Question 1

Part A

The logistic loss function for labels {-1,1} is the same as the logistic function for labels {0,1}. The proof is shown in [3] and summarized below.

We start with the given logistic loss equation:

$$P(y|\theta, x) = \frac{1}{1 + e^{-y\theta^T x}} \tag{1}$$

Given that the logistic function is:

$$P(y = 1|\theta, x) = \frac{1}{1 + e^{-\theta^T x}} \tag{2}$$

It can be shown that given Eq(2), we can show that:

$$P(-x) = 1 - P(x) \tag{3}$$

This means that:

$$P(y = 0|\theta, x) = \frac{e^{-\theta^T x}}{1 + e^{-\theta^T x}} \tag{4}$$

The equations are equivalent for $y = 1$. For $y = 0$, it can be shown that $P(y = 0|\theta, x)$ and $P(y = -1|\theta, x)$ are equivalent through Property 3. We could also show that this leads to the same decision boundary. We define the boundary for logistic regression as:

$$\begin{gathered} \frac{\frac{1}{1+e^{-\theta^T x}}}{\frac{e^{-\theta^T x}}{1+e^{-\theta^T x}}} > 1 \rightarrow y = 1 \\ \theta^T x > 0 \end{gathered} \tag{5}$$

Likewise the boundary for logistic loss can be written as:

$$\begin{gathered} \frac{\frac{1}{1+e^{-\theta^T x}}}{\frac{1}{1+e^{\theta^T x}}} > 1 \rightarrow y = 1 \\ \theta^T x > 0 \end{gathered} \tag{6}$$

Given the logistical loss function 1, we would try to maximize the probability of a vector of observed results y with the likelihood function:

$$\begin{aligned} L(\theta) &= p(\vec{y}|\theta; X) \\ &= \prod_{i=1}^{m} p(y^{(i)}|\theta; x^{(i)}) \\ &= \prod_{i=1}^{m} \frac{1}{1+e^{-y\theta^T x}} \end{aligned} \tag{7}$$

Taking the log likelihood and maximizing, we get the likelihood equation as:

$$l(\theta) = -\sum_{i=1}^{m} log(1+e^{-y\theta^T x}) \tag{8}$$

This is a similar form to the one provided in the question. Maximizing the log likelihood is also minimizing the loss function $\sum_{i=1}^{m} log(1+e^{-y\theta^T x})$, which can be found in [3].

Getting back to the question at hand, to show that the Hessian H is positive semidefinite, we differentiate $J(\theta)$ twice:

$$\begin{aligned} J(\theta) &= -\frac{1}{m}\sum_{i=1}^{m} log(1+e^{-y\theta^T x}) \\ &= \frac{1}{m}\sum_{i=1}^{m} log(g(y^{(i)}\theta^T x^{(i)}) \\ g(z) &= \frac{1}{1+e^{-z}} \\ g(y\theta^T x) &= \frac{1}{1+e^{-y^{(k)}\theta^T x^{(k)}}} \end{aligned} \tag{9}$$

$$
\begin{aligned}
\frac{\partial J'(\theta)}{\partial \theta_j} &= -\frac{1}{m}\sum_{i=1}^{m}\frac{1}{g(y^{(k)}\theta^T x^{(k)})}g'(y^{(k)}\theta^T x^{(k)})y^{(k)}x_j^{(k)} \\
&= -\frac{1}{m}\sum_{i=1}^{m}(1-g(y^{(k)}\theta^T x^{(k)}))y_j^{(k)}x_j^{(k)} \\
&= -\frac{1}{m}\sum_{i=1}^{m}y^{(k)}x_j^{(k)} - g(y^{(k)}\theta^T x^{(k)})y^{(k)}x_j^{(k)} \\
\frac{\partial J''(\theta)}{\partial \theta_j \partial \theta_i} &= -\frac{1}{m}\sum_{i=1}^{m} -y^{(k)}x_j^{(k)}g(y^{(k)}\theta^T x^{(k)})y^{(k)}x_i^{(k)} \\
&= \frac{1}{m}\sum_{i=1}^{m}y^2 x_i x_j[g(y\theta^T x)(1-g(y\theta^T x))]
\end{aligned} \tag{10}
$$

