This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

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

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

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

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

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

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

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

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

$$H =$$

$$\frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})(1-h(x^{(i)})))x^{(i)}x^{(i)T} =$$

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

$$z^T Hz =$$

$$z^T(\frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})(1-h(x^{(i)})))x^{(i)}x^{(i)T})z =$$

$$\frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})(1-h(x^{(i)}))z^T x^{(i)}x^{(i)T}z =$$

$$\frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})(1-h(x^{(i)}))(x^T z)^2 =$$

Now we know that
$h(x^{(i)})(1-h(x^{(i)})) \in [1,0]$
Also more generally it is clear that
$(x^T z)^2 \geqslant 0$
Hence for any vector z, it holds true that $z^T Hz \geqslant 0$

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:

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

$$
\begin{aligned}
&Let \theta = \Sigma^{-1}(\mu_1 - \mu_0) \\
&Let \theta_0 = \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1) + \log(\frac{1-\phi}{\phi}) \\
&p(y=1|x;\mathcal{H}) = \frac{1}{1+\exp(-(\theta^T x + \theta_0))}
\end{aligned}
$$

1.d

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

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

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

For ϕ:

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

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

For μ_0:

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

$$\begin{aligned}
\nabla_{\mu_0}\ell &= \\
&= -\frac{1}{2}\sum_{y^{(i)}=0} \nabla_{\mu_0}[(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)] \\
&= -\frac{1}{2}\sum_{y^{(i)}=0} \nabla_{\mu_0}[(x^{(i)T}\Sigma^{-1}x^{(i)}+\mu_0^T\Sigma^{-1}\mu_0-2\mu_0^T\Sigma^{-1}x^{(i)}] \\
&= -\frac{1}{2}\sum_{y^{(i)}=0} \nabla_{\mu_0}[\mu_0^T\Sigma^{-1}\mu_0-2\mu_0^T\Sigma^{-1}x^{(i)}] \\
&= -\frac{1}{2}\sum_{y^{(i)}=0} [2\Sigma^{-1}\mu_0-2\Sigma^{-1}x^{(i)}] = 0
\end{aligned}$$

$$0 = \sum_{y^{(i)}=0} [\Sigma^{-1}x^{(i)}-\Sigma^{-1}\mu_0]$$

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

$$\mu_0 = \frac{\sum_{i=1}^{n} 1\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)}=0\}}$$

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

For μ_1:

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

$$\nabla_{\mu_1}\ell =$$

$$= -\frac{1}{2}\sum_{y^{(i)}=1}\nabla_{\mu_1}[(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)]$$

$$= -\frac{1}{2}\sum_{y^{(i)}=1}\nabla_{\mu_1}[(x^{(i)T}\Sigma^{-1}x^{(i)}+\mu_1^T\Sigma^{-1}\mu_1-2\mu_1^T\Sigma^{-1}x^{(i)}]$$

$$= -\frac{1}{2}\sum_{y^{(i)}=1}\nabla_{\mu_1}[\mu_1^T\Sigma^{-1}\mu_1-2\mu_1^T\Sigma^{-1}x^{(i)}]$$

$$= -\frac{1}{2}\sum_{y^{(i)}=1}[2\Sigma^{-1}\mu_1-2\Sigma^{-1}x^{(i)}]=0$$

$$0=\sum_{y^{(i)}=1}[\Sigma^{-1}x^{(i)}-\Sigma^{-1}\mu_1]$$

$$\Sigma^{-1}\mu_1\sum_{i=1}^{n}1\{y^{(i)}=0\}=\Sigma^{-1}\sum_{i=1}^{n}1\{y^{(i)}=0\}x^{(i)}$$

$$\mu_1=\frac{\sum_{i=1}^{n}1\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{n}1\{y^{(i)}=1\}}$$

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

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:

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

$$\nabla_S b_i^TSb_i=\nabla_S tr\left(b_i^TSb_i\right)=\nabla_S tr\left(Sb_ib_i^T\right)=b_ib_i^T$$

$$\nabla_S \ell =$$

$$= -\frac{1}{2}\sum_{i=1}^{n} \nabla_S[-\log|S| + (x^{(i)} - \mu_{y^{(i)}})^T S(x^{(i)} - \mu_{y^{(i)}})]$$

$$= -\frac{1}{2}\sum_{i=1}^{n}[-S^{-1} + (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T]$$

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

$$0 = \sum_{i=1}^{n}\frac{1}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n}[(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T]$$

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

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

1.f

For Dataset 1
LR Accuracy obtained was 0.83
GDA Accuracy obtained was 0.81
It can be seen that Logistic regression performed better than GDA for Dataset 1
Also looking at the shape of the data points we can see that it doesn't follow Gaussian distribution and it would perform worse as demonstrated above.

1.g

(From pervious answer 1.h)
For Dataset 1
LR Accuracy obtained was 0.83
GDA Accuracy obtained was 0.81
It can be seen that Logistic regression performed better than GDA for Dataset 1
Also looking at the shape of the data points we can see that it doesn't follow Gaussian distribution and it would perform worse as demonstrated above.

(After running test case 1g-0-basic)
For Dataset 2
LR Accuracy obtained was 0.86
GDA Accuracy obtained was 0.86
It can be seen clearly that GDA performed worst than Logistic regression on Dataset1 (as against to Dataset2)
Looking at the shape of data points in Dataset2 it appears that data is more normally distributed and hence GDA did not perform worst then LR.

1.h

There can be various transformation that can be applied on Dataset1 so that GDA will perform better.
Some sample transformation to make data more gaussian like
1. Log transformation
2. Square root transformation
3. Box-Cox transformation
4. Yeo-Johnson transformation

Since data set contains negative values Yeo Johnson transformation will improve performance of GDA.

2.a

Lets start with poisson distribution parameterized by λ

$p(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$

$p(y;\lambda) = \exp(\log(\frac{e^{-\lambda}\lambda^y}{y!}))$

$p(y;\lambda) = \frac{1}{y!}\exp(y\log\lambda - \lambda)$

This can be written in exponential family as $b(y)\exp(\eta^T T(y) - a(\eta))$

So values of various terms will be

$b(y) = \frac{1}{y!}$

$T(y) = y$

$\eta = \log\lambda$

$a(\eta) = e^\eta = \lambda$

2.b

$g(\eta) = E[y; \eta]$
$g(\eta) = \lambda$
$g(\eta) = e^{\eta}$

2.c

The log-likelihood of an example $(x^{(i)}, y^{(i)})$ is defined as $\ell(\theta) = \log p(y^{(i)}|x^{(i)}; \theta)$. To derive the stochastic gradient ascent rule, use the results in part (a) and the standard GLM assumption that $\eta = \theta^T x$.

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \theta_j} &= \frac{\partial \log p(y^{(i)}|x^{(i)}; \theta)}{\partial \theta_j} \\
&= \frac{\partial \log \left(\frac{1}{y^{(i)!}} \exp(\eta^T y^{(i)} - e^{\eta}) \right)}{\partial \theta_j} \\
&= \frac{\partial \log \left(\frac{1}{y^{(i)!}} \exp((\theta^T x)^T y^{(i)} - e^{\theta^T x}) \right)}{\partial \theta_j} \\
&= \frac{\partial \log \left(\frac{1}{y^{(i)!}} \exp(x^{(i)T} \theta y^{(i)} - e^{x^{(i)T\theta}}) \right)}{\partial \theta_j} \\
&= \frac{\partial (-\log(y^{(i)!}) + x^{(i)T} \theta y^{(i)} - e^{x^{(i)T\theta}})}{\partial \theta_j} \\
&= x^{(i)T} y^{(i)} - x^{(i)T} e^{x^{(i)T\theta}} \\
&= (y^{(i)} - e^{\theta^T x^{(i)}}) x_j^{(i)}
\end{aligned}$$

Thus the stochastic gradient ascent update rule should be:

$$\theta_j := \theta_j + \alpha \frac{\partial \ell(\theta)}{\partial \theta_j},$$

which reduces here to: $\theta_j := \theta_j + \alpha (y^{(i)} - e^{\theta^T x^{(i)}}) x_j^{(i)}$