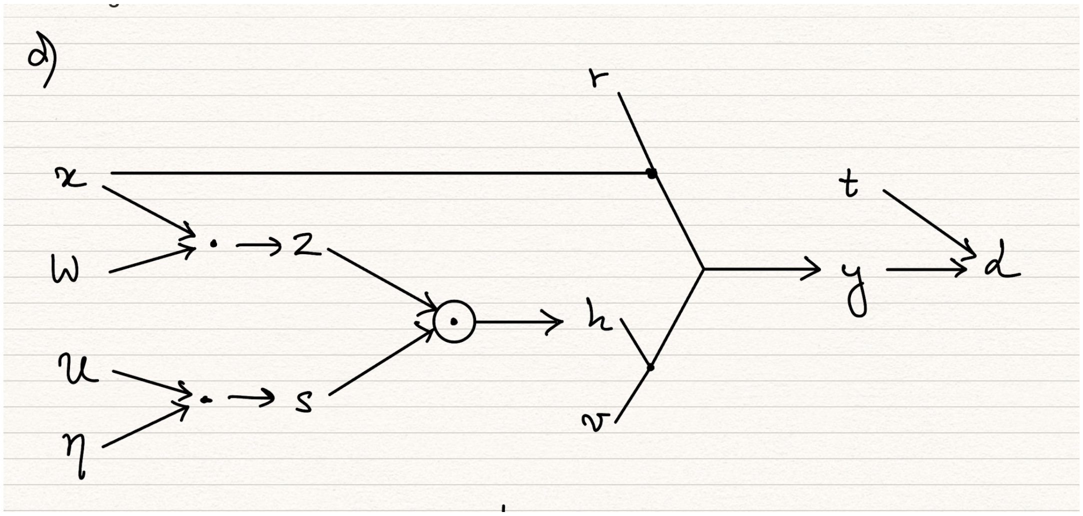
Question 1: Backpropagation

(a) Here's the computation graph for the given network:



(b) We proceed as follows, first computing both quantities and showing equality:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\sigma(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{1+e^{-x}} \\ &= -1 \cdot (1+e^{-x})^{-2} \cdot -1 \cdot e^{-x} \\ &= \frac{e^{-x}}{(1+e^{-x})^2}\end{aligned}$$

$$\begin{aligned}\sigma(x)(1-\sigma(x)) &= \frac{1}{1+e^{-x}}\left(1-\frac{1}{1+e^{-x}}\right) \\ &= \frac{1}{1+e^{-x}}\left(\frac{e^{-x}}{1+e^{-x}}\right) \\ &= \frac{e^{-x}}{(1+e^{-x})^2}\end{aligned}$$

showing that

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

(c) First, let's define the intermediate quantity

$$u = v^T h + r^T x$$

such that

$$y = \sigma(u)$$

Then, let's list all partial derivatives needed in the calculation:

$$\frac{\partial \mathcal{L}}{\partial y} = \frac{t}{y} - \frac{1-t}{1-y}$$
$$\frac{\partial \mathcal{L}}{\partial t} = \log(y) - \log(1-y)$$

$$\frac{\partial y}{\partial u} = \sigma(u)(1-\sigma(u))$$

$$\frac{\partial u}{\partial v} = h^T$$
$$\frac{\partial u}{\partial h} = v^T$$
$$\frac{\partial u}{\partial r} = x^T$$
$$\frac{\partial u}{\partial x} = r^T$$

$$\frac{\partial h}{\partial z} = \mathrm{diag}(s)$$
$$\frac{\partial h}{\partial s} = \mathrm{diag}(z)$$

$$\frac{\partial z}{\partial W} = x^T$$
$$\frac{\partial z}{\partial x} = W$$

$$\frac{\partial s}{\partial U} = \eta^T$$
$$\frac{\partial s}{\partial \eta} = U$$

Here are now the backprop formulas for the error signals.

$$\overline{\mathcal{L}} = 1$$

$$\begin{aligned}\overline{y} &= \overline{\mathcal{L}} \cdot \frac{\partial \mathcal{L}}{\partial y} \\ &= \frac{t}{y} - \frac{1-t}{1-y}\end{aligned}$$

$$\begin{aligned}\overline{t} &= \overline{\mathcal{L}} \cdot \frac{\partial \mathcal{L}}{\partial t} \\ &= \log(y) - \log(1-y)\end{aligned}$$

$$\begin{aligned}\overline{u} &= \overline{y} \cdot \frac{\partial y}{\partial u} \\ &= \overline{y} \cdot \sigma(u)(1-\sigma(u)) \\ &= \left(\frac{t}{y} - \frac{1-t}{1-y}\right) \cdot \sigma(u)(1-\sigma(u))\end{aligned}$$

$$\begin{aligned}\overline{r} &= \overline{u} \cdot \frac{\partial u}{\partial r} \\ &= \overline{u} \cdot x^T \\ &= \left(\frac{t}{y} - \frac{1-t}{1-y}\right) \cdot \sigma(u)(1-\sigma(u)) \cdot x^T\end{aligned}$$

$$\begin{aligned}\overline{v} &= \overline{u} \cdot \frac{\partial u}{\partial v} \\ &= \overline{u} \cdot h^T \\ &= \left(\frac{t}{y} - \frac{1-t}{1-y}\right) \cdot \sigma(u)(1-\sigma(u)) \cdot h^T\end{aligned}$$

$$\begin{aligned}\overline{h} &= \overline{u} \cdot \frac{\partial u}{\partial h} \\ &= \overline{u} \cdot v^T \\ &= \left(\frac{t}{y} - \frac{1-t}{1-y}\right) \cdot \sigma(u)(1-\sigma(u)) \cdot v^T\end{aligned}$$

$$\begin{aligned}\overline{z} &= \overline{h} \cdot \frac{\partial h}{\partial z} \\ &= \overline{h} \cdot \operatorname{diag}(s) \\ &= \left(\frac{t}{y} - \frac{1-t}{1-y}\right) \cdot \sigma(u)(1-\sigma(u)) \cdot v^T \cdot \operatorname{diag}(s)\end{aligned}$$

$$\begin{aligned}\overline{s} &= \overline{h} \cdot \frac{\partial h}{\partial s} \\ &= \overline{h} \cdot \operatorname{diag}(z) \\ &= \left(\frac{t}{y} - \frac{1-t}{1-y}\right) \cdot \sigma(u)(1-\sigma(u)) \cdot v^T \cdot \operatorname{diag}(z)\end{aligned}$$

$$
\begin{aligned}
\overline{\eta} &= \overline{s} \cdot \frac{\partial s}{\partial \eta} \\
&= \overline{s} \cdot U \\
&= \left(\frac{t}{y} - \frac{1-t}{1-y} \right) \cdot \sigma(u)(1-\sigma(u)) \cdot v^T \cdot \mathrm{diag}(z) \cdot U
\end{aligned}
$$

$$
\begin{aligned}
\overline{U} &= \overline{s} \cdot \frac{\partial s}{\partial U} \\
&= \overline{s} \cdot \eta^T \\
&= \left(\frac{t}{y} - \frac{1-t}{1-y} \right) \cdot \sigma(u)(1-\sigma(u)) \cdot v^T \cdot \mathrm{diag}(z) \cdot \eta^T
\end{aligned}
$$

$$
\begin{aligned}
\overline{W} &= \overline{z} \cdot \frac{\partial z}{\partial W} \\
&= \overline{z} \cdot x^T \\
&= \left(\frac{t}{y} - \frac{1-t}{1-y} \right) \cdot \sigma(u)(1-\sigma(u)) \cdot v^T \cdot \mathrm{diag}(s) \cdot x^T
\end{aligned}
$$

$$
\begin{aligned}
\overline{x} &= \overline{z} \cdot \frac{\partial z}{\partial x} + \overline{u} \cdot \frac{\partial u}{\partial x} \\
&= \overline{z} \cdot W + \overline{u} \cdot U^T \\
&= \left(\frac{t}{y} - \frac{1-t}{1-y} \right) \cdot \sigma(u)(1-\sigma(u)) \cdot v^T \cdot \mathrm{diag}(s) \cdot W + \left(\frac{t}{y} - \frac{1-t}{1-y} \right) \cdot \sigma(u)(1-\sigma(u)) \cdot r^T
\end{aligned}
$$

Question 2: Fitting a Naive Bayes Model

(a) First, let's compute $\log(\mathcal{D}|\theta, \pi)$ for the given data set. The MLE for θ and π are then the maximizers of this quantity.

$$\begin{aligned}
\log(\mathcal{D}|\theta, \pi) &= \log\left(\prod_{i=1}^{N} p(x^{(i)}, c^{(i)}|\theta, \pi)\right) \\
&= \sum_{i=1}^{N} \log\left(p(x^{(i)}, c^{(i)}|\theta, \pi)\right) \\
&= \sum_{i=1}^{N} \log\left(p(c^{(i)}|\pi) p(x^{(i)}|c^{(i)}, \theta, \pi)\right) \\
&= \sum_{i=1}^{N} \log\left(\pi_{c^{(i)}} \prod_{j=1}^{D} p(x_j^{(i)}|c^{(i)}, \theta_{j,c^{(i)}})\right) \\
&= \sum_{i=1}^{N} \left(\log(\pi_{c^{(i)}}) + \sum_{j=1}^{D} \log(p(x_j^{(i)}|c^{(i)}, \theta_{j,c^{(i)}}))\right) \\
&= \sum_{i=1}^{N} \left(\log(\pi_{c^{(i)}}) + \sum_{j=1}^{D} \log\left(\theta_{j,c^{(i)}}^{x_j^{(i)}} (1 - \theta_{j,c^{(i)}})^{1-x_j^{(i)}}\right)\right) \\
&= \sum_{i=1}^{N} \left(\log(\pi_{c^{(i)}}) + \sum_{j=1}^{D} \left(x_j^{(i)} \log(\theta_{j,c^{(i)}}) + (1 - x_j^{(i)}) \log(1 - \theta_{j,c^{(i)}})\right)\right)
\end{aligned}$$

We can then compute the MLE for θ and π by maximizing the above expression. We can do this by taking the partial derivative with respect to $\theta_{j,c}$ and π_c and setting them equal to zero. We will do this for $\theta_{j,c}$ first.

$$\begin{aligned}
\frac{\partial}{\partial \theta_{j,c}} \log(\mathcal{D}|\theta, \pi) &= \sum_{i=1}^{N} \left[\mathbb{I}(c^{(i)} = c) \left(\frac{x_j^{(i)}}{\theta_{j,c}} - \frac{1 - x_j^{(i)}}{1 - \theta_{j,c}}\right)\right] \\
&= \sum_{i=1}^{N} \left[\mathbb{I}(c^{(i)} = c) \left(\frac{x_j^{(i)}}{\theta_{j,c}}\right)\right] - \sum_{i=1}^{N} \left[\mathbb{I}(c^{(i)} = c) \left(\frac{1 - x_j^{(i)}}{1 - \theta_{j,c}}\right)\right] \\
&= \frac{1}{\theta_{j,c}} \cdot \sum_{i=1}^{N} \mathbb{I}(c^{(i)} = c) x_j^{(i)} - \frac{1}{1 - \theta_{j,c}} \cdot \sum_{i=1}^{N} \mathbb{I}(c^{(i)} = c)(1 - x_j^{(i)}) \\
&= \frac{N_{1,c}}{\theta_{j,c}} - \frac{N_{0,c}}{1 - \theta_{j,c}} = 0
\end{aligned}$$

where $N_{0,c}$ is the number of examples with class c and $x_j^{(i)} = 0$ and $N_{1,c}$ is the number of examples with class c and $x_j^{(i)} = 1$. We can then solve for $\theta_{j,c}$ to get

$$\theta_{j,c} = \frac{N_{1,c}}{N_{0,c} + N_{1,c}}$$

Which is the MLE for $\theta_{j,c}$. We can do the same for π_j as follows, first writing

$$\pi_{c^{(i)}} = \prod_{j=0}^{9} \pi_j^{t_j^{(i)}} = \left(\prod_{j=0}^{8} \pi_j^{t_j^{(i)}}\right) \cdot \left(1 - \sum_{j=0}^{8} pi_j^{t_j^{(i)}}\right)^{t_9^{(i)}}$$

and so

$$\begin{aligned}
\frac{\partial}{\partial \pi_j} \log(\mathcal{D}|\theta, \pi) &= \frac{\partial}{\partial \pi_j} \left(\sum_{j=0}^{8} t_{j'}^{(i)} \log(\pi_{j'}) + t_9^{(i)} \log(1 - \sum_{j'=0}^{8} \pi_{j'}) \right) \\
&= \sum_{i=1}^{N} \frac{t_j^{(i)}}{\pi_j} - \frac{t_9^{(i)}}{\pi_9} \qquad \text{for } j < 9 \\
&= \frac{N_{c=j}}{\pi_j} - \frac{N_{c=9}}{\pi_9} = 0
\end{aligned}$$

where $N_{c=j}$ is the number of classes with label j.

And so, solving for π_j, we get

$$\pi_j = \frac{N_{c=j}}{N_{c=9}} \pi_9$$

Since we know that $\sum_j \pi_j = 1$, we have

$$\begin{aligned}
\pi_9 + \left(\sum_{j=0}^{8} \frac{N_{c=j}}{N_{c=9}} \pi_9 \right) &= 1 \\
\pi_9 \left(1 + \frac{N - N_{c=9}}{N_{c=9}} \right) &= 1 \\
\pi_9 \left(\frac{N}{N_{c=9}} \right) &= 1 \\
\pi_9 &= \frac{N_{c=9}}{N}
\end{aligned}$$

and so, solving for the other pi_j, we get

$$\pi_j = \frac{N_{c=j}}{N_{c=9}} \frac{N_{c=9}}{N} = \frac{N_{c=j}}{N}$$

finally getting that the MLE estimators for θ and π are

$$\begin{aligned}
\theta_{j,c} &= \frac{N_{1,c}}{N_{0,c} + N_{1,c}} \\
\pi_k &= \frac{N_{c=k}}{N}
\end{aligned}$$

where $N_{0,c}$ is the number of examples with class c and $x_j^{(i)} = 0$ and $N_{1,c}$ is the number of examples with class c and $x_j^{(i)} = 1$ and $N_{c=k}$ is the number of examples with class $c = k$.

(b) The log-likelihood $\log p(t|x, \theta, \pi)$ is given by

$$
\begin{aligned}
\log p(t|x,\theta,\pi) &= \log\left(\frac{p(x,t|\theta,\pi)}{p(x|\theta,\pi)}\right) \\
&= \log\left(\pi_c \prod_{j=1}^{784} p(x_j|c,\theta_{j,c})\right) - \log p(x|\theta,\pi) \\
&= \log\left(\pi_c\right) + \sum_{j=1}^{784}\left[x_j \log(\theta_{j,c}) + (1-x_j)\log(1-\theta_{j,c})\right] - K
\end{aligned}
$$

where

$$K = \log(p(x|\theta,\pi)) = \log\left(\sum_{c=0}^{8} p(x,c|\theta,\pi)\right) = \log\left(\sum_{c=0}^{8} \pi_c \prod_{j=1}^{784} p(x_j|c,\theta_{j,c})\right)$$

is a constant with respect to c, and does not change the maximizer of the log-likelihood.

(c) I have fit θ, π using MLE on the training set, but reporting the average log-likelihood goes wrong. This is because there are a number of j, c for which $N_{1,c} = 0$, that is, the number of examples with $x_j^{(i)} = 1$ and $c^{(i)} = c$. This means that $\theta_{j,c} = 0$ and so $\log(\theta_{j,c}) = -\infty$.

This means that computing the log-likelihood of $p(t|x,\theta,\pi)$ for some examples yields a value of $-\infty$, causing the average log-likelihood to not make sense.

(d) Here's the picture:

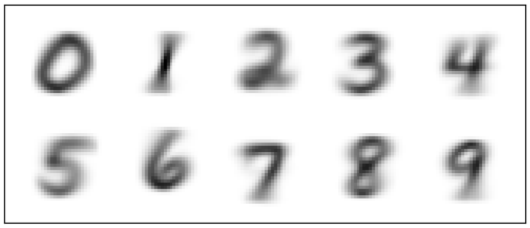


Figure 1: The MLE estimator θ for each class.

(e) Since we are to use a uniform prior for π, we have that

$$
\begin{aligned}
\text{argmax}_\pi \log(p(\pi|\mathcal{D})) &= \text{argmax}_\pi \log\left(\frac{p(\pi)p(\mathcal{D}|\pi)}{p(\mathcal{D}}\right) \\
&= \text{argmax}_\pi \log\big(p(\pi)\big) + \log\big(p(\mathcal{D}|\pi)\big) - \log(p(\mathcal{D}))
\end{aligned}
$$

Since $p(\mathcal{D})$ is a constant with respect to π, and $p(\pi)$ is also a constant with respect to π (uniform prior), we have that

$$\pi_{MAP} = \text{argmax}_\pi \log(p(\pi|\mathcal{D})) = \text{argmax}_\pi \log(p(\mathcal{D}|\pi)) = \pi_{MLE}$$

and so the MAP estimator for π is the same as the MLE estimator for π, which is given by

$$\pi_j = \frac{N_{c=k}}{N}$$

where $N_{c=k}$ is the number of examples with class $c = k$.
However, the MAP estimator for θ is different from the MLE estimator for θ. It is computed by

$$\begin{aligned}\text{argmax}_\theta \log(p(\theta|\mathcal{D})) &= \text{argmax}_\theta \log\left(\frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})}\right) \\ &= \text{argmax}_\theta \log\big(p(\theta)\big) + \log\big(p(\mathcal{D}|\theta)\big)\end{aligned}$$

Now, we know that

$$\begin{aligned}\log\big(p(\theta)\big) + \log\big(p(\mathcal{D}|\theta)\big) &= \log\left(\prod_{c,j} \theta_{j,c}^{\alpha-1}(1-\theta_{j,c})^{\beta-1}\right) + \log\big(p(\mathcal{D}|\theta)\big) \\ &= \sum_{c,j}\big[(\alpha-1)\log(\theta_{j,c}) + (\beta-1)\log(1-\theta_{j,c})\big] + \log\big(p(\mathcal{D}|\theta)\big)\end{aligned}$$

Taking the partial derivative with respect to $\theta_{j,c}$, we get

$$\begin{aligned}\frac{\partial}{\partial\theta_{j,c}} &= \frac{\alpha-1}{\theta_{j,c}} - \frac{\beta-1}{1-\theta_{j,c}} + \frac{\partial}{\partial\theta_{j,c}}\log\big(p(\mathcal{D}|\theta)\big) \\ &= \frac{\alpha-1}{\theta_{j,c}} - \frac{\beta-1}{1-\theta_{j,c}} + \frac{N_{1,c}}{\theta_{j,c}} - \frac{N_{0,c}}{1-\theta_{j,c}} \\ &= \frac{N_{1,c}+\alpha-1}{\theta_{j,c}} - \frac{N_{0,c}+\beta-1}{1-\theta_{j,c}} \\ &= 0\end{aligned}$$

and now solving for $\theta_{j,c}$, we get

$$\theta_{j,c} = \frac{N_{1,c}+\alpha-1}{N_{0,c}+N_{1,c}+\alpha+\beta-2}$$

(f) Done in the code. Here are the statistics:

```
Average log-likelihood for MAP is  -173.46659721464317  (not normalized)
Training accuracy for MAP is  0.8352166666666667
Test accuracy for MAP is  0.816
```

I am not normalizing the log-likelihood for each example i in class c by subtracting $\log(p(x|\theta,\pi))$, since this is a constant with respect to c and so does not change the predictions. Further, this is hard to compute, as it is given by

$$\log(p(x|\theta,\pi)) = \log\left(\sum_{c=1}^{k} p(x|\theta,\pi,c)p(c|\theta,\pi)\right)$$

which is a logarithm of a sum, which cannot be computed by the logarithms of the summands. In general, it was mentioned in lecture and tutorial that computing this is intractable. I have submitted, however, the code that computes it – it is only commented out.

