1.

a) Define set $E = \{i : h_t(x^{(i)}) \neq t^{(i)}\}$ and $E^c = \{i : h_t(x^{(i)}) = t^{(i)}\}$

$$err'_t = \frac{\sum_{i \in E} w_i'}{\sum_{i=1}^{N} w_i'}$$

By definition:

$$w_i' = \begin{cases} w_i e^{\alpha} & \text{if } i \in E \\ w_i e^{-\alpha} & \text{if } i \in E^c \end{cases}$$

$$\alpha = \frac{1}{2} \log \frac{1 - err_t}{err_t}$$

$$\therefore err'_t = \frac{\sum_{i \in E} w_i'}{\sum_{i \in E} w_i' + \sum_{i \in E^c} w_i'}$$

$$= \frac{e^{\alpha} \sum_{i \in E} w_i}{e^{\alpha} \sum_{i \in E} w_i + e^{-\alpha} \sum_{j \in E^c} w_i} = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + e^{-2\alpha}}$$

WTS $e^{-\alpha} \sum_{i \in E^c} w_i = e^{\alpha} \sum_{i \in E} w_i$

$$e^{-\alpha} \sum_{i \in E^c} w_i$$

$$= e^{-\frac{1}{2} \log \frac{1 - err_t}{err_t}} \sum_{j \in E^c} w_i$$

$$= e^{\frac{1}{2} \log \frac{err_t}{1 - err_t}} \sum_{i \in E^c} w_i$$

where $$\frac{err_t}{1 - err_t} = \frac{\frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \sum_{i \in E^c} w_i}}{1 - \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \sum_{i \in E^c} w_i}}$$

$$= \frac{\frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \sum_{i \in E^c} w_i}}{\frac{\sum_{i \in E^c} w_i}{\sum_{i \in E} w_i + \sum_{i \in E^c} w_i}} = \frac{\sum_{i \in E} w_i}{\sum_{i \in E^c} w_i} \quad (1)$$

$\therefore \; e^{-\alpha} \sum_{i \in E^c} w_i$

$$= e^{\frac{1}{2} \log \frac{\sum_{i \in E} w_i}{\sum_{i \in E^c} w_i}} \cdot \sum_{i \in E^c} w_i$$

$$= e^{\log \left(\frac{\sum_{i \in E} w_i}{\sum_{i \in E^c} w_i} \right)^{\frac{1}{2}}} \cdot \sum_{i \in E^c} w_i$$

$$= \left(\sum_{i \in E} w_i \right)^{\frac{1}{2}} \cdot \left(\sum_{i \in E^c} w_i \right)^{\frac{1}{2}}$$

$$= \left(\frac{\sum_{i \in E^c} w_i}{\sum_{i \in E} w_i} \right)^{\frac{1}{2}} \cdot \sum_{i \in E} w_i$$

$$= e^{\frac{1}{2} \log \frac{\sum_{i \in E^c} w_i}{\sum_{i \in E} w_i}} \cdot \sum_{i \in E} w_i$$

$$= e^{\frac{1}{2} \log \underbrace{\frac{1 - err_t}{err_t}}_{\text{by (1)}}} \cdot \sum_{i \in E} w_i = e^{\alpha} \sum_{i \in E} w_i$$

$$\therefore \; err_t' = \frac{e^{\alpha} \sum_{i \in E} w_i}{e^{\alpha} \sum_{i \in E} w_i + e^{\alpha} \sum_{i \in E} w_i} = \frac{1}{2}$$

b) $w_i e^{(-\alpha_t t^{(i)} h_t(x^{(i)}))}$

$= w_i e^{\alpha_t - \alpha_t (t^{(i)} h_t(x^{(i)}))} \cdot e^{-\alpha_t}$

$= w_i e^{\alpha_t (1 - t^{(i)} h_t(x^{(i)}))} \cdot e^{-\alpha_t}$

$= w_i e^{2\alpha_t \cdot \frac{1}{2}(1 - t^{(i)} h_t(x^{(i)}))} \cdot e^{-\alpha_t}$

$= w_i e^{2\alpha_t \mathbb{I}\{h_t(x^{(i)}) \neq t^{(i)}\}} \cdot e^{-\alpha_t}$

