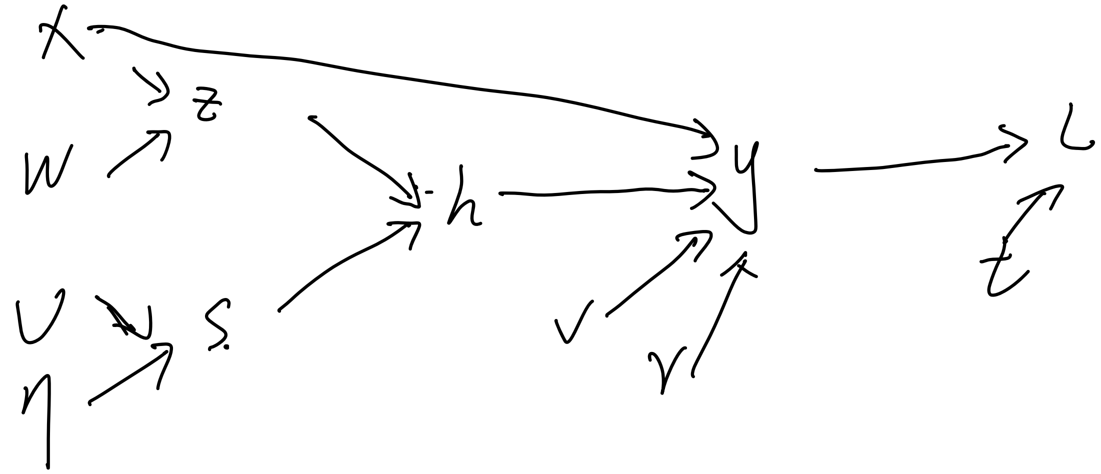
1. Backprop

a)



b) $\sigma(x) = \frac{1}{1+e^{-x}}$

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right)$$

$$= \sigma(x) \cdot (1 - \sigma(x))$$

c) $z = Wx$

$s = U\eta$

$h = z \odot s$

$y = \sigma(v^T h + r^T x)$

$L = t\log y + (1-t)\log(1-y)$

(by piazza, assume base e which is $\ln$)

$\frac{\partial L}{\partial L} = \bar{L} = 1$ $\qquad$ $\bar{y} = \frac{\partial L}{\partial y} = \frac{t}{y} - \frac{1-t}{1-y}$

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

$$\bar{v} = \bar{y} \cdot \frac{\partial y}{\partial v} = \bar{y} \frac{\partial \sigma(v^T h + r^T x)}{\partial v}$$

by part (b)

$$= \frac{y-t}{(y-1)y} \cdot \sigma(v^T h + r^T x)(1 - \sigma(v^T h + r^T x))$$

$$\cdot h^T$$

$$\bar{r} = \bar{y} \cdot \frac{\partial y}{\partial r} = \bar{y} \cdot \frac{\partial \sigma(v^T h + r^T x)}{\partial r}$$

$$= \bar{y} \cdot \sigma(v^T h + r^T x) \cdot (1 - \sigma(v^T h + r^T x)) \cdot x^T$$

$$\bar{h} = \bar{y} \cdot \frac{\partial \sigma(v^T h + r^T x)}{\partial h}$$

$$= \bar{y} \cdot \sigma(v^T h + r^T x)(1 - \sigma(v^T h + r^T x)) \cdot v^T$$

$$\bar{z} = \bar{h} \cdot S \qquad \bar{S} = \bar{h} \cdot z$$

$$\bar{U} = \bar{S} \cdot \eta^T \cdot \underline{I} \qquad \bar{\eta} = \bar{S} \cdot U$$

$$\bar{W} = \bar{z} \cdot x^T \cdot \underline{I}$$

$$\bar{x} = \bar{z} W + \bar{y} \cdot \frac{\partial y}{\partial x}$$

$$= \bar{z} W + \bar{y} \cdot \sigma(v^T h + r^T x) \cdot (1 - \sigma(v^T h + r^T x))$$

$$\cdot r^T$$

2.

a) $$L(\theta, \pi) = \prod_{i=1}^{N} P(x^{(i)}, c^{(i)} \mid \theta, \pi)$$

$$= \prod_{i=1}^{N} \left(P(c^{(i)} \mid \pi) \prod_{j=1}^{784} P(x_j^{(i)} \mid c^{(i)}, \theta_{jc} \right)$$

$$l(\theta, \pi) = \log(L(\theta, \pi))$$

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

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

For θ, $$\frac{\partial l}{\partial \theta_{jc}} = \sum_{i=1}^{N} t_c^{(i)} \left(\frac{x_j^{(i)}}{\theta_{jc}} - \frac{1 - x_j^{(i)}}{1 - \theta_{jc}} \right)$$

set $$\frac{\partial l}{\partial \theta_{jc}} = 0$$

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

N ~ image number

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

$j = 1, 2, \cdots, 784$

$c = 0, 1, \cdots, 9$

Rewrite:

$$l(\theta, \pi) = \sum_{i=1}^{N} \sum_{c=0}^{9} \left(t_c^{(i)} \log \pi_c + const \right)$$

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

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

$$\Rightarrow \sum_{i=1}^{N} \frac{t_j^{(i)}}{\hat{\pi}_j} = \sum_{i=1}^{N} \frac{t_9^{(i)}}{1 - \sum_{j=0}^{8} \hat{\pi}_j}$$

$$\Rightarrow \hat{\pi}_j = \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_q^{(i)}} \hat{\pi}_q$$

$$\hat{\pi}_q = 1 - \sum_{j=0}^{8} \pi_j$$

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

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

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

$$\Rightarrow \hat{\pi}_q \cdot \frac{N}{\sum_{i=1}^{N} t_q^{(i)}} = 1 \Rightarrow \hat{\pi}_q = \frac{\sum_{i=1}^{N} t_q^{(i)}}{N}$$

$$\pi_j^? = \frac{\sum_{i=1}^{N} t_j^{(i)}}{\sum_{i=1}^{N} t_q^{(i)}} \cdot \pi_q^?$$

$$= \frac{\sum_{i=1}^{N} t_j^{(i)}}{N} \quad \text{for } j = 0, 1, \dots, 9$$

b) $$p(t \mid x, \theta, \pi) = \frac{p(t, x, \theta, \pi)}{p(x, \theta, \pi)}$$

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

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

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

$$P(x|\theta,\pi) = \frac{P(x,\theta,\pi)}{P(\theta,\pi)} = \frac{\sum_{i=0}^{9} P(x,\theta,\pi,t_i)}{P(\theta,\pi)}$$

$$= \frac{\sum_{i=0}^{9} P(x|\theta,\pi,t_i)P(\theta,\pi,t_i)}{P(\theta,\pi)}$$

$$= \frac{\sum_{i=0}^{9} P(x|\theta,\pi,t_i)P(t_i|\theta,\pi)\cancel{P(\theta,\pi)}}{\cancel{P(\theta,\pi)}}$$

$$\Rightarrow P(t|x,\theta,\pi) = \frac{P(t|\pi)\prod_{j=1}^{784} P(x_j|t,\theta_{jc})}{\sum_{i=0}^{9} P(t_i|\theta,\pi)P(x|t_i,\theta,\pi)}$$

$$= \frac{\pi_c \cdot \prod_{j=1}^{784} \theta_{jc}^{x_j}(1-\theta_{jc})^{(1-x_j)}}{\sum_{i=0}^{9} P(t_i|\pi)\prod_{j=1}^{784} P(x_j|t,\theta)}$$

$$\Rightarrow \log p(t|x,\theta,\pi) = \log\left(\frac{\pi_c \cdot \prod_{j=1}^{784} \theta_{jc}^{x_j}(1-\theta_{jc})^{(1-x_j)}}{\sum_{i=0}^{9} P(t_i|\pi)\prod_{j=1}^{784} P(x_j|t,\theta)}\right)$$

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

$$\log\left(\sum_{i=0}^{9} p(t_i|\pi) \prod_{j=1}^{784} p(x_j|t,\theta)\right)$$

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

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

c) $\hat{\theta}_{jc} = \dfrac{\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{\sum_{i=1}^{N} t_c^{(i)}}$

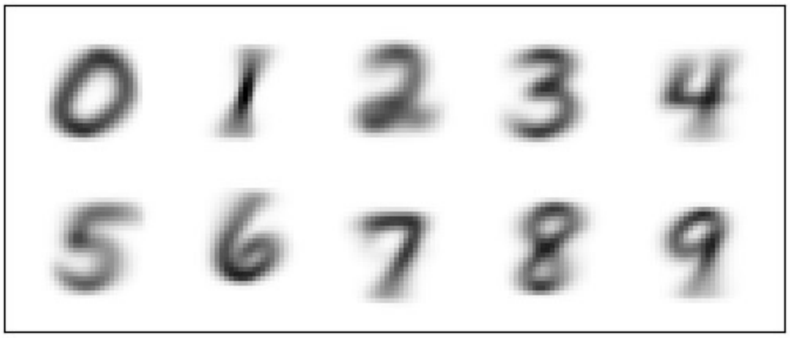
$\hat{\theta}_{jc}$ could be 0

$\Rightarrow \log\theta_{jc}$ would be undefined

which would go wrong, make average undefined.

Avg is nan.

d)



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

$\theta \sim Beta(3,3)$

$\log p(\theta_{jc}, 3, 3) = \log\left(\frac{\Gamma(6)}{\Gamma(3)\Gamma(3)} \theta_{jc}^{2}(1-\theta_{jc})^{2}\right)$

$= const + 2\log\theta_{jc} + 2\log(1-\theta_{jc})$

$\Rightarrow \log P(\theta) = \sum_{j=1}^{784}\sum_{c=0}^{9}(2\log\theta_{jc} + 2\log(1-\theta_{jc}))$.

(Ignore constant part)

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

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

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

$$\Rightarrow \log P(\theta, D) = \sum_{j=1}^{784} \sum_{c=0}^{9} (2\log\theta_{jc} + 2\log(1-\theta_{jc}))$$

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

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

$$\frac{\partial \log(P(\theta, D))}{\partial \theta_{jc}} = \frac{2 + \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}}{\theta_{jc}} - \frac{2 + \sum_{i=1}^{N} t_c^{(i)} (1-x_j^{(i)})}{1-\theta_{jc}} = 0$$

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

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

$$= \hat{\theta}_{jc}\left(4+\sum_{i=1}^{N} t_c^{(i)}\right)$$

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

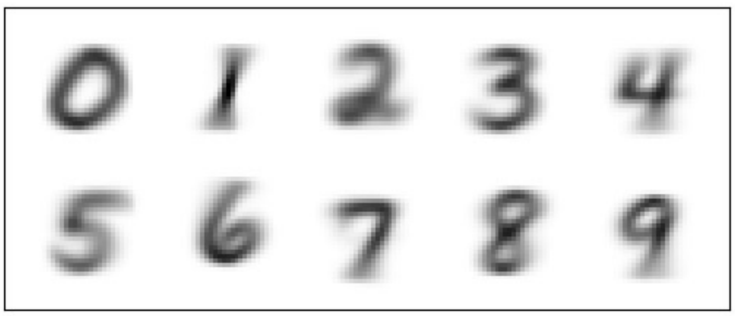
f)

Average log-likelihood for MAP: -3.35706

Training accuracy: 0.8352167

Test accuracy: 0.816

g)



h) it is reasonable that we can assume the features are conditionally independent. It is not reasonable since the probability of feature may be 0.

3.

a) $P(D|\theta) = \prod_{i=1}^{N} P(x^{(i)}|\theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_x^{x_k^{(i)}}$

$P(\theta) \propto \theta_1^{a_1-1} \cdots \theta_K^{a_K-1} = \prod_{i=1}^{K} \theta_i^{a_i-1}$

$P(\theta|D) \propto P(\theta) \cdot P(D|\theta)$

$\propto \left(\prod_{i=1}^{K} \theta_i^{\alpha_i-1}\right)\left(\prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{x_k^{(i)}}\right)$

$\propto \left(\prod_{i=1}^{K} \theta_i^{\alpha_i-1}\right)\left(\prod_{k=1}^{K} \theta_k^{\sum_{i=1}^{N} x_k^{(i)}}\right)$

$\propto \left(\prod_{i=1}^{K} \theta_i^{\alpha_i-1}\right)\left(\prod_{k=1}^{K} \theta_k^{N_k}\right)$

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

which is conjugate prior for categorical distribution

b) $l(\theta) = \log\left(\prod_{i=1}^{k} \theta_k^{\alpha_k - 1 + N_k} \right)$

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

$$= \sum_{i=1}^{k-1} \left[(\alpha_i + N_i - 1) \log \theta_i + (\alpha_k + N_k - 1) \log\left(1 - \sum_{j=1}^{k-1} \theta_j\right) \right]$$

$$\frac{d\,l(\theta)}{d\theta_i} = \frac{\alpha_i + N_i - 1}{\theta_i} - \frac{\alpha_k + N_k - 1}{\theta_K} = 0$$

$$\theta_i (\alpha_k + N_k - 1) = \theta_k (\alpha_i + N_i - 1)$$

$$\hat{\theta}_i = \frac{\hat{\theta}_K (\alpha_i + N_i - 1)}{\alpha_k + N_k - 1}$$

$$\theta_k + \sum_{i=1}^{k-1} \theta_i = 1$$

$$\theta_K + \sum_{i=1}^{K-1} \frac{\hat{\theta}_K (\alpha_i + N_i - 1)}{\alpha_K + N_K - 1} = 1$$

$$\sum_{i=1}^{K} \hat{\theta}_K \frac{\alpha_i + N_i - 1}{\alpha_K + N_K - 1} = 1$$

$$\hat{\theta}_K = \frac{\alpha_K + N_K - 1}{\sum_{i=1}^{K} (\alpha_i + N_i - 1)}$$

$$\hat{\theta}_i = \frac{\alpha_i + N_i - 1}{\alpha_K + N_K - 1} \cdot \frac{\alpha_K + N_K - 1}{\sum_{i=1}^{K} (\alpha_i + N_i - 1)}$$

$$= \frac{\alpha_i + N_i - 1}{\sum_{j=1}^{K} (\alpha_j + N_j - 1)}$$

c) $P(X_K^{N+1} < K) = \sum_{i=1}^{K-1} P(X_K^{N+1} = i)$

$P(X^{N+1} = i) = P(X_K^{N+1} = i \mid D)$

$= \int P(X^{N+1} = i \mid \theta) P(\theta \mid D) \, d\theta$

$= \int \theta_i \, P(\theta \mid D) d\theta$

$= E(\theta_i \mid D)$

$= E(\theta_i)$

since $\theta \sim \text{Dirichlet}(\alpha_i + N_i)$

so $P(X^{N+1} = i) = E(\theta_i) = \dfrac{\alpha_i + N_i}{\sum_{k'}(\alpha_{k'} + N_{k'})}$

$$\Rightarrow P(x^{N+1} < k) = \sum_{i=1}^{k-1} E(\theta_i)$$

$$= \sum_{i=1}^{k-1} \frac{a_i + N_i}{\sum_{k'} (a_{k'} + N_{k'})}$$

4.

a)

```
Avg conditional log-likelihood on training set is: -0.12462443666862932
Avg conditional log-likelihood on testing set is: -0.19667320325525448
```

b)

```
Accuracy on training set is: 0.9814285714285714
Accuracy on training set is: 0.97275
```

c) Avg log-likelihood for training set:

-1.230765422272908

Avg log-likelihood for testing set:

-1.2872603667558389

Accuracy on training set: 0.85

Accuracy on testing set: 0.84

The performance is worse by diagonal matrix compared to full-covariance matrix. Since diagonal matrix cannot model the pixel dependence.