Note in the last equation, we dropped the index k for clarity. Further since $z^T H z$ can be re-expressed as:

$$
\begin{aligned}
z^T H z &= \sum_i \sum_j (x_{ij} z_j) z_i \\
&= \sum_i \sum_j z_i[\frac{1}{m}\sum^{m} y^2 g(y\theta^T x)(1-g(y\theta^T x))]x_i x_j z_j
\end{aligned} \tag{11}
$$

Note that $\frac{1}{m}\sum^m y^2 g(y\theta^T x)(1-g(y\theta^T x))$ is always greater than zero since $y^2 \geq 0$ and $0 \leq g(y\theta^T x) \leq 1$. Therefore, we only need to prove that $\sum_i \sum_j z_i x_i x_j z_j > 0$. We can do this by:

$$
\begin{aligned}
\sum_i \sum_j z_i x_i x_j z_j &= \sum_i x_i z_i \sum_j x_j z_j \\
&= (x^T z)^2 \geq 0
\end{aligned} \tag{12}
$$

But why does proving semi-definite-ness prove convexity? We begin with the definition of convexity.

Definition 1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if its domain is a convex set and for all x_1, x_2 in its domain, and all $\lambda \in [0, 1]$, we have:

$$
f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \tag{13}
$$

This equation says that an equation f is convex in the range $[x_1, x_2]$ if you can draw a line between $f(x_1)$ and $f(x_2)$ such that the function f will always be smaller than that. A more detailed graphical interpretation can be found in [4].

We first prove a lemma.
Lemma 1: $f(x_1) \geq f(x_2) + \nabla f(x_2)^T(x_1 - x_2)$, then f is convex [1].
We let $z = \lambda x_1 + (1-\lambda)x_2$.

$$f(z) - f(x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2) - f(x_2) = \lambda f(x_1) - \lambda f(x_2) \tag{14}$$

We know that the gradient $\nabla f(x_2)^T(x_1 - x_2)$ can be re-expressed as [1]

$$\begin{aligned} \nabla f(x_2)^T(x_1 - x_2) &= \lim_{\lambda \to 0^+} \frac{f(x_2 + \lambda(x_1 - x_2)) - f(x)}{\lambda} \\ &= \lim_{\lambda \to 0^+} \frac{f(z) - f(x)}{\lambda} \leq f(x_2) - f(x_1) \end{aligned} \tag{15}$$

This implies that:

$$f(x_2) \geq f(z) + \nabla f(z)^T(x_2 - z) \tag{16}$$
$$f(x_1) \geq f(z) + \nabla f(z)^T(x_1 - z) \tag{17}$$
$$\tag{18}$$

Which if we multiply the first equation by λ and second by $(1-\lambda)$ and sum both equations, given that we had let $z = \lambda x_1 + (1-\lambda)x_2$, we would arrive at the definition of convexity in **Definition 1**.

Using **Lemma 1**, we can take the taylor expansion of $f(x_2)$:

$$\begin{aligned} f(x_2) = f(x_1) + \nabla f(x_1)^T(x_2 - x_1) + \frac{1}{2}((x_2 - x_1)^T H(z)(y - x)) \\ \implies f(x_2) \geq f(x_1) + \nabla f(x_1)^T(x_2 - x_1) \\ \text{if H is positive semidefinite} \end{aligned} \tag{19}$$

By extension of **Lemma 1**, semi-definiteness therefore proves convexity.

Part B

The co-efficients of the fit are [0.76037154 1.17194674 -2.6205116].

Several interesting caveats were discovered through mistakes. First, initially no parameter for y-intercept was learned. This produced pretty good results, but was 5% worse in accuracy than parameters that included a y-intercept (-2.6205116 above).

[1] Here $\nabla f(x_2)^T(x_1 - x_2)$ is a first-order directional derivative with expansion shown in [2]. It also shows why first/second-order expansion can be substituted by $z \in [x_1, x_2]$

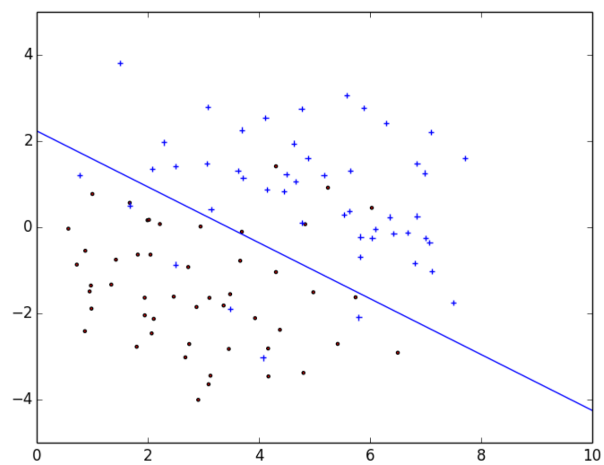

Figure 1: '+' indicate {1} labels and '.' indcate {-1} labels.

Second, a programming mistake in calculating gradient $\frac{\partial J'(\theta)}{\partial \theta_j}$ resulted in labels that were suppose to be -1 to be assigned as 1 and vice versa. Interestingly this produced the same boundary. The mistake was calculating $g(y^{(k)}\theta^T x)y_j^{(k)}x_j^{(k)}$ instead of the correct $(1 - g(y^{(k)}\theta^T x))y_j^{(k)}x_j^{(k)}$. At first, I thought this was because I may have assigned the labels to probability wrong, so I simply reversed the mappings to threshold. However, after working through the math a bit more, I realized I did the mapping to label correct, but somehow I was getting 12% accuracy when I assign $h(\theta^T x) > 0.5 y \rightarrow 1$. Upon closer inspection, the error was found.

Part C

See Figure 1.

Question 2

Part A

$$\begin{aligned}
p(y;\lambda) &= \frac{e^{-\lambda}\lambda^{y}}{y!} \\
&= \frac{1}{y!}e^{ylog\lambda-\lambda} \\
\\
T(y) &= y \\
\eta &= log\lambda \\
a(\eta) &= \lambda \\
b(y) &= \frac{1}{y!}
\end{aligned} \tag{20}$$

Part B

$$\begin{aligned}
g(\eta) = E[T(y);\eta] = E[y;\eta] \\
E[y;\lambda] = \lambda \text{ and } \eta = log\lambda \implies g(\eta) = E[y;e^{\eta}] = e^{\eta}
\end{aligned} \tag{21}$$

Part C

Note: y^i and x^i are reduced to y and x for simplicity and $\eta = \theta^T x \implies \lambda = e^{\theta^T x}$.

$$\begin{aligned}
L(\theta) &= \sum_i^m logp(y|x;\theta) \\
&= \sum_i^m log\frac{e^{-e^{\theta^T x}}(e^{\theta^T x})y}{y!} \\
&= \sum_i^m -e^{\theta^T x} + y\theta^T x - logy! \\
\frac{\partial L(\theta)}{\partial \theta_j} &= \sum_i^m -x_j e^{\theta^T x} + yx_j \\
&= \sum_i^m x_j(y - e^{\theta^T x})
\end{aligned} \tag{22}$$

$\theta_j = \theta_j + \alpha\nabla_\theta l(\theta) = \theta_j + \alpha(y - e^{\theta^T x})x_j$.

Part D

$$
\begin{aligned}
l(\theta) &= logb(y) + \eta y - a(\eta) \text{ with } \eta = \theta^T x \\
\frac{\partial l(\theta)}{\partial \theta_j} &= x_j(y - \frac{\partial a(\theta^T x)}{\partial(\theta^T x)})
\end{aligned} \tag{23}
$$

Since $\theta_j = \theta_j + \alpha \nabla_\theta l(\theta) = \theta_j - \alpha(\frac{\partial a(\theta^T x)}{\partial(\theta^T x)} - y)x_j$. We only need to show that $h(x)$, the canonical response function (or plain english the function we use to predict), is $\frac{\partial a(\theta^T x)}{\partial(\theta^T x)}$.

$$
\begin{aligned}
\int_y p(y|x;\theta)dy &= 1 \\
\int_y b(y)e^{\eta^T y - a(\eta)}dy &= 1 \\
\int_y b(y)e^{\eta^T y}dy &= e^{a(\eta)} \\
\frac{\partial}{\partial \eta}\int_y b(y)e^{\eta^T y}dy &= \frac{\partial}{\partial \eta}e^{a(\eta)} \\
\int_y yb(y)e^{\eta^T y}dy &= e^{a(\eta)}\frac{\partial a(\eta)}{\partial \eta} \\
\int_y ye^{\eta^T y - a(\eta)}dy &= \frac{\partial a(\eta)}{\partial \eta} \\
\int_y yp(y|x;\theta)dy &= \frac{\partial a(\eta)}{\partial \eta} \\
E[y|x;\theta] &= \frac{\partial a(\eta)}{\partial \eta} = h(x)
\end{aligned} \tag{24}
$$

Question 3

Part A

$$
p(y|x;\phi,\Sigma,\mu_{-1},\mu_1) = \frac{p(x|y)p(y)}{\sum p(x|y)p(y)} \tag{25}
$$

$$
\begin{aligned}
p(y=1|x) &= \frac{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^1(x-\mu_1)}\phi}{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^1(x-\mu_1)}\phi + \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})}(1-\phi)} \\
p(y=1|x) &= \frac{e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^1(x-\mu_1)}\phi}{e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^1(x-\mu_1)}\phi + e^{-\frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})}(1-\phi)} \\
p(y=1|x) &= \frac{1}{1+e^{x^T\Sigma^{-1}(\mu_{-1}-\mu_1)-\frac{1}{2}(\mu_{-1}^T\Sigma^{-1}\mu_{-1}+\mu_1^T\Sigma^{-1}\mu_1)+log\frac{1-\phi}{\phi}}}
\end{aligned}
\tag{26}
$$

$$
\begin{aligned}
p(y=-1|x) &= \frac{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu_{-1})^T\Sigma^1(x-\mu_{-1})}(1-\phi)}{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^1(x-\mu_1)}\phi + \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})}(1-\phi)} \\
p(y=-1|x) &= \frac{e^{-\frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})}(1-\phi)}{e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}\phi + e^{-\frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})}(1-\phi)} \\
p(y=-1|x) &= \frac{1}{1+e^{x^T\Sigma^{-1}(\mu_1-\mu_{-1})-\frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1+\mu_{-1}^T\Sigma^{-1}\mu_{-1})+log\frac{\phi}{1-\phi}}}
\end{aligned}
\tag{27}
$$

If we let $\theta = \Sigma^{-1}(\mu_1 - \mu_{-1})$ and $\theta_0 = log(\frac{\phi}{1-\phi}) + \frac{1}{2}(\mu_{-1}^T\Sigma^{-1}\mu_{-1} - \mu_1^T\Sigma^{-1}\mu_1)$, then:

$$
\begin{aligned}
p(y=1|x) &= \frac{1}{1+e^{-(\theta^Tx+\theta_0)}} \\
p(y=-1|x) &= \frac{1}{1+e^{\theta^Tx+\theta_0}}
\end{aligned}
\tag{28}
$$

More succinctly, this can be written as $p(y|x) = \frac{1}{1+e^{-y(\theta^Tx+\theta_0)}}$.

Part B

See Part C.

Part C

$$\begin{aligned}\frac{\partial l(\phi,\mu_{-1},\mu_1,\Sigma)}{\partial\phi} = &\sum_i 1\{y=1\}log\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}+\\ &1\{y=-1\}log\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}+\\ &1\{y=1\}(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))+\\ &1\{y=-1\}(-\frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1}))+\\ &1\{y=1\}log\phi+1\{y=-1\}log(1-\phi)\end{aligned} \tag{29}$$

For ϕ:

$$\begin{aligned}\frac{\partial l(\phi,\mu_{-1},\mu_1,\Sigma)}{\partial\phi} &= \sum_i 1\{y=1\}\frac{1}{\phi}-1\{y=-1\}\frac{1}{1-\phi}\\ (1-\phi)\sum_i 1\{y=1\} &= \phi\sum_i 1\{y=-1\}\\ \phi &= \frac{1}{m}\sum_i 1\{y=1\}\end{aligned} \tag{30}$$

For μ_{-1}, μ_1, and Σ, we use the following matrix gradient equations found in the notes:

$$\nabla_A trAB = B^T \tag{31}$$

$$\nabla_{A^T} f(A) = (\nabla_A f(A))^T \tag{32}$$

$$\nabla_A trABA^TC = CAB + C^TAB^T \tag{33}$$

$$\nabla_A|A| = |A|(A^{-1})^T \tag{34}$$

For μ_1:

$$\begin{aligned}
\frac{\partial l(\mu_1, \mu_{-1}, \mu_1, \Sigma)}{\partial \mu_1} &= \sum_i 1\{y=1\} * (-\frac{1}{2}) \frac{\partial}{\partial \mu_1} tr(x-\mu_1)^T \Sigma^{-1} (x-\mu_1) \\
&= \sum_i 1\{y=1\} * (-\frac{1}{2})((x-\mu_1)^T \Sigma^{-1} + (x-\mu_1)^T \Sigma^{-T}) \frac{\partial}{\partial \mu_1} (x-\mu_1)^T \\
&= \sum_i 1\{y=1\} (x-\mu_1)^T \Sigma^{-1} \\
\sum_i 1\{y=1\} x &= \sum_i 1\{y=1\} \mu_1 \\
\mu_1 &= \frac{\sum_i 1\{y=1\} x}{\sum_i 1\{y=1\}}
\end{aligned} \tag{35}$$

For μ_{-1}:

$$\begin{aligned}
\frac{\partial l(\mu_1, \mu_{-1}, \mu_1, \Sigma)}{\partial \mu_{-1}} &= \sum_i 1\{y=-1\} * (-\frac{1}{2}) \frac{\partial}{\partial \mu_{-1}} tr(x-\mu_{-1})^T \Sigma^{-1} (x-\mu_{-1}) \\
&= \sum_i 1\{y=-1\} * (-\frac{1}{2})((x-\mu_{-1})^T \Sigma^{-1} + (x-\mu_{-1})^T \Sigma^{-T}) \frac{\partial}{\partial \mu_{-1}} (x-\mu_{-1})^T \\
&= \sum_i 1\{y=-1\} (x-\mu_{-1})^T \Sigma^{-1} \\
\sum_i 1\{y=-1\} x &= \sum_i 1\{y=-1\} \mu_{-1} \\
\mu_{-1} &= \frac{\sum_i 1\{y=-1\} x}{\sum_i 1\{y=-1\}}
\end{aligned} \tag{36}$$

For Σ:

$$\begin{aligned}
\frac{\partial l(\mu_1, \mu_{-1}, \mu_1, \Sigma)}{\partial \Sigma^{-1}} &= -\frac{1}{2} \sum_i 1\{y=1\}\Sigma + 1\{y=-1\}\Sigma + \\
&\quad 1\{y=1\}(x-\mu_1)(x-\mu_1)^T + 1\{y=-1\}(x-\mu_{-1})(x-\mu_{-1})^T \\
m\Sigma &= \sum_i (x-\mu_1)(x-\mu_1)^T + (x-\mu_{-1})(x-\mu_{-1})^T \\
\Sigma &= \frac{1}{m} \sum_i (x-\mu_y)(x-\mu_y)^T
\end{aligned} \tag{37}$$

For Σ above, we needed to use several interesting properties:

1. determinant of matrix inverse is inverse of the matrix determinant $|\Sigma| = \frac{1}{|\Sigma^{-1}|}$
2. trace of matrix cyclically permutate $tr(ABC) = tr(BCA) = tr(CAB)$
3. matrix Σ is symmetric

Question 4

Part A

Newton's method is defined as $x^{i+1} = x^i - \frac{f'(x)}{f''(x)}$. For $g(z)$:

$$\begin{aligned}
z^{i+1} &= z^i - \frac{g'(z)}{g''(z)} \\
g'(z) &= f'(Az)\frac{\partial(Az)}{\partial z} = Af'(Az) \\
g''(z) &= f''(Az) = A^2 f''(Az) = A^2 g(z) \\
z^{i+1} &= z^i - \frac{Af'(Az)}{A^2 f''(Az)}
\end{aligned} \tag{38}$$

We assume $z^i = A^{-1}x^i$ and know that the base case z^0 is true. We only need to prove that $z^{i+1} = A^{-1}x^{i+1}$. Since $\frac{f'(x)}{f''(x)} = x^i - x^{i+1}$:

$$\begin{aligned}
z^{i+1} &= z^i - (x^i - x^{i+1})A^{-1} \\
&= A^{-1}x^{i+1}
\end{aligned} \tag{39}$$

Therefore Newton's method is invariant to linear re-parameterization.

Part B

Gradient descent is defined as $x^{i+1} = x^i - \alpha f'(x^i)$. Since $x^i = Az^i$ and $f'(x^i) = \frac{x^i - x^{i+1}}{\alpha}$:

$$\begin{aligned}
z^{i+1} &= z^i - \alpha A f'(Az^i) \\
z^{i+1} &= A^{-1}x^i - Ax^i + Ax^{i+1}
\end{aligned} \tag{40}$$

Gradient descent is not invariant to linear re-parameterization.

Question 5

Part A

i)

$$
\begin{aligned}
J(\theta) &= (X\theta - y)^T W (X\theta - y) \\
&= (\begin{bmatrix} X^0 \\ X^1 \\ \vdots \\ X^i \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} - \begin{bmatrix} y^0 \\ y^1 \\ \vdots \\ yi \end{bmatrix})^T (\begin{bmatrix} \frac{1}{2}w^0 & & & \\ & \frac{1}{2}w^1 & & \\ & & \ddots & \\ & & & \frac{1}{2}w^i \end{bmatrix})(\begin{bmatrix} X^0 \\ X^1 \\ \vdots \\ X^i \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} - \begin{bmatrix} y^0 \\ y^1 \\ \vdots \\ yi \end{bmatrix}) \\
&= \frac{1}{2} \begin{bmatrix} \theta^T x^0 & \theta^T x^1 & \cdots & \theta^T x^i \end{bmatrix} \begin{bmatrix} w^0(\theta^T x^0 - y^0) \\ w^1(\theta^T x^1 - y^1) \\ \vdots \\ w^i(\theta^T x^i - y^i) \end{bmatrix} \\
&= \frac{1}{2} \sum_i (\theta^T x^i - y^i) w^i (\theta^T x^i - y^i) \\
&= \frac{1}{2} \sum_i w^i (\theta^T x^i - y^i)^2
\end{aligned} \tag{41}
$$

ii)

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \frac{1}{2} \nabla_\theta (X\theta - y)^T W (X\theta - y) \\
&= \frac{1}{2} \nabla_\theta \theta^T X^T W X \theta - \theta^T X^T W y - y^T W X \theta + y^T W y \\
&= \frac{1}{2} ((\theta^T X^T W X)^T + \theta^T X^T W^T X - (X^T W y) - y^T W X)
\end{aligned} \tag{42}
$$

