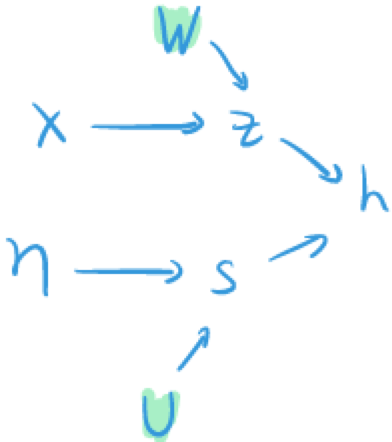
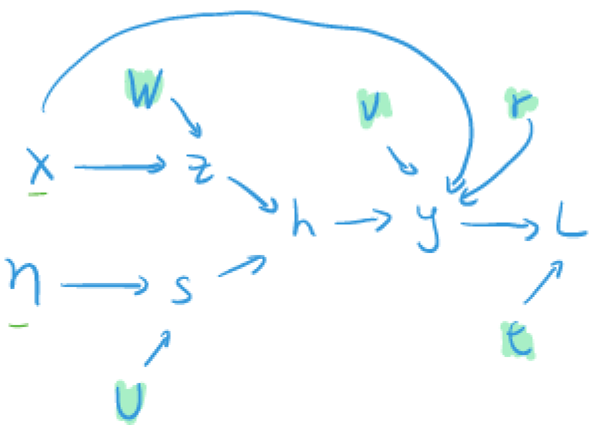
1)

1)

(a)



(b)



$\bar{L} = 1$

$\bar{y} = \bar{L}\frac{dL}{dy}$

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

$= (y-t)$

$\bar{h} = \bar{y}\frac{dy}{dh}$

$= \bar{y}v$

$\bar{x} = \bar{y}\frac{dy}{dx} + \bar{z}\frac{dz}{dx}$

$= \bar{y}r + \bar{z}W$

$\bar{z} = \bar{h}\frac{dh}{dz}$

$= \bar{h}\,1 \odot \sigma(s)$

$= \bar{h}\sigma(s)$

$\bar{s} = \bar{h}\frac{dh}{ds}$

$= \bar{h}\sigma'(s) \odot z$

$$\bar{x} = \bar{z}\frac{dz}{dx} \qquad n = \bar{s}\frac{ds}{dn}$$

$$= \bar{z}W \qquad = \bar{s}U$$

2)

2)

(a)

$$L(\theta, \pi) = p(x, c \mid \theta, \pi)$$

$$= p(c \mid \pi) \prod_{j=1}^{784} p(x_j \mid c, \theta_{jc})$$

$$= p(t \mid \pi) \prod_{j=1}^{784} p(x_j \mid c, \theta_{jc})$$

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

Assuming that there are k samples in the dataset, we have:

$$L(\theta, \pi \mid \{t^{(i)}, x^{(i)}\}_{i=1}^{k}) = \prod_{i=1}^{k} \prod_{c=0}^{9} \pi_c^{t_c^{(i)}} \prod_{j=1}^{784} \theta_{jc}^{x_j^{(i)}} (1-\theta_{jc})^{(1-x_j)^{(i)}}$$

as the likelihood function.

Taking the log-likelihood, we have:

$$\ell(\theta, \pi \mid \{t^{(i)}, x^{(i)}\}_{i=1}^{k} = \log\left[\prod_{i=1}^{k} \prod_{c=0}^{9} \pi_c^{t_c^{(i)}} \prod_{j=1}^{784} \theta_{jc}^{x_j^{(i)}} (1-\theta_{jc})^{(1-x_j)^{(i)}}\right]$$

