}\sum_{j=0}^{9} t_j^{(i)} \log \pi_j = \sum_{i=1}^{N}\left(\sum_{j=0}^{8}\left(t_j^{(i)}\log\pi_j\right) + t_9^{(i)}\log\left(1-\sum_{j=0}^{8}\pi_j\right)\right)$$

$$\frac{\partial \ell}{\partial \pi_j} = \sum_{i=1}^{N}\left(\frac{t_j^{(i)}}{\pi_j} - \frac{t_9^{(i)}}{1-\sum_{j=0}^{8}\pi_j}\right) \quad \text{for } j = 0,1,\ldots 8$$

$$= \sum_{i=1}^{N}\left(\frac{t_j^{(i)}}{\pi_j} - \frac{t_9^{(i)}}{\pi_9}\right) = \sum_{i=1}^{N}\frac{t_j^{(i)}}{\pi_j} - \sum_{i=1}^{N}\frac{t_9^{(i)}}{\pi_9}$$

$$= \frac{1}{\pi_j}\sum_{i=1}^{N} t_j^{(i)} - \frac{1}{\pi_9}\sum_{i=1}^{N} t_9^{(i)} \overset{\text{set}}{=} 0 \quad \text{to find MLE.}$$

$$\Rightarrow \frac{1}{\pi_j}\sum_{i=1}^{N} t_j^{(i)} = \frac{1}{\pi_9}\sum_{i=1}^{N} t_9^{(i)}$$

$$\Rightarrow \frac{\pi_j}{\pi_9} = \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}}$$

$$\Rightarrow \pi_j = \left(\frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}}\right)\pi_9 \quad , j = 0,1,\ldots 9$$

also, we have $\sum_{j=0}^{8}\pi_j + \pi_9 = 1$ $\quad \Rightarrow \sum_{j=0}^{8}\left(\frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}}\right)\pi_9 + \pi_9 = 1$

$$\Rightarrow \left(\frac{\pi_9}{\sum_{i=1}^{N} t_9^{(i)}}\right)\sum_{j=0}^{8}\sum_{i=1}^{N} t_j^{(i)} + \pi_9 = 1$$

$$\Rightarrow \pi_9\left(\frac{1}{\sum_{i=1}^{N} t_9^{(i)}}\sum_{j=0}^{8}\sum_{i=1}^{N} t_j^{(i)} + 1\right) = 1$$

$$\Rightarrow \pi_9 = \frac{1}{\left(\frac{1}{\sum_{i=1}^{N} t_9^{(i)}}\right)\sum_{j=0}^{8}\sum_{i=1}^{N} t_j^{(i)} + 1} = \frac{\sum_{i=1}^{N} t_9^{(i)}}{\sum_{j=0}^{8}\sum_{i=1}^{N} t_j^{(i)} + \sum_{i=1}^{N} t_9^{(i)}} = \frac{\sum_{i=1}^{N} t_9^{(i)}}{\sum_{i=1}^{N}\left(\sum_{j=0}^{8} t_j^{(i)} + t_9^{(i)}\right)} = \frac{\sum_{i=1}^{N} t_9^{(i)}}{\sum_{i=1}^{N}\sum_{j=0}^{9} t_j^{(i)}}$$

combined

$$\Rightarrow \pi_j = \left(\frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}}\right)\left(\frac{\sum_{i=1}^{N} t_9^{(i)}}{\sum_{i=1}^{N}\sum_{j=0}^{9} t_j^{(i)}}\right) = \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N}\sum_{j=0}^{9} t_j^{(i)}} \quad \text{for } j = 0,1,\ldots 8$$

$$\boxed{\Rightarrow \pi_{j_{MLE}} = \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N}\sum_{j=0}^{9} t_j^{(i)}} = \frac{\sum_{i=1}^{N} t_j^{(i)}}{N} = \frac{\text{\# images with class} = j \text{ in dataset}}{\text{total \# images in dataset}} \quad \text{for } j = 0,1,\ldots 9}$$

$$\vec{\pi} = [\pi_0 \quad \ldots \quad \pi_9]$$

To optimize for each θ_{jc}, we maximize $\sum_{i=1}^{N} \log p(x_j^{(i)} | c^{(i)}, \theta_{jc^{(i)}})$

$$\ell = \sum_{i=1}^{N} \log p(x_j^{(i)} | c^{(i)}, \theta_{jc^{(i)}}) = \sum_{i=1}^{N} \log\left(\theta_{jc^{(i)}}^{x_j^{(i)}} (1-\theta_{jc^{(i)}})^{(1-x_j^{(i)})}\right)$$

$$= \sum_{i=1}^{N} \left[x_j^{(i)} \log \theta_{jc^{(i)}} + (1-x_j^{(i)}) \log(1-\theta_{jc^{(i)}})\right]$$

$$\frac{d\ell}{d\theta_{jc}} = \sum_{i=1}^{N} \mathbb{1}(c^{(i)} = c) \left[\frac{x_j^{(i)}}{\theta_{jc}} - \frac{(1-x_j^{(i)})}{1-\theta_{jc}}\right]$$

$$= \sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)} = c) \left(\frac{(1-\theta_{jc}) x_j^{(i)} - (1-x_j^{(i)})\theta_{jc}}{\theta_{jc}(1-\theta_{jc})}\right)\right]$$

$$= \sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)} = c) \left(\frac{x_j^{(i)} - x_j^{(i)}\theta_{jc} - \theta_{jc} + x_j^{(i)}\theta_{jc}}{\theta_{jc}(1-\theta_{jc})}\right)\right]$$

$$= \sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)} = c) \frac{(x_j^{(i)} - \theta_{jc})}{\theta_{jc}(1-\theta_{jc})}\right]$$

$$= \frac{1}{\theta_{jc}(1-\theta_{jc})} \sum_{i=1}^{N} \left(\mathbb{1}(c^{(i)} = c)(x_j^{(i)} - \theta_{jc})\right)$$

$$= \frac{1}{\theta_{jc}(1-\theta_{jc})} \sum_{i=1}^{N} \left(\mathbb{1}(c^{(i)} = c) x_j^{(i)} - \mathbb{1}(c^{(i)} = c)\theta_{jc}\right)$$

$$= \frac{1}{\theta_{jc}(1-\theta_{jc})} \left(\sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)} = c) x_j^{(i)}\right] - \sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)} = c)\theta_{jc}\right]\right)$$

$$= \frac{1}{\theta_{jc}(1-\theta_{jc})} \left(\sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)} = c) x_j^{(i)}\right] - \theta_{jc} \sum_{i=1}^{N} \mathbb{1}(c^{(i)} = c)\right)$$

Set $= 0$ to find MLE

$$\Rightarrow \sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)} = c) x_j^{(i)}\right] - \theta_{jc} \sum_{i=1}^{N} \mathbb{1}(c^{(i)} = c) = 0$$

$$\Rightarrow \theta_{jc} \sum_{i=1}^{N} \mathbb{1}(c^{(i)} = c) = \sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)} = c) x_j^{(i)}\right]$$

$$\Rightarrow \boxed{\theta_{jc_{MLE}} = \frac{\sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)} = c) x_j^{(i)}\right]}{\sum_{i=1}^{N} \mathbb{1}(c^{(i)} = c)} = \frac{\sum_{i=1}^{N} \mathbb{1}\left[(c^{(i)} = c) \,\&\, (x_j^{(i)} = 1)\right]}{\sum_{i=1}^{N} \mathbb{1}(c^{(i)} = c)} = \frac{\#\text{ pixel } j \text{ with value } 1 \text{ in class } c}{\#\text{ class } c \text{ in dataset}}}$$

$$\Rightarrow \vec{\theta}_{MLE} = \begin{bmatrix} \frac{\sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)} = 0) x_1^{(i)}\right]}{\sum_{i=1}^{N} \mathbb{1}(c^{(i)} = 0)} & \cdots & \frac{\sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)} = 0) x_{784}^{(i)}\right]}{\sum_{i=1}^{N} \mathbb{1}(c^{(i)} = 0)} \\ \vdots & & \\ \frac{\sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)} = 9) x_1^{(i)}\right]}{\sum_{i=1}^{N} \mathbb{1}(c^{(i)} = 9)} & \cdots & \frac{\sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)} = 9) x_{784}^{(i)}\right]}{\sum_{i=1}^{N} \mathbb{1}(c^{(i)} = 9)} \end{bmatrix}^T$$

b) $L = p(\vec{t}\,|\,\vec{x},\vec{\theta},\vec{\pi}) = p(c\,|\,\vec{x},\vec{\theta},\vec{\pi})$

$$= \frac{p(c|\vec{\pi})\,p(\vec{x}|c,\vec{\theta})}{\sum_{c'=0}^{9} p(c'|\vec{\pi})\,p(\vec{x}|c',\vec{\theta})} = \frac{p(c|\vec{\pi})\prod_{j=1}^{784} p(x_j|c,\theta_{jc})}{\sum_{c'=0}^{9} p(c'|\vec{\pi})\prod_{j=1}^{784} p(x_j|c',\theta_{jc})}$$

$$\ell = \log\left(p(c|\vec{\pi})\prod_{j=1}^{784} p(x_j|c,\theta_{jc})\right) - \log\left(\sum_{c'=0}^{9}\left(p(c'|\vec{\pi})\prod_{j=1}^{784} p(x_j|c',\theta_{jc})\right)\right)$$

$$= \log p(c|\vec{x}) + \sum_{j=1}^{784}\log p(x_j|c,\theta_{jc}) - \log\left(\sum_{c'=0}^{9}\left(p(c'|\vec{\pi})\prod_{j=1}^{784} p(x_j|c',\theta_{jc})\right)\right)$$

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

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

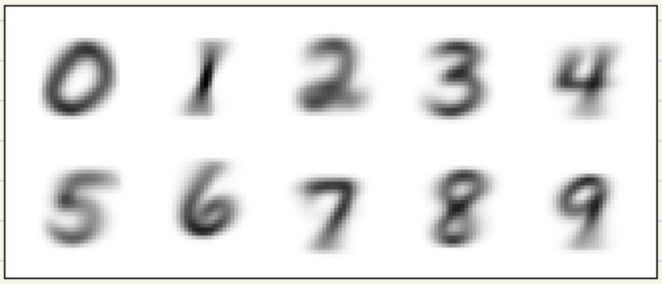
$$\boxed{= \log \pi_c + \sum_{j=1}^{784} x_j\log\theta_{jc} + \sum_{j=1}^{784}(1-x_j)\log(1-\theta_{jc}) - \log\left(\sum_{c'=0}^{9}\exp\left(\log\pi_{c'} + \sum_{j=1}^{784} x_j\log\theta_{jc'} + \sum_{j=1}^{784}(1-x_j)\log(1-\theta_{jc'})\right)\right)}$$

c) $\ell = \log\vec{\pi} + X\log(\vec{\theta}) + (1-X)\log(1-\vec{\theta}) - \log\left(\underset{\text{axis}=0}{\text{sum}}\left(\exp\left(\log\vec{\pi} + X\log(\vec{\theta}) + (1-X)\log(1-\vec{\theta})\right)\right)\right)$

```
/Users/lisayu/Downloads/code_and_data/naive_bayes.py:120: RuntimeWarning: divide by zero encountered in log
  a = np.log(pi) + images @ np.log(theta) + (1 - images) @ np.log(1 - theta)
/Users/lisayu/Downloads/code_and_data/naive_bayes.py:120: RuntimeWarning: invalid value encountered in matmul
  a = np.log(pi) + images @ np.log(theta) + (1 - images) @ np.log(1 - theta)
Average log-likelihood for MLE is  nan
```

This gives an error since we are taking log (0) for some elements of $\vec{\theta}_{MLE}$. A lot of the elements of $\vec{\theta}_{MLE}$ are 0 because of the sparsity of the data.

d)



$\hat{\theta}_{MLE}$ for each class c.

e) $P(\theta_{jc}) = \frac{\Gamma(3+3)}{\Gamma(3)\Gamma(3)} \theta_{jc}^{2} (1-\theta_{jc})^{2}$

$P(\theta | X, \vec{c}) \propto p(\theta) p(X, \vec{c} | \theta) = p(\theta, X, \vec{c})$

$\propto \prod_{c=0}^{9} \prod_{j=1}^{784} p(\theta_{cj}) \prod_{i=1}^{N} \left(p(c^{(i)} | \vec{\pi}) \prod_{j=1}^{784} p(x_j^{(i)} | c^{(i)}, \theta_{jc^{(i)}}) \right)$

* note X is $N \times 784$ matrix
$\vec{c}$ is vector of target classes
* we also assum $\vec{\pi}$ is fixed.

$$\underset{\theta}{\operatorname{argmax}}\, p(\theta | X, \vec{c}) = \underset{\theta}{\operatorname{argmax}} \left[\sum_{c=0}^{9} \sum_{j=1}^{784} \log P(\theta_{cj}) + \sum_{i=1}^{N} \log p(c^{(i)} | \vec{\pi}) + \sum_{j=1}^{784} \sum_{i=1}^{N} \log p(x_j^{(i)} | c^{(i)}, \theta_{jc^{(i)}}) \right]$$

$$= \underset{\theta}{\operatorname{argmax}} \left[\sum_{c=0}^{9} \sum_{j=1}^{784} \left(2 \log \theta_{jc} + 2 \log (1-\theta_{jc}) \right) + \sum_{i=1}^{N} \log p(c^{(i)} | \vec{\pi}) + \sum_{j=1}^{784} \sum_{i=1}^{N} \left[x_j^{(i)} \log \theta_{jc^{(i)}} + (1-x_j^{(i)}) \log (1-\theta_{jc^{(i)}}) \right] \right]$$

we maximize each θ_{jc}:

$$\frac{d}{d\theta_{jc}} \log p(\theta, X, \vec{c}) = \frac{2}{\theta_{jc}} - \frac{2}{1-\theta_{jc}} + \frac{1}{\theta_{jc}(1-\theta_{jc})} \left(\sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)}=c)\, x_j^{(i)} \right] - \theta_{jc} \sum_{i=1}^{N} \mathbb{1}(c^{(i)}=c) \right)$$

$$= \frac{2(1-\theta_{jc}) - 2(\theta_{jc}) + \sum_{i=1}^{N} \left[\mathbb{1}(c^{(i)}=c)\, x_j^{(i)} \right] - \theta_{jc} \sum_{i=1}^{N} \mathbb{1}(c^{(i)}=c)}{\theta_{jc}(1-\theta_{jc})}$$

$$= \frac{2 - 4\theta_{jc} - \theta_{jc} \sum_{i=1}^{N} \mathbb{1}(c^{(i)}=c) + \sum_{i=1}^{N} \mathbb{1}(c^{(i)}=c)\, x_j^{(i)}}{\theta_{jc}(1-\theta_{jc})}$$

$\overset{set}{=} 0 \quad \Rightarrow 2 - 4\theta_{jc} - \theta_{jc} \sum_{i=1}^{N} \mathbb{1}(c^{(i)}=c) + \sum_{i=1}^{N} \mathbb{1}(c^{(i)}=c)\, x_j^{(i)} = 0$

$\Rightarrow 4\theta_{jc} + \theta_{jc} \sum_{i=1}^{N} \mathbb{1}(c^{(i)}=c) = \sum_{i=1}^{N} \mathbb{1}(c^{(i)}=c)\, x_j^{(i)} + 2$

$\Rightarrow \theta_{jc} \left(4 + \sum_{i=1}^{N} \mathbb{1}(c^{(i)}=c) \right) = \sum_{i=1}^{N} \mathbb{1}(c^{(i)}=c)\, x_j^{(i)} + 2$

$$\Rightarrow \theta_{jc} = \boxed{\frac{\sum_{i=1}^{N} \mathbb{1}(c^{(i)}=c)\, x_j^{(i)} + 2}{\sum_{i=1}^{N} \mathbb{1}(c^{(i)}=c) + 4} = \theta_{jc\,MAP}}$$

f) to find π_{MAP}: assume unform prior for π_c

$\Rightarrow P(\pi_c) = 1$

$\Rightarrow P(\vec{\pi} | X, \vec{c}) \propto p(\vec{\pi}) p(X, \vec{c} | \vec{\pi}) = p(\vec{\pi}, X, \vec{c})$

$= \prod_{c=0}^{9} p(\pi_c) \prod_{i=1}^{N} \left(p(c^{(i)} | \vec{\pi}) \prod_{j=1}^{784} p(x_j^{(i)} | c^{(i)}, \theta_{jc^{(i)}}) \right)$

$= 1 \cdot \prod_{i=1}^{N} \left(p(c^{(i)} | \vec{\pi}) \prod_{j=1}^{784} p(x_j^{(i)} | c^{(i)}, \theta_{jc^{(i)}}) \right)$

optmizing for $\vec{\pi}$ is the same as in 2 a),

$$\text{so} \quad \boxed{\pi_{j\,MAP} = \pi_{j\,MLE} = \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} \sum_{j=0}^{9} t_j^{(i)}} = \frac{\sum_{i=1}^{N} t_j^{(i)}}{N} \quad \frac{\#\text{ images with class}=j\text{ in dataset}}{\text{total \# images in dataset}} \quad \text{for } j=0,1,\dots 9}$$

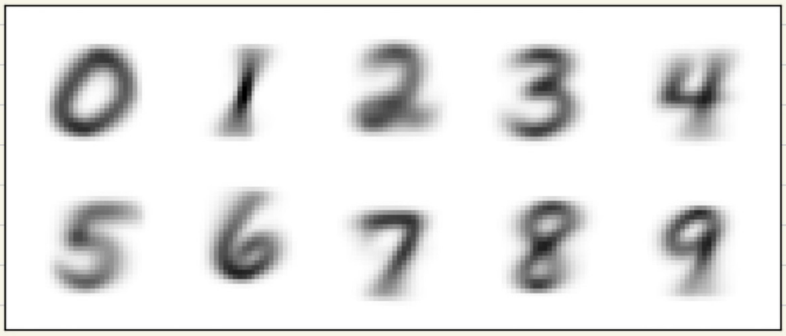
```
Average log-likelihood for MAP is  -3.3570631378602847
Training accuracy for MAP is  0.8352166666666667
Test accuracy for MAP is  0.816
```

The average log-likelihood for MAP is -3.357

The training accuracy for MAP is $0.835 = 83.5\%$

The test accuracy for MAP is $0.816 = 81.6\%$

g)



$\hat{\theta}_{MAP}$ for each class c.

Q3) a) $p(\theta) \propto \theta_1^{a_1-1} \dots \theta_K^{a_K-1}$

$$p(\theta|D) = \frac{p(\theta)\,p(D|\theta)}{\int p(\theta')\,p(D|\theta')\,d\theta}$$

$$\propto \left(\theta_1^{a_1-1} \dots \theta_K^{a_K-1}\right)\left(\prod_{i=1}^{N}\prod_{k=1}^{K} \theta_k^{x_k^{(i)}}\right)$$

$$= \left(\theta_1^{a_1-1} \dots \theta_K^{a_K-1}\right)\left(\prod_{k=1}^{K}\prod_{i=1}^{N} \theta_k^{x_k^{(i)}}\right)$$

$$= \theta_1^{a_1-1} \dots \theta_K^{a_K-1} \prod_{k=1}^{K} \theta_k^{\sum_{i=1}^{N} \mathbb{1}(x_k^{(i)}=1)}$$

$$\boxed{= \prod_{k=1}^{K} \theta_k^{a_k-1+\sum_{i=1}^{N} x_k^{(i)}}}$$

$$= \theta_1^{a_1+\sum_{i=1}^{N} x_1^{(i)}-1} + \theta_2^{a_2+\sum_{i=1}^{N} x_2^{(i)}-1} + \dots + \theta_K^{a_K+\sum_{i=1}^{N} x_K^{(i)}-1}$$

which is a Dirichlet distribution with new parameters $a_k' = a_k + \sum_{i=1}^{N} x_k^{(i)}$

$\boxed{\text{So Yes, the Dirichlet distribution is a conjugate prior for the categorical distribution.}}$

b) $$\log\left(\prod_{k=1}^{K} \theta_k^{a_k-1+\sum_{i=1}^{N} x_k^{(i)}}\right) = \sum_{k=1}^{K}\left[\left(a_k-1+\sum_{i=1}^{N} x_k^{(i)}\right)\log\theta_k\right] = \sum_{k=1}^{K-1}\left[\left(a_k-1+\sum_{i=1}^{N} x_k^{(i)}\right)\log\theta_k\right] + \left(a_K-1+\sum_{i=1}^{N} x_K^{(i)}\right)\log\theta_K$$

$$= \sum_{k=1}^{K-1}\left[\left(a_k-1+\sum_{i=1}^{N} x_k^{(i)}\right)\log\theta_k\right] + \left(a_K-1+\sum_{i=1}^{N} x_K^{(i)}\right)\log\left(1-\sum_{k=1}^{K-1}\theta_k\right) \quad \text{since } 1-\sum_{k=1}^{K-1}\theta_k = \theta_K$$

taking derivative w.r.t. θ_k for $k \in \{1,2,\dots K-1\}$

$$\frac{d}{d\theta_k}\left[\sum_{k=1}^{K-1}\left[\left(a_k-1+\sum_{i=1}^{N} x_k^{(i)}\right)\log\theta_k\right] + \left(a_K-1+\sum_{i=1}^{N} x_K^{(i)}\right)\log\left(1-\sum_{k=1}^{K-1}\theta_k\right)\right]$$

$$= \frac{\left(a_k-1+\sum_{i=1}^{N} x_k^{(i)}\right)}{\theta_k} - \frac{a_K-1+\sum_{i=1}^{N} x_K^{(i)}}{1-\sum_{k=1}^{K-1}\theta_k} \overset{\text{set}}{=} 0$$

$$\Rightarrow \frac{\left(a_k-1+\sum_{i=1}^{N} x_k^{(i)}\right)}{\theta_k} = \frac{a_K-1+\sum_{i=1}^{N} x_K^{(i)}}{1-\sum_{k=1}^{K-1}\theta_k} = \frac{a_K-1+\sum_{i=1}^{N} x_K^{(i)}}{\theta_K}$$

$$\Rightarrow \theta_k = \left(\frac{a_k-1+\sum_{i=1}^{N} x_k^{(i)}}{a_K-1+\sum_{i=1}^{N} x_K^{(i)}}\right)(\theta_K) \quad \text{for } k = \{1,2,\dots .K-1\}$$

also, since $\sum_{k=1}^{K-1}\theta_k + \theta_K = 1$

$$\Rightarrow \sum_{k=1}^{K-1}\left(\frac{a_k-1+\sum_{i=1}^{N} x_k^{(i)}}{a_K-1+\sum_{i=1}^{N} x_K^{(i)}}\right)(\theta_K) + \theta_K = 1$$

$$\Rightarrow \frac{\theta_K}{a_K-1+\sum_{i=1}^{N} x_K^{(i)}} \sum_{k=1}^{K-1}\left(a_k-1+\sum_{i=1}^{N} x_k^{(i)}\right) + \theta_K = 1$$

$$\Rightarrow \theta_K\left(\frac{\sum_{k=1}^{K-1}\left(a_k-1+\sum_{i=1}^{N} x_k^{(i)}\right)}{a_K-1+\sum_{i=1}^{N} x_K^{(i)}} + 1\right) = 1$$

$$\Rightarrow \theta_K = \frac{1}{\dfrac{\sum_{k=1}^{K-1}\left(a_k - 1 + \sum_{i=1}^{N} x_k^{(i)}\right)}{a_K - 1 + \sum_{i=1}^{N} x_K^{(i)}} + 1} = \frac{a_K - 1 + \sum_{i=1}^{N} x_K^{(i)}}{\sum_{k=1}^{K-1}\left(a_k - 1 + \sum_{i=1}^{N} x_k^{(i)}\right) + a_K - 1 + \sum_{i=1}^{N} x_K^{(i)}}$$

$$= \frac{a_K - 1 + \sum_{i=1}^{N} x_K^{(i)}}{\sum_{k=1}^{K}\left(a_k - 1 + \sum_{i=1}^{N} x_k^{(i)}\right)}$$

$\Rightarrow$ for $k = \{1, 2, \ldots, K-1\}$,

$$\theta_k = \left(\frac{a_k - 1 + \sum_{i=1}^{N} x_k^{(i)}}{a_K - 1 + \sum_{i=1}^{N} x_K^{(i)}}\right)\left(\theta_K\right) = \left(\frac{a_k - 1 + \sum_{i=1}^{N} x_k^{(i)}}{a_K - 1 + \sum_{i=1}^{N} x_K^{(i)}}\right)\left(\frac{a_K - 1 + \sum_{i=1}^{N} x_K^{(i)}}{\sum_{k=1}^{K}\left(a_k - 1 + \sum_{i=1}^{N} x_k^{(i)}\right)}\right)$$

$$= \frac{a_k - 1 + \sum_{i=1}^{N} x_k^{(i)}}{\sum_{k=1}^{K}\left(a_k - 1 + \sum_{i=1}^{N} x_k^{(i)}\right)}$$