We set $\nabla_\theta J(\theta) = 0$, and since W is a diagonal matrix $W^T = W$:

$$
\begin{aligned}
2X^T W X \theta &= 2X^T W y \\
\theta &= (X^T W X)^{-1} X^T W^T y
\end{aligned} \tag{43}
$$

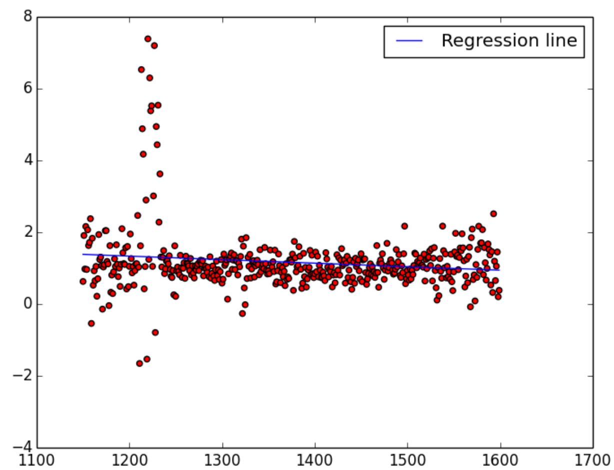

Figure 2: Linear Regression of first training sample of quasar data. $y = \theta^T x$ where $\theta = (X^T X)^{-1} X^T y$

iii)

$$
\begin{aligned}
l(\theta) &= log \prod_i \frac{1}{\sqrt{2\pi}\sigma^i} e^{-\frac{(y^i-\theta^T x^i)^2}{2(\sigma^i)^2}} \\
&= -\sum_i log\sqrt{2\pi}\sigma^i - \frac{(y^i-\theta^T x^i)^2}{2(\sigma^i)^2} \qquad (44) \\
\frac{\partial l'(\theta)}{\partial\theta} &= \frac{\partial l'(\theta)}{\partial\theta} \frac{1}{2} \sum_i \frac{1}{\sigma^{(i)2}} (\theta^T x^i - y^i)^2
\end{aligned}
$$

In this case, the problem of normal distributed samples with differing variances reduce to a weighted linear regression problem where $w^i = \frac{1}{(\sigma^i)^2}$.

Part B

i)

See Figure 2.

ii)

See Figure 3.

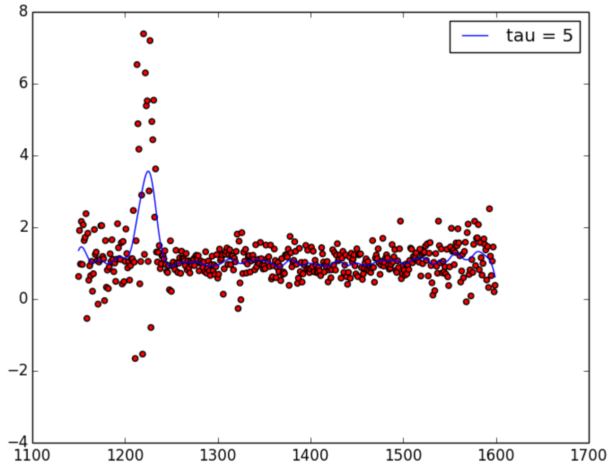

Figure 3: Weighted Linear Regression of first training sample of quasar data. $y = \theta^T(x)x$ where $\theta = (X^TWX)^{-1}X^TW^Ty$

iii)

See Figure 4. The higher the value τ, the closer the estimated curve $y = h(x)$ tracks the data points.

Part C

i)

See Q5.py.

ii)

See Q5.py. Training set error 1.0664.

iii)

See Q5.py and Figure 5 and 6. Test set error 2.7100.

