1.a

Since $g'(z) = g(z)(1-g(z))$ and $h(x) = g(\theta^T x)$, it follows that $\partial h(x)/\partial \theta_k = h(x)(1-h(x))x_k$. Letting $h_\theta(x^{(i)}) = g(\theta^T x^{(i)}) = 1/(1+\exp(-\theta^T x^{(i)}))$, we have

$$\begin{aligned}
\frac{\partial \log h_\theta(x^{(i)})}{\partial \theta_k} &= \frac{1}{h_\theta(x^{(i)})} \cdot h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))x_k \\
&= 1 - h_\theta(x^{(i)})x_k \\
\frac{\partial \log(1-h_\theta(x^{(i)}))}{\partial \theta_k} &= \frac{1}{1-h_\theta(x^{(i)})} \cdot -h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))x_k \\
&= -h_\theta(x^{(i)})x_k.
\end{aligned}$$

Substituting into our equation for $J(\theta)$, we have

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta_k} &= -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)}(1-h_\theta(x^{(i)}))x_k - (1-y^{(i)})h_\theta(x^{(i)})x_k\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left((h_\theta(x^{(i)}) - y^{(i)})x_k\right).
\end{aligned}$$

Consequently, the (k, l) entry of the Hessian is given by

$$\begin{aligned}
H_{kl} = \frac{\partial^2 J(\theta)}{\partial \theta_k \partial \theta_l} &= \frac{\partial}{\partial \theta_l}\frac{\partial J(\theta)}{\partial \theta_k} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))x_k x_l\right).
\end{aligned}$$

Using the fact that $X_{ij} = x_i x_j$ if and only if $X = xx^T$, we have

$$H = \frac{1}{n}\sum_{i=1}^{n}\left(h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))xx^T\right).$$

To prove that H is positive semi-definite, show $z^T H z \geq 0$ for all $z \in \mathbb{R}^d$.

$$\begin{aligned}
z^T H z &= \frac{1}{n}\sum_{i=1}^{n}\left(h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))z^T x x^T z\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))(x^T z)^2\right).
\end{aligned}$$

Since $h_\theta(x^{(i)})) \in [0, 1]$ and $(x^T z) \geq 0$, we have that $z^T H z \geq 0$ and conclude that H is PSD.

1.c

For shorthand, we let $\mathcal{H} = \{\phi, \Sigma, \mu_0, \mu_1\}$ denote the parameters for the problem. Since the given formulae are conditioned on y, use Bayes rule to get:

$$\begin{aligned}
p(y=1|x;\mathcal{H}) &= \frac{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H})}{p(x;\mathcal{H})} \\
&= \frac{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H})}{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H}) + p(x|y=0;\mathcal{H})p(y=0;\mathcal{H})}
\end{aligned}$$

Taking the equation above, we note that we can reformulate the posterior distribution to look more like the desired exponential function before substituting terms,

$$\begin{aligned}
p(y=1|x;\mathcal{H}) &= \frac{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H})}{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H}) + p(x|y=0;\mathcal{H})p(y=0;\mathcal{H})} \\
&= \frac{1}{1 + \frac{p(x|y=0;\mathcal{H})p(y=0;\mathcal{H})}{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H})}}.
\end{aligned}$$

Consider the fraction in the denominator and substitute for the parameters of the model,

$$\begin{aligned}
\frac{p(x|y=0;\mathcal{H})p(y=0;\mathcal{H})}{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H})} &= \frac{\frac{1-\phi}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))}{\frac{\phi}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))} \\
&= \left(\frac{1-\phi}{\phi}\right)\exp\left(\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right).
\end{aligned}$$

We can then take advantage of the fact that Σ^{-1} is symmetric to employ the following trick to simplify the argument of the exponential function,

$$\begin{aligned}
&= \left(\frac{1-\phi}{\phi}\right)\exp\left(\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1+\mu_0-\mu_0) - \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right) \\
&= \left(\frac{1-\phi}{\phi}\right)\exp\left(-\frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right) \\
&= \left(\frac{1-\phi}{\phi}\right)\exp\left(-(\mu_1-\mu_0)^T\Sigma^{-1}(x-\frac{(\mu_1+\mu_0)}{2})\right).
\end{aligned}$$

We desire a term of the form $\exp(-(\theta^Tx+\theta_0))$ to solve for θ and θ_0. Hence, we incorporate all terms into the exponential function and simplify,

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

This shows that the posterior distribution can be written as

$$p(y=1|x;\phi,\mu_0,\mu_1,\Sigma) = \frac{1}{1+\exp(-(\theta^Tx+\theta_0))},$$

where $\theta = (\mu_1-\mu_0)^T\Sigma^{-1}$ and $\theta_0 = -\frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0) - \log(\frac{1-\phi}{\phi})$.

