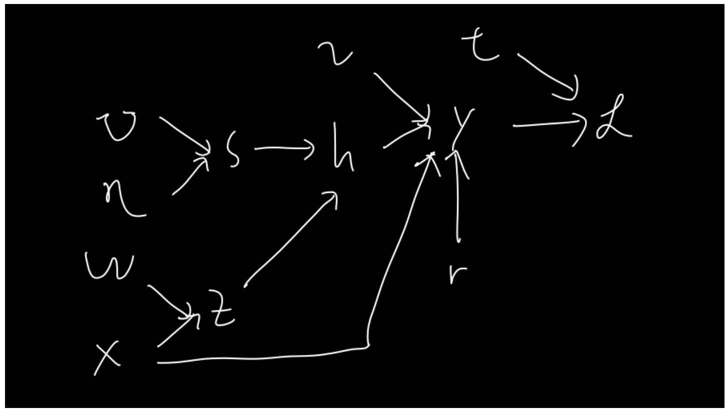
1. (a) here is the computation graph



(b) compute all the intermediate quantities,

$$\overline{\mathcal{L}} = 1,\, \overline{y} = \overline{\mathcal{L}}(y-t) = (y-t)$$

$$\overline{v_i} = \overline{y}\, h_i,\, \overline{h_i} = \overline{y}\, v_i,\, \overline{r_i} = \overline{y}\, x_i,\, \frac{dy}{dx_i} = \overline{y}\, r_i$$

$$\overline{z_i} = \overline{h_i}\, \sigma(s_i),\, \overline{s_i} = \overline{h_i}\, z_i\, \sigma'(s_i)$$

$$\overline{W_{ij}} = \overline{z_i}\, x_j,\, \frac{dz_j}{dx_i} = W_{ji}$$

$$\overline{U_{ij}} = \overline{s_i}\, \eta_j,\, \overline{\eta_i} = \sum_j \overline{s_j}\, U_{ji}$$

$$\overline{x_i} = \overline{y}\, r_i + \sum_j \overline{z_j}\, \frac{dz_j}{dx_i} = \overline{y}\, r_i + \sum_j \overline{z_j}\, W_{ji}$$

after vectorization, we obtain,

$$\overline{\mathcal{L}} = 1,\, \overline{y} = \overline{\mathcal{L}}(y-t) = (y-t)$$

$$\overline{\mathbf{v}} = \overline{y}\, \mathbf{h},\, \overline{\mathbf{h}} = \overline{y}\, \mathbf{v},\, \overline{\mathbf{r}} = \overline{y}\, \mathbf{x},\, \frac{dy}{d\mathbf{x}} = \overline{y}\, \mathbf{r}$$

$$\overline{\mathbf{z}} = \overline{\mathbf{h}} \odot \sigma(\mathbf{s}),\, \overline{\mathbf{s}} = \overline{\mathbf{h}} \odot \mathbf{z} \odot \sigma'(\mathbf{s})$$

$$\overline{\mathbf{W}} = \overline{\mathbf{z}} \cdot \mathbf{x}^T,\, \frac{dz_j}{d\mathbf{x}} = W_j^T$$

$$\overline{\mathbf{U}} = \overline{\mathbf{s}} \cdot \boldsymbol{\eta}^T,\, \overline{\boldsymbol{\eta}} = \mathbf{U}^T \cdot \overline{\mathbf{s}}$$

$$\overline{\mathbf{x}} = \overline{y}\, \mathbf{r} + \sum_j \overline{z_j}\, \frac{dz_j}{d\mathbf{x}} = \overline{y}\, \mathbf{r} + \mathbf{W}^T \cdot \overline{\mathbf{z}}$$

2. (a) the likelihood function of θ, π is

$$\begin{aligned}
\ell(\boldsymbol{\theta}, \boldsymbol{\pi}) &= \sum_{i=1}^{N} \log p(\mathbf{t}^{(i)}, \mathbf{x}^{(i)} \,|\, \boldsymbol{\theta}, \boldsymbol{\pi}) \\
&= \sum_{i=1}^{N} \log(p(\mathbf{t}^{(i)} \,|\, \boldsymbol{\pi}) p(\mathbf{x}^{(i)} \,|\, \mathbf{t}^{(i)}, \boldsymbol{\theta}, \boldsymbol{\pi})) \\
&= \sum_{i=1}^{N} \log p(\mathbf{t}^{(i)} \,|\, \boldsymbol{\pi}) + \sum_{i=1}^{N} \sum_{j=1}^{784} \log p(\mathbf{x}_j^{(i)} \,|\, \mathbf{t}^{(i)}, \boldsymbol{\theta})
\end{aligned}$$

