**CIS5200: Machine Learning** Spring 2025

Homework 2

*Release Date: February 15, 2025* *Due Date: February 28, 2025*

**Name**: Haoze Wu
**PennKey**: haozewu
**Collaborators**: None

## Problem 1: Gradient Descent

**1.1** Answer:
To determine whether the unregularized objective $\hat{R}(w)$ is strongly convex or not, we need to calculate the Hessian matrix of $\hat{R}(w)$ is greater than or equal to $\mu I$ where $I$ is an identity matrix and $\mu > 0$. For the second order derivative of the empirical risk function $\hat{R}(w)$, we may start with the first order derivative of it, which is:

$$\begin{aligned}\nabla_w \hat{R}(w) &= \frac{1}{m}\sum_{i=1}^{m} \frac{\partial \log(1+\exp{(-y_i w^\top x_i)})}{\partial w} \\ &= \frac{1}{m}\sum_{i=1}^{m} \frac{1}{1+\exp{(-y_i w^\top x_i)}} \cdot (-y_i x_i) \cdot (\exp{(-y_i w^\top x_i)} \\ &= \frac{1}{m}\sum_{i=1}^{m} (1 - \frac{1}{1+\exp{(-y_i w^\top x_i)}})(-y_i x_i)\end{aligned} \tag{1}$$

Thus, we can calculate the second order derivative:

$$\begin{aligned}\nabla_w^2 \hat{R}(w) &= \frac{\partial \frac{1}{m}\sum_{i=1}^{m}(1 - \frac{1}{1+\exp{(-y_i w^\top x_i)}})(-y_i x_i)}{\partial w} \\ &= \frac{1}{m}\sum_{i=1}^{m} (-y_i x_i) \cdot (-\frac{1}{(1+\exp{(-y_i w^\top x_i)}^2}) \cdot \exp{(-y_i w^\top x_i)} \cdot (-y_i x_i) \\ &= \frac{1}{m}\sum_{i=1}^{m} x_i^\top x_i (\frac{1}{(1+\exp{(-y_i w^\top x_i)}} - \frac{1}{(1+\exp{(-y_i w^\top x_i)}^2})\end{aligned} \tag{2}$$

, which is equivalent to $X^T M X$, where $M$ is a $m \times m$ dimension diagonal matrix where $M_{ii} = \frac{1}{(1+\exp{(-y_i w^\top x_i)}} - \frac{1}{(1+\exp{(-y_i w^\top x_i)}^2}$ for all $i$ from 0 to $m$. Consider the value of $M_{ii}$, which is assembled by a sigmoid function. The value range of any sigmoid function is $(0, 1)$.

Then, if $\hat{R}(x)$ here is a $\mu-$strongly convex function, we then must have:

$$X^T M X \succeq \mu I \tag{3}$$

, which also means that for any vector $v$, we have:

$$v^T X^T M X v \succeq \mu ||v||_2^2 \tag{4}$$

, which is not true for those vectors $v'$ in the null space of $X$ since $Xv' = 0$. Thus, the unregularized objective $\hat{R}(w)$ is not strongly convex.

**1.2** Answer:
Consider the Hessian matrix calculated in part 1.1, which is

$$X^T MX \tag{5}$$

where $M$ is a $m \times m$ dimension diagonal matrix where $M_{ii} = \frac{1}{(1+\exp{(-y_i w^\top x_i)}} - \frac{1}{(1+\exp{(-y_i w^\top x_i)^2}}$ for all $i$ from 0 to $m$. Since each diagonal element of $M$ is assembled by sigmoid functions, whose value range is $(0, 1)$, we can calculate the value range of $M_{ii}$ by:

$$M_{ii} = \sigma(y_i w^\top x_i) - \sigma(y_i w^\top x_i)^2 \tag{6}$$

, where $\sigma$ is the sigmoid function. Since the value range of sigmoid function is $(0, 1)$, we the value range of $M_{ii}$ is $(0, 0.25]$. Thus, the smallest eigenvalue of $X^T MX$ is 0 and the largest eigenvalue of $X^T MX$ is 0.25 given that $||x_i||_2 \leq 1$ for all $i$ from 0 to $m$. Then, to prove that the Hessian matrix is L-smooth, we need to prove that the largest eigenvalue of $X^T MX$ is less than or equal to $L$. Since the largest possible value of $M_{ii}$ is 0.25, we can calculate the largest eigenvalue of $X^T MX$ by:

$$\lambda_{\max}(X^T MX) = \lambda_{\max}(X^T X) \cdot \lambda_{\max}(M) \tag{7}$$

, where $\lambda_{\max}(X^T X)$ is the largest eigenvalue of $X^T X$ and $\lambda_{\max}(M)$ is the largest eigenvalue of $M$. Since $||x_i||_2 \leq 1$ for all $i$ from 0 to $m$, we have $\lambda_{\max}(X^T X) \leq 1$. Thus, the largest eigenvalue of $X^T MX$ is 0.25, which is less than 1, and thus the Hessian matrix is $1-$smooth.

**1.3** Answer:
Recall that the definition of a L-smooth function is given by:

$$F(w') \leq F(w) + \nabla F(w)^\top (w' - w) + \frac{L}{2}||w' - w||_2^2 \tag{8}$$

Then, recall that the update rule of the gradient descent algoorithm is given by:

$$w_{t+1} = w_{(t)} - \eta \nabla_w \hat{R}(w_t) \tag{9}$$

For the empirical risk function $\hat{R}(w)$, we have:

$$\begin{aligned}
\hat{R}(w') &\leq \hat{R}(w) + \nabla \hat{R}(w)^\top (w' - w) + \frac{L}{2}||w' - w||_2^2 \\
\hat{R}(w_{t+1}) &\leq \hat{R}(w_t) + \nabla \hat{R}(w_t)^\top (w_{t+1} - w_t) + \frac{L}{2}||w_{t+1} - w_t||_2^2 \\
\hat{R}(w_{t+1}) &\leq \hat{R}(w_t) - \eta \nabla \hat{R}(w_t)^\top \nabla \hat{R}(w_t) + \frac{L}{2}||\eta \nabla \hat{R}(w_t)||_2^2 \\
\hat{R}(w_{t+1}) &\leq \hat{R}(w_t) - \eta ||\nabla \hat{R}(w_t)||_2^2 + \frac{L\eta^2}{2}||\nabla \hat{R}(w_t)||_2^2
\end{aligned} \tag{10}$$

To satisfy the requirement such that the objective value of the objective function is non-increasing each iteration, we must have:

$$\begin{aligned} \hat{R}(w_{t+1}) &\leq \hat{R}(w_t) \\ \hat{R}(w_t) - \eta||\nabla\hat{R}(w_t)||_2^2 + \frac{L\eta^2}{2}||\nabla\hat{R}(w_t)||_2^2 &\leq \hat{R}(w_t) \\ -\eta||\nabla\hat{R}(w_t)||_2^2 + \frac{L\eta^2}{2}||\nabla\hat{R}(w_t)||_2^2 &\leq 0 \\ \frac{L\eta^2}{2}||\nabla\hat{R}(w_t)||_2^2 &\leq \eta||\nabla\hat{R}(w_t)||_2^2 \\ \frac{L\eta}{2} &\leq 1 \\ \eta &\leq \frac{2}{L} \end{aligned} \tag{11}$$

Thus, the learning rate $\eta$ must be less than or equal to $\frac{2}{L}$ so that the objective value of the objective function is non-increasing each iteration.

**1.4** Answer:
To show the convergence rate of the gradient descent algorithm on this unregularized problem, we need to use the Theorem 7 fron the lecture notes that for a L-smooth function $F(w)$ with a global minimum $w^\star$, there is:

$$F(w_{T+1}) - F(w^\star) \leq \frac{L||w_1 - w^\star||_2^2}{2T} \tag{12}$$

By using this theorem, we have:

$$\hat{R}(w_{T+1}) - \hat{R}(w^\star) \leq \frac{L||w_1 - w^\star||_2^2}{2T} \tag{13}$$

If we initialize $w_1$ to be 0, we have:

$$\hat{R}(w_{T+1}) - \hat{R}(w^\star) \leq \frac{L||w^\star||_2^2}{2T} \tag{14}$$

To ensure the requirement stated in the problem that

$$\hat{R}(w_{T+1}) - \hat{R}(w^\star) \leq \epsilon \tag{15}$$

, we must have:

$$\begin{aligned} \frac{L||w^\star||_2^2}{2T} &\leq \epsilon \\ T &\geq \frac{L||w^\star||_2^2}{2\epsilon} \end{aligned} \tag{16}$$

If we scale the weight vector such that $||w^\star||_2 = 1$, we have:

$$T \geq \frac{L}{2\epsilon} \tag{17}$$

This implies that after $T = \frac{L}{2\epsilon}$ iterations, we get that $\hat{R}(w_{T+1}) - \hat{R}(w^\star) \leq \epsilon$. Thus, the convergence rate of the gradient descent algorithm on this unregularized problem is $O(\frac{1}{T})$.

**1.5** Answer:
First, we shall calculate the second order derivative of the regularized objective function, which is equivalent to adding the second order derivative of the regularizer to the Hessian matrix of the unregularized function calculated in part 1.1. For the regularizer, we have:

$$\begin{aligned}\nabla_{w_i} t(w) &= \sum_{j=1}^{d} \frac{\partial \lambda_j w_j^2}{\partial w_i} \\ &= 2\lambda_i w_i\end{aligned} \tag{18}$$

For the second order derivative of the regularizer, we may write the element of its Hessian matrix as:

$$H_{ij} = \frac{\partial^2 \lambda_j w_j^2}{\partial w_i \partial w_j} = 2\lambda_j \tag{19}$$

for any $i = j$ and 0 otherwise. Then, the Hessian matrix of the regularized objective function is:

$$\nabla_w^2(\hat{R}(w) + t(w)) = X^T M X + \text{diag}(2\lambda_1, 2\lambda_2, ..., 2\lambda_d) \tag{20}$$

From the Hessian matrix, we can tell that the eigenvalue of its is at least $\min_{j\in[d]} 2\lambda_j$, which means that the regularized objective function is $\min_{j\in[d]} 2\lambda_j-$strongly convex since

$$\nabla_w^2(\hat{R}(w) + t(w)) \succeq \min_{j\in[d]} 2\lambda_j I = \mu I \tag{21}$$

Then, consider the L-smooth aspect. We have proved that If $g$ is $\alpha$-smooth, and $f$ is $\beta$-smooth then $f + g$ is $\alpha + \beta$-smooth during the third recitation. For the regularizer term, we have:

$$\nabla_w^2 t(w) = \text{diag}(2\lambda_1, 2\lambda_2, ..., 2\lambda_d) \tag{22}$$

, which means that the regularizer term is $2\max_{j\in[d]} \lambda_j-$smooth since

$$\nabla_w^2 t(w) \preceq 2\max_{j\in[d]} \lambda_j I = LI \tag{23}$$

We have shown that the unregularized objective function is $1-$smooth in part 1.2. Thus, the regularized objective function is $1 + 2\max_{j\in[d]} \lambda_j-$smooth.

**1.6** Answer:
As proved before, the regularized objective function is $\min_{j\in[d]} 2\lambda_j-$strongly convex and $1 + 2\max_{j\in[d]} \lambda_j$-smooth. Then, we have:

$$||w_{T+1} - w_\star||_2^2 \leq (1 - \frac{\mu}{L})^T ||w_1 - w_\star||_2^2 \tag{24}$$

To achieve the goal such that $||w_{T+1} - w_\star||_2 \leq \epsilon$, we then must have:

$$\sqrt{(1 - \frac{\mu}{L})^T} ||w_1 - w_\star||_2 \leq \epsilon \tag{25}$$

We then have:

$$\begin{aligned}(1-\frac{\mu}{L})^T &\leq \frac{\epsilon^2}{||w_1-w_\star||_2^2} \\ \log\left((1-\frac{\mu}{L})^T\right) &\leq \log\left(\frac{\epsilon^2}{||w_1-w_\star||_2^2}\right) \\ T\log(1-\frac{\mu}{L}) &\leq 2\log\left(\frac{\epsilon}{||w_1-w_\star||_2}\right) \\ T &\leq \frac{2\log\left(\frac{\epsilon}{||w_1-w_\star||_2}\right)}{\log(1-\frac{\mu}{L})}\end{aligned} \tag{26}$$

Consider that $\mu = 2\min_{j\in[d]}\lambda_j$ and $L = 1+2\max_{j\in[d]}\lambda_j$, the value of $\frac{\mu}{L}$ is less than 1 and it can approach 0 if $\max_{j\in[d]}\lambda_j$ is large enough or $\min_{j\in[d]}\lambda_j$ is small enough, which leads to the approximation that

$$\log(1-\frac{\mu}{L}) \approx -\frac{\mu}{L} \tag{27}$$

Thus, with the approximation, we then have

$$\begin{aligned}T &\leq \frac{2\log\left(\frac{\epsilon}{||w_1-w_\star||_2}\right)}{\log(1-\frac{\mu}{L})} \\ T &\leq -2\frac{L}{\mu}\log\left(\frac{\epsilon}{||w_1-w_\star||_2}\right) \\ T &\leq 2\frac{L}{\mu}\log\left(\frac{||w_1-w_\star||_2}{\epsilon}\right)\end{aligned} \tag{28}$$

If we initialize at $w_1 = 0$ and $||w_\star|| = 1$, we than have:

$$\begin{aligned}T &\leq 2\frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right) \\ T &\leq 2\frac{1+2\max_{j\in[d]}\lambda_j}{2\min_{j\in[d]}\lambda_j}\log\left(\frac{1}{\epsilon}\right)\end{aligned} \tag{29}$$

, which means the convergence rate is $O(\frac{L}{\mu}\log(\frac{1}{\epsilon}))$, i.e., $O(\frac{1+2\max_{j\in[d]}\lambda_j}{2\min_{j\in[d]}\lambda_j}\log(\frac{1}{\epsilon}))$.

**1.7** Answer:
From 1.6, we know that

$$T \leq 2\frac{1+2\max_{j\in[d]}\lambda_j}{2\min_{j\in[d]}\lambda_j}\log\left(\frac{1}{\epsilon}\right) \tag{30}$$

When we choose $\lambda$ such that $\lambda_1 = \lambda_2 = \cdots = \lambda_d = \lambda$ for some $\lambda > 0$, we then have:

$$\begin{aligned}T &\leq 2\frac{1+2\lambda}{2\lambda}\log\left(\frac{1}{\epsilon}\right) \\ T &\leq (2+\frac{1}{\lambda})\log\left(\frac{1}{\epsilon}\right)\end{aligned} \tag{31}$$

Notice that part of the expression $\frac{1}{\lambda}$ is decreasing with the increment of $\lambda$, which implies that the convergence time of the gradient descent algorithm would be faster for a larger $\lambda$.

But this does not mean that a larger $\lambda$ would guarantee better performance since if the $\lambda$ is so large that it dominates the training process, the model would focus on training on the regularizer rather than the empirical risk.

## Problem 2: MLE for Linear Regression

**2.1** Answer:
Given that $\epsilon \sim \mathcal{N}(0, \sigma^2)$, for the expectation and variance of $y|x$, we have:

$$\begin{aligned} \mathbb{E}[y|x] &= \mathbb{E}[w^\top x + \epsilon] = w^\top x \\ \mathrm{Var}[y|x] &= \mathrm{Var}[w^\top x + \epsilon] = \sigma^2 \end{aligned} \tag{32}$$

, which indicates that $y|x \sim \mathcal{N}(w^\top x, \sigma^2)$. Thus, by the probability density function of the normal distribution, we have:

$$p(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp{(-\frac{(y - w^\top x)^2}{2\sigma^2})} \tag{33}$$

**2.2** Answer:
Given that

$$R(f) = \mathbb{E}_{x,y}[(y - f(x))^2] \tag{34}$$

, we than have:

$$\begin{aligned} R(f) &= \mathbb{E}_{x,y}[(y - f(x))^2] \\ &= \mathbb{E}_{x,y}[(y - w^\top x)^2] \\ &= \mathbb{E}_{x,y}[(w^\top x + \epsilon - w^\top x)^2] \\ &= \mathbb{E}_{x,y}[\epsilon^2] \\ &= \mathrm{Var}[\epsilon] - (\mathbb{E}[\epsilon])^2 \\ &= \sigma^2 \end{aligned} \tag{35}$$

**2.3** Answer:
To calculate the log conditional likelihood, we have:

$$\begin{aligned} \log \hat{L}(w, \sigma) &= \log p(y_1, \cdots, y_m | x_1, \cdots, x_m) \\ &= \log \prod_{i=1}^{m} p(y_i|x_i) \\ &= \sum_{i=1}^{m} \log p(y_i|x_i) \\ &= \sum_{i=1}^{m} \log{(\frac{1}{\sqrt{2\pi}\sigma} \exp{(-\frac{(y_i - w^\top x_i)^2}{2\sigma^2})})} \\ &= -m \log{(\sqrt{2\pi}\sigma)} - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - w^\top x_i)^2 \end{aligned} \tag{36}$$

**2.4** Answer:
Calculate the first order derivative of $\log \hat{L}(w,\sigma)$ with respect to $w$, we have:

$$\begin{aligned}\frac{\partial \log \hat{L}(w,\sigma)}{\partial w} &= \frac{\partial(-m\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - w^\top x_i)^2)}{\partial w} \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{m} 2(y_t - w^\top x_i)(-x_i) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{m}(y_t - w^\top x_i)(x_i)\end{aligned} \tag{37}$$

To maximize this log conditional likelihood, we need to calculate when its first order derivative is 0:

$$\begin{aligned}\frac{1}{\sigma^2}\sum_{i=1}^{m}(y_t - w^\top x_i)(x_i) = 0 \\ \sum_{i=1}^{m}(y_t - w^\top x_i)(x_i) = 0 \\ \sum_{i=1}^{m} y_i x_i = \sum_{i=1}^{m} x_i x_i^\top w\end{aligned} \tag{38}$$

By solving this equation can we get the optimal $w$ to maximize the log conditional likelihood.

Then, for the empirical risk, we have:

$$\hat{R}(w) = \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top x_i)^2 \tag{39}$$

whose first order derivative is:

$$\begin{aligned}\nabla_w \hat{R}(w) &= \frac{\partial \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top x_i)^2}{\partial w} \\ &= \frac{1}{m}\sum_{i=1}^{m} 2(y_i - w^\top x_i)(-x_i)\end{aligned} \tag{40}$$

To minimize $\hat{R}(w)$, we set its derivative to 0:

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m} 2(y_i - w^\top x_i)(-x_i) = 0 \\ \sum_{i=1}^{m}(y_i - w^\top x_i)(x_i) = 0 \\ \sum_{i=1}^{m} y_i x_i = \sum_{i=1}^{m} x_i x_i^\top w\end{aligned} \tag{41}$$

, which is the same as the equation (38). The solution to minimize the empirical risk is the same as maximize the log conditional likelihood. Thus, maximizing the log conditional likelihood is equivalent to minimizing the empirical risk.