Machine Learning Over Networks
Homework Assignment 2(a)

Student: José Mairton B. da Silva Jr.

Problem 2(a)

Consider Human Activity Recognition Using Smartphones dataset $\{(x_i, y_i)\}_{i\in[N]}$, with:
inputs: accelerometer and gyroscope sensors output: moving (e.g., walking , running, dancing) or not (sitting or standing)

Consider logistic ridge regression: minimize$_{\mathbf{w}}$ $\frac{1}{N}\sum_{i\in[N]} f_i(\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2$ where $f_i(\mathbf{w}) = \log\left(1 + \exp\left(-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i\right)\right)$. For classification, we can use the solution $\mathbf{w}^\star$ and compute $\operatorname{sign}({\mathbf{w}^\star}^{\mathrm{T}}\mathbf{x})$.

1. Is f Lipschitz continuous? If so, find a small B?
2. Is f_i smooth? If so, find a small L for f_i? What about f?
3. Is f strongly convex? If so, find a high μ?

Proof. For the proofs herein, we use the facts that:

- $\|\mathbf{w}\|_2 \le D$;
- If f and h are L_1 and L_2 Lipschitz continuous, then $f+h$ is also Lipschitz continuous with constant $L_1 + L_2$.

1) Let us define $g(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2$ and $h(\mathbf{w}) = \frac{1}{N}\sum_{i\in[N]} f_i(\mathbf{w})$. For $g(\mathbf{w})$, we have that

$$
\begin{aligned}
g(\mathbf{w}_1) - g(\mathbf{w}_2) &= \lambda\left(\|\mathbf{w}_1\|_2^2 - \|\mathbf{w}_2\|_2^2\right), && \text{(1a)}\\
&= \lambda\left(\mathbf{w}_1 - \mathbf{w}_2\right)\left(\mathbf{w}_1 + \mathbf{w}_2\right), \text{apply abs. value} && \text{(1b)}\\
|g(\mathbf{w}_1) - g(\mathbf{w}_2)| &= \left|\lambda\left(\mathbf{w}_1 - \mathbf{w}_2\right)\left(\mathbf{w}_1 + \mathbf{w}_2\right)\right|, \text{use Cauchy-Schwartz} && \text{(1c)}\\
&\le \lambda\|\mathbf{w}_1 - \mathbf{w}_2\|_2\|\mathbf{w}_1 + \mathbf{w}_2\|_2, \text{use triangle ineq.} && \text{(1d)}\\
&\le \lambda\|\mathbf{w}_1 - \mathbf{w}_2\|_2\left(\|\mathbf{w}_1\|_2 + \|\mathbf{w}_2\|_2\right), \text{use bounds} && \text{(1e)}\\
&\le 2D\lambda\|\mathbf{w}_1 - \mathbf{w}_2\|_2. && \text{(1f)}
\end{aligned}
$$

Therefore, $L_g = 2D\lambda$.

For $h(\mathbf{w})$, we have the following:

$$h(\mathbf{w}_1) - h(\mathbf{w}_2) = \frac{1}{N} \sum_{i \in [N]} \log \left(\frac{1 + \exp\left(-y_i {\mathbf{w}_1}^{\mathrm{T}} \mathbf{x}_i\right)}{1 + \exp\left(-y_i {\mathbf{w}_2}^{\mathrm{T}} \mathbf{x}_i\right)} \right), \text{apply abs. value} \tag{2a}$$

$$|h(\mathbf{w}_1) - h(\mathbf{w}_2)| = \left| \frac{1}{N} \sum_{i \in [N]} \log \left(\frac{1 + \exp\left(-y_i {\mathbf{w}_1}^{\mathrm{T}} \mathbf{x}_i\right)}{1 + \exp\left(-y_i {\mathbf{w}_2}^{\mathrm{T}} \mathbf{x}_i\right)} \right) \right|, \text{use triangle ineq.} \tag{2b}$$

$$\leq \frac{1}{N} \sum_{i \in [N]} \left| \log \left(\frac{1 + \exp\left(-y_i {\mathbf{w}_1}^{\mathrm{T}} \mathbf{x}_i\right)}{1 + \exp\left(-y_i {\mathbf{w}_2}^{\mathrm{T}} \mathbf{x}_i\right)} \right) \right|, \text{use } \log(1+x) \geq \log(x), \tag{2c}$$

$$\leq \frac{1}{N} \sum_{i \in [N]} \left| \log \left(\frac{\exp\left(-y_i {\mathbf{w}_1}^{\mathrm{T}} \mathbf{x}_i\right)}{\exp\left(-y_i {\mathbf{w}_2}^{\mathrm{T}} \mathbf{x}_i\right)} \right) \right|, \tag{2d}$$

$$= \frac{1}{N} \sum_{i \in [N]} \left| \log \left(\exp \left(-y_i (\mathbf{w}_1 - \mathbf{w}_2)^{\mathrm{T}} \mathbf{x}_i \right) \right) \right|, \tag{2e}$$

$$= \frac{1}{N} \sum_{i \in [N]} \left| \left(-y_i (\mathbf{w}_1 - \mathbf{w}_2)^{\mathrm{T}} \mathbf{x}_i \right) \right|, \text{use Cauchy-Schwart} \tag{2f}$$

$$\leq \frac{1}{N} \sum_{i \in [N]} |y_i| \, \|\mathbf{x}_i\|_2 \, \|\mathbf{w}_1 - \mathbf{w}_2\|_2 \,, \tag{2g}$$

$$= \left(\frac{1}{N} \sum_{i \in [N]} |y_i| \, \|\mathbf{x}_i\|_2 \right) \|\mathbf{w}_1 - \mathbf{w}_2\|_2 \,. \tag{2h}$$

Thus, $L_h = \frac{1}{N} \sum_{i \in [N]} |y_i| \, \|\mathbf{x}_i\|_2$. Therefore, $B = L_g + L_h$.

For the dataset, we assumed that $D = 1$ and varied λ from 0 to 100. Moreover, we used the training set available in the dataset. The result is present in Table 1. Since D is small, the role of λ is much smaller in promoting a low norm of the desired vector $\mathbf{w}$.

λ	0	1	5	100
B	200.4618	202.4618	210.4618	400.4618

Table 1: Regularization parameter λ and Lipschitz constant B

2) To prove that f_i is smooth, we need to prove that the gradient is Lipschitz, which can be proved also by showing the gradient is bounded as $\|\nabla f_i(\mathbf{w})\|_2 \leq L$. With this, we have the following:

$$\nabla f_i(\mathbf{w}) = \frac{1}{1 + \exp\left(-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i\right)} \left(-y_i \mathbf{x}_i \exp\left(-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i\right) \right), \tag{3a}$$

$$= \left[\frac{1}{1 + \exp\left(-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i\right)} - 1 \right] y_i \mathbf{x}_i, \text{aply norm} \tag{3b}$$

$$\|\nabla f_i(\mathbf{w})\|_2 = \left\| \left[\frac{1}{1 + \exp\left(-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i\right)} - 1 \right] y_i \mathbf{x}_i \right\|_2, \text{use Cauchy-Schwartz} \tag{3c}$$

$$\leq |y_i| \, \|\mathbf{x}_i\|_2 \left\| \frac{1}{1 + \exp\left(-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i\right)} - 1 \right\|_2 . \tag{3d}$$

Using Cauchy-Schwartz, note that

$$-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i \leq |y_i| \, \|\mathbf{w}\|_2 \, \|\mathbf{x}\|_2 \leq D \, |y_i| \, \|\mathbf{x}\|_2 \,, \text{use exp is monotonic} \tag{4a}$$

$$\exp\left(-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i\right) \leq \exp\left(D \, |y_i| \, \|\mathbf{x}\|_2\right), \tag{4b}$$

$$\frac{1}{1 + \exp\left(D \, |y_i| \, \|\mathbf{x}\|_2\right)} - 1 \leq \frac{1}{1 + \exp\left(-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i\right)} - 1, \tag{4c}$$

If we substitute the inequality above in Eq. (3d), we have that

$$\|\nabla f_i(\mathbf{w})\|_2 \leq |y_i| \|\mathbf{x}_i\|_2 \left\| \frac{1}{1+\exp\left(D\,|y_i|\,\|\mathbf{x}\|_2\right)} - 1 \right\|_2, \text{rearranging terms} \tag{5a}$$

$$\|\nabla f_i(\mathbf{w})\|_2 \leq |y_i| \|\mathbf{x}_i\|_2 \left\| \frac{\exp\left(D\,|y_i|\,\|\mathbf{x}\|_2\right)}{1+\exp\left(D\,|y_i|\,\|\mathbf{x}\|_2\right)} \right\|_2, \tag{5b}$$

which implies that $L_i = |y_i| \|\mathbf{x}_i\|_2 \left\| \frac{\exp\left(D|y_i|\|\mathbf{x}\|_2\right)}{1+\exp\left(D|y_i|\|\mathbf{x}\|_2\right)} \right\|_2$. Therefore, f_i is smooth with constant L_i.

For f, note that

$$\nabla f(\mathbf{w}) = \frac{1}{N} \sum_{i\in[N]} \nabla f_i(\mathbf{w}) + 2\lambda\mathbf{w}, \text{apply norm} \tag{6a}$$

$$\|\nabla f(\mathbf{w})\|_2 = \left\| \frac{1}{N} \sum_{i\in[N]} \nabla f_i(\mathbf{w}) + 2\lambda\mathbf{w} \right\|_2, \text{use triangle ineq.} \tag{6b}$$

$$\leq \left\| \frac{1}{N} \sum_{i\in[N]} \nabla f_i(\mathbf{w}) \right\|_2 + 2\lambda \|\mathbf{w}\|_2, \text{use triangle ineq.} \tag{6c}$$

$$\leq \frac{1}{N} \sum_{i\in[N]} \|\nabla f_i(\mathbf{w})\|_2 + 2D\lambda, \tag{6d}$$

$$\leq \underbrace{\frac{1}{N} \sum_{i\in[N]} |y_i| \|\mathbf{x}_i\|_2 \left\| \frac{\exp\left(D\,|y_i|\,\|\mathbf{x}\|_2\right)}{1+\exp\left(D\,|y_i|\,\|\mathbf{x}\|_2\right)} \right\|_2 + 2D\lambda}_{L_f}. \tag{6e}$$

Therefore, f is also smooth and has Lipschitz constant L_f.

For the dataset, we assumed that $D = 1$ and varied λ from 0 to 100. The result is present in Table 2. Notice that the values for L_f and B in problem 1 are equal. For the functions we analysed it happens to be true both statements, smoothness and Lipschitz continuity of the function, but it cannot be generalized.

λ	0	1	5	100
L_f	200.4618	202.4618	210.4618	400.4618

Table 2: Regularization parameter λ and Lipschitz constant L_f

3)

The function f is strongly convex because f_i is convex and $\lambda \|\mathbf{w}\|_2^2$ is strongly convex with a possible $\mu = 2\lambda$. To obtain a higher μ, we can increase the penalty parameter λ. In a similar manner, we can select μ based on the Hessian matrix of $f_i(\mathbf{w})$. Since $f_i(\mathbf{w})$ is convex, its Hermitian matrix is positive semidefinite which implies that all its eigenvalues are positive. Therefore, we can pick the smallest nonzero eigenvalue $\sigma_{\min} = \sigma_{\min}\left(\frac{1}{N}\sum_{i\in[N]} \nabla^2 f_i(\mathbf{w})\right)$ and majorize the Hessian matrix as $\nabla^2\left(\frac{1}{N}\sum_{i\in[N]} f_i(\mathbf{w})\right) \succeq \sigma_{\min}\mathbf{I}$. With this, we can increase μ by summing it up with $\sigma_{\min}$.

$\square$

Homework Assignment #2

Problem (b): From

$$\text{Var}[g(\omega_k,\zeta_k)] = \mathbb{E}_{\zeta_k}[\parallel g(\omega_k,\zeta_k) \parallel_2^2] - \parallel \mathbb{E}_{\zeta_k}[g(\omega_k,\zeta_k)] \parallel_2^2,$$

we have

$$\mathbb{E}_{\zeta_k}[\parallel g(\omega_k,\zeta_k) \parallel_2^2] = \text{Var}[g(\omega_k,\zeta_k)] + \parallel \mathbb{E}_{\zeta_k}[g(\omega_k,\zeta_k)] \parallel_2^2,$$

As $\text{Var}[g(\omega_k,\zeta_k)] \leq M + M_v \parallel \bigtriangledown f(\omega_k) \parallel_2^2$, $\parallel \mathbb{E}_{\zeta_k}[g(\omega_k,\zeta_k)] \parallel_2 \leq c_0 \parallel \bigtriangledown f(\omega_k) \parallel_2$, we have

$$\mathbb{E}_{\zeta_k}[\parallel g(\omega_k,\zeta_k) \parallel_2^2] \leq M + (M_v + c_0^2) \parallel \bigtriangledown f(\omega_k) \parallel_2^2,$$

thus $\alpha = M$, $\beta = M_v + c_0^2$

HW 2(c)

Group 3

February 16, 2019

Given hint by instructor (personnel email)

> *start from (4), assume that α_k sequence can satisfy $\alpha_k LM_G < mu$, which is reasonable due to the property of $alpha_k$ sequence. Then sum both sides of (4) for $k = 1 : K$ and conclude.*

The equation (4) mentioned in the above email is as follows

$$\mathbb{E}[f(w_{k+1})] - f(w_k) \leq -(c - \frac{1}{2}\alpha_k LM_G)\alpha_k \|\nabla f(w_k)\|_2^2 \tag{1}$$

Summing it over k as given in the hint.

$$\mathbb{E}[F(w_{k+1})] - \mathbb{E}[f(w_1)] \leq -\frac{1}{2}c\sum_{k=1}^{K}\alpha_k \mathbb{E}[\|\nabla f(w_k)\|_2^2] + \frac{1}{2}LM_G\sum_{k=1}^{K}\alpha_k^2 \tag{2}$$

Dividing the above by $c/2$ and rearranging the terms we have

$$\sum_{k=1}^{K}\alpha_k \mathbb{E}[\|\nabla f(w_k)\|_2^2] \leq 2\frac{\mathbb{E}[f(w_1)] - \mathbb{E}[F(w_{k+1})]}{c} + \frac{LM_G}{c}\sum_{k=1}^{K}\alpha_k^2 \tag{3}$$

Here the given assumption is as follows

$$\sum_k \alpha_k = \infty \tag{4}$$

$$\sum_k \alpha_k^2 < \infty \tag{5}$$

Based on 5, we can conclude that left hand side of 5 is finite. Hence the lemma