The constant factor is $e^{-\alpha_t}$

2.

a) The likelihood function w.r.t. θ, π can be expressed as

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

$$l(\theta, \pi) = \underbrace{\log(p(c|\pi))}_{(1)} + \sum_{j=1}^{784} \underbrace{\log(p(x_j | c, \theta_{jc}))}_{(2)}$$

(1): Log-likelihood of labels (2): Log-likelihood of features

We can find the partial derivatives with the 2 terms separately.

$$(1): \log(p(c|\pi)) = \sum_{i=1}^{N} \sum_{j=0}^{9} \log(\pi_j^{t_j^{(i)}}) = \sum_{i=1}^{N} \sum_{j=0}^{9} t_j^{(i)} \log(\pi_j)$$

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

$\forall j \neq 9$

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

$$\therefore \quad \sum_{i=1}^{N} -t_9^{(i)} \pi_j + t_j^{(i)} \pi_9 = 0$$

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

$$\frac{\hat{\pi}_j}{\hat{\pi}_9} = \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}} = \frac{\sum_{i=1}^{N} I(t_j^{(i)} = 1)}{\sum_{i=1}^{N} I(t_9^{(i)} = 1)}$$

$$= \frac{\#\text{ of data with label } j}{\#\text{ of data with label } 9}$$

Since we know by law of probability $\sum_{j=0}^{9} \pi_j = 1$

Therefore $\sum_{j=0}^{9} \hat{\pi}_j = 1$

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

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

$$= \frac{\hat{\pi}_q}{\sum_{i=1}^{N} I(t_q^{(i)} = 1)} \sum_{j=0}^{9} \sum_{i=1}^{N} I(t_j^{(i)} = 1)$$

$$= \frac{\hat{\pi}_q}{\sum_{i=1}^{N} I(t_q^{(i)} = 1)} N \qquad \therefore \hat{\pi}_q = \frac{\sum_{i=1}^{N} I(t_q^{(i)} = 1)}{N}$$

and $\hat{\pi}_j = \frac{\sum_{i=1}^{N} I(t_q^{(i)} = 1)}{N} \cdot \frac{\sum_{i=1}^{N} I(t_j^{(i)} = 1)}{\sum_{i=1}^{N} I(t_q^{(i)} = 1)} = \frac{\sum_{i=1}^{N} I(t_j^{(i)} = 1)}{N} \quad \forall j \in \{0-8\}$

And we can generalize the formula to $\hat{\pi}_q$ so

$\forall j \in \{0-9\}$ $\hat{\pi}_j = \frac{\sum_{i=1}^{N} I(t_j^{(i)} = 1)}{N} = \frac{\#\text{ of data with label } j}{N}$

(2): $\sum_{j=1}^{784} \log(p(x_j \mid c, \theta_{jc}))$

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

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

$$\frac{\partial \ell^{(2)}}{\partial \theta_{jc}} = \sum_{i=1}^{N} t_c^{(i)} \cdot \left[\frac{x_j^{(i)}}{\theta_{jc}} + \frac{-(1 - x_j^{(i)})}{1 - \theta_{jc}} \right] = 0$$

$$= \sum_{i=1}^{N} t_c^{(i)} \cdot \frac{x_j^{(i)}(1 - \theta_{jc}) - (1 - x_j^{(i)})\theta_{jc}}{\theta_{jc}(1 - \theta_{jc})} = 0$$

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

$$\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)} = \theta_{jc} \sum_{i=1}^{N} t_c^{(i)}$$

$$\therefore \hat{\theta}_{jc} = \frac{\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{\sum_{i=1}^{N} t_c^{(i)}} = \frac{\sum_{i=1}^{N} I(x_j^{(i)} = 1 \ \& \ t_c^{(i)} = 1)}{\sum_{i=1}^{N} I(t_c^{(i)} = 1)}$$

$$= \frac{\#\text{ of data with feature } j \text{ appears with label } c}{\#\text{ of data with label } c.}$$

b) $\log(p(t \mid x, \theta, \pi))$

$$= \log\left(\frac{p(t \mid \theta, \pi)\, p(x \mid t, \theta, \pi)}{p(x \mid \theta, \pi)}\right)$$

$$= \log\left(\frac{p(t \mid \pi)\, p(x \mid t_c, \theta_c)}{p(x \mid \theta_c)}\right)$$

$$= \log(p(t \mid \pi)) + \log(p(x \mid t_c, \theta_c)) - \log(p(x \mid \theta_c))$$

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

$$= \log(\pi_c) + \sum_{j=1}^{784}\left[x_j \log\theta_{jc} + (1-x_j)\log(1-\theta_{jc})\right] - \log\left(\sum_{i=0}^{9} \pi_i \cdot \prod_{j=1}^{784} \theta_{ji}^{x_j} (1-\theta_{ji})^{1-x_j}\right)$$

$$= \log(\pi_c) + \sum_{j=1}^{784}\left[x_j \log\theta_{jc} + (1-x_j)\log(1-\theta_{jc})\right] - \log\left(\sum_{i=0}^{9} \pi_i \cdot e^{\log \prod_{j=1}^{784} \theta_{ji}^{x_j}(1-\theta_{ji})^{1-x_j}}\right)$$

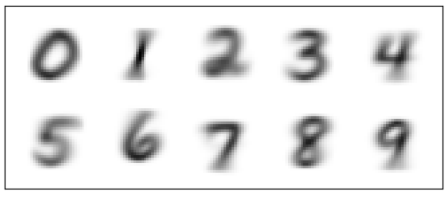
$$= \log(\pi_c) + \sum_{j=1}^{784}\left[x_j \log\theta_{jc} + (1-x_j)\log(1-\theta_{jc})\right] - \log\left(\sum_{i=0}^{9} \pi_i \cdot e^{\sum_{j=1}^{784} x_j \log\theta_{ji} + (1-x_j)\log(1-\theta_{ji})}\right)$$

* this expansion is for vectorization purpose in implementation.

* In reality, we can drop this constant term if we don't need to find avg log-likelihood.

c) The avg-log-likelihood is undefined because many elements in $\hat{\theta}$ are zero and $\log(\hat{\theta}_{jc})$ is undefined when $\hat{\theta}_{jc} = 0$

d)



e) By given prior, $p(D|\theta) = \frac{\theta^{3-1}(1-\theta)^{3-1}}{B(3,3)} = \frac{\theta^2(1-\theta)^2}{B(3,3)} \propto \theta^2(1-\theta)^2$

We also know by Bayes rule that

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

For θ_{jc}

$$= \prod_{i=1}^{N} I(t_c^{(i)}=1)\cdot\theta_{jc}^{x_j^{(i)}}(1-\theta_{jc})^{1-x_j^{(i)}} \cdot \theta_{jc}^2(1-\theta_{jc})^2$$

We can use N_c to denote the number of data with label c and use N_{jc} to denote the number of data with label c and feature j.

$$= \theta_{jc}^{N_{jc}}(1-\theta_{jc})^{N_c-N_{jc}} \cdot \theta_{jc}^2 \cdot (1-\theta_{jc})^2$$

$$P(\theta_{jc}|D) = \theta_{jc}^{N_{jc}+2}(1-\theta)^{N_c-N_{jc}+2}$$

$$\therefore \frac{\partial \log(\theta_{jc}|D)}{\partial \theta_{jc}} = \frac{N_{jc}+2}{\theta_{jc}} + \frac{-(N_c-N_{jc}+2)}{1-\theta_{jc}} = 0$$

$$(N_{jc}+2)(1-\theta_{cj}) - \theta_{cj}(N_c-N_{cj}+2) = 0$$

$$N_{jc} - N_{jc}\theta_{jc} + 2 - 2\theta - N_c\theta + N_{jc}\theta_{jc} - 2\theta = 0$$

$$N + 2 = (N_c+4)\theta$$

$$\hat{\theta} = \frac{N_{jc}+2}{N_c+4}$$

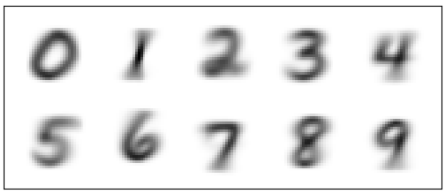
f) Training avg log-likelihood: -3.357

Training accuracy: 0.835

Test avg log-likelihood: -3.449

Test accuracy: 0.816

g)



3.

a) True

b) False

c)