ANLY 561 HW

Name:Yuqi Wang
NetID:yw545

Problem 1

Part (a)

$$\ell(\tilde{\beta}) = -\sum_{k=1}^{N} \log \text{logit}(y_i \cdot (\tilde{x}^{(i)})^T \tilde{\beta})$$

$$f(\tilde{\beta}) = X\tilde{\beta} \text{ for } X = \begin{pmatrix} (\tilde{x}^{(1)})^T \\ (\tilde{x}^{(2)})^T \\ \vdots \\ (\tilde{x}^{(N)})^T \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,n} \\ 1 & x_{2,1} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,n} \end{pmatrix}$$

$$g(\mathbf{q}) = \begin{pmatrix} -\log \text{logit}(y_1 q_1) \\ -\log \text{logit}(y_2 q_2) \\ \vdots \\ -\log \text{logit}(y_N q_N) \end{pmatrix}$$

$$s(\mathbf{v}) = \mathbf{1}_N^T \mathbf{v}$$

So

$$f(\tilde{\beta}) = \begin{pmatrix} (\tilde{x}^{(1)})^T \tilde{\beta} \\ (\tilde{x}^{(2)})^T \tilde{\beta} \\ \vdots \\ (\tilde{x}^{(N)})^T \tilde{\beta} \end{pmatrix}$$

Then

$$g(f(\tilde{\beta})) = \begin{pmatrix} -\log \text{logit}(y_1 \cdot (\tilde{x}^{(1)})^T \tilde{\beta}) \\ -\log \text{logit}(y_2 \cdot (\tilde{x}^{(2)})^T \tilde{\beta}) \\ \vdots \\ -\log \text{logit}(y_N \cdot (\tilde{x}^{(N)})^T \tilde{\beta}) \end{pmatrix}$$

Finally, we have

$$s(g(f(\tilde{\beta}))) = \mathbf{1}_N^T g(f(\tilde{\beta})) = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} -\log \text{logit}(y_1 \cdot (\tilde{x}^{(1)})^T \tilde{\beta}) \\ -\log \text{logit}(y_2 \cdot (\tilde{x}^{(2)})^T \tilde{\beta}) \\ \vdots \\ -\log \text{logit}(y_N \cdot (\tilde{x}^{(N)})^T \tilde{\beta}) \end{pmatrix}$$

Thus

$$\begin{aligned} s(g(f(\tilde{\beta}))) &= -\log \text{logit}(y_1 \cdot (\tilde{x}^{(1)})^T \tilde{\beta}) - \log \text{logit}(y_2 \cdot (\tilde{x}^{(2)})^T \tilde{\beta}) - \cdots - \log \text{logit}(y_N \cdot (\tilde{x}^{(N)})^T \tilde{\beta}) \\ &= -\sum_{k=1}^{N} \log \text{logit}(y_i \cdot (\tilde{x}^{(i)})^T \tilde{\beta}) = \ell(\tilde{\beta}) \end{aligned}$$

Part (b)

Firstly, we can get the following properties by some simple calculations,

$$\begin{aligned}
\operatorname{logit}(x) &= \frac{1}{1+e^{-x}}, \\
(\log \operatorname{logit}(x))' &= \operatorname{logit}(-x)
\end{aligned}$$

And we define

$$h(\mathbf{q}) = \begin{pmatrix} -y_1 \operatorname{logit}(-y_1 q_1) \\ -y_2 \operatorname{logit}(-y_2 q_2) \\ \vdots \\ -y_N \operatorname{logit}(-y_N q_N) \end{pmatrix}$$

From part (a), we have $\ell(\tilde{\beta}) = s(g(f(\tilde{\beta})))$, so

$$\nabla \ell(\tilde{\beta}) = \nabla(s \circ g \circ f)(\tilde{\beta}) = \nabla(s \circ (g \circ f))(\tilde{\beta})$$

Then we use multivariate chain rule,

$$\nabla(s \circ (g \circ f))(\tilde{\beta}) = D(g \circ f)(\tilde{\beta})^T \nabla s(g \circ f(\tilde{\beta}))$$

We use multivariate chain rule again to get $D(g \circ f)(\tilde{\beta}) = Dg(f(\tilde{\beta}))Df(\tilde{\beta})$, so

$$D(g \circ f)(\tilde{\beta})^T \nabla s(g \circ f(\tilde{\beta})) = Df(\tilde{\beta})^T Dg(f(\tilde{\beta}))^T \nabla s(g(f(\tilde{\beta})))$$

Since $f(\tilde{\beta}) = X\tilde{\beta}$, so

$$Df(\tilde{\beta}) = X,$$
$$Df(\tilde{\beta})^T = X^T$$

Then since $g(\mathbf{q}) = \begin{pmatrix} -\log \text{logit}(y_1 q_1) \\ -\log \text{logit}(y_2 q_2) \\ \vdots \\ -\log \text{logit}(y_N q_N) \end{pmatrix}$ and $(\log \text{logit}(x))' = \text{logit}(-x)$, so

$$Dg(\mathbf{q}) = \begin{pmatrix} -\text{logit}(-y_1 q_1) y_1 & 0 & \cdots & 0 \\ 0 & -\text{logit}(-y_2 q_2) y_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\text{logit}(-y_N q_N) y_N \end{pmatrix}$$

Thus

$$Dg(f(\tilde{\beta})) = \begin{pmatrix} -\text{logit}(-y_1 (\tilde{x}^{(1)})^T \tilde{\beta}) y_1 & 0 & \cdots & 0 \\ 0 & -\text{logit}(-y_2 (\tilde{x}^{(2)})^T \tilde{\beta}) y_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\text{logit}(-y_N (\tilde{x}^{(N)})^T \tilde{\beta}) y_N \end{pmatrix} = Dg(f(\tilde{\beta}))^T$$

Since $s(\mathbf{v}) = \mathbf{1}_N^T \mathbf{v}$, we have

$$\nabla s(\mathbf{v}) = \mathbf{1}_N$$

Also,

$$\nabla s(g(f(\tilde{\beta}))) = \mathbf{1}_N$$

Finally,

$$\begin{aligned}
\nabla \ell(\tilde{\beta}) &= Df(\tilde{\beta})^T Dg(f(\tilde{\beta}))^T \nabla s(g(f(\tilde{\beta}))) \\
&= X^T \begin{pmatrix} -\text{logit}(-y_1(\tilde{x}^{(1)})^T\tilde{\beta})y_1 & 0 & \cdots & 0 \\ 0 & -\text{logit}(-y_2(\tilde{x}^{(2)})^T\tilde{\beta})y_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\text{logit}(-y_N(\tilde{x}^{(N)})^T\tilde{\beta})y_N \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \\
&= X^T \begin{pmatrix} -y_1\text{logit}(-y_1(\tilde{x}^{(1)})^T\tilde{\beta}) \\ -y_2\text{logit}(-y_2(\tilde{x}^{(2)})^T\tilde{\beta}) \\ \vdots \\ -y_N\text{logit}(-y_N(\tilde{x}^{(N)})^T\tilde{\beta}) \end{pmatrix}
\end{aligned}$$

Since $h(\mathbf{q}) = \begin{pmatrix} -y_1\text{logit}(-y_1q_1) \\ -y_2\text{logit}(-y_2q_2) \\ \vdots \\ -y_N\text{logit}(-y_Nq_N) \end{pmatrix}$, so

$$h(f(\tilde{\beta})) = \begin{pmatrix} -y_1\text{logit}(-y_1(\tilde{x}^{(1)})^T\tilde{\beta}) \\ -y_2\text{logit}(-y_2(\tilde{x}^{(2)})^T\tilde{\beta}) \\ \vdots \\ -y_N\text{logit}(-y_N(\tilde{x}^{(N)})^T\tilde{\beta}) \end{pmatrix}$$

Therefore,

$$\nabla \ell(\tilde{\beta}) = X^T h(f(\tilde{\beta}))$$

Part (c)

Firstly, we can get the following properties by some simple calculations,

$$\mathrm{logit}(x) = \frac{1}{1+e^{-x}},$$
$$\mathrm{logit}'(x) = \mathrm{logit}(x)\mathrm{logit}(-x)$$

And we define

$$G(\mathbf{q}) = \begin{pmatrix} \mathrm{logit}(q_1)\mathrm{logit}(-q_1) \\ \mathrm{logit}(q_2)\mathrm{logit}(-q_2) \\ \vdots \\ \mathrm{logit}(q_N)\mathrm{logit}(-q_N) \end{pmatrix}$$

$$\mathrm{diag}(d) = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_N \end{pmatrix}$$

Observe that

$$\nabla^2 \ell(\tilde{\beta}) = \nabla\nabla\ell(\tilde{\beta}) = D\nabla\ell(\tilde{\beta}) = D(X^T h(f(\tilde{\beta})))$$
$$= X^T D(h \circ f)(\tilde{\beta}) = X^T D(h(f(\tilde{\beta}))D(f(\tilde{\beta})) \text{ (Using multivariate chain rule)}$$
$$= X^T D(h(f(\tilde{\beta}))X \text{ (In part b, we have showed that } D(f(\tilde{\beta})) = X)$$

Since $h(\mathbf{q}) = \begin{pmatrix} -y_1 \text{logit}(-y_1 q_1) \\ -y_2 \text{logit}(-y_2 q_2) \\ \vdots \\ -y_N \text{logit}(-y_N q_N) \end{pmatrix}$, so

$$Dh(\mathbf{q}) = \begin{pmatrix} y_1^2 \text{logit}(y_1 q_1)\text{logit}(-y_1 q_1) & 0 & \cdots & 0 \\ 0 & y_2^2 \text{logit}(y_2 q_2)\text{logit}(-y_2 q_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_N^2 \text{logit}(y_N q_N)\text{logit}(-y_N q_N) \end{pmatrix}$$

and $h(f(\tilde{\beta})) = \begin{pmatrix} -y_1 \text{logit}(-y_1 (\tilde{x}^{(1)})^T \tilde{\beta}) \\ -y_2 \text{logit}(-y_2 (\tilde{x}^{(2)})^T \tilde{\beta}) \\ \vdots \\ -y_N \text{logit}(-y_N (\tilde{x}^{(N)})^T \tilde{\beta}) \end{pmatrix}$, so

$$D(h(f(\tilde{\beta}))$$
$$= \begin{pmatrix} y_1^2 \text{logit}(y_1 (\tilde{x}^{(1)})^T \tilde{\beta})\text{logit}(-y_1 (\tilde{x}^{(1)})^T \tilde{\beta}) & 0 & \cdots & 0 \\ 0 & y_2^2 \text{logit}(y_2 (\tilde{x}^{(2)})^T \tilde{\beta})\text{logit}(-y_2 (\tilde{x}^{(2)})^T \tilde{\beta}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_N^2 \text{logit}(y_N (\tilde{x}^{(N)})^T \tilde{\beta})\text{logit}(-y_N (\tilde{x}^{(N)})^T \tilde{\beta}) \end{pmatrix}$$

Since $y_i \in \{1, -1\}$, so $y_i^2 = 1$ and $\text{logit}(y_i (\tilde{x}^{(1)})^T \tilde{\beta})\text{logit}(-y_i (\tilde{x}^{(1)})^T \tilde{\beta}) = \text{logit}((\tilde{x}^{(1)})^T \tilde{\beta})\text{logit}(-(\tilde{x}^{(1)})^T \tilde{\beta})$ for all i.

$$D(h(f(\tilde{\beta})) = \begin{pmatrix} \text{logit}((\tilde{x}^{(1)})^T\tilde{\beta})\text{logit}(-(\tilde{x}^{(1)})^T\tilde{\beta}) & 0 & \cdots & 0 \\ 0 & \text{logit}((\tilde{x}^{(2)})^T\tilde{\beta})\text{logit}(-(\tilde{x}^{(2)})^T\tilde{\beta}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \text{logit}((\tilde{x}^{(N)})^T\tilde{\beta})\text{logit}(-(\tilde{x}^{(N)})^T\tilde{\beta}) \end{pmatrix}$$

Since $G(\mathbf{q}) = \begin{pmatrix} \text{logit}(q_1)\text{logit}(-q_1) \\ \text{logit}(q_2)\text{logit}(-q_2) \\ \vdots \\ \text{logit}(q_N)\text{logit}(-q_N) \end{pmatrix}$ and $\text{diag}(d) = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_N \end{pmatrix}$, so

$$\begin{aligned} \text{diag}(G(f(\tilde{\beta}))) &= \begin{pmatrix} \text{logit}((\tilde{x}^{(1)})^T\tilde{\beta})\text{logit}(-(\tilde{x}^{(1)})^T\tilde{\beta}) & 0 & \cdots & 0 \\ 0 & \text{logit}((\tilde{x}^{(2)})^T\tilde{\beta})\text{logit}(-(\tilde{x}^{(2)})^T\tilde{\beta}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \text{logit}((\tilde{x}^{(N)})^T\tilde{\beta})\text{logit}(-(\tilde{x}^{(N)})^T\tilde{\beta}) \end{pmatrix} \\ &= D(h(f(\tilde{\beta})) \end{aligned}$$

Therefore,

$$\nabla^2 \ell(\tilde{\beta}) = X^T \text{diag}(G(f(\tilde{\beta})))X$$

Part (d)

Take a $\mathbf{v} \in \mathbb{R}^{n+1}$, then we have

$$\mathbf{v}^T \nabla^2 \ell(\tilde{\beta})\mathbf{v} = \mathbf{v}^T X^T \text{diag}(G(f(\tilde{\beta})))X\mathbf{v} = \mathbf{V}^T \text{diag}(G(f(\tilde{\beta})))\mathbf{V} \text{ where } \mathbf{V} = X\mathbf{v}$$

Since this is in a sum form like $\sum_{i=1}^N d_i V_i^2$ where d_i is the ith diagonal element in $\text{diag}(G(f(\tilde{\beta})))$ and V_i is the ith element in $\mathbf{V}$. We know that $\text{logit}(x) > 0$ for all $x \in \mathbb{R}$ and $V_i^2 \geq 0$, so $d_i V_i^2 \geq 0$ for all i. Thus, $\nabla^2 \ell(\tilde{\beta})$ is positive semidefinite.

Part (e)

From part (d), we know that $d_i > 0$ for all i. Then we have to make $V_i > 0$ for all i. It means that $V \neq 0$ if $v \neq 0$. Since $V = Xv$, we just need that the null space of X is a set contains only $\mathbf{0}$ vector. Therefore, X needs to have full rank to make sure that $\nabla^2 \ell(\tilde{\beta})$ is positive definite.

Part (f)

Part (d) and Part (e) imply the minimization program is always convex and the program is strictly convex if X has full rank. Thus, all critical points are solutions to this program. When X has full rank, there will be a unique solution to this program.