Putting the above θ_K and $\theta_{k=\{1,\ldots,K-1\}}$ together, we get:

$$\Rightarrow \theta_k = \boxed{\frac{a_k - 1 + \sum_{i=1}^{N} x_k^{(i)}}{\sum_{k=1}^{K}\left(a_k - 1 + \sum_{i=1}^{N} x_k^{(i)}\right)} = \theta_{k\,MAP} \quad \text{for } k = \{1, 2, \ldots, K\}}$$

c) $P(x_k^{(N+1)} = 1 \mid D) = p(\vec{x}^{(N+1)} \mid D)$, where $x_i^{(N+1)} = \begin{cases} 0 & \text{if } i \neq k \\ 1 & \text{if } i = k \end{cases}$

$$= \int p(\vec{x}^{(N+1)} \mid \vec{\theta})\, p(\vec{\theta} \mid D)\, d\vec{\theta}$$

$$= \int \left(\prod_{i=1}^{K} \theta_i^{x_i}\right) p(\vec{\theta} \mid D)\, d\vec{\theta}$$

shown in 3 a)

$$= \int \theta_k\, p(\vec{\theta} \mid D)\, d\vec{\theta} \qquad \vec{\theta} \sim \text{Dirichlet}(a_1', a_2', \ldots a_K'), \text{ where } a_j' = a_j + \sum_{i=1}^{N} x_j^{(i)}$$

$$= \mathbb{E}[\theta_k] = \frac{a_k'}{\sum_{k'} a_{k'}'} = \frac{a_k + \sum_{i=1}^{N} x_k^{(i)}}{\sum_{k'=1}^{K}\left(a_{k'} + \sum_{i=1}^{N} x_{k'}^{(i)}\right)} = \frac{a_k + \sum_{i=1}^{N} x_k^{(i)}}{\sum_{k'=1}^{K} a_{k'} + \sum_{k'=1}^{K}\sum_{i=1}^{N} x_{k'}^{(i)}} = \boxed{\frac{a_k + \sum_{i=1}^{N} x_k^{(i)}}{\sum_{k'=1}^{K} a_{k'} + N}}$$

4) a)

$$\vec{\mu}_k = \frac{\sum_{i=1}^{N} \mathbb{1}(y^{(i)}=k)\, \vec{x}^{(i)}}{\sum_{i=1}^{N} \mathbb{1}(y^{(i)}=k)}$$

$$\Sigma_k = \frac{\sum_{i=1}^{N} \mathbb{1}(y^{(i)}=k)\,(\vec{x}^{(i)}-\vec{\mu}_k)(\vec{x}^{(i)}-\vec{\mu}_k)^T}{\sum_{i=1}^{N} \mathbb{1}(y^{(i)}=k)}$$

conditional_likelihood:

$$\log p(y|\vec{x}) = \log\left(\frac{p(y)\,p(\vec{x}|y)}{\sum_{y'} p(y')\,p(\vec{x}|y')}\right)$$

$$= \left(\log p(y) + \log p(\vec{x}|y) - \log\left(\sum_{y'} p(y')\,p(\vec{x}|y')\right)\right.$$

$$= \log\left(\tfrac{1}{10}\right) + \log p(\vec{x}|y) - \log\left(\sum_{y'}\left(\tfrac{1}{10}\right) p(\vec{x}|y')\right)$$

$$= \cancel{\log\left(\tfrac{1}{10}\right)} + \log p(\vec{x}|y) - \cancel{\log\tfrac{1}{10}} - \log\sum_{y'} p(\vec{x}|y')$$

$$= \log p(\vec{x}|y) - \log\sum_{y'} p(\vec{x}|y')$$

$$= \log p(\vec{x}|y) - \sum_{y'} \log p(\vec{x}|y')$$

```
Avg conditional log-likelihood for train: -0.12462443666863043
Avg conditional log-likelihood for test: -0.19667320325525608
```

b)

```
Training accuracy: 98.14285714285714%
Testing accuracy: 97.275%
```

c) Leading eigenvectors for each class plotted side by side: