EP3260: Fundamentals of Machine Learning over Networks
Homework Assignment 2, Due Date: February 26, 2020
Group 4: H. Hellström, M. Tavana, F. Vannella, and M. Zeng

Problem 1. Consider Human Activity Recognition Using Smartphones dataset $\{(\boldsymbol{x}_i, y_i)\}_{i\in[N]}$, with inputs defined as the accelerometer and gyroscope sensors, and outputs defined as moving (e.g., walking , running, dancing) or not (sitting or standing). Consider the logistic ridge regression loss function

$$\text{minimize} \qquad f(\boldsymbol{w}) = \frac{1}{N}\sum_{i\in[N]} f_i(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_2^2, \tag{1}$$

where $f_i(\boldsymbol{w}) = \log\left(1+\exp\left\{-y_i \boldsymbol{w}^T \boldsymbol{x}_i\right\}\right)$. Then, address the following questions:

(a) Is f Lipschitz continuous? If so, find a small B?

Proof. We need to show that

$$\|\boldsymbol{w}\|_2 \le D \Rightarrow \|\nabla f(\boldsymbol{w})\|_2 \le B.$$

We know that for $h : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$, we have

$$\nabla h(g(\boldsymbol{z})) = h'(g(\boldsymbol{z}))\nabla g(\boldsymbol{z}). \tag{2}$$

By substituting $h(z) = \log(z)$ and $g(\boldsymbol{w}) = 1 + \exp\left\{-y_i\boldsymbol{w}^T\boldsymbol{x}_i\right\}$ in (2), we have

$$\begin{aligned}\nabla f_i(\boldsymbol{w}) &= h'(g(\boldsymbol{w}))\nabla g(\boldsymbol{w}) \\ &= \frac{-y_i}{1+\exp\{y_i\boldsymbol{w}^T\boldsymbol{x}_i\}}\boldsymbol{x}_i. \end{aligned}\tag{3}$$

Therefore, for $\|\boldsymbol{w}\|_2 \le D$, we have

$$\|\nabla f_i(\boldsymbol{w})\|_2 = \frac{|y_i|\,\|\boldsymbol{x}_i\|_2}{1+\exp\{y_i\boldsymbol{w}^T\boldsymbol{x}_i\}} \le \frac{|y_i|\,\|\boldsymbol{x}_i\|_2}{1+\exp\{-D|y_i|\,\|\boldsymbol{x}_i\|_2\}} \triangleq B_{f_i}$$

For $r(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|_2^2$ and $\|\boldsymbol{w}\|_2 \le D$, we have

$$\|\nabla r(\boldsymbol{w})\|_2 = 2\lambda\|\boldsymbol{w}\|_2 \le 2\lambda D \triangleq B_r$$

Theorem 1. *Suppose that $f_1, \dots, f_n$ are Lipschitz continuous on I with Lipschitz constants $B_1, \dots, B_n$, respectively. Then the linear combination $c_1f_1 + \cdots + c_nf_n$ is Lipschitz continuous on I with Lipschitz constant $|c_1|B_1 + \cdots + |c_n|B_n$.*

According to (1), for $\|\boldsymbol{w}\|_2 \le D$, $f(\boldsymbol{w})$ is a linear combination of Lipschitz continuous functions. Therefore, we have

$$\|\nabla f(\boldsymbol{w})\|_2 \le \frac{1}{N}\sum_{i\in[N]}\frac{|y_i|\,\|\boldsymbol{x}_i\|_2}{1+\exp\{-D|y_i|\,\|\boldsymbol{x}_i\|_2\}} + 2\lambda D \triangleq B$$

Assume that $|y_i|\,\|\boldsymbol{x}_i\|_2 \le C$ for all $i \in [N]$, then

$$\|\nabla f(\boldsymbol{w})\|_2 \le \frac{C}{1+\exp\{-DC\}} + 2\lambda D \triangleq B$$

$\square$

(b) Is f_i smooth? If so, find a small L for f_i? What about f?

Proof. We know that Hessian of a function is the Jacobian of its gradient. Assuming $J(\cdot)$ is the Jacobian operator, we have

$$\begin{aligned}
\nabla^2 f_i(\boldsymbol{w}) = J(\nabla f_i(\boldsymbol{x})) &= \boldsymbol{x}_i \boldsymbol{x}_i^T \frac{y_i^2 \exp\{y_i \boldsymbol{w}^T \boldsymbol{x}_i\}}{(1+\exp\{y_i \boldsymbol{w}^T \boldsymbol{x}_i\})^2} \\
&\preceq \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T y_i^2}{4} && \text{(4a)} \\
&\preceq \sigma_{\max}\left(\frac{\boldsymbol{x}_i \boldsymbol{x}_i^T y_i^2}{4}\right) I && \text{(4b)} \\
&= \frac{y_i^2 \sigma_{\max}(\boldsymbol{x}_i^T \boldsymbol{x}_i)}{4} I && \text{(4c)} \\
&= \frac{y_i^2 \|\boldsymbol{x}_i\|_2^2}{4} I, && \text{(4d)}
\end{aligned}$$

where in (4a) we used the inequality $\frac{\exp\{x\}}{(1+\exp\{x\})^2} \leq \frac{1}{4}$. In (4b), $\sigma_{\max}(X)$ is the largest eigenvalue of a (symmetric positive semidefinite) matrix X. In (4c), we used the facts that $\operatorname{eig}(AB) = \operatorname{eig}(BA)$. Hence, f_i is L_i-smooth with

$$L_i \leq \frac{y_i^2 \|\boldsymbol{x}_i\|_2^2}{4}.$$

Similarly for f we have

$$\begin{aligned}
\nabla^2 f(\boldsymbol{w}) = J(\nabla f(\boldsymbol{x})) &= \frac{1}{N} \sum_{i \in [N]} \boldsymbol{x}_i \boldsymbol{x}_i^T \frac{y_i^2 \exp\{y_i \boldsymbol{w}^T \boldsymbol{x}_i\}}{(1+\exp\{y_i \boldsymbol{w}^T \boldsymbol{x}_i\})^2} + 2\lambda I && \text{(5a)} \\
&\preceq \frac{1}{4N} \sum_{i \in [N]} \boldsymbol{x}_i \boldsymbol{x}_i^T y_i^2 + 2\lambda I && \text{(5b)} \\
&= \frac{1}{4N} A A^T + 2\lambda I && \text{(5c)} \\
&\preceq \sigma_{\max}\left(\frac{1}{4N} A A^T + 2\lambda I\right) I && \text{(5d)} \\
&= \frac{1}{4N}\left(\sigma_{\max}(A^T A) + 2\lambda\right) I, && \text{(5e)}
\end{aligned}$$

where $A \triangleq [y_1 \boldsymbol{x}_1, y_2 \boldsymbol{x}_2, \ldots, y_N \boldsymbol{x}_N]$. In (5e), we used the facts that $\operatorname{eig}(X + \lambda I) = \operatorname{eig}(X) + \lambda$ and $\operatorname{eig}(AB) = \operatorname{eig}(BA)$. Hence, f is L-smooth with

$$L \leq \frac{1}{4N} \sigma_{\max}(A^T A) + 2\lambda.$$

□

(c) Is f strongly convex? If so, find a high μ?

Proof. From (5a) and for all $\boldsymbol{v} \in \mathbb{R}^d$, we have

$$\begin{aligned}
\boldsymbol{v}^T\left(\nabla^2 f(\boldsymbol{w}) - 2\lambda I\right)\boldsymbol{v} &= \frac{1}{N} \sum_{i \in [N]} \frac{y_i^2 \exp\{y_i \boldsymbol{w}^T \boldsymbol{x}_i\}}{(1+\exp\{y_i \boldsymbol{w}^T \boldsymbol{x}_i\})^2} \boldsymbol{v}^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{v} \\
&= \frac{1}{N} \sum_{i \in [N]} \frac{y_i^2 \exp\{y_i \boldsymbol{w}^T \boldsymbol{x}_i\}}{(1+\exp\{y_i \boldsymbol{w}^T \boldsymbol{x}_i\})^2} \left\|\boldsymbol{v}^T \boldsymbol{x}_i\right\|_2^2 && \text{(6)} \\
&\geq 0. && \text{(7)}
\end{aligned}$$

