1 (a)

$$\frac{\partial E(w)}{\partial w_j} = \sum_{i=1}^{N} (x_j)^{(i)} y^{(i)} (1 - \sigma(w^T x^{(i)})) - (x_j)^{(i)} \sigma(w^T x^{(i)}) + (x_j)^{(i)} y^{(i)} \sigma(w^T x^{(i)})$$

$$= \sum_{i=1}^{N} (x_j)^{(i)} y^{(i)} - (x_j)^{(i)} \sigma(w^T x^{(i)})$$

$$\frac{\partial^2 E(w)}{\partial (w_j)^2} = - \sum_{i=1}^{N} (x_j)^{(i)} \sigma(w^T x^{(i)})(1 - \sigma(w^T x^{(i)}))(x_j)^{(i)}$$

$$\frac{\partial^2 E(w)}{\partial w_j w_k} = - \sum_{i=1}^{N} (x_j)^{(i)} \sigma(w^T x^{(i)})(1 - \sigma(w^T x^{(i)}))(x_k)^{(i)}$$

Let $X \in \mathbb{R}^{n \times m}$ be the design matrix and $w \in \mathbb{R}^m$ be the weight vector, where n is the number of observations and m is the number of features. Let $\odot$ express the Hadamard product of its operands. We can express the second-order partial derivatives in matrix form as

$$X^T diag[-\sigma(Xw) \odot (1 - \sigma(Xw))]X$$

which gives us our Hessian matrix.

1 (b) Let $x, z \in \mathbb{R}^m$. Then

$$(x^T z)(x^T z) = \sum_{i=1}^{N} z_i x_i \sum_{j=1}^{N} x_j z_j = \sum_{i=1}^{N} \sum_{j=1}^{N} z_i x_i x_j z_j = (x^T z)^2$$

Consider

$$z^T X^T diag[\sigma(Xw) \odot (1 - \sigma(Xw))]Xz$$

Let D represent $X^T diag[\sigma(Xw)(1 - \sigma(Xw))]X$. $D \in \mathbb{R}^{m \times m}$. Let D_i be a column of D and $D^{(j)}$ be a row of D. we can now express the above as

$$- \sum_{i=1}^{m} [\sum_{j=1}^{m} z_j (D_i)^{(j)}] z_i$$

$$= - \sum_{i=1}^{m} [\sum_{j=1}^{m} z_j ((D_i)^{(j)})^{\frac{1}{2}} ((D_i)^{(j)})^{\frac{1}{2}}] z_i = - \sum_{i=1}^{m} \sum_{j=1}^{m} z_j ((D_i)^{(j)})^{\frac{1}{2}} ((D_i)^{(j)})^{\frac{1}{2}} z_i$$

Since in general $\sum_{i=1}^{N}\sum_{j=1}^{N} z_i x_i x_j z_j = (x^T z)^2 \geq 0$,

$$\sum_{i=1}^{m}\sum_{j=1}^{m} z_j((D_i)^{(j)})^{\frac{1}{2}}((D_i)^{(j)})^{\frac{1}{2}} z_i \geq 0$$

therefore

$$-\sum_{i=1}^{m}[\sum_{j=1}^{m} z_j (D_i)^{(j)}] z_i \leq 0 \, \Box$$

The Hessian matrix is negative semi-definite, and our original log-likelihood function is concave.

1 (c) We can invert the sign of the log-lihelihood function to get a loss function. This gives us the Hessian matrix $X^T diag[\sigma(Xw) \odot (1-\sigma(Xw))]X$. For the gradient of the loss function with respect to w, we can take the partial derivative expression from **1 (a)**

$$\sum_{i=1}^{N}(x_j)^{(i)}y^{(i)} - (x_j)^{(i)}\sigma(w^T x^{(i)})$$

invert the sign, and express it in matrix form:

$$\nabla_w E = X^T\sigma(Xw) - X^T y$$

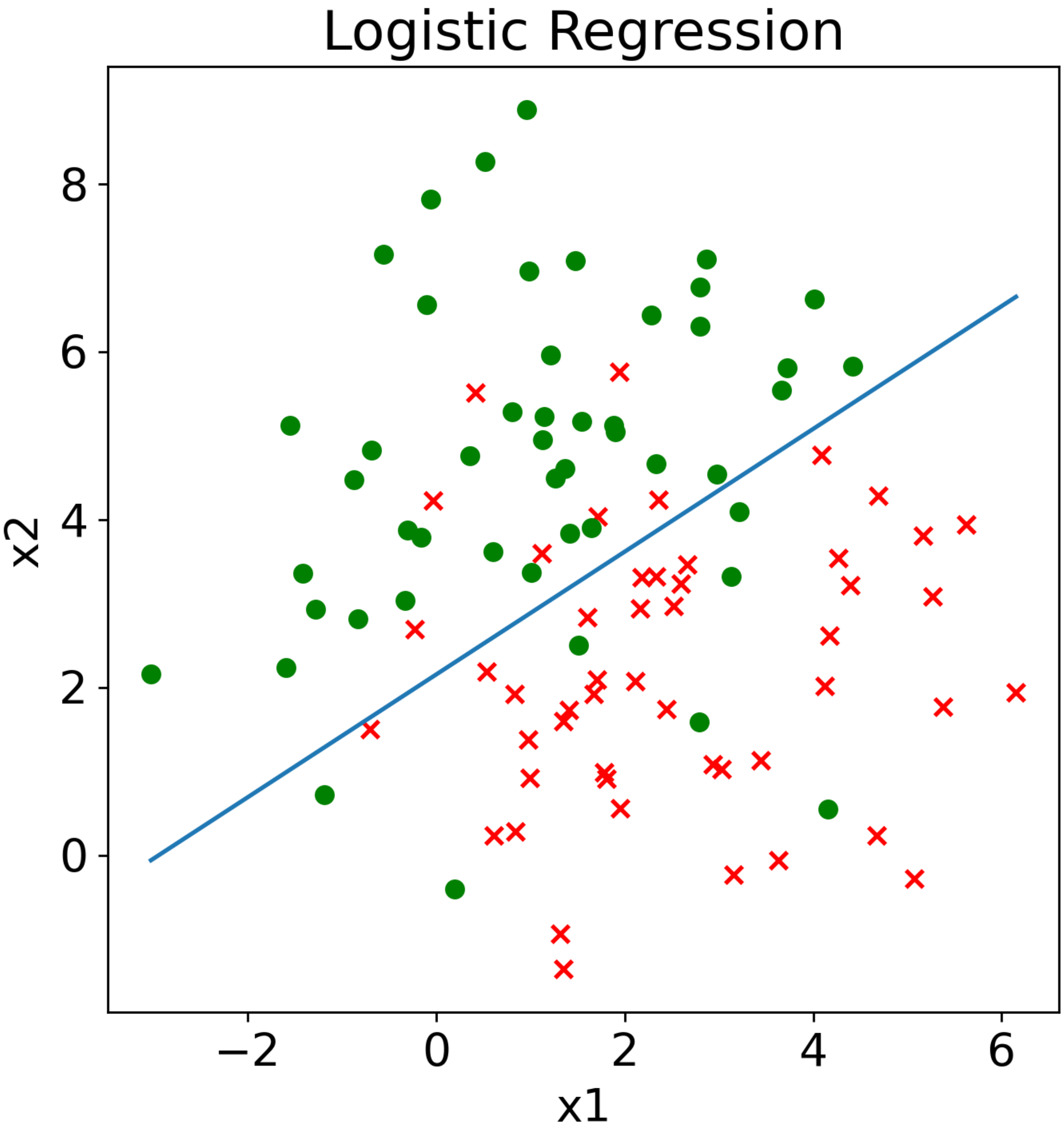
So the update for Newton's method is

$$w_{new} = w_{old} - H^{-1}\nabla_w E$$

$$w_{new} = w_{old} - \{X^T diag[\sigma(Xw) \odot (1-\sigma(Xw))]X\}^{-1}[X^T\sigma(Xw) - X^T y]$$

1 (e) The weights from the logistic regression are
w_0: -1.84922892 w_1: -0.62814188 w_2: 0.85846843

1 (f)



2 (a) If we consider only the case where $k = w_m$, we can express the log-likelihood as

$$\sum_{i=1}^{N} log([p(y^{(i)} = m|x^{(i)}, w)]^{\mathbb{I}\{y^{(i)}=m\}})$$

$$\sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = m\}log([\frac{exp[w_m^T\phi(x^{(i)})]}{\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})]}])$$

If we look only at this case, then the corresponding part of $\nabla w_m l(w)$ is

$$\sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = m\}\frac{\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})]}{exp[w_m^T\phi(x^{(i)})]}[\frac{exp[w_m^T\phi(x^{(i)})]\phi(x^{(i)})}{\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})]} - \frac{(exp[w_m^T\phi(x^{(i)})])^2\phi(x^{(i)})}{(\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})])^2}]$$

$$= \sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = m\}[\frac{\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})]exp[w_m^T\phi(x^{(i)})]\phi(x^{(i)})}{exp[w_m^T\phi(x^{(i)})]\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})]}$$

$$-\frac{\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})](exp[w_m^T\phi(x^{(i)})])^2\phi(x^{(i)})}{exp[w_m^T\phi(x^{(i)})](\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})])^2}]$$

$$= \sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = m\}[\phi(x^{(i)}) - \frac{(exp[w_m^T\phi(x^{(i)})])\phi(x^{(i)})}{(\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})])}]$$

$$= \sum_{i=1}^{N} \phi(x^{(i)})\mathbb{I}\{y^{(i)} = m\}[1 - \frac{(exp[w_m^T\phi(x^{(i)})])}{(\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})])}]$$

Now, let us look at the case where $k \neq m$. Here, the corresponding part of $\nabla w_m l(w)$ is

$$\sum^{k\neq m} \mathbb{I}\{y^{(i)} = k\}\frac{\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})]}{exp[w_k^T\phi(x^{(i)})]}[-\frac{exp[w_k^T\phi(x^{(i)})]exp[w_m^T\phi(x^{(i)})]\phi(x^{(i)})}{[\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})]]^2}]$$

$$= -\sum^{k\neq m} \mathbb{I}\{y^{(i)} = k\}[\frac{exp[w_m^T\phi(x^{(i)})]\phi(x^{(i)})}{[\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})]]}]$$

If we incorporate both cases, we get

$$\sum_{i=1}^{N} \phi(x^{(i)})\mathbb{I}\{y^{(i)} = m\}[1 - \frac{(exp[w_m^T\phi(x^{(i)})])}{(\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})])}] - \sum^{k\neq m} \mathbb{I}\{y^{(i)} = k\}[\frac{exp[w_m^T\phi(x^{(i)})]\phi(x^{(i)})}{[\sum_{j=1}^{K} exp[w_j^T\phi(x^{(i)})]]}]$$

$$= \sum_{i=1}^{N} \phi(x^{(i)})\mathbb{I}\{y^{(i)} = m\}[1 - \frac{(exp[w_m^T \phi(x^{(i)})])}{(\sum_{j=1}^{K} exp[w_j^T \phi(x^{(i)})])}] - [1 - \mathbb{I}\{y^{(i)} = m\}][\frac{exp[w_m^T \phi(x^{(i)})]\phi(x^{(i)})}{[\sum_{j=1}^{K} exp[w_j^T \phi(x^{(i)})]]}]$$

$$= \sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = m\}\phi(x^{(i)}) - \frac{\mathbb{I}\{y^{(i)} = m\}exp[w_m^T \phi(x^{(i)})])(\phi(x^{(i)})}{(\sum_{j=1}^{K} exp[w_j^T \phi(x^{(i)})])}]$$

$$-[\frac{exp[w_m^T \phi(x^{(i)})]\phi(x^{(i)})}{[\sum_{j=1}^{K} exp[w_j^T \phi(x^{(i)})]]}] + \frac{\mathbb{I}\{y^{(i)} = m\}]exp[w_m^T \phi(x^{(i)})]\phi(x^{(i)})}{[\sum_{j=1}^{K} exp[w_j^T \phi(x^{(i)})]]}$$

$$= \sum_{i=1}^{N} \phi(x^{(i)})[\mathbb{I}\{y^{(i)} = m\} - [\frac{exp[w_m^T \phi(x^{(i)})]}{[\sum_{j=1}^{K} exp[w_j^T \phi(x^{(i)})]]}]]$$

Because w_K is fixed at 0, we can write this as

$$\sum_{i=1}^{N} \phi(x^{(i)})[\mathbb{I}\{y^{(i)} = m\} - [\frac{exp[w_m^T \phi(x^{(i)})]}{[1 + \sum_{j=1}^{K-1} exp[w_j^T \phi(x^{(i)})]]}]] \square$$

So the update rule is

$$w_m* = w_m + \alpha \sum_{i=1}^{N} \phi(x^{(i)})[\mathbb{I}\{y^{(i)} = m\} - \frac{(exp[w_m^T \phi(x^{(i)})])}{[1 + \sum_{j=1}^{K-1} exp[w_j^T \phi(x^{(i)})]]}]$$

2 (c) Accuracy on the test set is 94%.

3 (a) With Bayes' theorem, we can express the posterior probability of the Gaussian discriminant analysis for $y = 1$ as

$$p(y=1|x;\phi,\Sigma,m_0,m_1) = \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1)+p(x|y=0)p(y=0)}$$

$$= \frac{\phi\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}exp[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]}{\phi\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}exp[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]+(1-\phi)\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}exp[-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)]}$$

$$= \frac{\phi\ exp[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]}{\phi\ exp[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]+(1-\phi)\ exp[-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)]}$$

$$= \frac{1}{1+\frac{(1-\phi)\ exp[-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)]}{\phi\ exp[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]}}$$

Consider

$$\frac{(1-\phi)\ exp[-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)]}{\phi\ exp[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]}$$

$$= \frac{1-\phi}{\phi}exp[-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)+\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]$$

$$= \frac{1-\phi}{\phi}exp[-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)+\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]$$

$$= \frac{1-\phi}{\phi}exp[x^T\Sigma^{-1}\mu_0-\frac{1}{2}\mu_0\Sigma^{-1}\mu_0-x^T\Sigma^{-1}\mu_1+\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1]$$

$$= exp[log(\frac{1-\phi}{\phi})]exp[x^T\Sigma^{-1}\mu_0-\frac{1}{2}\mu_0\Sigma^{-1}\mu_0-x^T\Sigma^{-1}\mu_1+\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1]$$

$$= exp[log(\frac{1-\phi}{\phi})+x^T\Sigma^{-1}\mu_0-\frac{1}{2}\mu_0\Sigma^{-1}\mu_0-x^T\Sigma^{-1}\mu_1+\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1]$$

$$= exp[-(x^T(\Sigma^{-1}\mu_1-\Sigma^{-1}\mu_0)+\frac{1}{2}(\mu_0\Sigma^{-1}\mu_0-\mu_1^T\Sigma^{-1}\mu_1)-log(\frac{1-\phi}{\phi}))]$$

By incorporating this into our expression for $p(y=1|x;\phi,\Sigma,m_0,m_1)$ above, we get

$$\frac{1}{1+exp[-(x^T(\Sigma^{-1}\mu_1-\Sigma^{-1}\mu_0)+\frac{1}{2}(\mu_0\Sigma^{-1}\mu_0-\mu_1^T\Sigma^{-1}\mu_1)-log(\frac{1-\phi}{\phi}))]}$$

If we let w represent a function of ϕ, Σ, μ_0, and μ_1, then

$$p(y=1|x;\phi,\Sigma,m_0,m_1) = \frac{1}{1+exp[-w^T x]} \ \square$$

3 (b) $\ell(\phi, m_0, m_1, \Sigma) =$

$$\sum_{i=1}^{N} log[\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}} exp[-\frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})]\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}})]$$

$$\frac{\partial}{\partial\phi}\ell(\phi,\mu_0,\mu_1,\Sigma) =$$

$$\sum_{i=1}^{N} \frac{1}{\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}} exp[-\frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})]\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}}}$$

$$\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}} exp[-\frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})]$$

$$[(1-\phi)^{1-y^{(i)}} y^{(i)} \phi^{y^{(i)}-1} - \phi^{y^{(i)}}(1-y^{(i)})(1-\phi)^{-y^{(i)}}]$$

$$= \sum_{i=1}^{N} y^{(i)}\phi^{-1} - \frac{(1-y^{(i)})(1-\phi)^{-y^{(i)}}}{(1-\phi)^{1-y^{(i)}}}$$

$$= \sum_{i=1}^{N} \frac{y^{(i)}}{\phi} - \frac{1-y^{(i)}}{1-\phi}$$

$$= \sum_{i=1}^{N} \frac{y^{(i)}(1-\phi)}{\phi(1-\phi)} - \frac{\phi - y^{(i)}\phi}{\phi(1-\phi)}$$

$$0 = \sum_{i=1}^{N} \frac{y^{(i)}}{\phi(1-\phi)} - \frac{\phi}{\phi(1-\phi)}$$

$$N\phi = \sum_{i=1}^{N} y^{(i)}$$

$$\phi = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = 1\} \ \square$$

$$\frac{\partial}{\partial \mu_0}\ell(\phi, \mu_0, \mu_1, \Sigma) =$$

$$\sum_{i=1}^{N} \frac{1}{\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}} exp[-\frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})]\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}}}$$

$$\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}} exp[-\frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})]\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}}$$

$$\frac{\partial}{\partial \mu_{y_{(i)}}} - \frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})$$

$$= \frac{\partial}{\partial \mu_{y_{(i)}}} \sum_{i=1}^{N} -\frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})$$

$$= \frac{\partial}{\partial \mu_{y_{(i)}}} \sum_{i=1}^{N} -\frac{1}{2}[{x^{(i)}}^T \Sigma^{-1} x^{(i)} - {x^{(i)}}^T \Sigma^{-1} \mu_{y^{(i)}} - \mu_{y_{(i)}}^T \Sigma^{-1} x^{(i)} + \mu_{y_{(i)}}^T \Sigma^{-1} \mu_{y_{(i)}}]$$

$$\frac{\partial}{\partial \mu_0} = 0 = \sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = 0\}{x^{(i)}}^T \Sigma^{-1} - \mathbb{I}\{y^{(i)} = 0\}\mu_{y_{(i)}}^T \Sigma^{-1}$$

$$\sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = 0\}\mu_0^T = \sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = 0\}{x^{(i)}}^T$$

$$\mu_0 = \frac{\sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = 0\}} \square$$

Similarly,

$$\mu_1 = \frac{\sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = 1\}} \square$$

3 (c) $\ell(\phi, m_0, m_1, \Sigma) =$

$$\sum_{i=1}^{N} log[\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}} exp[-\frac{1}{2}(x^{(i)}-\mu_{y_{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y_{(i)}})]\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}})]$$

$$\frac{\partial}{\partial\Sigma}\ell(\phi, \mu_0, \mu_1, \Sigma) =$$

$$\sum_{i=1}^{N} \frac{1}{\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}} exp[-\frac{1}{2}(x^{(i)}-\mu_{y_{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y_{(i)}})]\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}}}$$

$$[\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}} exp[-\frac{1}{2}(x^{(i)}-\mu_{y_{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y_{(i)}})]\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}}\frac{\partial}{\partial\Sigma}-\frac{1}{2}(x^{(i)}-\mu_{y_{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y_{(i)}})$$

$$+exp[-\frac{1}{2}(x^{(i)}-\mu_{y_{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y_{(i)}})]\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}}\frac{\partial}{\partial\Sigma}\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}]$$

$$= \sum_{i=1}^{N} \frac{\partial}{\partial\Sigma} - \frac{1}{2}(x^{(i)}-\mu_{y_{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y_{(i)}}) + \frac{1}{\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}}\frac{\partial}{\partial\Sigma}\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}$$

$$= \sum_{i=1}^{N} \frac{\partial}{\partial\Sigma} - \frac{1}{2}x^{(i)T}\Sigma^{-1}x^{(i)} + x^{(i)T}\Sigma^{-1}\mu_{y_{(i)}} - \frac{1}{2}\mu_{y_{(i)}}^T\Sigma^{-1}\mu_{y_{(i)}} + \frac{1}{\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}}\frac{\partial}{\partial\Sigma}\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}$$

$$= \sum_{i=1}^{N} \frac{1}{2}x^{(i)T}\Sigma^{-2}x^{(i)} - x^{(i)T}\Sigma^{-2}\mu_{y_{(i)}} + \frac{1}{2}\mu_{y_{(i)}}^T\Sigma^{-2}\mu_{y_{(i)}} + \frac{1}{\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}}\frac{-1}{(2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}})^2}\frac{\pi^{\frac{M}{2}}}{\Sigma^{\frac{1}{2}}}$$

$$= \sum_{i=1}^{N} \frac{1}{2}x^{(i)T}\Sigma^{-2}x^{(i)} - x^{(i)T}\Sigma^{-2}\mu_{y_{(i)}} + \frac{1}{2}\mu_{y_{(i)}}^T\Sigma^{-2}\mu_{y_{(i)}} - \frac{1}{2\Sigma}$$

$$N\Sigma = \sum_{i=1}^{N} x^{(i)T}x^{(i)} - 2x^{(i)T}\mu_{y_{(i)}} + \mu_{y_{(i)}}^T\mu_{y_{(i)}}$$

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)}-\mu_{y_{(i)}})^2 \ \square$$

3 (d) We can reuse much of the proof in **3 (c)** that did not require that $M = 1$. $\ell(\phi, m_0, m_1, \Sigma) =$

$$\sum_{i=1}^{N} log[\frac{1}{2\pi^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}} exp[-\frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})]\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}})]$$

Here, however, we can isolate $|\Sigma|$:

$$\sum_{i=1}^{N} log(\frac{1}{|\Sigma|^{\frac{1}{2}}}) + log[\frac{1}{2\pi^{\frac{M}{2}}} exp[-\frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})]\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}})]$$

$$= \sum_{i=1}^{N} log(1) - \frac{1}{2}log(|\Sigma|) + log[\frac{1}{2\pi^{\frac{M}{2}}} exp[-\frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})]\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}})]$$

Now, we can differentiate as before, but express $\nabla_\Sigma log|\Sigma|$ as Σ^{-1}.

$$\frac{\partial}{\partial\Sigma}\ell(\phi, \mu_0, \mu_1, \Sigma) = \sum_{i=1}^{N} -\frac{1}{2}\Sigma^{-1} + \frac{1}{\frac{1}{2\pi^{\frac{M}{2}}} exp[-\frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})]\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}}}$$

$$[\frac{1}{2\pi^{\frac{M}{2}}} exp[-\frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})]\phi^{y^{(i)}}(1-\phi)^{1-y^{(i}} \frac{\partial}{\partial\Sigma} - \frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})]$$

$$= \sum_{i=1}^{N} -\frac{1}{2}\Sigma^{-1} + \frac{\partial}{\partial\Sigma} - \frac{1}{2}(x^{(i)} - \mu_{y_{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y_{(i)}})$$

$$= \sum_{i=1}^{N} -\frac{1}{2}\Sigma^{-1} + \frac{\partial}{\partial\Sigma} - \frac{1}{2}x^{(i)T}\Sigma^{-1}x^{(i)} + x^{(i)T}\Sigma^{-1}\mu_{y_{(i)}} - \frac{1}{2}\mu_{y_{(i)}}^T\Sigma^{-1}\mu_{y_{(i)}}$$

$$0 = \sum_{i=1}^{N} -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}x^{(i)}x^{(i)T}\Sigma^{-1} - \Sigma^{-1}x^{(i)}\mu_{y_{(i)}}^T\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\mu_{y_{(i)}}\mu_{y_{(i)}}^T\Sigma^{-1}$$

$$N\Sigma^{-1} = \sum_{i=1}^{N} \Sigma^{-1}x^{(i)}x^{(i)T}\Sigma^{-1} - \Sigma^{-1}2x^{(i)}\mu_{y_{(i)}}^T\Sigma^{-1} + \Sigma^{-1}\mu_{y_{(i)}}\mu_{y_{(i)}}^T\Sigma^{-1}$$

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}x^{(i)T} - 2x^{(i)}\mu_{y_{(i)}}^T + \mu_{y_{(i)}}\mu_{y_{(i)}}^T$$

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N} (x^{(i)} - \mu_{y_{(i)}})(x^{(i)} - \mu_{y_{(i)}})^T \; \square$$

4 (a) We can start with the joint log-likelihood expression from lecture, modifying it to apply to $i = 1, ..., K$ classes, as expressed for data points $\{(x^{(n)}, y^{(n)}); n = 1, ..., N\}$:

$$\sum_{n}^{y^{(n)}=i} log[P(x^{(n)}, y^{(n)})] = \sum_{n}^{y^{(n)}=i} log(\phi_{y^{(n)}}) + \sum_{j=1}^{M} log(\mu_j^{y^{(n)}})$$

If we make this an MAP estimate by incorporating our Dirichlet distribution prior, we get

$$\sum_{n}^{y^{(n)}=i} log(\phi_{y^{(n)}}) + \sum_{j=1}^{M} log(\mu_j^{y^{(n)}}) + \sum_{k=1}^{K} log[\frac{1}{Z}\prod_{j=1}^{M}(\mu_j^k)^\alpha]$$

$$= \sum_{n}^{y^{(n)}=i} log(\phi_{y^{(n)}}) + \sum_{j=1}^{M} log(\mu_j^{y^{(n)}}) + K\log[\frac{1}{Z}] + \sum_{k=1}^{K}\alpha\sum_{j=1}^{M} log(\mu_j^k)$$

$$= \sum_{i=1}^{K} N^i log(\phi_i) + \sum_{j=1}^{M} N_j^i log(\mu_j^i) + \log[\frac{1}{Z}] + \alpha\sum_{j=1}^{M} log(\mu_j^i)$$

Since we know that $\sum_i^K \phi_i = 1$, $\sum_{i'}^{i'\neq i} \phi_{i'} = 1 - \phi_i$ for any $i \in \{1, ..., K\}$. So

$$\frac{\partial}{\partial\phi_i} = N^i\frac{1}{\phi_i} - N^{C_{i'}}\frac{1}{1-\phi_i} = 0$$

$$N^{C_{i'}}\frac{1}{1-\phi_i} = N^i\frac{1}{\phi_i}$$

$$\phi_i N^{C_{i'}} = N^i - N^i\phi_i$$

$$\phi_i = \frac{N^i}{N^{i'} + N^i} = \frac{N^i}{\sum_{i'}^{i'\in K} N^{i'}} \square$$

Similarly,

$$\sum_{j=1}^{M} N_j^i log(\mu_j^i) = \sum_{j=1}^{M-1} N_j^i log(\mu_j^i) + N_M^i log(1 - \sum_{j=1}^{M-1} log(\mu_j^i))$$

so we can write our posterior probability expression as

$$\sum_{i=1}^{K} N^i log(\phi_i) + \sum_{j=1}^{M-1} N_j^i log(\mu_j^i) + N_M^i log(1 - \sum_{j=1}^{M-1} log(\mu_j^i))$$

$$+ \ \log[\frac{1}{Z}] + \alpha \sum_{j=1}^{M-1} log(\mu_j^i) + \alpha \ log(1 - \sum_{j=1}^{M-1} log(\mu_j^i))$$

$$\frac{\partial l}{\partial \mu_j^i} = \frac{N_j^i}{\mu_j^i} - \frac{N_M^i}{1 - \sum_{j=1}^{M-1} \mu_j^i} + \frac{\alpha}{\mu_j^i} - \frac{\alpha}{1 - \sum_{j=1}^{M-1} \mu_j^i}$$

$$\frac{N_j^i + \alpha}{\mu_j^i} - \frac{N_M^i + \alpha}{1 - \sum_{j=1}^{M-1} \mu_j^i} = 0$$

Similarly to what we saw in lecture, this means that $\frac{N_j^i + \alpha}{\mu_j^i}$ must be a constant $\forall j$. Let this constant be A. Since $\sum_{j=1}^{M} \mu_j^i = 1$ and $\mu_j^i = \frac{N_j^i + \alpha}{A}$,

$$\sum_{j=1}^{M} \frac{N_j^i + \alpha}{A} = 1$$

$$A = \sum_{j=1}^{M} (N_j^i + \alpha) = \sum_{j=1}^{M} N_j^i + M\alpha$$

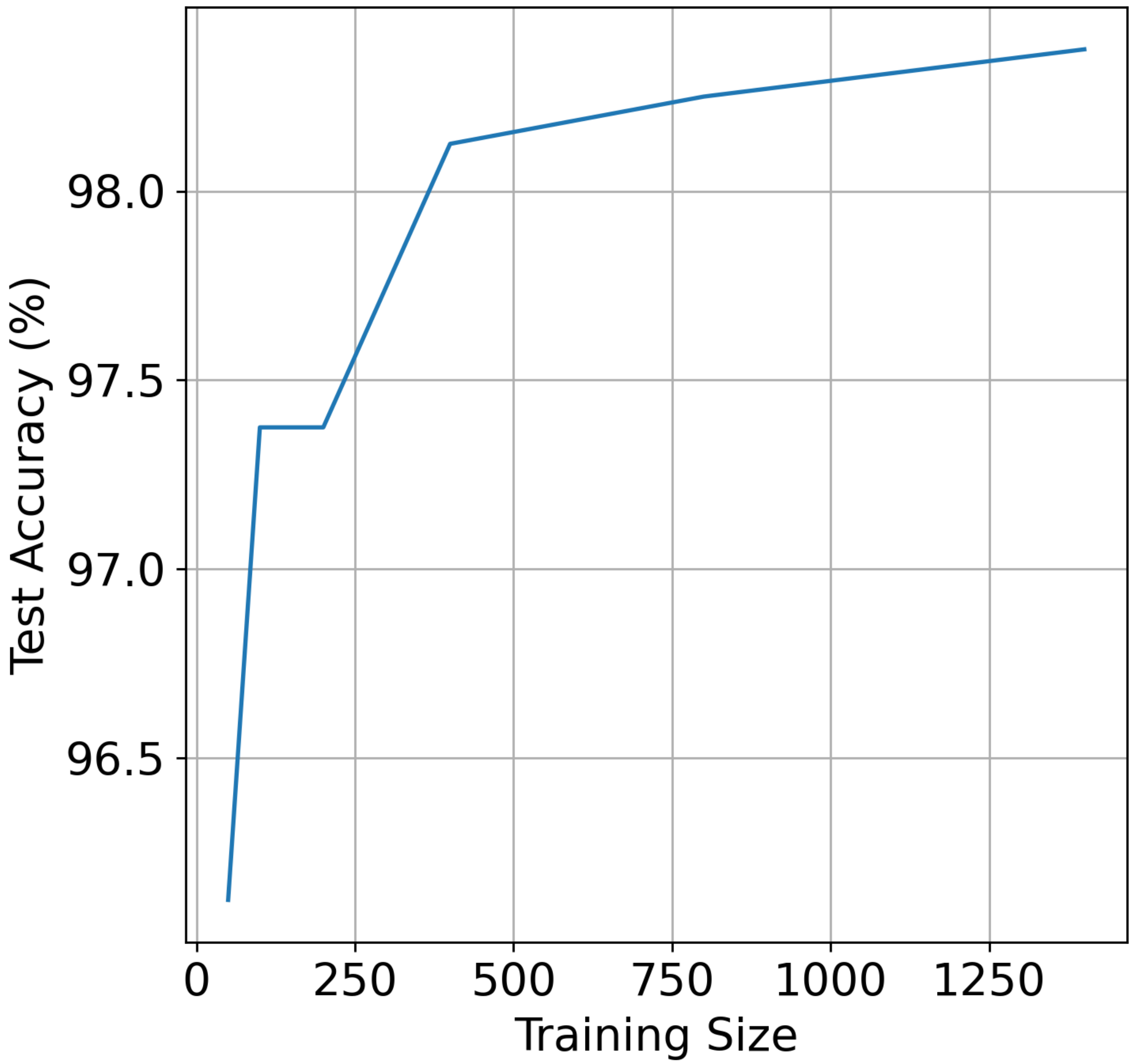
$$\frac{N_j^i + \alpha}{\mu_j^i} = \sum_{j=1}^{M} N_j^i + M\alpha$$

$$\mu_j^i = \frac{N_j^i + \alpha}{\sum_{j=1}^{M} N_j^i + M\alpha} \ \square$$

4 (b) ii. According to the metric given in the homework, the top 5 most signicant works for classifying an email as spam are 'httpaddr', 'spam', 'unsubscrib', 'ebai', and 'valet'.

4 (b) iii. Accuracy for 50 mail data (data/q4_data/MATRIX.TRAIN.50): 96.1250%
Accuracy for 100 mail data (data/q4_data/MATRIX.TRAIN.100): 97.3750%
Accuracy for 200 mail data (data/q4_data/MATRIX.TRAIN.200): 97.3750%
Accuracy for 400 mail data (data/q4_data/MATRIX.TRAIN.400): 98.1250%
Accuracy for 800 mail data (data/q4_data/MATRIX.TRAIN.800): 98.2500%
Accuracy for 1400 mail data (data/q4_data/MATRIX.TRAIN.1400): 98.3750%.

4 (b) iv.



4 (b) v. It seems that the best classification accuracy comes from the largest training set with 1400 data points. This makes sense: in general, overfitting is not as much of a worry with naive Bayes as with other machine learning algorithms.