(g) Here's the picture:

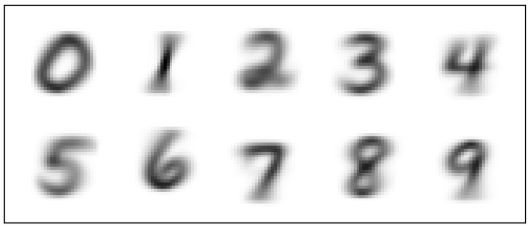


Figure 2: The MAP estimator θ for each class.

(h) An advantage of the Naive Bayes approach is that since parameters are assumed to be independent of each other, we can optimize them independently of each other. In this case, we were able to compute a closed-form solution for θ and π for both MLE and MAP, producing very fast training times.

A disadvantage of the Naive Bayes approach is that it assumes that the features are independent of each other, which is not the case in this problem. For example, in the case of the MNIST dataset, the pixels are not independent of each other, as it is likely that a pixel is white given that neighbouring pixels are also white. This is not captured by the Naive Bayes approach.

Question 3: Logistic Regression with Gaussian Prior

(a) First, let's write down an expression for the log likelihood of θ.

$$\begin{aligned}
ll(\theta) &= \log(p(\mathcal{D}|\theta)) \\
&= \log\left(\prod_{i=1}^{N} \sigma((x^{(i)})^T\theta)^{y^{(i)}}(1-\sigma((x^{(i)})^T\theta))^{1-y^{(i)}}\right) \\
&= \sum_{i=1}^{N} y^{(i)}\log(\sigma((x^{(i)})^T\theta)) + (1-y^{(i)})\log(1-\sigma((x^{(i)})^T\theta))
\end{aligned}$$

This is optimized by gradient descent by taking the gradient with respect to θ to perform gradient descent. We get

$$\begin{aligned}
\nabla_\theta ll(\theta) &= \sum_{i=1}^{N} y^{(i)}(1-\sigma((x^{(i)})^T\theta))x^{(i)} - (1-y^{(i)})\sigma((x^{(i)})^T\theta)x^{(i)} \\
&= \sum_{i=1}^{N}(y^{(i)}-\sigma((x^{(i)})^T\theta))x^{(i)}
\end{aligned}$$

and so we optimize (maximize) by performing gradient ASCENT with the update rule

$$\theta_{t+1} = \theta_t + \eta\sum_{i=1}^{N}(y^{(i)}-\sigma((x^{(i)})^T\theta))x^{(i)}$$

(b) The MAP log-likelihood of θ is given by

$$\begin{aligned}
ll_{MAP}(\theta) &= \log(p(\theta|D)) \\
&= C_1 + \log(p(\theta)p(D|\theta)) \\
&= C_1 + \log\left(\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}\theta^T\Sigma^{-1}\theta\right)\right) + \log(p(D|\theta)) \\
&= C_1 + C_2 + ll_{MLE}(\theta) + \log\left(\exp\left(-\frac{1}{2}\theta^T\left(\sigma^2 I\right)^{-1}\theta\right)\right) \\
&= C_1 + C_2 + ll_{MLE}(\theta) - \frac{1}{2\sigma^2}\theta^T\theta \\
&= C_1 + C_2 + ll_{MLE}(\theta) - \frac{1}{2\sigma^2}||\theta||^2 \\
&= C_1 + C_2 + \sum_{i=1}^{N} y^{(i)}\log(\sigma((x^{(i)})^T\theta)) + (1-y^{(i)})\log(1-\sigma((x^{(i)})^T\theta)) - \frac{1}{2\sigma^2}||\theta||^2
\end{aligned}$$

where $C_1 = \log(p(D))$ and $C_2 = \log\left(\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\right)$.