(1.a)

$$\frac{\partial J}{\partial \theta} = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\frac{1}{h_\theta(x^{(i)})}h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))(-x^{(i)})$$

$$+ (1-y^{(i)})\frac{1}{1-h_\theta(x^{(i)})}h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))x^{(i)}$$

$$= \frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))x^{(i)}$$

$$H = \partial\frac{\partial J}{\partial \theta} = \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))x^{(i)}x^{(i)T}$$

To prove H is positive semidefinite, we show $z^T H z \geq 0$ for all z

$$\begin{aligned}
z^T H z &= \frac{1}{m} z^T \left(\sum_{i=1}^{m} h\left(x^{(i)}\right)\left(1 - h\left(x^{(i)}\right)\right) x^{(i)} x^{(i)T} \right) z \\
&= \frac{1}{m}\sum_{i=1}^{m} h\left(x^{(i)}\right)\left(1 - h\left(x^{(i)}\right)\right) z^T x^{(i)} x^{(i)T} z \\
&= \frac{1}{m}\sum_{i=1}^{m} h\left(x^{(i)}\right)\left(1 - h\left(x^{(i)}\right)\right) \left(z^T x^{(i)}\right)^2 \geq 0
\end{aligned}$$

(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:

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

Now, we expand and rearrange the difference of quadratic terms in the preceding expression, finding that

$$
\begin{aligned}
&(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)-(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\\
&=x^T\Sigma^{-1}x-\mu_0^T\Sigma^{-1}x-x^T\Sigma^{-1}\mu_0+\mu_0^T\Sigma^{-1}\mu_0-x^T\Sigma^{-1}x+\mu_1^T\Sigma^{-1}x+x^T\Sigma^{-1}\mu_1-\mu_1^T\Sigma^{-1}\mu_1\\
&=-2\mu_0^T\Sigma^{-1}x+\mu_0^T\Sigma^{-1}\mu_0+2\mu_1^T\Sigma^{-1}x-\mu_1^T\Sigma^{-1}\mu_1\\
&=2(\mu_1-\mu_0)^T\Sigma^{-1}x+\mu_0^T\Sigma^{-1}\mu_0-\mu_1^T\Sigma^{-1}\mu_1
\end{aligned}
$$

Thus, we have

$$
p(y=1\mid x;\mathcal{H})=\frac{1}{1+\exp\left(\log\frac{1-\phi}{\phi}+\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1-\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0+(\mu_0-\mu_1)^T\Sigma^{-1}x\right)}
$$

and setting

$$
\theta=\Sigma^{-1}(\mu_1-\mu_0)\text{ and }\theta_0=\frac{1}{2}\left(\mu_0^T\Sigma^{-1}\mu_0-\mu_1^T\Sigma^{-1}\mu_1\right)-\log\frac{1-\phi}{\phi}
$$

gives that

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

(1.d)

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

$$
\begin{aligned}
\frac{\partial\ell}{\partial\phi}&=\sum_{i=1}^m\left[\frac{y^{(i)}}{\phi}-\frac{1-y^{(i)}}{1-\phi}\right]\\
&=\frac{\sum_{i=1}^m 1\left\{y^{(i)}=1\right\}}{\phi}-\frac{m-\sum_{i=1}^m 1\left\{y^{(i)}=1\right\}}{1-\phi}
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{\mu_0} \ell &= -\frac{1}{2} \sum_{i:y^{(i)}=-1} \nabla_{\mu_0} \left(x^{(i)} - \mu_0\right)^T \Sigma^{-1} \left(x^{(i)} - \mu_0\right) \\
&= -\frac{1}{2} \sum_{i:y^{(i)}=-1} \nabla_{\mu_0} \left[\mu_0^T \Sigma^{-1} \mu_0 - x^{(i)^T} \Sigma^{-1} \mu_0 - \mu_0^T \Sigma^{-1} x^{(i)}\right] \\
&= -\frac{1}{2} \sum_{i:y^{(i)}=-1} \nabla_{\mu_0} \operatorname{tr}\left[\mu_0^T \Sigma^{-1} \mu_0 - x^{(i)^T} \Sigma^{-1} \mu_0 - \mu_0^T \Sigma^{-1} x^{(i)}\right] \\
&= -\frac{1}{2} \sum_{i:y^{(i)}=-1} \left[2\Sigma^{-1} \mu_0 - 2\Sigma^{-1} x^{(i)}\right]
\end{aligned}
$$

