1 Problem 1

(a)

$$\begin{aligned}\frac{\partial \mathcal{L}(\vec{w})}{\partial \vec{w}} &= \frac{\partial}{\partial \vec{w}}\left[(\mathbf{X}^T\vec{w}-\vec{y})^T(\mathbf{X}^T\vec{w}-\vec{y})+\lambda\vec{w}^T\vec{w}\right] \\ &= \frac{\partial}{\partial \vec{w}}(\vec{w}^T\mathbf{X}\mathbf{X}^T\vec{w}-2\vec{w}^T\mathbf{X}\vec{y}+\vec{y}^T\vec{y}+\lambda\vec{w}^T\vec{w}) \\ &= 2\mathbf{X}\mathbf{X}^T\vec{w}-2\mathbf{X}\vec{y}+2\lambda\vec{w}\end{aligned}$$

Thus,

$$\vec{w}_{t+1} = \vec{w}_t - c(2\mathbf{X}\mathbf{X}^T\vec{w}-2\mathbf{X}\vec{y}+2\lambda\vec{w}) \qquad c>0$$

(b)

$$\begin{aligned}\frac{\partial \mathcal{L}(\vec{w})}{\partial \vec{w}} &= \frac{\partial}{\partial \vec{w}}\left[(\mathbf{X}^T\vec{w}-\vec{y})^T(\mathbf{X}^T\vec{w}-\vec{y})+\lambda\|\vec{w}\|_1\right] \\ &= 2\mathbf{X}\mathbf{X}^T\vec{w}-2\mathbf{X}\vec{y}+\lambda sign(\vec{w})\end{aligned}$$

Where $sign()$ represents a element-wise sign function here.
If $\vec{w}[i]>0$, then $sign(\vec{w})[i]=1$. If $\vec{w}[i]<0$, then $sign(\vec{w})[i]=-1$
Thus,

$$\vec{w}_{t+1} = \vec{w}_t - c(2\mathbf{X}\mathbf{X}^T\vec{w}-2\mathbf{X}\vec{y}+\lambda sign(\vec{w})) \qquad c>0$$

(c)

$$\frac{\partial \mathcal{L}(\vec{w})}{\partial \vec{w}} = \sum_{i=1}^{n}\frac{-y_i exp(-y_i\vec{w}^T\vec{x_i})\vec{x_i}}{1+exp(-y_i\vec{w}^T\vec{x_i})}$$

Thus,

$$\vec{w}_{t+1} = \vec{w}_t - c\sum_{i=1}^{n}\frac{-y_i exp(-y_i\vec{w}^T\vec{x_i})\vec{x_i}}{1+exp(-y_i\vec{w}^T\vec{x_i})}$$

(d)

$$\frac{\partial \mathcal{L}(\vec{w})}{\partial \vec{w}} = C\sum_{i=1}^{n}[\![1-y_i\vec{w}^T\vec{x_i}>0]\!](-y_i\vec{x_i})+2\vec{w}$$

where $[\![$ and $]\!]$ is the Iverson bracket:
The Iverson bracket, named after Kenneth E. Iverson, is a notation that denotes a number that is 1 if the condition in square brackets is satisfied, and 0 otherwise. resource: https://oeis.org/wiki/Iverson_bracket
Thus,

$$\vec{w}_{t+1} = \vec{w}_t - c(C\sum_{i=1}^{n}[\![1-y_i\vec{w}^T\vec{x_i}>0]\!](-y_i\vec{x_i})+2\vec{w})$$

2 Problem 2

(a) Say the old label $y_{old} \in \{-1, 1\}$ Then for Logistics Regression, the objective function should be:

$$\min_{\vec{w}} \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_{old\,i}(\vec{w}^T \vec{x_i})})$$

Let $y = \frac{y_{old}+1}{2}$, then when $y_{old} = 1, y = 1$; when $y_{old} = -1, y = 0$.

$$y_{old} = 2y - 1$$

Thus the new objective function is:

$$\min_{\vec{w}} \sum_{i=1}^{n} \log(1 + e^{-(2y_i-1)(\vec{w}^T \vec{x_i})})$$

$$\begin{aligned}
&\min_{\vec{w}} \sum_{i=1}^{n} \log(1 + e^{-(2y_i-1)(\vec{w}^T \vec{x_i})}) \\
&= \min_{\vec{w}} \sum_{i=1}^{n} \log(1 + e^{-y_i(\vec{w}^T \vec{x_i})} e^{-(y_i-1)(\vec{w}^T \vec{x_i})}) \\
&= \min_{\vec{w}} \sum_{y=1} \log(1 + e^{-\vec{w}^T \vec{x_i}}) + \sum_{y=0} \log(1 + e^{\vec{w}^T \vec{x_i}}) \\
&= \min_{\vec{w}} \sum_{y=1} \log\left(sigm(\vec{w}^T \vec{x_i})^{-1}\right) + \sum_{y=0} \log\left(sigm(-\vec{w}^T \vec{x_i})^{-1}\right) \\
&= \min_{\vec{w}} - \sum_{y=1} \log\ sigm(\vec{w}^T \vec{x_i}) - \sum_{y=0} \log\left(1 - sigm(\vec{w}^T \vec{x_i})\right) \\
&= \min_{\vec{w}} - \sum_{i=1}^{n} \left[y_i \log\left(sigm(\vec{w}^T \vec{x_i})\right) + (1 - y_i) \log\left(1 - sigm(\vec{w}^T \vec{x_i})\right)\right]
\end{aligned}$$

(b)

$$\begin{aligned}
\frac{\partial L(\vec{w})}{\partial \vec{w}} &= - \sum_{i=1}^{n} \left[y_i \frac{sigm(\vec{w}^T \vec{x_i})(1 - sigm(\vec{w}^T \vec{x_i}))}{sigm(\vec{w}^T \vec{x_i})} \vec{x_i} + (1 - y_i) \frac{-sigm(\vec{w}^T \vec{x_i})(1 - sigm(\vec{w}^T \vec{x_i}))}{1 - sigm(\vec{w}^T \vec{x_i})} \vec{x_i} \right] \\
&= - \sum_{i=1}^{n} \left[y_i \left(1 - sigm(\vec{w}^T \vec{x_i})\right)\vec{x_i} - (1 - y_i)\ sigm(\vec{w}^T \vec{x_i})\vec{x_i}\right] \\
&= - \sum_{i=1}^{n} \left[y_i - y_i\ sigm(\vec{w}^T \vec{x_i})) -\ sigm(\vec{w}^T \vec{x_i}) + y_i\ sigm(\vec{w}^T \vec{x_i})\right] \vec{x_i} \\
&= - \sum_{i=1}^{n} \left[y_i -\ sigm(\vec{w}^T \vec{x_i})\right] \vec{x_i}
\end{aligned}$$

(c)

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = -\mathbf{X}(\vec{y} - sigm(\mathbf{X}^T\vec{w}))$$

$$\frac{\partial^2 L(\vec{w})}{\partial \vec{w}\,\partial \vec{w}^T} = \mathbf{X}\,diag(sigm(\mathbf{X}^T\vec{w})(1 - sigm(\mathbf{X}^T\vec{w})))\,\mathbf{X}^T$$

Thus, $\quad \mathbf{H} = \mathbf{X}\mathbf{W}\mathbf{X}^T$

As $\mathbf{W}_{ii} = sigm(\vec{w}^T\vec{x_i})(1 - sigm(\vec{w}^T\vec{x_i}))$, when $\vec{w}^T\vec{x_i}$ is far away from 0, $\mathbf{W}_{ii}$ is small. When $\vec{w}^T\vec{x_i}$ is near to 0, $\mathbf{W}_{ii}$ is large.

(d) Let $\vec{v}$ be an arbitrary $d*1$ vector. Then

$$\begin{aligned}\vec{v}^T\mathbf{X}\mathbf{W}\mathbf{X}^T\vec{v} &= \sum_{j=1}^{n}(\sum_{i=1}^{d}(\vec{v}_i\mathbf{X}_{ij}\mathbf{W}_{jj}\mathbf{X}_{ij}\vec{v}_i)) \\ &= \sum_{j=1}^{n}(\sum_{i=1}^{d}(\mathbf{W}_{jj}{\mathbf{X}_{ij}}^2\vec{v}_i^2))\end{aligned}$$

$$\begin{aligned}&Since && \mathbf{W}_{jj} = sigm(\vec{w}^T\vec{x_j})(1 - sigm(\vec{w}^T\vec{x_j})), where\ 0 < sigm(\vec{w}^T\vec{x_j}) < 1 \\ &Thus && \mathbf{W}_{jj} > 0 \\ &Since && {\mathbf{X}_{ij}}^2 \geq 0\ and\ \vec{v}_i^2 \geq 0 \\ &Thus && \sum_{j=1}^{n}(\sum_{i=1}^{d}(\mathbf{W}_{jj}{\mathbf{X}_{ij}}^2\vec{v}_i^2)) \geq 0\end{aligned}$$

Thus, the Hessian H is positive-semidefinite.

(e) The step s would be:

$$\vec{s} = \arg\min_{\vec{s}}\,(l(\vec{w}) + g(\vec{w})^T\vec{s} + \frac{1}{2}\vec{s}^T\mathbf{H}\vec{s})$$

$$\begin{aligned}\vec{s} &= -\mathbf{H}^{-1}g(\vec{w}) \\ &= -(\mathbf{X}\mathbf{W}\mathbf{X}^T)^{-1}(-\mathbf{X}(\vec{y} - sigm(\mathbf{X}^T\vec{w})))\end{aligned}$$

$$\begin{aligned}\vec{w_{new}} &= \vec{w} + \vec{s} \\ &= \vec{w} - (\mathbf{X}\mathbf{W}\mathbf{X}^T)^{-1}(-\mathbf{X}(\vec{y} - sigm(\mathbf{X}^T\vec{w}))) \\ &= \vec{w} + (\mathbf{X}\mathbf{W}\mathbf{X}^T)^{-1}(\mathbf{X}(\vec{y} - sigm(\mathbf{X}^T\vec{w}))) \\ &= \mathbf{I}\vec{w} + (\mathbf{X}\mathbf{W}\mathbf{X}^T)^{-1}(\mathbf{X}\mathbf{I}(\vec{y} - sigm(\mathbf{X}^T\vec{w}))) \\ &= (\mathbf{X}\mathbf{W}\mathbf{X}^T)^{-1}(\mathbf{X}\mathbf{W}\mathbf{X}^T)\vec{w} + (\mathbf{X}\mathbf{W}\mathbf{X}^T)^{-1}(\mathbf{X}\mathbf{W}\mathbf{W}^{-1}(\vec{y} - sigm(\mathbf{X}^T\vec{w}))) \\ &= (\mathbf{X}\mathbf{W}\mathbf{X}^T)^{-1}(\mathbf{X}\mathbf{W})(\mathbf{X}^T\vec{w} + \mathbf{W}^{-1}(\vec{y} - sigm(\mathbf{X}^T\vec{w}))) \\ &= (\mathbf{X}\mathbf{W}\mathbf{X}^T)^{-1}(\mathbf{X}\mathbf{W}\vec{z})\end{aligned}$$

$$where\ \vec{z} = \mathbf{X}^T\vec{w} + \mathbf{W}^{-1}(\vec{y} - sigm(\mathbf{X}^T\vec{w}))$$

3 Problem3

(a)

$$l(\vec{w}) = (\mathbf{X}^T\vec{w} - \vec{y})^T\mathbf{P}(\mathbf{X}^T\vec{w} - \vec{y}) + \lambda\vec{w}^T\vec{w}$$

(b)

$$\begin{aligned}\frac{\partial}{\partial\vec{w}}l(\vec{w}) &= \frac{\partial}{\partial\vec{w}}(\vec{w}^T\mathbf{XPX}^T\vec{w} - 2\vec{w}^T\mathbf{XP}\vec{y} + \vec{y}^T\mathbf{P}\vec{y} + \lambda\vec{w}^T\vec{w}) \\ &= 2\mathbf{XPX}^T\vec{w} - 2\mathbf{XP}\vec{y} + 2\lambda\vec{w}\end{aligned}$$

To derive the closed form solution for $\vec{w}$, we need to solve

$$2\mathbf{XPX}^T\vec{w} - 2\mathbf{XP}\vec{y} + 2\lambda\vec{w} = 0$$

Thus,

$$\vec{w} = (\mathbf{XPX}^T - \lambda\mathbf{I})^{-1}\mathbf{XP}\vec{y}$$

(c) When $\lambda = 0$, the update weights of **Problem 2(e)** Newton's Method shares the same format with the closed form solution of $\vec{w}$ in **Problem 2(b)**. The reason why that algorithm is call *Iteratively Reweighted Least Squares* is that that algorithm (algorithm in Problem 2) is using the analytical optimal result of weighted Least Square algorithm, with $\mathbf{P} = \mathbf{W}$ in this case, iteratively to approach the real optimal solution of Logistic Regression. Reweighted refers to in each iteration, we recompute the error weights. reference: `https://cnx.org/contents/krkDdys0@12/Iterative-Reweighted-Least-Squares`

4 Problem4

(a) Assume that all data are generated independently.

$$\begin{aligned}
&maximize && P(D_n|\mu,\sigma^2) \\
&maximize && \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
&maximize && \log(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}) \\
&maximize && \sum_{i=1}^{n} \log(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}) \\
&maximize && \sum_{i=1}^{n} (\log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{(x-\mu)^2}{2\sigma^2})
\end{aligned}$$

$$\frac{\partial}{\partial\mu}(\sum_{i=1}^{n}(\log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{(x-\mu)^2}{2\sigma^2}) = \sum_{i=1}^{n} \frac{2(x_i-\mu)}{2\sigma^2} = 0$$

$$\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\sum_{i=1}^{n} x_i - n\mu = 0$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\frac{\partial}{\partial\sigma}(\sum_{i=1}^{n}(\log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{(x-\mu)^2}{2\sigma^2}) = \sum_{i=1}^{n}(\sigma\sqrt{2\pi}\frac{-1}{\sigma^2\sqrt{2\pi}} + \frac{(x_i-\mu)^2}{\sigma^3}) = 0$$

$$\sum_{i=1}^{n}(-\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{\sigma^3}) = 0$$

$$\sum_{i=1}^{n}(-\sigma^2 + (x_i-\mu)^2) = 0$$

$$\sigma^2 = \sum_{i=1}^{n}(x_i-\mu)^2$$

(b) Assume that all data are generated independently.

$$
\begin{aligned}
&maximize && P(D_n|\lambda) \\
&maximize && \prod_{i=1}^{n} \lambda e^{-\lambda x} \\
&maximize && \log(\prod_{i=1}^{n} \lambda e^{-\lambda x}) \\
&maximize && \sum_{i=1}^{n} \log(\lambda e^{-\lambda x}) \\
&maximize && \sum_{i=1}^{n} (\log \lambda - \lambda x)
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial \lambda}(\sum_{i=1}^{n}(\log \lambda - \lambda x)) = \sum_{i=1}^{n}(\frac{1}{\lambda} - x_i) = 0 \\
&\lambda = \frac{n}{\sum_{i=1}^{n} x_i}
\end{aligned}
.
$$

(c) Assume that all data are generated independently.

$$
\begin{aligned}
&maximize && P(D_n|p) \\
&maximize && \prod_{i=1}^{n} (p(1-p)^{x_i}) \\
&maximiza && \log(\prod_{i=1}^{n} (p(1-p)^{x_i})) \\
&maximiza && \sum_{i=1}^{n} \log(p(1-p)^{x_i}) \\
&maximiza && \sum_{i=1}^{n} (\log p + x_i \log(1-p))
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial p}(\sum_{i=1}^{n}(\log p + x_i \log(1-p))) = \sum_{i=1}^{n}(\frac{1}{p} - \frac{x_i}{1-p}) = 0 \\
&\frac{n}{p} = \frac{\sum_{i=1}^{n} x_i}{1-p} \\
&n - np = p\sum_{i=1}^{n} x_i \\
&p = \frac{n}{\sum_{i=1}^{n} x_i + n}
\end{aligned}
.
$$