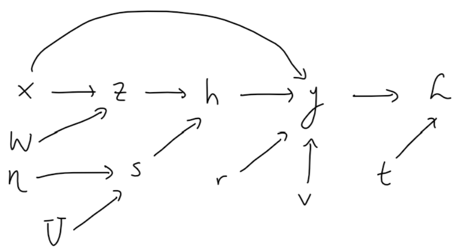
1. Backprop.

a) Below is the computation graph relating **x, z, s, h** and η and the model parameters.



b) Below are the backprop formulas for all error signals and model parameters.

$$
\begin{aligned}
\bar{\mathcal{L}} &= 1 \\
\bar{\mathbf{y}} &= \bar{\mathcal{L}}(\mathbf{y} - \mathbf{t}) \\
\bar{\mathbf{v}} &= \bar{\mathbf{y}}\mathbf{h} \\
\bar{\mathbf{h}} &= \bar{\mathbf{y}}\mathbf{v} \\
\bar{\mathbf{r}} &= \bar{\mathbf{y}}\mathbf{x} \\
\bar{\mathbf{z}} &= \bar{\mathbf{h}}\ \sigma(\mathbf{s}) \\
\bar{\mathbf{s}} &= \bar{\mathbf{h}}\mathbf{z}\ \sigma'(\mathbf{s}) \\
\bar{\mathbf{U}} &= \bar{\mathbf{s}}\eta^T \\
\bar{\eta} &= \bar{\mathbf{s}}\mathbf{U} \\
\bar{\mathbf{W}} &= \bar{\mathbf{z}}\bar{\mathbf{x}}^T \\
\bar{\mathbf{x}} &= \bar{\mathbf{y}}\mathbf{r} + \bar{\mathbf{z}}\bar{\mathbf{W}}
\end{aligned}
$$

2. Fitting a Naïve Bayes Model.

a) The log-likelihood decomposes into independent terms for each feature, and thus we can optimize them separately to compute the *maximum likelihood estimates*.

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log p(c^{(i)} \mid \boldsymbol{\pi}) \prod_{j=1}^{784} p\left(x_j^{(i)} \mid c^{(i)}, \theta_{jc}\right) \\ &= \sum_{i=1}^{N} \log p(c^{(i)} \mid \boldsymbol{\pi}) \; + \; \sum_{j=1}^{784}\sum_{i=1}^{N} p\left(x_j^{(i)} \mid c^{(i)}, \theta_{jc}\right)\end{aligned}$$

For the prior, we maximize $\sum_{i=1}^{N} \log p(c^{(i)} \mid \boldsymbol{\pi})$ to obtain $\hat{\boldsymbol{\pi}}_{\text{MLE}}$.

$$\begin{aligned}\sum_{i=1}^{N} \log p(c^{(i)} \mid \boldsymbol{\pi}) &= \sum_{i=1}^{N} \log \prod_{j=0}^{9} \pi_j^{t_j^{(i)}} \\ &= \sum_{i=1}^{N}\sum_{j=0}^{9} \log \pi_j^{t_j^{(i)}} \\ &= \sum_{i=1}^{N}\sum_{j=0}^{8} t_j^{(i)} \log(\pi_j) + t_9^{(i)} \log \pi_9 \\ &= \sum_{i=1}^{N}\sum_{j=0}^{8} t_j^{(i)} \log(\pi_j) + \log\left(1 - \sum_{j=0}^{8} \pi_j\right) \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \pi_j} &= \sum_{i=1}^{N} \frac{t_j^{(i)}}{\pi_j} - \frac{t_9^{(i)}}{(1 - \sum_{j=0}^{8} \pi_j)} \\ &= \frac{1}{\pi_j}\sum_{i=1}^{N} t_j^{(i)} - \frac{1}{\pi_9}\sum_{i=1}^{N} t_9^{(i)} = 0\end{aligned}$$

Denote $\sum_{i=1}^{N} t_j^{(i)} = \sum_{i=1}^{N} \mathbb{1}(t_j^{(i)} = 1) = N_j$ to be the number of labels of class j and $\sum_{i=1}^{N} t_9^{(i)} = \sum_{i=1}^{N} \mathbb{1}(t_9^{(i)} = 1) = N_9$ to be the number of labels of class 9.