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

$$= \left[\sum_{i=1}^{k} \sum_{c=0}^{9} t_c^{(i)} \log \pi_c \right] + \left[\sum_{i=1}^{k} \sum_{c=0}^{9} \left(\sum_{j=1}^{784} x_j^{(i)} \log \theta_{jc} + (1-x_j^{(i)}) \log(1-\theta_{jc}) \right) \right]$$

taking the derivative w.r.t. θ_{jc}, we have

$$\rightarrow \frac{\partial}{\partial \theta_{jc}} \sum_{i=1}^{k} \sum_{c=0}^{9} \left(\sum_{j=1}^{784} x_j^{(i)} \log \theta_{jc} + (1-x_j^{(i)}) \log(1-\theta_{jc}) \right)$$

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

$$= \sum_{i=1}^{k} t_c^{(i)} \left(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{(1-x_j^{(i)})}{(1-\theta_{jc})} \right)$$

$$= \sum_{i=1}^{k} t_c^{(i)} \left(\frac{x_j^{(i)}(1-\theta_{jc}) - (1-x_j^{(i)})\theta_{jc}}{\theta_{jc}(1-\theta_{jc})} \right)$$

$$= \sum_{i=1}^{k} t_c^{(i)} \left(x_j^{(i)}(1-\theta_{jc}) - (1-x_j^{(i)})\theta_{jc} \right)$$

$$= \sum_{i=1}^{k} t_c^{(i)} \left(x_j^{(i)} - \cancel{x_j^{(i)}\theta_{jc}} - \theta_{jc} + \cancel{x_j^{(i)}\theta_{jc}} \right)$$

$$= \sum_{i=1}^{k} t_c^{(i)} \left(x_j^{(i)} - \theta_{jc} \right)$$

$$\therefore \sum_{i=1}^{k} t_c^{(i)} x_j^{(i)} - t_c^{(i)} \theta_{jc} = 0$$

$\therefore$ MLE for θ_{jc} is:

$$\hat{\theta}_{jc} = \frac{\sum_{i=1}^{k} t_c^{(i)} x_j^{(i)}}{\sum_{i=1}^{k} t_c^{(i)}}$$

The likelihood function for π is as follows

$\ell(\theta, \pi \mid \{t^{(i)}, x^{(i)}\}_{i=1}^{k}$

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

$$= \sum_{i=1}^{k} \left(\sum_{c=0}^{8} t_c^{(i)} \log \pi_c + t_9^{(i)} \log \left(1 - \sum_{c=0}^{8} \pi_c \right) + \right.$$

$$\left. \left(\sum_{c=0}^{9} \sum_{j=1}^{784} \left(x_j^{(i)} \log \theta_{jc} + (1 - x_j^{(i)}) \log (1 - \theta_{jc}) \right) \right) \right) \text{ \# using hint 2}$$

$$\therefore \frac{\partial \ell}{\partial \pi_c} = \sum_{i=1}^{k} \left(\sum_{c=0}^{8} \frac{t_c^{(i)}}{\pi_c} - \frac{t_9^{(i)}}{1 - \sum_{c=0}^{8} \pi_c} \right)$$

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

$\therefore$ setting $\frac{\partial \ell}{\partial \pi_c} = 0$, we have

$$\sum_{i=1}^{k} \left(\frac{t_c^{(i)}}{\pi_c} - \frac{t_q^{(i)}}{\pi_q} \right) = 0$$

$$\sum_{i=1}^{k} \frac{t_c^{(i)}}{\pi_c} - \sum_{i=1}^{k} \frac{t_q^{(i)}}{\pi_q} = 0$$

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

$$\therefore \quad \frac{\hat{\pi}_c}{\hat{\pi}_q} = \frac{\sum_{i=1}^{k} t_c^{(i)}}{\sum_{i=1}^{k} t_q^{(i)}}$$

Since $\sum_{i=0}^{q} \pi_i = 1$, we have $\frac{\pi_1}{\pi_q} + \frac{\pi_2}{\pi_q} + \dots + \frac{\pi_q}{\pi_q} = \frac{1}{\pi_q}$

$$\therefore \quad \frac{1}{\hat{\pi}_q} = \frac{\sum_{i=1}^{k} t_c^{(i)}}{\sum_{i=1}^{k} t_q^{(i)}}$$

$$\hat{\pi}_q = \frac{\sum_{i=1}^{k} t_q^{(i)}}{\sum_{i=1}^{k} t_c^{(i)}}$$

Thus, if we were to use the same argument for each $\hat{\pi}_1, \hat{\pi}_2, \ldots \hat{\pi}_8$, we will obtain:

$$\hat{\pi}_j = \frac{\sum_{i=1}^{k} t_j^{(i)}}{k}$$

(b)

Using Bayes rule we have:

$$p(c|x) = \frac{p(c)\,p(x|c)}{\sum_{c'} p(c')\,p(x|c')}$$

$$\therefore \quad p(t|x,\theta,\pi) = \frac{p(t|\pi)\,p(x|t,\theta)}{\sum_{t'} p(t'|\pi)\,p(x|t',\theta_{jt'})}$$

$$\log\left(p(t|x,\theta,\pi)\right) = \log\left(\frac{p(t|\pi)\,p(x|t,\theta)}{\sum_{t'} p(t'|\pi)\,p(x|t',\theta_{jt'})}\right)$$

$$= \log\left(p(t|\pi)\prod_{j=1}^{784} p(x_j|t,\theta)\right) - \log\left(\sum_{t'} p(t'|\pi)\,p(x|t',\theta_{jt'})\right)$$

$$= \log p(t|\pi) + \sum_{j=1}^{784} \log p(x_j|t,\theta) - \log \sum_{t'=0}^{9} \pi_{t'} \prod_{j=1}^{784} \theta_{jt'}^{x_j}(1-\theta_{jt'})^{(1-x_j)}$$

$$= \log \pi_t + \sum_{j=1}^{784} \log \theta_{jt}^{x_j}(1-\theta_{jt})^{(1-x_j)} - \log \sum_{t'=0}^{9} \pi_{t'} \prod_{j=1}^{784} \theta_{jt'}^{x_j}(1-\theta_{jt'})^{(1-x_j)}$$

$$= \log \pi_t + \sum_{j=1}^{784} \left(x_j \log \theta_{jt} + (1-x_j) \log (1-\theta_{jt}) \right)$$

$$- \left(\log \sum_{t'=0}^{9} \pi_{t'} + \sum_{j=1}^{784} \left(x_j \log \theta_{jt'} + (1-x_j) \log (1-\theta_{jt'}) \right) \right)$$

$$= \log \pi_t + \sum_{j=1}^{784} \left(x_j \log \theta_{jt} + (1-x_j) \log (1-\theta_{jt}) \right)$$

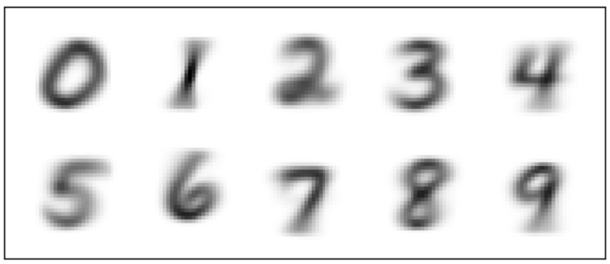
$$- \log \left[\sum_{t'=0}^{9} \exp \left(\log \pi_{t'} + \sum_{j=1}^{784} \left(x_j \log \theta_{jt'} + (1-x_j) \log (1-\theta_{jt'}) \right) \right) \right]$$

(c)

```
Average log-likelihood for MLE is  nan
```

When computing the log-likelihood for MLE, we could encounter a divide by zero error if the count of a particular class is equal to 0 in the data.

(d)



(e) $\hat{\theta}_{MAP} = \underset{\theta}{\arg\max}\, p(\theta | D)$

$$= \underset{\theta}{\arg\max} \log p(\theta) + \log p(D|\theta)$$

$$= \underset{\theta}{\arg\max} \log p(\theta; 3, 3) + \log p(D|\theta; 3,3)$$

$$= \underset{\theta}{\arg\max} \log \theta_{jc}^{3-1}(1-\theta_{jc})^{3-1} + \log \prod_{i=1}^{k} p(c|\pi) \prod_{j=1}^{784} p(x_j | c, \theta_{jc})$$

$$= \underset{\theta}{\arg\max} \log \theta_{jc}^{2}(1-\theta_{jc})^{2} + \log\left[\prod_{i=1}^{k}\prod_{c=0}^{9} \pi_c^{t_c^{(i)}} \prod_{j=1}^{784} \theta_{jc}^{x_j^{(i)}}(1-\theta_{jc})^{(1-x_j)^{(i)}}\right]$$

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

$$= \underset{\theta}{\arg\max}\, 2\log\theta_{jc} + 2\log(1-\theta_{jc}) + \left[\sum_{i=1}^{k}\sum_{c=0}^{9} t_c^{(i)}\log\pi_c\right] + \left[\sum_{i=1}^{k}\sum_{c=0}^{9}\left(\sum_{j=1}^{784} x_j^{(i)}\log\theta_{jc} + (1-x_j^{(i)})\log(1-\theta_{jc})\right)\right]$$

Taking the derivative w.r.t. θ_{jc} we have:

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

$$= \frac{2}{\theta_{jc}} - \frac{2}{1-\theta_{jc}} + \sum_{i=1}^{k} t_c^{(i)}\left(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{(1-x_j^{(i)})}{(1-\theta_{jc})}\right) \quad \text{\# from 1(a)}$$

having the derivative equal to 0 to solve for $\hat{\theta}_{MAP}$:

$$\frac{2}{\theta_{jc}} - \frac{2}{1-\theta_{jc}} + \sum_{i=1}^{k} t_c^{(i)} \left(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{(1-x_j^{(i)})}{(1-\theta_{jc})} \right) = 0$$

$$\frac{2 - 2\theta_{jc} - 2\theta_{jc}}{\theta_{jc}(1-\theta_{jc})} + \sum_{i=1}^{k} t_c^{(i)} \left(\frac{x_j^{(i)}(1-\theta_{jc}) - (1-x_j^{(i)})\theta_{jc}}{\theta_{jc}(1-\theta_{jc})} \right) = 0$$

$$2 - 2\theta_{jc} - 2\theta_{jc} + \sum_{i=1}^{k} t_c^{(i)} x_j^{(i)} - t_c^{(i)} \theta_{jc} = 0 \quad \text{\# from 2(a)}$$

$$2 + \sum_{i=1}^{k} t_c^{(i)} x_j^{(i)} = \sum_{i=1}^{k} t_c^{(i)} \theta_{jc} + 4\theta_{jc}$$

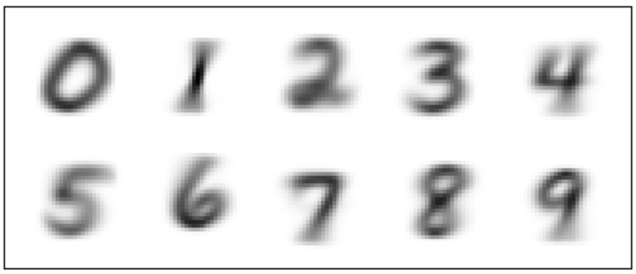
$$2 + \sum_{i=1}^{k} t_c^{(i)} x_j^{(i)} = \theta_{jc} \left(\sum_{i=1}^{k} t_c^{(i)} + 4 \right)$$

$$\hat{\theta}_{MAP} = \frac{2 + \sum_{i=1}^{k} t_c^{(i)} x_j^{(i)}}{4 + \sum_{i=1}^{k} t_c^{(i)}}$$

(f)

```
Average log-likelihood for MAP is  -3.357063137860285
Training accuracy for MAP is  0.8352166666666667
Test accuracy for MAP is  0.816
```

(g)



3)

3)

(a)

To determine the posterior distribution, we have:

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

$$\therefore \; p(\theta|D) \propto p(\theta)p(D|\theta)$$

We know that $p(\theta) \propto \theta_1^{\alpha_1 - 1} \dots \theta_1^{\alpha_K - 1}$

$$p(\theta) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Determining $p(D|\theta)$:

$$p(D|\theta) = \prod_{i=1}^{N} p(x^{(i)}|\theta) \quad \text{\# independently distributed}$$

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

$\therefore \quad p(\theta|D) \propto p(\theta)\, p(D|\theta)$

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

$$= \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \prod_{k=1}^{K} \theta_k^{\sum_{i=1}^{N} x_k^{(i)}} \quad \# \prod_{i=1}^{I} \exp x_i = \exp\left(\sum_{i=1}^{I} x_i\right)$$

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

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

Yes. This is because that the prior $p(\theta)$ & the likelihood $p(D|\theta)$ have the same functional form.

(b)

$$\hat{\theta}_{MAP} = \underset{\theta}{\arg\max}\; p(\theta | D)$$

$$= \underset{\theta}{\arg\max} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1 + N_k}$$

$$= \underset{\theta}{\arg\max} \log\left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1 + N_k} \right)$$

$$= \underset{\theta}{\arg\max} \sum_{k=1}^{K} \log\left(\theta_k^{\alpha_k - 1 + N_k} \right)$$

$$= \underset{\theta}{\arg\max} \sum_{k=1}^{K} (\alpha_k - 1 + N_k) \log(\theta_k)$$

Setting up the lagrangian, we have:

$$L = \sum_{k=1}^{K} (\alpha_k - 1 + N_k) \log(\theta_k) + \lambda\left(1 - \sum_{k} \theta_k\right)$$

$$\frac{dL}{d\theta} = \frac{\alpha_k - 1 + N_k}{\theta_k} - \lambda = 0$$

$$\frac{\alpha_k - 1 + N_k}{\theta_k} = \lambda$$

$$\theta_k = \frac{\alpha_k - 1 + N_k}{\lambda} \longrightarrow (1)$$

$$\frac{dL}{d\lambda} = 1 - \sum_k \theta_k \longrightarrow (2)$$

subbing (1) into (2) we have:

$$1 - \sum_k \theta_k = 1 - \sum_k \frac{\alpha_k - 1 + N_k}{\lambda} = 0$$

$$\Leftrightarrow \sum_k \frac{\alpha_k - 1 + N_k}{\lambda} = 1$$

$$\lambda = \sum_k \alpha_k - 1 + N_k$$

$$\lambda = \sum_k \alpha_k - K + N$$

$$\therefore \hat{\theta}_{kMAP} = \frac{\alpha_k - 1 + N_k}{\sum_k \alpha_k - K + N}$$

(c)

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

$$= \int \theta_k^{x_k^{(N+1)}}\, p(\theta_k \mid D)\, d\theta_k$$

$$= \int \theta_k\, p(\theta_k \mid D)\, d\theta_k$$

since probability of each category is θ_k

$$= E[\theta_k \mid D]$$

$E(x) = \int x f(x)\, dx$

$$= \frac{a_k + N_k}{\sum_{k'} a_{k'} + N_{k'}}$$

4)

(a)

```
Average Conditional Log Likelihood for Training Data: -0.12462443666863034
Average Conditional Log Likelihood for Testing Data: -0.19667320325525578
```

(b)

```
Training Accuracy: 0.9814285714285714
Testing Accuracy: 0.97275
```

(c)

Top Left to Right: 0, 1, 2, 3, 4

Bottom Left to Right: 5, 6, 7, 8, 9