Problem 1

(a) The log-likelihood function is

$$\ell(\theta) = \sum_{i=1}^{m} \left[y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) \right], \tag{1}$$

where

$$h(x^{(i)}) = \frac{1}{1 + e^{-\theta_\mu x_\mu^{(i)}}}. \tag{2}$$

Note that I have introduced a compact summation notation over features

$$\theta^T x^{(i)} \equiv \sum_{\mu=0}^{n} \theta_\mu x_\mu^{(i)} \equiv \theta_\mu x_\mu. \tag{3}$$

From here, and throughout the rest of this problem set I will stick to the convention that features are labelled by Greek letters, with implicit summations over repeated greek letter indices. I find this to considerably simplify much of the algebra involved.
A matrix element of the Hessian, $H_{\alpha\beta}$ of this function is given by

$$H_{\alpha\beta} = \frac{\partial^2 \ell(\theta)}{\partial\theta_\alpha \partial\theta_\beta} \tag{4}$$

We can then take successive partial derivatives as follows:

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial\theta_\alpha} &= \frac{\partial}{\partial\theta_\alpha} \sum_{i=1}^{m} \left[-y^{(i)} \log\left(1 + e^{-\theta_\mu x_\mu^{(i)}}\right) - (1 - y^{(i)}) \log\left(1 + e^{\theta_\mu x_\mu}\right) \right] \\
&= \sum_{i=1}^{m} \left[y^{(i)} \frac{x_\alpha^{(i)} e^{-\theta_\mu x_\mu^{(i)}}}{1 + e^{-\theta_\mu x_\mu^{(i)}}} - (1 - y^{(i)}) \frac{x_\alpha^{(i)} e^{\theta_\mu x_\mu^{(i)}}}{1 + e^{\theta_\mu x_\mu^{(i)}}} \right] \\
&= \sum_{i=1}^{m} \left[y^{(i)} \frac{x_\alpha^{(i)}}{1 + e^{\theta_\mu x_\mu^{(i)}}} - (1 - y^{(i)}) \frac{x_\alpha^{(i)}}{1 + e^{-\theta_\mu x_\mu^{(i)}}} \right] \\
\implies \frac{\partial^2 \ell(\theta)}{\partial\theta_\beta \theta_\alpha} &= \sum_{i=1}^{m} \left[-y^{(i)} \frac{x_\alpha^{(i)} x_\beta^{(i)} e^{\theta_\mu x_\mu^{(i)}}}{(1 + e^{\theta_\mu x_\mu^{(i)}})^2} - (1 - y^{(i)}) \frac{x_\alpha^{(i)} x_\beta^{(i)} e^{-\theta_\mu x_\mu^{(i)}}}{(1 + e^{-\theta_\mu x_\mu^{(i)}})^2} \right] \\
&= \sum_{i=1}^{m} \left[-y^{(i)} \frac{x_\alpha^{(i)} x_\beta^{(i)}}{(e^{-(\theta_\mu x_\mu)/2} + e^{(\theta_\mu x_\mu^{(i)})/2})^2} - (1 - y^{(i)}) \frac{x_\alpha^{(i)} x_\beta^{(i)}}{(e^{-(\theta_\mu x_\mu)/2} + e^{(\theta_\mu x_\mu^{(i)})/2})^2} \right] \\
&= -\sum_{i=1}^{m} \frac{x_\alpha^{(i)} x_\beta^{(i)}}{4\cosh^2\left(\frac{\theta_\mu x_\mu}{2}\right)}
\end{aligned}$$

Thus we find that a matrix element of the Hessian is

$$H_{\alpha\beta} = -\sum_{i=1}^{m} \frac{x_\alpha^{(i)} x_\beta^{(i)}}{4\cosh^2\left(\frac{\theta^T x}{2}\right)} \tag{5}$$

Now the proof that this is non-positive is easy. As described in the problem, we must first show that for a matrix Λ which can be written as the outer (tensor) product of identical vectors x, i.e. $\Lambda_{\alpha\beta} = x_\alpha x_\beta$, then we have the property

$$z^T \Lambda z = \sum_\alpha \sum_\beta z_\alpha \Lambda_{\alpha\beta} z_\beta = \sum_{\alpha,\beta} z_\alpha x_\alpha x_\beta z_\beta = \sum_\alpha (z_\alpha x_\alpha) \sum_\beta (x_\beta z_\beta) = (x^T z)^2 \geq 0, \tag{6}$$

i.e. $z^T \Lambda z$ is non negative. For the case of the Hessian we have just defined, the Hessian is now the Kronecker product of two vectors $\tilde{x}^{(i)}$ whose α'th component is $x_\alpha^{(i)}/2\cosh(\theta^T x/2)$. Thus we find

$$z^T H z = -\sum_{i=1}^{m} \sum_{\alpha,\beta} z_\alpha \frac{x_\alpha^{(i)} x_\beta^{(i)}}{4\cosh^2\left(\frac{\theta^T x}{2}\right)} z_\beta = -\sum_{i=1}^{m} \left[\left(\tilde{x}^{(i)}\right)^T z \right]^2 \leq 0 \tag{7}$$

i.e. $z^T H z$ is the negative the sum of the squares of real numbers and is guaranteed to be non-positive.
(b) The code for Newton's method is attached. The coefficients resulting from my fit are

$$\theta_0 = -2.6205$$
$$\theta_1 = 0.7604$$
$$\theta_2 = 1.1719$$

(c) The training data and decision boundary is shown in Fig. 1

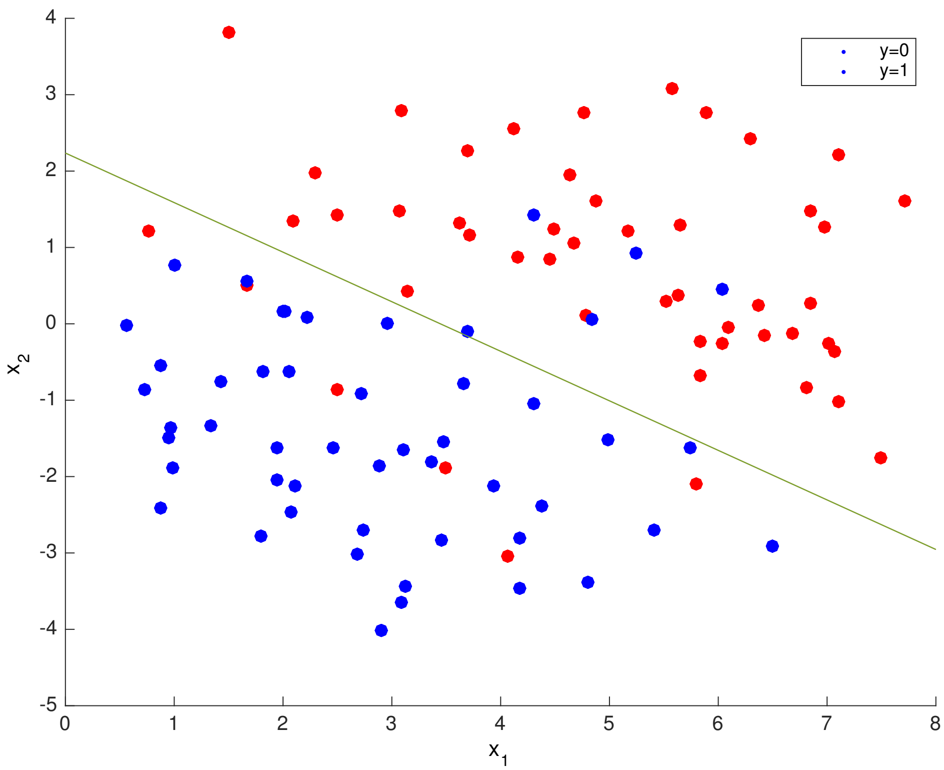

Figure 1: The training data set. Points in blue are for $y = 0$ while points in red correspond to $y = 1$. The decision boundary is shown by the green line

Problem 2

(a) We are given the function

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m} w^{(i)} \left(\theta_\mu x_\mu^{(i)} - y^{(i)}\right)^2 \tag{8}$$

For clarity, we define the following quantities:

$$X = \begin{pmatrix} x_1^{(1)} & \dots & x_n^{(1)} \\ x_1^{(2)} & \dots & x_n^{(2)} \\ \vdots & \vdots & \vdots \\ x_1^{(m)} & \dots & x_n^{(m)} \end{pmatrix}, \qquad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix} \qquad \vec{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix} \tag{9}$$

With these definitions, we find that $X\theta - \vec{y}$ is a column vector of length m (see Problem 1 for notation $\theta_\mu x_\mu$):

$$X\theta - \vec{y} = \begin{pmatrix} \theta_\mu x_\mu^{(1)} - y^{(1)} \\ \theta_\mu x_\mu^{(2)} - y^{(2)} \\ \vdots \\ \theta_\mu x_\mu^{(m)} - y^{(m)} \end{pmatrix} \tag{10}$$

We now introduce the $m \times m$ diagonal matrix

$$W = \frac{1}{2}\begin{pmatrix} w^{(1)} & & & \\ & w^{(2)} & & \\ & & \ddots & \\ & & & w^{(m)} \end{pmatrix} \tag{11}$$

with which it should be clear that the function defined above can be written equivalently in matrix notation as

$$J(\theta) = (X\theta - \vec{y})^T \, W \, (X\theta - \vec{y}) \tag{12}$$

(b) We take the derivative with respect to a component θ_α:

$$\begin{aligned} \frac{\partial J(\theta)}{\partial \theta_\alpha} &= \frac{1}{2}\sum_{i=1}^{m} w^{(i)} x_\alpha^{(i)} \left(\theta_\mu x_\mu^{(i)} - y^{(i)}\right) = 0 \\ & \frac{1}{2}\sum_{i=1}^{m} x_\alpha^{(i)} w^{(i)} \left(x_\mu^{(i)} \theta_\mu - y^{(i)}\right) = 0 \end{aligned} \tag{13}$$

where I have reordered the terms suggestively in the second line (note that each term written above is just a number, so the ordering is irrelevant). This equation holds for each component α, and so may be written in matrix notation as

$$X^T W X\theta - X^T W \vec{y} = 0 \tag{14}$$

which is the corresponding normal equation for a locally weighted linear regression.

(c) The likelihood is

$$L(\theta) = \prod_{i=1}^{m} p\left(y^{(i)} | x^{(i)}; \theta\right) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma^{(i)}} \exp\left(-\frac{\left(y^{(i)} - \theta_\mu x_\mu^{(i)}\right)^2}{2\left(\sigma^{(i)}\right)^2}\right), \tag{15}$$

so that the log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{m} \left[\log\left(\frac{1}{\sqrt{2\pi}\sigma^{(i)}} \right) - \frac{\left(y^{(i)} - \theta_\mu x_\mu^{(i)} \right)^2}{2\left(\sigma^{(i)}\right)^2} \right]. \tag{16}$$

Maximizing the log-likelihood function amounts to maximizing the function

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \frac{1}{\left(\sigma^{(i)}\right)^2} \left(y^{(i)} - \theta_\mu x_\mu^{(i)} \right)^2 \tag{17}$$

i.e. we find $w^{(i)} = 1/\left(\sigma^{(i)}\right)^2$.

(d)(i)

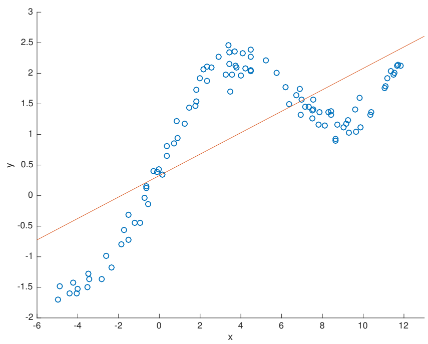

Figure 2: The result of a linear (unweighted) regression. The fitted line is $y = 0.3277 + 0.1753x$

(d)(ii)

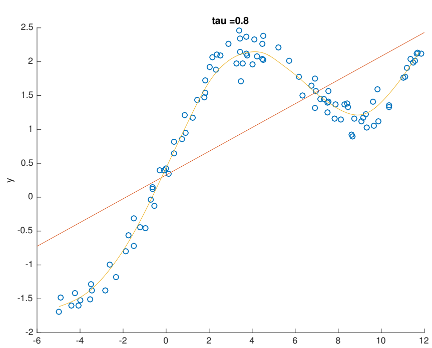

Figure 3: The result of the locally weighted regression with width parameter $\tau = 0.8$. The unweighted regression line is shown for comparison.

(d)(iii) The results of the locally weighted linear regression upon varying τ are shown in Figure 4. It is clear that too small a value of τ results in overfitting - the resulting curve is 'jittery', and probably dominated by noise. Conversely, as $\tau \to \infty$, the regression approaches the simple linear regression of d(i).

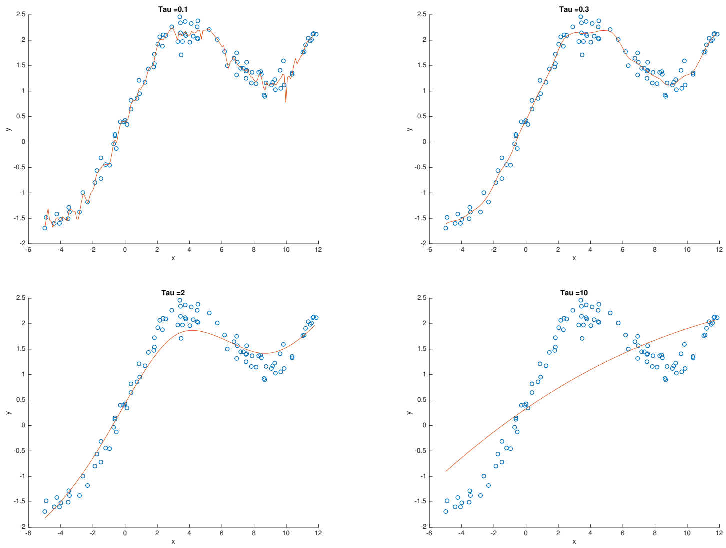

Figure 4: The result of the locally weighted regression with varying width parameters τ

Problem 3

(a) The Poisson distribution takes the form

$$\begin{aligned} p(y;\lambda) &= \frac{e^{-\lambda}\lambda^y}{y!} \\ &= \frac{1}{y!}e^{[y\log(\lambda)-\lambda]} \\ &\equiv b(y)e^{[\eta T(y)-a(\eta)]} \end{aligned} \tag{18}$$

and so is a member of the exponential family, with the identifications:

$$\begin{aligned} b(y) &= \frac{1}{y!} \\ \eta &= \log\lambda \\ T(y) &= y \\ a(\eta) &= \lambda = e^{\eta} \end{aligned} \tag{19}$$

(b) The target variable y is given by the expectation value of y given x parametrized by θ:

$$\begin{aligned} h_\theta(x) &= E[y|x;\theta] \\ &= \lambda \\ &= e^{\eta} \end{aligned}$$

Therefore the canonical response function is $g(\eta) = e^{\eta}$.

(c) The assumption of generalized linear models is that the natural parameter η and inputs x are related by

$\eta = \theta^T x = \theta_\mu x_\mu$. We therefore find the likelihood function

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{m} p\left(y^{(i)}|x^{(i)};\theta\right) && (20) \\ &= \prod_{i=1}^{m} \frac{1}{y^{(i)}!} \exp\left[\theta_\mu x_\mu^{(i)} y^{(i)} - e^{\theta_\mu x_\mu^{(i)}}\right] \\ \implies \ell(\theta) &= \sum_{i=1}^{m} \log\left(\frac{1}{y^{(i)}!}\right) + \left[\theta_\mu x_\mu^{(i)} y^{(i)} - e^{\theta_\mu x_\mu^{(i)}}\right] && (21) \end{aligned}$$

Maximizing the log-likelihood function, we must maximize

$$\begin{aligned} \tilde{J}(\theta) &= \sum_{i=1}^{m} \left(\theta_\mu x_\mu^{(i)} y^{(i)} - e^{\theta_\mu x_\mu^{(i)}}\right) && (22) \\ \implies \frac{\partial \tilde{J}(\theta)}{\partial \theta_\alpha} &= \sum_{i=1}^{m} \left[y^{(i)} - e^{\theta_\mu x_\mu^{(i)}}\right] x_\alpha^{(i)} && (23) \end{aligned}$$

The stochastic gradient ascent rule then takes the form (recall $\theta_\mu x_\mu^{(i)} \equiv \theta^T x^{(i)}$ in my notation):

for $i = 1$ *to* m **do**
  for $\beta = 0$ *to* n **do**
    $\theta_\beta := \theta_\beta + \alpha \left(y^{(i)} - e^{\theta_\mu x_\mu^{(i)}}\right) x_\beta^{(i)}$
  end
end

(d) For the general exponential family distribution with $T(y) = y$, we have

$$\begin{aligned} p(y;\eta) &= b(y) \exp\left[\eta y - a(\eta)\right] \\ \implies p(y|x;\theta) &= b(y) \exp\left[\theta_\mu x_\mu y - a(\theta_\mu x_\mu)\right] && (24) \end{aligned}$$

Following similar procedures we find the log-likelihood function

$$\ell = \sum_{i=1}^{m} \log b(y) + \left[\theta_\mu x_\mu^{(i)} y^{(i)} - a(\theta_\mu x_\mu^{(i)})\right] \tag{25}$$

which in turn means we must find θs which maximize

$$\tilde{J}(\theta) = \sum_{i=1}^{m} \left(\theta_\mu x_\mu^{(i)} y^{(i)} - a(\theta_\mu x_\mu^{(i)})\right).$$

This results in

$$\begin{aligned} \frac{\partial \tilde{J}(\theta)}{\partial \theta_\alpha} &= \sum_{i=1}^{m} \left(x_\alpha^{(i)} y^{(i)} - \frac{\partial a(\eta)}{\partial \eta} \frac{\partial \eta}{\partial \theta_\alpha}\right) \\ &= \sum_{i=1}^{m} \left(y^{(i)} - \frac{\partial a(\eta)}{\partial \eta}\right) x_\alpha^{(i)} \end{aligned}$$

and so the stochastic update rule is $\theta_\mu := \theta_\mu - \alpha \left(\frac{\partial a(\eta)}{\partial \eta}|_{\eta=\theta^T x} - y^{(i)}\right) x_\mu^{(i)}$. We must therefore show that $h(x) = \frac{\partial a(\eta)}{\partial \eta}|_{\eta=\theta^T x}$. This is simple to do; note that because of the normalization of probability distributions

Problem 3 continued on next page...

we have

$$\begin{aligned}
E[y|\eta] &= \int dyyp(y;\eta) \\
&= \int dy\, yb(y) \exp\left[\eta y - a(\eta)\right] \\
&= \int dy \left(y - \frac{\partial a}{\partial \eta} + \frac{\partial a}{\partial \eta} \right) b(y) \exp\left[\eta y - a(\eta)\right] \\
&= \frac{\partial}{\partial \eta} \left(\int dyb(y) \exp\left[\eta y - a(\eta)\right] \right) + \left(\frac{\partial a}{\partial \eta} \right) \int dyb(y) \exp\left[\eta y - a(\eta)\right] \\
&= \frac{\partial}{\partial \eta} [1] + \frac{\partial a}{\partial \eta} [1] \\
&= 0 + \frac{\partial a}{\partial \eta}
\end{aligned}$$

so that we find

$$h_\theta(x) = E[y|\eta = \theta^T x] = \frac{\partial a}{\partial \eta}|_{\eta = \theta_\mu x_\mu} \tag{26}$$

So that the stochastic gradient ascent rule is $\theta_\mu := \theta_\mu - \alpha \left(h_\theta(x) - y \right) x_\mu$.

Problem 4

(a) We use the following identity,

$$\begin{aligned} p\left(y=1|x;\phi;\Sigma;\mu_0,\mu_1\right) &= \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1)+p(x|y=0)p(y=0)} \\ &= \frac{1}{1+\frac{p(x|y=0)p(y=0)}{p(x|y=1)p(y=1)}} \end{aligned} \tag{27}$$

Substituting for the probability distributions given in the first part of the problem we find:

$$\begin{aligned} &p\left(y=1|x;\phi;\Sigma;\mu_0,\mu_1\right) \\ &= \frac{1}{1+\exp\left[-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)+\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right]\left(\frac{1-\phi}{\phi}\right)} \\ &= \frac{1}{1+\exp\left\{\frac{1}{2}\left[\mu_0^T\Sigma^{-1}x-\mu_1^T\Sigma^{-1}x+x^T\Sigma^{-1}\mu_0-x^T\Sigma^{-1}\mu_1-\mu_0^T\Sigma^{-1}\mu_0+\mu_1^T\Sigma^{-1}\mu_1+2\log\left(\frac{1-\phi}{\phi}\right)\right]\right\}} \end{aligned} \tag{28}$$

Now note that because Σ is a symmetric matrix, Σ^{-1} is also symmetric.[1] This in turn means that $(\mu_i\Sigma^{-1}x)^T = (x^T\Sigma^{-1}\mu_i) = \mu_i\Sigma^{-1}x$, where the second equality follows because these are simply scalars, and are necessarily equal. We therefore find,

$$p\left(y=1|x;\phi;\Sigma;\mu_0,\mu_1\right) = \frac{1}{1+\exp\left\{\left[\Sigma^{-1}\left(\mu_0-\mu_1\right)\right]^T x-\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0+\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1+\log\left(\frac{1-\phi}{\phi}\right)\right\}} \tag{29}$$

$$= \frac{1}{1+\exp\left[\theta_0x_0+\sum_{\alpha=1}^n\theta_\alpha x_\alpha\right]} \tag{30}$$

$$= \frac{1}{1+\exp\left[\sum_{\alpha=0}^n\theta_\alpha x_\alpha\right]} \tag{31}$$

where we have defined $x_0 = 1$, and

$$\theta_0 = -\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0+\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1+\log\left(\frac{1-\phi}{\phi}\right) \tag{32}$$

$$\theta_\alpha = \Sigma^{-1}(\mu_0-\mu_1) \quad \text{for } \alpha\neq 0 \tag{33}$$

(b) and (c) Let us prove these identities for the general scenario of n-dimensional features. The log-likelihood of the data is

$$\ell(\phi,\mu_0,\mu_1,\Sigma) = \log\prod_{i=1}^m p(x^{(i)}|y^{(i)};\mu_0;\mu_1,\Sigma)p(y^{(i)};\phi) \tag{34}$$

$$= \sum_{i=1}^m \log(p(x^{(i)}|y^{(i)};\mu_0;\mu_1,\Sigma))+\log p(y^{(i)};\phi) \tag{35}$$

$$= \sum_{i=1}^m -\frac{1}{2}\log|\Sigma|-\frac{1}{2}(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}\frac{1}{2}(x^{(i)}-\mu_{y^{(i)}})+y^{(i)}\log\phi+(1-y^{(i)})\log(1-\phi) \tag{36}$$

[1] The proof is straightforward: $\Sigma^{-1}\Sigma = I \implies \Sigma^T\left(\Sigma^{-1}\right)^T = I \implies \Sigma\left(\Sigma^{-1}\right)^T = I \implies \left(\Sigma^{-1}\right)^T = \Sigma^{-1}$.

Problem 4 continued on next page...

where in the final line we have ignored the additive constant, $-\log(2\pi)^{n/2}$. First we maximize w.r.t. ϕ:

$$\begin{aligned}
\frac{\partial \ell}{\partial \phi} &= \sum_{i=1}^{m} \frac{y^{(i)}}{\phi} - \frac{1-y^{(i)}}{1-\phi} = 0 \\
\implies \frac{(1-\phi)}{\phi} \sum_{i=1}^{m} y^{(i)} &= \sum_{i=1}^{m} (1-y^{(i)}) \\
\frac{1}{\phi} \sum_{i=1}^{m} y^{(i)} &= \sum_{i=1}^{m} (1) \\
\frac{\sum_{i=1}^{m} y^{(i)}}{\sum_{i=1}^{m} (1)} &= \phi \\
\implies \phi &= \frac{1}{m} \sum_{i=1}^{m} 1\{y^{(i)} = 1\}
\end{aligned} \tag{37}$$

Next we maximize w.r.t μ_a, for $a = 0, 1$. Writing the α'th component of μ_a as $[\mu_a]_\alpha$, and using the identity $\frac{\partial [\mu_{y^{(i)}}]_\beta}{\partial [\mu_a]_\alpha} = \delta_{\beta,\alpha} 1\{y^{(i)} = a\}$, where $\delta_{\alpha,\beta}$ is the Kronecker delta, we have

$$\begin{aligned}
\frac{\partial \ell}{\partial [\mu_a]_\alpha} &= \sum_{i=1}^{m} \frac{1}{2} \left(x^{(i)} - \mu_{y^{(i)}}\right)_\beta \Sigma^{-1}_{\beta\gamma} \delta_{\gamma,\alpha} 1\{y^{(i)} = a\} + \frac{1}{2} 1\{y^{(i)} = a\} \delta_{\alpha,\beta} \Sigma^{-1}_{\beta\gamma} \left(x^{(i)} - \mu_{y^{(i)}}\right)_\gamma = 0 \\
\implies \frac{1}{2} \sum_{i=1}^{m} 1\{y^{(i)} = a\} &\left[\left(\Sigma^{-1}_{\alpha\beta}\right)^T \left(x^{(i)} - \mu_{y^{(i)}}\right)_\beta + \Sigma^{-1}_{\alpha\gamma} \left(x^{(i)} - \mu_{y^{(i)}}\right)_\gamma \right] = 0 \\
\frac{1}{2} \sum_{i=1}^{m} 1\{y^{(i)} = a\} &\left[\left(\Sigma^{-1}_{\alpha\beta}\right) \left(x^{(i)} - \mu_{y^{(i)}}\right)_\beta + \Sigma^{-1}_{\alpha\beta} \left(x^{(i)} - \mu_{y^{(i)}}\right)_\beta \right] = 0 \\
\sum_{i=1}^{m} 1\{y^{(i)} = a\} &\left(x^{(i)} - \mu_{y^{(i)}}\right)_\beta = 0
\end{aligned}$$

where in going from the first to the second line I contracted the Kronecker -delta function by summing over repeated indices (i.e. $\gamma \to \alpha$ in first term and $\beta \to \alpha$ in the second term), while in going from the second to the third term I have replaced the dummy (summation) index γ with β in the second term, and used the fact that Σ^{-1} is a symmetric matrix. We thus find the final expression:

$$\implies \mu_a = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = a\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = a\}} \tag{38}$$

From which substituting $a = 0$ or $a = 1$ gives the desired results
For the final part of the problem, let us discuss the properties of matrix derivatives. We have

$$\frac{\partial \Sigma_{\alpha\beta}}{\partial \Sigma_{\gamma\delta}} = \delta_{\alpha,\gamma} \delta_{\beta,\delta} \tag{39}$$

Now using $\Sigma^{-1}\Sigma = 1$, we find

$$\frac{\partial \Sigma^{-1}}{\partial x} \Sigma + \Sigma^{-1} \frac{\partial \Sigma}{\partial x} = 0 \tag{40}$$

$$\implies \frac{\partial \Sigma^{-1}}{\partial x} = -\Sigma^{-1} \frac{\partial \Sigma}{\partial x} \Sigma^{-1} \tag{41}$$

So combining the above result we find

$$\frac{\partial \Sigma^{-1}_{\alpha\beta}}{\partial \Sigma_{\gamma\delta}} = -\Sigma^{-1}_{\alpha\mu} \frac{\partial \Sigma_{\mu\nu}}{\partial \Sigma_{\gamma\delta}} \Sigma^{-1}_{\nu\beta} = -\Sigma^{-1}_{\alpha\gamma} \Sigma^{-1}_{\delta\beta} \tag{42}$$

For the next identity, note that

$$\frac{\partial |\Sigma|}{\partial \Sigma} = |\Sigma|\Sigma^{-1}$$
$$\implies \frac{\partial \log |\Sigma|}{\partial \Sigma_{\gamma\delta}} = \frac{1}{|\Sigma}|\Sigma|\Sigma^{-1}_{\delta\gamma} = \Sigma^{-1}_{\delta\gamma} \tag{43}$$

Thus we can finally tackle this maximization:

$$\frac{\partial \ell}{\partial \Sigma_{\gamma\delta}} = -\frac{1}{2}\sum_{i=1}^{m} \Sigma^{-1}_{\delta\gamma} - (x^{(i)} - \mu_{y^{(i)}})_\alpha (x^{(i)} - \mu_{y^{(i)}})_\beta \Sigma^{-1}_{\alpha\gamma}\Sigma^{-1}_{\delta\beta} \tag{44}$$

Multiplying both terms by $\Sigma_{\alpha\delta}\Sigma_{\gamma\beta}$ we find

$$\sum_{i=1}^{m} \Sigma_{\alpha\beta} - (x^{(i)} - \mu_{y^{(i)}})_\alpha (x^{(i)} - \mu_{y^{(i)}})_\beta = 0$$
$$\implies \Sigma_{\alpha\beta} = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})_\alpha (x^{(i)} - \mu_{y^{(i)}})_\beta \tag{45}$$

Note the the right hand side is the exterior product of two $n+1$ dimensional vectors.

Problem 5

(a) Let me slightly change notation and say that the vector $x \to \vec{x}$ etc. Note that we have $g(\vec{z}) = f(A\vec{z}) = f(\vec{x})$. Now compare the updates in Newton's method for x vs. z. For x, we have

$$\vec{x}^{j+1} = \vec{x}^j - (\nabla_x^2 f(\vec{x}^j))^{-1} \nabla_x f(\vec{x}^j) \tag{46}$$

while for z we have

$$\vec{z}^{j+1} = \vec{z}^j - (\nabla_z^2 g(\vec{z}^j))^{-1} \nabla_z g(\vec{z}^j) \tag{47}$$

Now under the linear transformation $\vec{z} = A^{-1}x$, we have that the $i'th$ component of the gradient of g is

$$\begin{aligned}
[\nabla g(\vec{z})]_i &= \frac{\partial g(\vec{z})}{\partial z_i} \\
&= \frac{\partial g(\vec{z})}{\partial x_j} \frac{\partial x_j}{\partial z_i} \\
&= \frac{\partial f(A\vec{z})}{\partial x_j} A_{ji} \\
&= \frac{\partial f(\vec{x})}{\partial x_j} A_{ji} \\
\implies \nabla_z g(\vec{z}) &= A^T \nabla_x f(\vec{x})
\end{aligned}$$

where I am doing sums over repeated indices. Similarly, for the inverse Hessian, we have

$$\begin{aligned}
H_{ij}^{-1} &= \frac{\partial^2 g(\vec{z})}{\partial z_i \partial z_j} \\
&= \frac{\partial^2 f(A\vec{z})}{\partial x_k x_l} \frac{\partial x_k}{\partial z_i} \frac{\partial x_l}{\partial z_j} \\
&= \frac{\partial^2 f(\vec{x})}{\partial x_k x_l} A_{ki} A_{lj} \\
\implies \nabla_z^2 g(\vec{z}) &= A^T \left(\nabla_x^2 f(x) \right) A \\
\implies (\nabla_z^2 g(\vec{z}))^{-1} &= A^{-1} \left(\nabla_x^2 f(x) \right)^{-1} (A^T)^{-1}
\end{aligned}$$

With these two results we find

$$\begin{aligned}
\vec{z}^{j+1} &= \vec{z}^j - (\nabla_z^2 g(\vec{z}^j))^{-1} \nabla g(\vec{z}^j) \\
&= A^{-1}(\vec{x}^j) - A^{-1} \left(\nabla_x^2 f(x) \right)^{-1} (A^T)^{-1} A^T \nabla_x f(\vec{x}) \\
\implies \vec{z}^{j+1} &= A^{-1} \left[\vec{x}^j - (\nabla_x^2 f(\vec{x}^j))^{-1} \nabla_x f(\vec{x}^j) \right] && (48) \\
\text{i.e.} \quad \vec{z}^{j+1} &= A^{-1} \vec{x}^{j+1} && (49)
\end{aligned}$$

from Equation 46. This completes the proof - we have shown that provided we identify $\vec{z}^j = A^{-1}\vec{x}^j$, then it follows that $\vec{z}^{j+1} = A^{-1}\vec{x}^{j+1}$.

(b) It should be clear that gradient descent **is not** invariant to linear re-paramterizations. As we calculated above, the gradient term transforms like

$$\nabla_z g(\vec{z}) = A^T \nabla_x f(\vec{x}) \tag{50}$$

So under

$$\vec{x}^{j+1} = \vec{x}^j - \alpha \nabla_x f(\vec{x}^j),$$

with $\vec{z}^j = A^{-1}\vec{x}^j$ we instead get

$$\vec{z}^{j+1} = \vec{z}^j - \alpha \nabla_z g(\vec{z}^j) = A^{-1}\vec{x}^j - \alpha A^T \nabla_x f(\vec{x}^j) \neq A^{-1}\vec{x}^{j+1}$$