$$\begin{aligned}\Rightarrow \; \frac{N_j}{\pi_j} &= \frac{N_9}{\pi_9} \\ \frac{\hat{\pi}_j}{\hat{\pi}_9} &= \frac{N_j}{N_9}\end{aligned}$$

By given definition of π_j we have $\sum_{j=0}^{9} \pi_j = 1$ and $\sum_{j=0}^{9} \pi_j = 1$ and $\sum_{j=0}^{9} N_j = N$.

$$\begin{aligned}\sum_{j=0}^{9} \frac{\hat{\pi}_j}{\hat{\pi}_9} &= \sum_{j=0}^{9} \frac{N_j}{N_9} \\ \frac{1 - \hat{\pi}_9}{\hat{\pi}_9} &= \frac{1 - N_9}{N_9} \\ N_9 - N_9\hat{\pi}_9 &= N\hat{\pi}_9 - N_9\hat{\pi}_9 \\ \hat{\pi}_9 &= \frac{N_9}{N} = \frac{\mathbb{1}(t_9^{(i)} = 1)}{N} \\ \Rightarrow \hat{\pi}_j &= \frac{\mathbb{1}(t_j^{(i)} = 1)}{N}\end{aligned}$$

Since each θ_{jc} can be treated separately, we will maximize $\sum_{i=1}^{N} p\left(x_j^{(i)} \mid c^{(i)}, \theta_{jc}\right)$ to obtain $\hat{\theta}_{jc\mathrm{MLE}}$.

$$\begin{aligned}
\sum_{i=1}^{N} p\left(x_j^{(i)} \mid c^{(i)}, \theta_{jc}\right) &= \sum_{i=1}^{N} \log\left[\prod_{k=0}^{9} \pi_k^{t_k^{(i)}} \theta_{jc}^{x_j^{(i)}} \left(1-\theta_{jc}\right)^{1-x_j^{(i)}}\right] \\
&= \sum_{i=1}^{N}\sum_{k=0}^{9} t_k^{(i)} \log \pi_k + x_j^{(i)} \log \theta_{jc} + \left(1 - x_j^{(i)}\right) \log\left(1-\theta_{jc}\right) \\
\frac{\partial \ell(\theta)}{\partial \theta_{jc}} &= \sum_{i=1}^{N}\sum_{k=0}^{9} \frac{x_j^{(i)}}{\theta_{jc}} - \frac{1-x_j^{(i)}}{1-\theta_{jc}} \\
&= \sum_{i=1}^{N}\sum_{k=0}^{9} \frac{x_j\left(1-\theta_{ic}\right) - \theta_{jc}\left(1-x_j^{(i)}\right)}{\theta_{jc}\left(1-\theta_{jc}\right)}
\end{aligned}$$

Let us introduce an indicator function $\mathbb{1}(c^{(i)} = c)$ to return 1 when our i'th sample matches the given class.

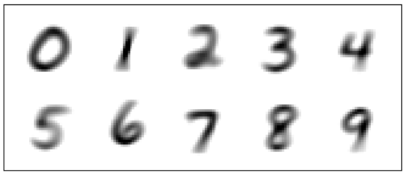
$$\begin{aligned}
&\Rightarrow \sum_{i=1}^{N} \mathbb{1}\left(c^{(i)} = c\right)\left[x_j^{(i)}\left(1-\theta_{jc}\right) - \theta_{jc}\left(1 - x_j^{(i)}\right)\right] = 0 \\
&\Rightarrow \left(1-\theta_{jc}\right)\sum_{i=1}^{N} \mathbb{1}\left(c^{(i)} = c\right) x_j^{(i)} - \theta_{jc}\sum_{i=1}^{N} \mathbb{1}\left(c^{(i)} = c\right)\left(1 - x_j^{(i)}\right) = 0 \\
&\Rightarrow \sum_{i=1}^{N} \mathbb{1}\left(c^{(i)} = c\right) x_j^{(i)} - \theta_{jc}\left[\sum_{i=1}^{N} \mathbb{1}\left(c^{(i)} = c\right) x_j^{(i)} + 1 - x_j^{(n)}\right] = 0 \\
&\Rightarrow \hat{\theta}_{jc} = \frac{\sum_{i=1}^{N} \mathbb{1}\left(c^{(i)} = c \;\&\; x_j^{(i)} = 1\right)}{\sum_{i=1}^{N} \mathbb{1}\left(c^{(i)} = c\right)}
\end{aligned}$$

b) Below is the derivation for the log-likelihood for a single training image.

$$\begin{aligned}
\ell(\theta) &= \log(p(\mathbf{t}|\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\pi})) \\
&= \log\left(\frac{p(c|\boldsymbol{\pi})(p(\mathbf{x}|c, \boldsymbol{\theta}, \boldsymbol{\pi})}{p(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\pi})}\right) \\
&= \log\left(\frac{p(c|\boldsymbol{\pi}) p(\mathbf{x} \mid c, \theta, \pi)}{\sum_{c'=0}^{9} p(c'|\boldsymbol{\pi}) p\left(\mathbf{x} \mid c', \theta, \pi\right)}\right) \\
&\propto \log\left(\pi_c\, p(\mathbf{x} \mid c, \theta, \pi)\right) \\
&= \log\left(p(c|\boldsymbol{\pi}) \prod_{j=1}^{784} \theta_{jc}^{x_j}\left(1-\theta_{jc}\right)^{1-x_j}\right) \\
&= \log p(c|\boldsymbol{\pi}) + \sum_{j=1}^{784}\left(x_j \log\, \theta_{jc} + \left(1-x_j\right)\log(1-\theta_{jc})\right)
\end{aligned}$$

c) It seems like some of the probabilities in $\hat{\theta}_{MLE}$ are very close in value to zero and thus are approximated as zeroes. This becomes a problem during our log-likelihood computations as the logarithm of 0 is undefined.

d) Below is the plot for the MLE estimator $\hat{\boldsymbol{\theta}}$.



e) We will derive the MAP estimate for θ similar to how we derived the MLE estimate earlier.

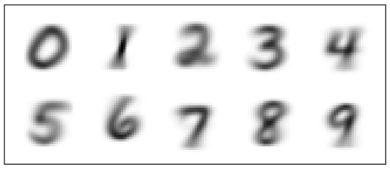
$$
\begin{aligned}
\sum_{i=1}^{N} p\left(x_j^{(i)} \mid c^{(i)}, \theta_{jc}\right) &= \sum_{i=1}^{N} \log \left[\prod_{k=0}^{9} \pi_k^{t_k^{(i)}} \theta_{jc}^{x_j^{(i)}} \left(1-\theta_{jc}\right)^{1-x_j^{(i)}} \times Beta(3,3)\right] \\
&= \sum_{i=1}^{N} \log \left[\prod_{k=0}^{9} \pi_k^{t_k^{(i)}} \theta_{jc}^{x_j^{(i)}} \left(1-\theta_{jc}\right)^{1-x_j^{(i)}} \theta_{jc}^{3-1}\left(1-\theta_{jc}\right)^{3-1}\right] \\
&= \sum_{i=1}^{N} \log \left[\prod_{k=0}^{9} \pi_k^{t_k^{(i)}} \theta_{jc}^{x_j^{(i)}+2} \left(1-\theta_{jc}\right)^{1-x_j^{(i)}+2}\right] \\
&= \sum_{i=1}^{N} \sum_{k=0}^{9} t_k^{(i)} \log \pi_k + (x_j^{(i)}+2) \log \theta_{jc} + \left(3-x_j^{(i)}\right) \log\left(1-\theta_{jc}\right) \\
\frac{\partial \ell(\theta)}{\partial \theta_{jc}} &= \sum_{i=1}^{N} \sum_{k=0}^{9} \frac{x_j^{(i)}+2}{\theta_{jc}} - \frac{3-x_j^{(i)}}{1-\theta_{jc}} \\
&= \sum_{i=1}^{N} \sum_{k=0}^{9} \frac{(x_j^{(i)}+2)\left(1-\theta_{jc}\right) - \theta_{jc}\left(3-x_j^{(i)}\right)}{\theta_{jc}\left(1-\theta_{jc}\right)}
\end{aligned}
$$

Let us introduce an indicator function $\mathbb{1}(c^{(i)}=c)$ to return 1 when our i'th sample matches the given class.

$$
\begin{aligned}
&\Rightarrow \sum_{i=1}^{N} \mathbb{1}\left(c^{(i)}=c\right) \left[(x_j^{(i)}+2)\left(1-\theta_{jc}\right) - \theta_{jc}\left(3-x_j^{(i)}\right)\right] = 0 \\
&\Rightarrow \left(1-\theta_{jc}\right) \sum_{i=1}^{N} \mathbb{1}\left(c^{(i)}=c\right)(x_j^{(i)}+2) - \theta_{jc} \sum_{i=1}^{N} \mathbb{1}\left(c^{(i)}=c\right)\left(3-x_j^{(i)}\right) = 0 \\
&\Rightarrow \sum_{i=1}^{N} \mathbb{1}\left(c^{(i)}=c\right)(x_j^{(i)}+2) - \theta_{jc}\left[\sum_{i=1}^{N} \mathbb{1}\left(c^{(i)}=c\right) x_j^{(i)} + 1 + 3 - x_j^{(n)}\right] = 0 \\
&\Rightarrow \hat{\theta}_{jc} = \frac{\sum_{i=1}^{N} \mathbb{1}\left(c^{(i)}=c \;\&\; x_j^{(i)}=1\right)+2}{\sum_{i=1}^{N} \mathbb{1}\left(c^{(i)}=c\right)+4}
\end{aligned}
$$

f) The average log-likelihood per data point was -3.36, the training accuracy was 80.1%, and the test accuracy was 79.3%.

g) Below is the plot for the MLE estimator $\hat{\boldsymbol{\theta}}$.



3. Categorical Distribution.

a) Below is the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$.

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &\propto p(\boldsymbol{\theta}p(\mathcal{D}|\boldsymbol{\theta}) \\ &= \prod_{j=1}^{K} \theta_j^{a_j-1} \prod_{i=1}^{N}\prod_{j=1}^{K} \theta_j^{x_j^{(i)}} \\ &= \prod_{j=1}^{K} \theta_j^{(\sum_{i=1}^{N} x_j^{(i)}+a_j)-1} \end{aligned}$$

Yes, the Dirichlet distribution is a conjugate prior for the categorical distribution since the prior has the same functional form as the likelihood.

b) Below is the derivation for the MAP estimate of the parameter vector $\boldsymbol{\theta}$.

$$\begin{aligned} \log(p(\boldsymbol{\theta}|\mathcal{D})) &= \prod_{j=1}^{K} \theta_j^{(\sum_{i=1}^{N} x_j^{(i)}+a_j)-1} \\ &= \sum_{j=1}^{K}\left(\sum_{i=1}^{N} x_j^{(i)} + a_j - 1\right)\log(\theta_j) \\ &= \sum_{j=1}^{K-1}\left(\sum_{i=1}^{N} x_j^{(i)} + a_j - 1\right)\log(\theta_j) + \left(\sum_{i=1}^{N} x_K^{(i)} + a_K - 1\right)\log(\theta_K) \end{aligned}$$

Since $\sum_{j=1}^{K}\theta_j = 1$, we have the following expression:

$$\sum_{j=1}^{K-1}\left(\sum_{i=1}^{N} x_j^{(i)} + a_j - 1\right)\log(\theta_j) + \left(\sum_{i=1}^{N} x_K^{(i)} + a_K - 1\right)\log\left(1 - \sum_{j=1}^{K-1}\theta_j\right) \tag{1}$$

We will now differentiate expression (1) w.r.t θ_j and set it for 0 to find the max for all $j = 1, 2, ..., K-1$.

$$\begin{aligned}
\frac{\partial(1)}{\partial\theta_j} &= \frac{\sum_{i=1}^N x_j^{(i)} + a_j - 1}{\theta_j} - \frac{\sum_{i=1}^N x_K^{(i)} + a_K - 1}{1 - \sum_{j=1}^{K-1}\theta_j} \\
&\Rightarrow \frac{\sum_{i=1}^N x_j^{(i)} + a_j - 1}{\theta_j} - \frac{\sum_{i=1}^N x_K^{(i)} + a_K - 1}{\theta_K} = 0 \\
&\Rightarrow \frac{\hat\theta_j}{\hat\theta_k} = \frac{\sum_{i=1}^N x_j^{(i)} + a_j - 1}{\sum_{i=1}^N x_K^{(i)} + a_K - 1}
\end{aligned}$$

Let $N_j = \sum_{i=1}^N x_j^{(i)}$ denote the number of samples of class j and let $N_K = \sum_{i=1}^N x_K^{(i)}$ denote the number of samples of class K. Also note that $\sum_{j=1}^{K-1}\frac{\hat\theta_j}{\hat\theta_k} = \frac{1-\hat\theta_K}{\hat\theta_k}$ by previous assumption.

$$\begin{aligned}
&\sum_{j=1}^{K-1}\frac{\hat\theta_j}{\hat\theta_k} = \sum_{j=1}^{K-1}\frac{N_j + a_j - 1}{N_K + a_K - 1} \\
&\Rightarrow \frac{1-\hat\theta_K}{\hat\theta_k} = \frac{\sum_{j=1}^{K-1}(N_j + a_j - 1)}{N_K + a_K - 1} \\
&\Rightarrow \frac{1-\hat\theta_j}{\hat\theta_k} = \frac{N - N_K + \sum_{j=1}^{K-1} a_j - K + 1}{N_K + a_K - 1} \\
&\Rightarrow N_K + a_K - 1 + \hat\theta_K(N_K + a_K - 1) = \hat\theta_K(N - N_K + \sum_{j=1}^{K-1} a_j - K + 1) \\
&\Rightarrow \hat\theta_K(N - N_K + \sum_{j=1}^{K-1} a_j - K + 1 + N_K + a_K - 1) = N_K + a_K - 1 \\
&\Rightarrow \hat\theta_K = \frac{N_K + a_K - 1}{\sum_{j=1}^K a_j + N - K} \\
&\Rightarrow \hat\theta_{kMAP} = \frac{N_k + a_k - 1}{\sum_{j=1}^K a_j + N - K}
\end{aligned}$$

c) We will calculate the probability that the $N+1$ outcome was k under the posterior predictive distribution.

$$\begin{aligned}
p(\mathbf{x}^{(N+1)} \mid \mathcal{D}) &= \int p(\mathbf{x}^{(N+1)} \mid \boldsymbol\theta)p(\boldsymbol\theta \mid \mathcal{D})d\boldsymbol\theta \\
&= \int \prod_{j=1}^K \theta_j^{x_j^{(N+1)}} \prod_{j=1}^K \theta_j^{(\sum_{i=1}^N x_j^{(i)} + a_j) - 1} d\boldsymbol\theta \\
&= \int \theta_k \prod_{j=1}^K \theta_j^{(\sum_{i=1}^N x_j^{(i)} + a_j) - 1} d\boldsymbol\theta \qquad (x_k^{(N+1)} = 1) \\
&= \mathbb{E}[\theta_k]
\end{aligned}$$

Since our expression in the integrand follows the Dirichlet distribution with $\boldsymbol\alpha$ parameter as,

$$\boldsymbol\alpha = \left(\sum_{i=1}^N x_1^{(i)} + a_1, \sum_{i=1}^N x_2^{(i)} + a_2, \ldots, \sum_{i=1}^N x_K^{(i)} + a_K\right)$$

we have the following expression:

$$\begin{aligned}\mathbb{E}[\theta_k] &= \frac{\sum_{i=1}^{N} x_k^{(i)} + a_k}{\sum_{j=1}^{K} \sum_{i=1}^{N} x_j^{(i)} + a_j} \\ &= \frac{\sum_{i=1}^{N} x_k^{(i)} + a_k}{N + \sum_{j=1}^{K} a_j}\end{aligned}$$

4. Gaussian Discriminant Analysis.

c)