we can maximize these two term seperately, to get $\hat{\pi}_j$

$$\begin{aligned}
\sum_{i=1}^{N} \log p(\mathbf{t}^{(i)} \,|\, \boldsymbol{\pi}) = \sum_{i=1}^{N} \log \prod_{j=0}^{9} \pi_j^{t_j^{(i)}} &= \sum_{i=1}^{N} \sum_{j=0}^{9} t_j^{(i)} \log \pi_j \\
&= \sum_{i=1}^{N} (\sum_{j=0}^{8} t_j^{(i)} \log \pi_j) + t_9^{(i)} \log(1 - \sum_{j=0}^{8} \pi_j)
\end{aligned}$$

differentiate with respect to π_k for $k \in \{0, \dots, 8\}$, we get

$$\begin{aligned}
\frac{1}{\pi_k} \sum_{i=1}^{N} t_k^{(i)} - \frac{1}{\pi_9} \sum_{i=1}^{N} t_9^{(i)} &\overset{\text{set}}{=} 0 \\
\implies \frac{\hat{\pi}_k}{\hat{\pi}_9} &= \frac{\sum_{i=1}^{N} t_k^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}}
\end{aligned}$$

since $\hat{\pi}_i$'s should sum up to 1, as hinted,

$$\begin{aligned}
\hat{\pi}_9 + \sum_{i=0}^{8} \hat{\pi}_9 \frac{\sum_{i=1}^{N} t_k^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}} &= 1 \\
\hat{\pi}_9 \, (1 + \frac{1}{\sum_{i=1}^{N} t_9^{(i)}} \sum_{j=0}^{8} \sum_{i=1}^{N} t_j^{(i)}) &= 1 \\
\hat{\pi}_9 \, (1 + \frac{1}{\sum_{i=1}^{N} t_9^{(i)}} (N - \sum_{i=1}^{N} t_9^{(i)})) &= 1 \\
\implies \hat{\pi}_9 &= \frac{\sum_{i=1}^{N} t_9^{(i)}}{N}
\end{aligned}$$

$\forall j \neq 9. \, \hat{\pi}_j = \hat{\pi}_9 \frac{\sum_{i=1}^{N} t_k^{(i)}}{\sum_{i=1}^{N} t_9^{(i)}} = \frac{1}{N} \sum_{i=1}^{N} t_j^{(i)}$.
Therefore, $\forall j$, the MLE of π_j is

$$\hat{\pi}_j = \frac{1}{N} \sum_{i=1}^{N} t_j^{(i)} = \frac{\text{no. of data with label } i}{N}.$$

Use the other term to maximize $\boldsymbol{\theta}$,

$$\sum_{i=1}^{N}\sum_{j=1}^{784} \log p(\mathbf{x}_j^{(i)} \mid \mathbf{t}^{(i)}, \boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j=1}^{784}\sum_{c=0}^{9} t_c^{(i)} \log p(\mathbf{x}_j^{(i)} \mid \theta_{jc})$$

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

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

differentiate with respect to θ_{mn}, we obtain

$$\sum_{i=1}^{N} t_n^{(i)} x_m^{(i)} \frac{1}{\theta_{mn}} - t_n^{(i)}(1-x_m^{(i)})\frac{1}{1-\theta_{mn}} \stackrel{\text{set}}{=} 0$$

$$\frac{1}{\theta_{mn}}\sum_{i=1}^{N} t_n^{(i)} x_m^{(i)} = \frac{1}{1-\theta_{mn}}\sum_{i=1}^{N} t_n^{(i)}(1-x_m^{(i)})$$

$$(\frac{1}{\theta_{mn}} - 1)\sum_{i=1}^{N} t_n^{(i)} x_m^{(i)} = \sum_{i=1}^{N} t_n^{(i)}(1-x_m^{(i)})$$

$$\frac{1}{\theta_{mn}}\sum_{i=1}^{N} t_n^{(i)} x_m^{(i)} = \sum_{i=1}^{N} t_n^{(i)}$$

$$\Longrightarrow \hat{\theta}_{mn} = \frac{\sum_{i=1}^{N} t_n^{(i)} x_m^{(i)}}{\sum_{i=1}^{N} t_n^{(i)}} = \frac{\text{no. of data with label } n \text{ and feature } m}{\text{no. of data with label } n}$$

(b) By Bayes rule, $p(\mathbf{t} \mid \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\pi}) = \dfrac{p(\mathbf{x}, \mathbf{t} \mid \boldsymbol{\theta}, \boldsymbol{\pi})}{\sum_{c=0}^{9} p(c)p(\mathbf{x} \mid c, \boldsymbol{\theta})} = \dfrac{p(\mathbf{t} \mid \boldsymbol{\pi})p(\mathbf{x} \mid \mathbf{t}, \boldsymbol{\theta})}{\sum_{c=0}^{9} p(c)p(\mathbf{x} \mid c, \boldsymbol{\theta})}$ so the log likelihood is

$$\log p(\mathbf{t} \mid \boldsymbol{\pi}) + \sum_{j=1}^{784}\sum_{c=0}^{9} t_c \log p(x_j \mid \theta_{jc}) - \log(\sum_{c=0}^{9} p(c) \prod_{j=1}^{784} p(x_j \mid c, \theta_{jc}))$$

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

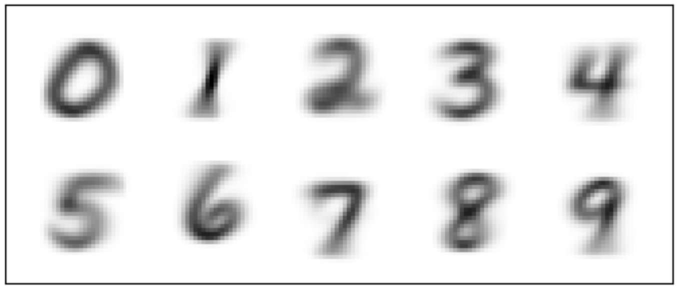
$$= \log\pi_c + \sum_{j=1}^{784}\sum_{c=0}^{9} t_c(x_j \log\theta_{jc} + (1-x_j)\log(1-\theta_{jc}))$$

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

(Note: the last line is just for vectorization in coding)

(c) since $\hat{\theta}_{jc}$ could be numerically zero after fitting, $\log(\hat{\theta}_{jc})$ causes numerical error and the average log-likelihood could not be computed.

(d) here is the plot of the MLE estimator $\hat{\boldsymbol{\theta}}$ as 10 separate greyscale images



(e) According to MAP estimator,

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

where for each θ_{jc}, $\theta_{jc} \sim \text{Beta}(3, 3)$

$$\arg\max_{\boldsymbol{\theta}} \log \prod_{j,c} p(\theta_{jc}) + \log \prod_{i=1}^{N} \prod_{j=1}^{784} p(x_j^{(i)} \,|\, t^{(i)}, \boldsymbol{\theta})$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{j,c} \log \frac{\gamma(3+3)}{\gamma(3)\gamma(3)} + (3-1)\log\theta_{jc} + (3-1)\log(1-\theta_{jc}) + \sum_{i=1}^{N}\sum_{j=1}^{784} \log p(x_j^{(i)} \,|\, t^{(i)}, \boldsymbol{\theta})$$

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

differentiate with respect to θ_{mn}, we get

$$\frac{2}{\theta_{mn}} - \frac{2}{1-\theta_{mn}} + \sum_{i=1}^{N} t_n^{(i)} x_m^{(i)} \frac{1}{\theta_{mn}} - t_n^{(i)}(1-x_m^{(i)})\frac{1}{1-\theta_{mn}} \overset{\text{set}}{=} 0$$

$$2(1-\theta_{mn}) - 2\theta_{mn} + \sum_{i=1}^{N} t_n^{(i)} x_m^{(i)}(1-\theta_{mn}) - t_n^{(i)}(1-x_m^{(i)})\theta_{mn} = 0$$

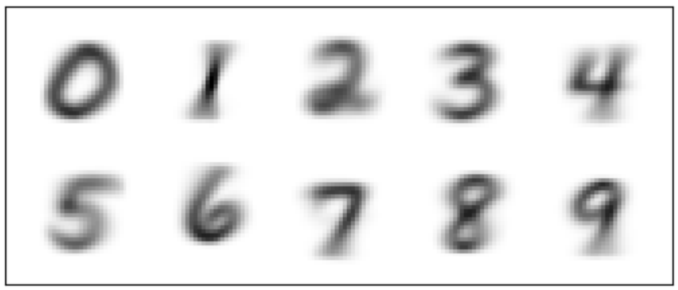
$$2 - 4\theta_{mn} + \sum_{i=1}^{N} t_n^{(i)} x_m^{(i)} - t_n^{(i)}\theta_{mn} = 0$$

$$\hat{\theta}_{mn} = \frac{2 + \sum_{i=1}^{N} t_n^{(i)} x_m^{(i)}}{4 + \sum_{i=1}^{N} t_n^{(i)}} = \frac{2 + \text{no. of data with label } n \text{ and feature } m}{4 + \text{no. of data with label } n}$$

(f) here is the report

```
Average log-likelihood for MAP is  -3.3570631378602855
Training accuracy for MAP is  0.8352166666666667
Test accuracy for MAP is  0.816
```

(g) here is the plot of the MAP estimator $\hat{\boldsymbol{\theta}}$ as 10 separate greyscale images



3. (a) The posterior $p(\boldsymbol{\theta} \mid \mathcal{D}) \propto p(\boldsymbol{\theta}) p(\mathcal{D} \mid \boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{a_k - 1} \prod_{i=1}^{N} p(x^{(i)} \mid \boldsymbol{\theta})$

$$
\begin{aligned}
&= \prod_{k=1}^{K} \theta_k^{a_k-1} \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{x_k^{(i)}} = \prod_{k=1}^{K} \theta_k^{a_k-1} \prod_{k=1}^{K} \prod_{i=1}^{N} \theta_k^{x_k^{(i)}} \\
&= \prod_{k=1}^{K} \theta_k^{a_k-1} \prod_{k=1}^{K} \theta_k^{\sum_{i=1}^{N} x_k^{(i)}} = \prod_{k=1}^{K} \theta_k^{a_k-1} \prod_{k=1}^{K} \theta_k^{N_k} \\
&= \prod_{k=1}^{K} \theta_k^{a_k-1+N_k}
\end{aligned}
$$

therefore, $\boldsymbol{\theta} \mid \mathcal{D} \sim \text{Dirichlet}(a_1 + N_1, \ldots, a_K + N_K)$ and Dirichlet distribution is a conjugate prior.

(b) the log-likelihood function of $\boldsymbol{\theta}$ is

$$
\begin{aligned}
\ell(\theta) = \log p(\mathcal{D} \mid \theta) &= \log \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{x_k^{(i)}} \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} x_k^{(i)} \log \theta_k \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K-1} x_k^{(i)} \log \theta_k + x_K^{(i)} \log(1 - \sum_{k=1}^{K-1} \theta_k)
\end{aligned}
$$

for $j \neq K$, differentiate with respect to θ_j, we get

$$
\begin{aligned}
\sum_{i=1}^{N} x_j^{(i)} \frac{1}{\theta_j} - x_K^{(i)} \frac{1}{\theta_K} &\overset{\text{set}}{=} 0 \\
\implies \hat{\theta}_j = \hat{\theta}_K \frac{\sum_{i=1}^{N} x_j^{(i)}}{\sum_{i=1}^{N} x_K^{(i)}} &= \hat{\theta}_K \frac{N_j}{N_K}
\end{aligned}
$$

since $\hat{\theta}_i$'s should sum up to one, we know

$$
\begin{aligned}
\hat{\theta}_K + \hat{\theta}_K \sum_{j=1}^{K-1} \frac{N_j}{N_K} &= 1 \\
\hat{\theta}_K (1 + \frac{1}{N_K} \sum_{j=1}^{K-1} N_j &= 1) \\
\hat{\theta}_K (1 + \frac{1}{N_K} (N - N_K)) &= 1 \\
\implies \hat{\theta}_K = \frac{N_K}{N} \text{ and } \hat{\theta}_j &= \frac{N_j}{N}
\end{aligned}
$$

thus, the MAP estimate of $\hat{\theta}_i = \frac{N_i}{N}$ for all i.

(c) by the posterior predictive distribution,

$$
\begin{aligned}
p(x_k^{N+1} = 1 \,|\, \mathcal{D}) &= \int p(x_k^{N+1} = 1 \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \mathcal{D}) \, d\boldsymbol{\theta} \\
&= \int \theta_k \prod_{k=1}^{K} \theta_k^{a_k - 1 + N_k} \, d\boldsymbol{\theta} \quad (*)
\end{aligned}
$$

given $\boldsymbol{\theta} \sim \text{Dirichlet}(a_1, \ldots, a_K)$, calculate the expectation of θ_k,

$$
\mathbb{E}(\theta_k) = \int \theta_k p(\boldsymbol{\theta}) = \int \theta_k \prod_{k=1}^{K} \theta_k^{a_k - 1} = \frac{a_k}{\sum_{k'} a_{k'}}
$$

let $\boldsymbol{\theta} \sim \text{Dirichlet}(a_1 + N_1, \ldots, a_K + N_K)$, observe that

$$
(*) = \mathbb{E}(\theta_k) = \frac{a_k + N_k}{\sum_{k'} a_{k'} + N_{k'}} = \frac{a_k + N_k}{N + \sum_{k'} a_{k'}}
$$

4. (a) here is the report of average conditional log-likelihood:

```
the average conditional log-likelihood on training set is -0.12462443666863039
the average conditional log-likelihood on test set is -0.19667320325525578
```

(b) here is the report of accuracy:

```
the accuracy on training set is 0.9814285714285714
the accuracy on training set is 0.97275
```

(c) here is the report of eigenvectors: