Quiz 2

1. The cross entropy loss function is $L(w) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(g_i) + (1-y_i)\log(1-g_i)$ where $g_i = \sigma(w^T x_i) = 1/(1+e^{-w^T x_i})$ for $(x_1, y_1) \ldots (x_n, y_n)$ in $\mathbb{R}^n \times \{0,1\}$.
To show that this is convex, set $t = w^T x_i$.
$\partial/\partial t\, g_i = \partial/\partial t\, (1+e^{-t})^{-1} = e^{-t}(1+e^{-t})^{-2} = g_i(1-g_i)$
$\partial \log(g_i)/\partial w^T = 1/g_i \; \partial g_i/\partial w^T = 1/g_i \; \partial g_i/\partial t \; \partial t/\partial w^T = (1-g_i)x_i$
$\partial \log(1-g_i)/\partial w^T = 1/g_i \; \partial(1-g_i)/\partial w^T = -g_i x_i$
Summation component $l_i(w) = -y_i \log(g_i) - (1-y_i)\log(1-g_i)$
$\nabla l_i(w) = -y_i x_i (1-g_i) + (1-y_i) x_i g_i = x_i(g_i - y_i)$.
$\nabla^2 l_i(w) = x_i x_i^T g_i(1-g_i) = \frac{1}{n}\sum_{i=1}^{m} \nabla^2 l_i(w) = \frac{1}{n}\sum_{i=1}^{m} x_i x_i^T g_i(1-g_i) \equiv XDX^T$
where D is a diagonal matrix with all entries $D_{ii} = g_i(1-g_i) > 0$.
So $\nabla^2 L(w)$ is positive semidefinite and $L(w)$ is convex.
$L(w)$ is not strongly convex. So min of a convex function must achieve global min.
For gradient descent method $x_{k+1} = x_k - s_k \nabla L_{B_k}(x_k)$, $w_{k+1} = w_k - s_k \nabla L(w_k)$.
$L(w) \in C^2$, $|\lambda_i| \le M$ for all λ_i of $\nabla^2 L(w)$. Use backtracking line search to find s_k. # iterations $= O(1/\varepsilon)$ for gradient decent to reach $|L(w_k) - L(w^*)| < \varepsilon$.
Adding an l_2 regularization term we get cross-entropy loss $L(w)$ due to convexity. So converge rate becomes $O(\log(1/\varepsilon))$ for reaching $|L(w_k) - L(w^*)| < \varepsilon$.

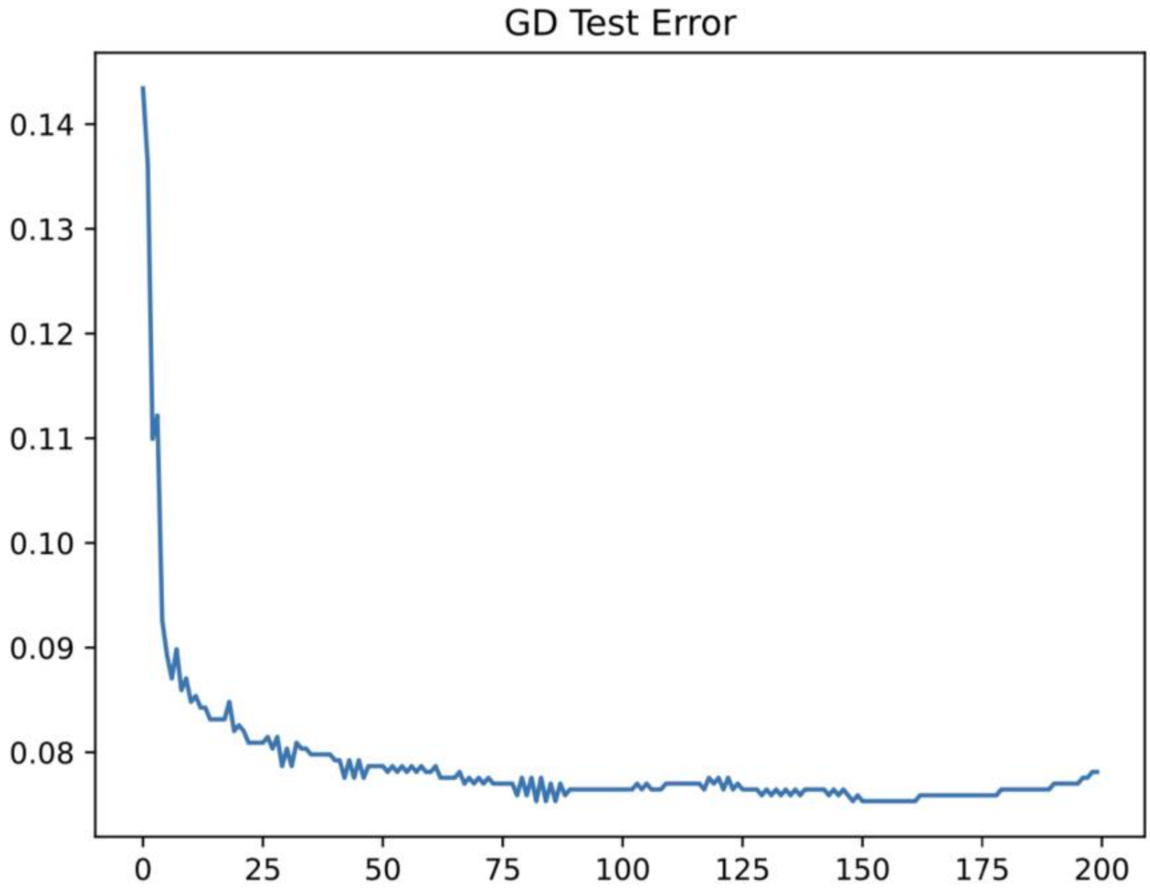
2. (b) With $L(w) = \frac{1}{n}\sum_{i=1}^{m} l_i(w)$,

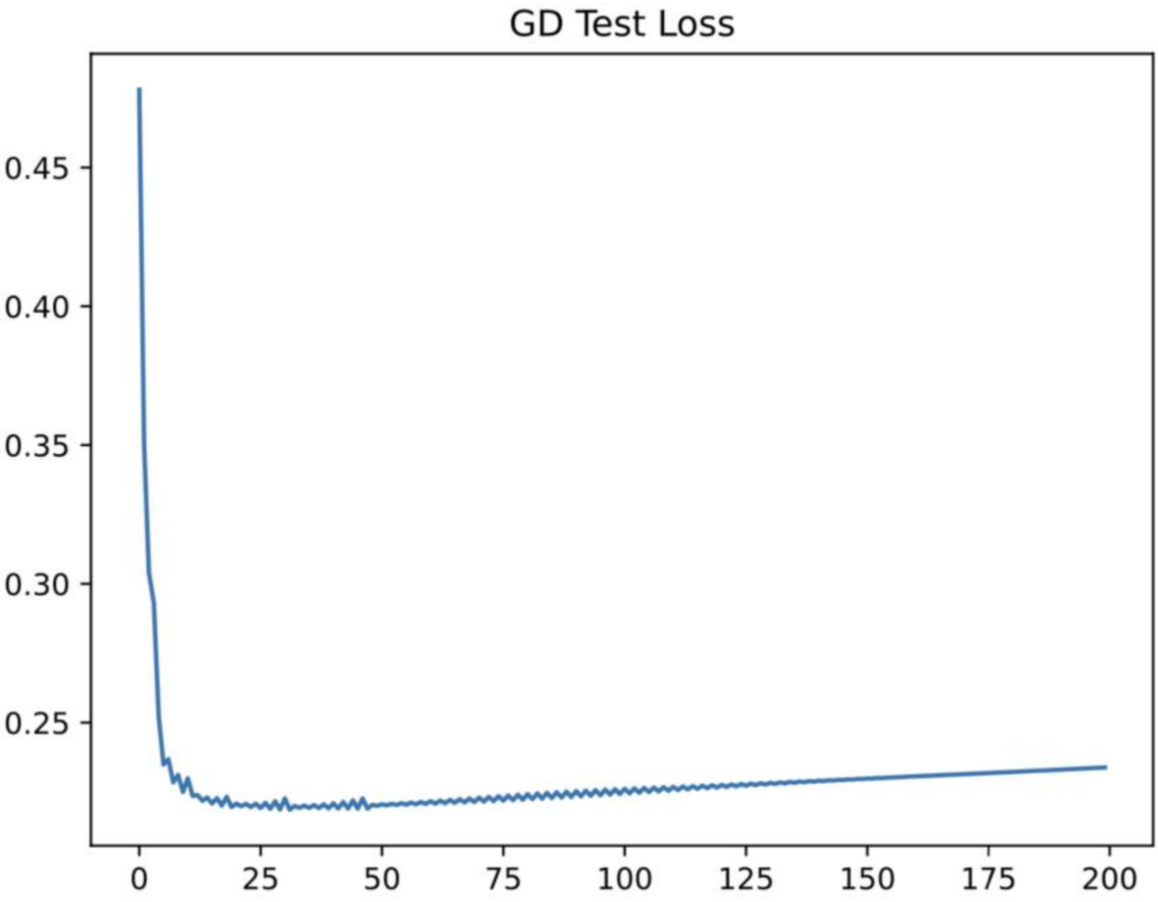
$\nabla l_i(w) = -y_i x_i (1-g_i) + (1-y_i) x_i g_i = x_i(g_i - y_i)$

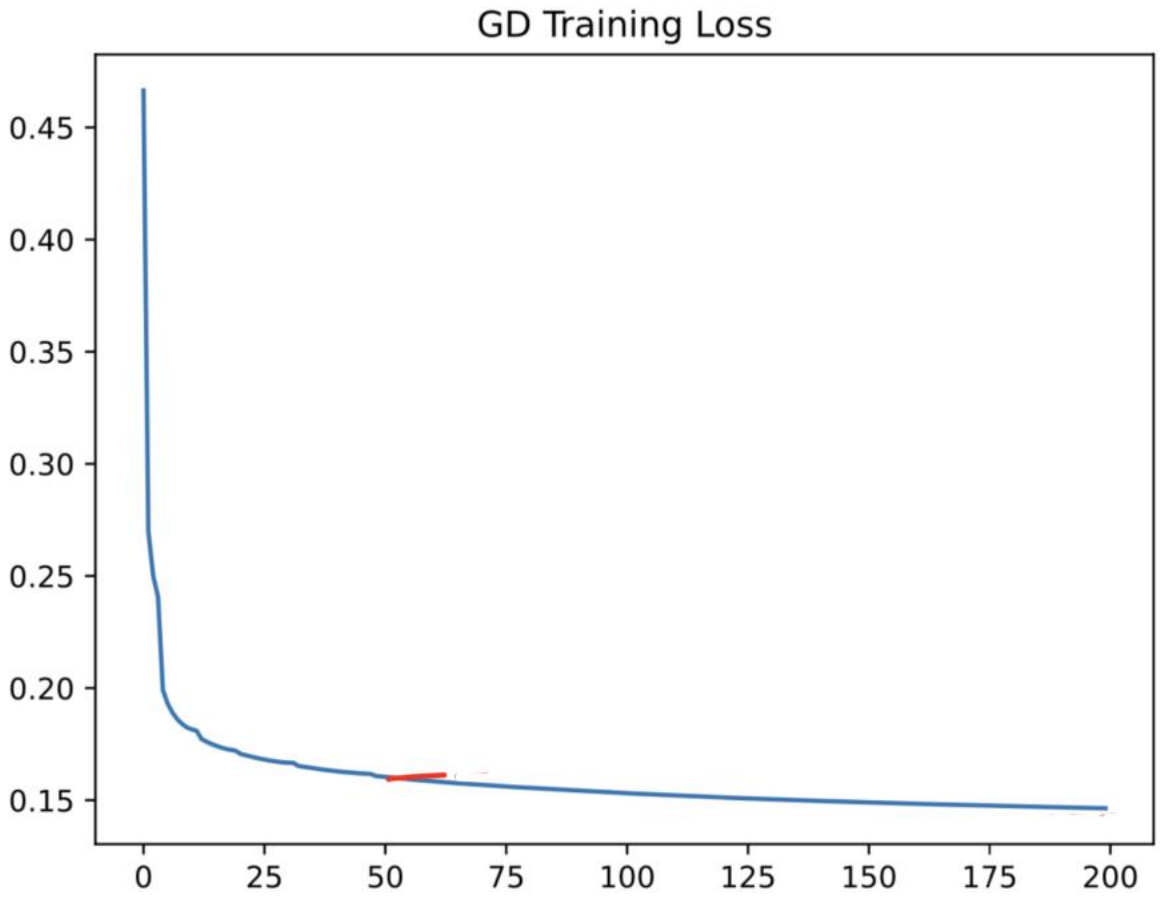
$= \frac{1}{n}\sum_{i=1}^{m} x_i(g_i - y_i)$ where $g_i = (1+e^{-w^T x_i})^{-1}$

(e) From the output data, the resulting loss becomes stable or converges at roughly 50 iterations. This is in accordance with $O(1/\epsilon)$ for $|L(w_k) - L(w^*)| < \epsilon$.

Problem 2 Outputs:






GD Training Loss
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0
25
50
75
100
125
150
175
200