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_m found in this way would correspond to the actual maximum likelihood estimate Σ_m as $S_m^{-1} = \Sigma_m$

$$
\begin{aligned}
\nabla_S \ell &= \sum_{i=1}^{m} \nabla_S [\frac{1}{2} \log |S| - \frac{1}{2} \underbrace{\left(x^{(i)} - \mu_{y^{(i)}}\right)^T}_{b_i^T} S \underbrace{(x^{(i)} - \mu_{y^{(i)}})}_{b_i}] \\
&= \sum_{i=1}^{m} \left[\frac{1}{2|S|} \nabla_S |S| - \frac{1}{2} \nabla_S b_i^T S b_i\right]
\end{aligned}
$$

But, we have the following identities:

$$
\begin{aligned}
\nabla_S |S| &= |S| \left(S^{-1}\right)^T \\
\nabla_S b_i^T S b_i &= \nabla_S \operatorname{tr}\left(b_i^T S b_i\right) = \nabla_S \operatorname{tr}\left(S b_i b_i^T\right) = b_i b_i^T
\end{aligned}
$$

In the above, we again used matrix calculus identities, and also the commutatitivity of the trace operator for square matrices. Putting these into the original equation, we get:

$$
\begin{aligned}
\nabla_S \ell &= \sum_{i=1}^{m} \left[\frac{1}{2} S^{-1} - \frac{1}{2} b_i b_i^T\right] \\
&= \frac{1}{2} \sum_{i=1}^{m} \left[\Sigma - b_i b_i^T\right]
\end{aligned}
$$

(1.g)

The dataset is more concentrated and unseparable.

(1.h)

Use logarithmic function to preprocess the data.

(2.a)

Because:

$$p\left(y^{(i)}=1 \mid t^{(i)}=1, x^{(i)}\right)=p\left(y^{(i)}=1 \mid t^{(i)}=1\right)$$

then:

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

So:

$$\begin{gathered}
p\left(t^{(i)}=1 | x^{(i)}\right)=\frac{p\left(y^{(i)}=1 | x^{(i)}\right)}{p\left(y^{(i)}=1 | t^{(i)}=1\right)}=\frac{p\left(y^{(i)}=1 | x^{(i)}\right)}{\alpha} \\
\alpha=p\left(y^{(i)}=1 | t^{(i)}=1\right)
\end{gathered}$$

(2.b)

$$h(x^i) \approx p\left(y^{(i)}=1 | x^{(i)}\right)=p\left(y^{(i)}=1 | t^{(i)}=1\right) p\left(t^{(i)}=1 | x^{(i)}\right)=\alpha$$

(3.a)

Rewrite the distribution function as:

$$\begin{aligned}
p(y ; \lambda) &= \frac{e^{-\lambda} e^{y \log \lambda}}{y!} \\
&= \frac{1}{y!} \exp (y \log \lambda-\lambda)
\end{aligned}$$

Comparing with the standard form for the exponential family:

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

(3.b)

The canonical response function for the GLM model will be:

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

(3.c)

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

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

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:

$$
\theta_j:=\theta_j+\alpha\left(y^{(i)}-e^{\theta^T x}\right) x_j^{(i)}
$$

(4.a)

$$
\begin{aligned}
\frac{\partial}{\partial \eta} \int p(y ; \eta) d y &= \int \frac{\partial}{\partial \eta} p(y ; \eta) d y \\
&= \int b(y)\left(y-a'(\eta)\right) exp(\eta y-a(\eta)) d y \\
&= \int y b(y) exp(\eta y-a(\eta)) d y-\int b(y) a'(\eta) exp(\eta y-a(\eta)) d y \\
&= \mathbb{E}_y(y \mid \eta)-a'(\eta) \int b(y) exp(\eta y-a(\eta)) d y \\
&= \mathbb{E}_y(y \mid \eta)-a'(\eta)=0
\end{aligned}
$$

(4.b)

$$
\begin{aligned}
\frac{\partial}{\partial^2\eta}\int p(y;\eta)dy &= \int \frac{\partial}{\partial^2\eta}p(y;\eta)dy \\
&= \int \frac{\partial}{\partial\eta}b(y)(y-a'(\eta))exp(\eta y - a(\eta))dy \\
&= \int \frac{\partial}{\partial\eta}yb(y)exp(\eta y - a(\eta))dy - \int \frac{\partial}{\partial\eta}b(y)a'(\eta)exp(\eta y - a(\eta))dy \\
&= \mathbb{E}_y[y^2] - a'(\eta)\mathbb{E}[y] - \frac{\partial}{\partial\eta}\int b(y)a'(\eta)exp(\eta y - a(\eta))dy \\
&= \mathbb{E}[y^2] - \mathbb{E}[y]^2 - a''(\eta) = 0
\end{aligned}
$$

so : $var[y|\eta] = a''(\eta)$

(4.c)

$$
NLL = -\log b(y)exp(\eta y - a(\eta)) = -\log b(y) - (\eta y - a(\eta))
$$

And $\eta = \theta^T x$:

$$
\begin{aligned}
\frac{\partial NLL}{\partial\theta} &= \frac{\partial a(\eta)}{\partial\theta} - xy \\
&= \frac{\partial a(\eta)}{\partial\eta}\frac{\partial\eta}{\partial\theta} - xy \\
&= a'(\eta)x - xy
\end{aligned}
$$

Then:

$$
\begin{aligned}
\frac{\partial NLL}{\partial^2\theta} &= \frac{\partial a'(\eta)}{\partial\eta}\frac{\partial\eta}{\partial\theta} = x^2 a''(\eta) \\
&= x^2[\mathbb{E}[y^2] - \mathbb{E}[y]^2] = x^2\sigma^2 \geq 0
\end{aligned}
$$

(5.a)(i)

Let $W_{ii} = \frac{1}{2}w^{(i)}, W_{ij} = 0$ for $i \neq j$, let $\vec{z} = X\theta - \vec{y}$, i.e. $z_i = \theta^T x^{(i)} - y^{(i)}$ Then we have:

$$
\begin{aligned}
(X\theta - \vec{y})^T W (X\theta - \vec{y}) &= \vec{z}^T W \vec{z} \\
&= \frac{1}{2}\sum_{i=1}^{m} w^{(i)} z_i^2 \\
&= \frac{1}{2}\sum_{i=1}^{m} w^{(i)}\left(\theta^T x^{(i)} - y^{(i)}\right)^2 \\
&= J(\theta)
\end{aligned}
$$

(5.a)(ii)

$$
\nabla_\theta J(\theta) = \nabla_\theta\left(\theta^T X^T W X\theta + \vec{y}^T W \vec{y} - 2\vec{y}^T W X\theta\right) = 2\left(X^T W X\theta - X^T W \vec{y}\right)
$$

so we have $\nabla_\theta J(\theta) = 0$ if and only if

$$X^T W X \theta = X^T W \vec{y}$$

These are the normal equations, from which we can get a closed form formula for θ

$$\theta = \left(X^T W X\right)^{-1} X^T W$$

(5.a)(iii)

$$\begin{aligned}
\arg\max_\theta \prod_{i=1}^{m} p\left(y^{(i)} \mid x^{(i)}; \theta\right) &= \arg\max_\theta \sum_{i=1}^{m} \log p\left(y^{(i)} \mid x^{(i)}; \theta\right) \\
&= \arg\max_\theta \sum_{i=1}^{m} \left(\log \frac{1}{\sqrt{2\pi}\sigma^{(i)}} - \frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\left(\sigma^{(i)}\right)^2} \right) \\
&= \arg\max_\theta - \sum_{i=1}^{m} \frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\left(\sigma^{(i)}\right)^2} \\
&= \arg\min_\theta \frac{1}{2} \sum_{i=1}^{m} \frac{1}{\left(\sigma^{(i)}\right)^2} \left(y^{(i)} - \theta^T x^{(i)}\right)^2 \\
&= \arg\min_\theta \frac{1}{2} \sum_{i=1}^{m} w^{(i)} \left(y^{(i)} - \theta^T x^{(i)}\right)^2
\end{aligned}$$

where in the last step, we substituted: $w^{(i)} = \frac{1}{\left(\sigma^{(i)}\right)^2}$ to get the linear regression form.