1.d

First, derive the expression for the log-likelihood of the training data:

$$\begin{aligned}
\ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{n} p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi) \\
&= \sum_{i=1}^{n} \log p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^{n} \log p(y^{(i)}; \phi) \\
&= \sum_{i=1}^{n} \left(\log \left(\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \right) - \frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) + y^{(i)} \log \phi + (1 - y^{(i)}) \log(1 - \phi) \right)
\end{aligned}$$

Now, the likelihood is maximized by setting the derivative (or gradient) with respect to each of the parameters to zero.

For ϕ:

$$\frac{\partial \ell}{\partial \phi} = \sum_{i=1}^{n} \left(\frac{y^{(i)}}{\phi} - \frac{1 - y^{(i)}}{1 - \phi} \right).$$

Setting this equal to zero and solving for ϕ gives the maximum likelihood estimate.

$$\begin{aligned}
0 &= \sum_{i=1}^{n} \left(\frac{y^{(i)}}{\phi} - \frac{1 - y^{(i)}}{1 - \phi} \right) \\
&= \sum_{i=1}^{n} \left(\frac{y^{(i)} - \phi}{\phi(1 - \phi)} \right) \implies \phi = \frac{1}{n} \sum_{i=1}^{n} 1(y^{(i)} = 1).
\end{aligned}$$

For μ_0:
Hint: Remember that Σ (and thus Σ^{-1}) is symmetric.

$$\nabla_{\mu_0} \ell = \Sigma^{-1} \sum_{i=1}^{n} \left(1(y^{(i)} = 0)(x^{(i)} - \mu_0) \right)$$

Setting this gradient to zero gives the maximum likelihood estimate for μ_0.

$$\begin{aligned}
0 &= \Sigma^{-1} \sum_{i=1}^{n} \left(1(y^{(i)} = 0)(x^{(i)} - \mu_0) \right) \\
\implies \mu_0 &= \frac{\sum_{i=1}^{n} 1(y^{(i)} = 0) x^{(i)}}{\sum_{i=1}^{n} 1(y^{(i)} = 0)}.
\end{aligned}$$

For μ_1:
Hint: Remember that Σ (and thus Σ^{-1}) is symmetric.

$$\nabla_{\mu_1} \ell = \Sigma^{-1} \sum_{i=1}^{n} \left(1(y^{(i)} = 1)(x^{(i)} - \mu_1) \right)$$

Setting this gradient to zero gives the maximum likelihood estimate for μ_1.

$$\begin{aligned}
0 &= \Sigma^{-1} \sum_{i=1}^{n} \left(1(y^{(i)} = 1)(x^{(i)} - \mu_1) \right) \\
\implies \mu_1 &= \frac{\sum_{i=1}^{n} 1(y^{(i)} = 1) x^{(i)}}{\sum_{i=1}^{n} 1(y^{(i)} = 1)}.
\end{aligned}$$

For Σ, we find the gradient with respect to $S = \Sigma^{-1}$ rather than Σ just to simplify the derivation (note that $|S| = \frac{1}{|\Sigma|}$). You should convince yourself that the maximum likelihood estimate S_n found in this way would correspond to the actual maximum likelihood estimate Σ_n as $S_n^{-1} = \Sigma_n$.

Hint: You may need the following identities:

$$\nabla_S |S| = |S|(S^{-1})^T$$

$$\nabla_S b_i^T S b_i = \nabla_S tr\left(b_i^T S b_i\right) = \nabla_S tr\left(S b_i b_i^T\right) = b_i b_i^T$$

$$\begin{aligned}
\nabla_S \ell &= \nabla_S \sum_{i=1}^{n} \left(\log\left(\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\right) - \frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}}) + y^{(i)} \log\phi + (1 - y^{(i)})\log(1-\phi) \right) \\
&= \nabla_S \sum_{i=1}^{n} \frac{1}{2}\log|S| + \nabla_S \sum_{i=1}^{n} -\frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T S (x^{(i)} - \mu_{y^{(i)}}) \\
&= \frac{1}{2}\sum_{i=1}^{n}(S^{-1})^T - \frac{1}{2}\sum_{i=1}^{n} \nabla_S tr((x^{(i)} - \mu_{y^{(i)}})^T S (x^{(i)} - \mu_{y^{(i)}})) \\
&= \frac{1}{2}\sum_{i=1}^{n}(S^{-1})^T - \frac{1}{2}\sum_{i=1}^{n} \nabla_S tr(S(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T) \\
&= \frac{1}{2}\sum_{i=1}^{n}(S^{-1})^T - \frac{1}{2}\sum_{i=1}^{n}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \\
&= \frac{1}{2}\sum_{i=1}^{n}(S^{-1})^T - (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.
\end{aligned}$$

Next, substitute $\Sigma = S^{-1}$. Setting this gradient to zero gives the required maximum likelihood estimate for Σ.

$$\begin{aligned}
0 &= \frac{1}{2}\sum_{i=1}^{n}\Sigma^T - (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \\
\implies \Sigma &= \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.
\end{aligned}$$

1.f

Based on the validation set plots, it appears that outliers in the dataset have a greater influence on GDA than logistic regression. This suggests that GDA is more susceptible to error when the dataset fails to meet the method assumptions to compute θ.

1.g

Dataset 2 resulted in comparable performance between logistic regression and GDA. As we alluded to in 1(f), Gaussian Discriminant Analysis underperforms on Dataset 1 compared to logistic regression. This is likely because the data, x_1, x_2, are not roughly Gaussian in Dataset 1. Without this assumption met, GDA risks inaccuracy.

1.h

If we force the data from Dataset 1 to take on a more Gaussian distribution, this would improve the performance of GDA. We can do so according to the transformation

$$x^{(i)} \to \frac{(x^{(i)})^{\lambda} - 1}{\lambda}$$

where λ is a scalar that we can tune according to the data. I searched this transformation online to understand better ways of adding Gaussian traits to a dataset beyond using the mean and standard deviation. This transformation is known as the Box-Cox transform and seems to commonly used for this express purpose.

2.c

We employ Bayes rule to rewrite the conditional expression

$$\begin{aligned} p(t^{(i)} = 1|y^{(i)} = 1, x^{(i)}) &= \frac{p(y^{(i)} = 1|t^{(i)} = 1, x)p(t^{(i)} = 1|x^{(i)})}{p(y^{(i)} = 1|t^{(i)}, x^{(i)})} \\ &= \frac{p(y^{(i)} = 1|t^{(i)} = 1, x)p(t^{(i)} = 1|x^{(i)})}{p(y^{(i)} = 1|t^{(i)} = 1, x)p(t^{(i)} = 1|x^{(i)}) + p(y^{(i)} = 1|t^{(i)} = 0, x)p(t^{(i)} = 0|x^{(i)})}. \end{aligned}$$

We assume that $p(y^{(i)} = 1|t^{(i)} = 1, x) = \alpha$ and $p(y^{(i)} = 1|t^{(i)} = 0, x) = 0$. We can therefore reduce to arrive at the desired observation:

$$\begin{aligned} p(t^{(i)} = 1|y^{(i)} = 1, x^{(i)}) &= \frac{p(y^{(i)} = 1|t^{(i)} = 1, x)p(t^{(i)} = 1|x^{(i)})}{p(y^{(i)} = 1|t^{(i)} = 1, x)p(t^{(i)} = 1|x^{(i)}) + p(y^{(i)} = 1|t^{(i)} = 0, x)p(t^{(i)} = 0|x^{(i)})} \\ &= \frac{\alpha \cdot p(t^{(i)} = 1|x^{(i)})}{\alpha \cdot p(t^{(i)} = 1|x^{(i)}) + 0 \cdot p(t^{(i)} = 0|x^{(i)})} \\ &= 1. \end{aligned}$$

2.d

We employ a rearranged form of Bayes rule to expand the conditional probability statement and reduce using the given assumptions,

$$\begin{aligned} p(y^{(i)} = 1 | x^{(i)}) &= p(y^{(i)} = 1 | t^{(i)} = 1, x^{(i)}) p(t^{(i)} = 1 | x^{(i)}) + p(y^{(i)} = 1 | t^{(i)} = 0, x^{(i)}) p(t^{(i)} = 0 | x^{(i)}) \\ &= \alpha \cdot p(t^{(i)} = 1 | x^{(i)}) + 0 \cdot p(t^{(i)} = 0 | x^{(i)}) \\ &= \alpha \cdot p(t^{(i)} = 1 | x^{(i)}). \end{aligned}$$

Therefore, $p(t^{(i)} = 1 | x^{(i)}) = \frac{1}{\alpha} p(y^{(i)} = 1 | x^{(i)})$.

2.e

We assume that $h(x^{(i)}) = p(y^{(i)} = 1|x^{(i)})$. Using the same expansion as 2(d), we have

$$\begin{aligned} h(x^{(i)}) &= p(y^{(i)} = 1|x^{(i)}) \\ &= \alpha \cdot p(t^{(i)} = 1|x^{(i)}). \end{aligned}$$

Since we also assume that $p(t^{(i)} = 1|x^{(i)}) \in [0, 1]$, we are able to evaluate the equation above for $y^{(i)} = 1$ and $y^{(i)} = 0$. Starting with $y^{(i)} = 1$, we take the result of problem 2(c) to get

$$h(x^{(i)}|y^{(i)} = 1) = \alpha \cdot p(t^{(i)} = 1|y^{(i)} = 1, x^{(i)}) = \alpha(1) = \alpha.$$

For $y^{(i)} = 0$, we have a trivial statement since $h(x^{(i)}) = p(y^{(i)} = 1|x^{(i)}) = 0$ must be true when y is binary. The statement $h(x^{(i)}) = \alpha$ when $y^{(i)} = 1$ and $h(x^{(i)}) = 0$ when $y^{(i)} = 0$ equivalently proves that

$$\alpha = \mathbb{E}[h(x^{(i)})|y^{(i)} = 1].$$

3.ai

We consider a binary classification problem where we assume the number of positive examples is much smaller than the number of negative examples. That is, $\rho << 1-\rho$. We posit that there exists a trivial classifier with accuracy *at least* $1-\rho$ that always predicts the majority class label (i.e. the negative case given our assumption).

If the classifier always predicts negative, then the accuracy of the classifier is given by

$$\hat{A} = \frac{TN}{TN + FN}$$

since there still exists a fraction of positive examples with incorrect negative predictions. The total fraction of negative examples is given by

$$1 - \rho = \frac{TN + FP}{TP + TN + FP + FN}.$$

We seek to show that $\hat{A} \geq 1 - \rho$. Given that the classifier will always predict negative, this means that $TP = FP = 0$. A tautology therefore follows

$$\begin{aligned} \hat{A} &\geq 1 - \rho \\ \frac{TN}{TN + FN} &\geq \frac{TN + FP}{TP + TN + FP + FN} \\ \frac{TN}{TN + FN} &\geq \frac{TN}{TN + FN}. \end{aligned}$$

We conclude that the trivial classifier has accuracy at least $1 - \rho$.

3.aii

We define ρ as the total fraction of positively labeled data in our set. This includes the number of true positives and false negatives over the total number of examples in the dataset, or $\rho = \frac{TP+FN}{TP+TN+FP+FN}$. The accuracy is equal to the total number of correct predictions over the total number of examples. That is, $A = \frac{TP+TN}{TP+TN+FP+FN}$. We are also given $A_0 = \frac{TN}{TN+FP}$ and $A_1 = \frac{TP}{TP+FN}$. With these definitions in mind, we can rearrange A as a linear combination of A_0 and A_1,

$$\begin{aligned} A &= \frac{TP + TN}{TP + TN + FP + FN} \\ &= \frac{TP}{TP + TN + FP + FN} + \frac{TN}{TP + TN + FP + FN} \\ &= \frac{TP}{TP + TN + FP + FN} \cdot \frac{TP + FN}{TP + FN} + \frac{TN}{TP + TN + FP + FN} \cdot \frac{TN + FP}{TN + FP} \\ &= \frac{TP + FN}{TP + TN + FP + FN} \cdot \frac{TP}{TP + FN} + \frac{TN + FP}{TP + TN + FP + FN} \cdot \frac{TN}{TN + FP} \\ &= \rho A_1 + (1 - \rho) A_0. \end{aligned}$$

We conclude that the accuracy and balanced accuracy, $\bar{A}$ are both linear combinations of A_0 and A_1 with different weighting.

3.aiii

We return to the trivial classifier from part (i) where we stated that the model will always give a majority class, negative prediction. This naturally means that there are zero positive predictions, or equivalently $TP = FP = 0$. With respect to the balanced classifier, we substitute for A_0 and A_1 and simplify:

$$\begin{aligned}
\bar{A} &= \frac{1}{2}(A_0 + A_1) \\
&= \frac{1}{2}(\frac{TN}{TN+FP} + \frac{TP}{TP+FN}) \\
&= \frac{1}{2}(\frac{TN}{TN+0} + \frac{0}{0+FN}) \\
&= \frac{1}{2}.
\end{aligned}$$

Therefore, the trivial classifier has a balanced accuracy of 50%.

3.c

Documentation: I did not collaborate with anyone on this assignment.