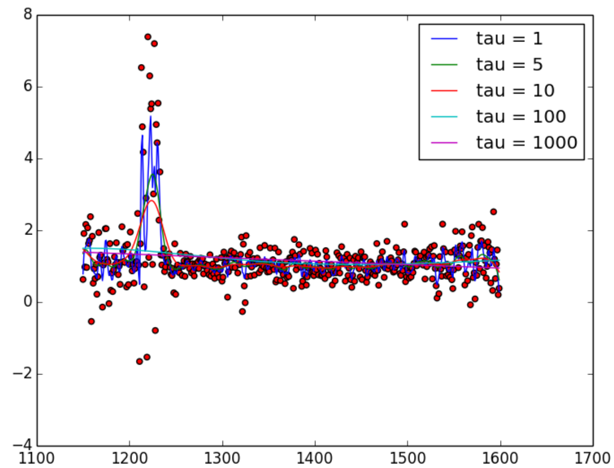

Figure 4: Weighted Linear Regression of first training sample of quasar data. $y = \theta^T(x)x$ where $\theta = (X^TWX)^{-1}X^TW^Ty$. $\tau = 1, 10, 100, 1000$.

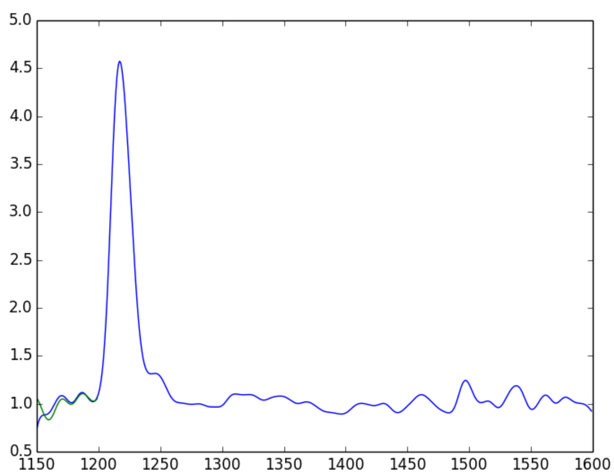

Figure 5: Weighted Linear Regression of test sample 1 of quasar data.

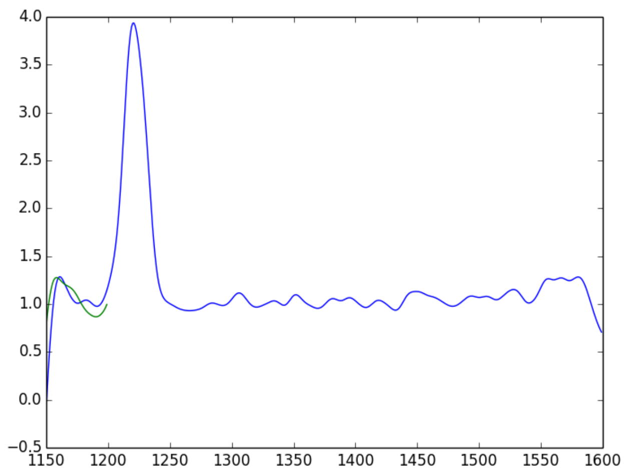

Figure 6: Weighted Linear Regression of test sample 1 of quasar data.

References

[1] Characterization of convex functions. http://people.seas.harvard.edu/ yaron/AM221/lecture_notes/AM221_lecture10.pdf. Accessed: 2018-2-08.

[2] Characterization of convex functions. http://people.seas.harvard.edu/ yaron/AM221/lecture_notes/AM221_lecture9.pdf. Accessed: 2018-2-08.

[3] Notes on logistic loss function. http://www.hongliangjie.com/wp-content/uploads/2011/10/logistic.pdf. Accessed: 2018-1-31.

[4] Theory of convex functions. http://www.princeton.edu/ amirali/Public/Teaching/ORF523/S16/ORF523_S16_Lec7_gh.pdf. Accessed: 2018-2-08.