Therefore, $\nabla^2 f(\boldsymbol{w}) - 2\lambda I \succeq 0$ or equivalently $\nabla^2 f(\boldsymbol{w}) \succeq 2\lambda I$. Hence, f is strongly convex with $\mu = 2\lambda$. In summary we have

$$\underbrace{2\lambda I}_{\mu I} \preceq \nabla^2 f(\boldsymbol{w}) \preceq \underbrace{\left(\frac{1}{4N}\sigma_{\max}\left(A^T A\right) + 2\lambda\right) I}_{LI},$$

where $A \triangleq [y_1 \boldsymbol{x}_1, y_2 \boldsymbol{x}_2, \ldots, y_N \boldsymbol{x}_N]$. $\square$

Problem 2. Let us assume that there exist scalars $c_0 \geq c > 0$ such that for all $k \in \mathbb{N}$

$$\nabla f\left(\boldsymbol{w}_k\right)^T \mathbb{E}_{\zeta_k}\left[g\left(\boldsymbol{w}_k;\zeta_k\right)\right] \geq c \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2, \tag{8a}$$

$$\left\|\mathbb{E}_{\zeta_k}\left[g\left(\boldsymbol{w}_k;\zeta_k\right)\right]\right\|_2 \leq c_0 \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2. \tag{8b}$$

Furthermore, let us assume that there exist scalars $M \geq 0$ and $M_V \geq 0$ such that for all $k \in \mathbb{N}$

$$\mathrm{Var}_{\zeta_k}\left[g\left(\boldsymbol{w}_k;\zeta_k\right)\right] \leq M + M_V \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2. \tag{9}$$

For the convergence proof of SGD with an L-smooth convex objective function (see slides), prove that

$$\mathbb{E}_{\zeta_k}\left[\left\|g\left(\boldsymbol{w}_k;\zeta_k\right)\right\|_2^2\right] \leq \alpha + \beta \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2. \tag{10}$$

Proof. By assumption 9 we know that

$$\begin{aligned}\mathrm{Var}_{\zeta_k}\left[g\left(\boldsymbol{w}_k;\zeta_k\right)\right] &= \mathbb{E}_{\zeta_k}\left[\left\|g\left(\boldsymbol{w}_k;\zeta_k\right)\right\|_2^2\right] - \left\|\mathbb{E}_{\zeta_k}\left[g\left(\boldsymbol{w}_k;\zeta_k\right)\right]\right\|_2^2 \\ &\leq M + M_V \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2.\end{aligned} \tag{11}$$

This is equivalent to

$$\mathbb{E}_{\zeta_k}\left[\left\|g\left(\boldsymbol{w}_k;\zeta_k\right)\right\|_2^2\right] \leq M + M_V \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2 + \left\|\mathbb{E}_{\zeta_k}\left[g\left(\boldsymbol{w}_k;\zeta_k\right)\right]\right\|_2^2. \tag{12}$$

Also, by assumption 8b we have

$$\begin{aligned}\mathbb{E}_{\zeta_k}\left[\left\|g\left(\boldsymbol{w}_k;\zeta_k\right)\right\|_2^2\right] &\leq M + M_V \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2 + c_0^2 \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2 \\ &= M + \left(M_V + c_0^2\right) \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2\end{aligned} \tag{13}$$

By comparing Equation 13 to Equation 10 we observe that $\alpha = M$, and $\beta = (M_V + c_0^2)$. $\square$

Problem 3. For the SGD with non-convex objective function, prove that with square summable but not summable step-size, we have for any $K \in \mathbb{N}$

$$\mathbb{E}\left[\sum_{k\in[K]} \alpha_k \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2\right] < \infty, \tag{14}$$

and therefore

$$\mathbb{E}\left[\frac{1}{\sum_{k\in[K]}\alpha_k}\sum_{k\in[K]} \alpha_k \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2\right] \xrightarrow{K\to\infty} 0. \tag{15}$$

Proof. Generic SG algorithm on L-smooth function satisfies

$$\mathbb{E}\left[f(\boldsymbol{w}_{k+1})\right] - f(\boldsymbol{w}_k) \leq -\left(c - 0.5\alpha_k L M_G\right)\alpha_k \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2 + 0.5\alpha_k^2 LM. \tag{16}$$

By recursively adding them up over $k \in [K]$, we obtain

$$\begin{aligned} f_{\inf} - f(\boldsymbol{w}_1) &\leq \mathbb{E}\left[f(\boldsymbol{w}_{k+1})\right] - f(\boldsymbol{w}_1) && \text{(17a)}\\ &\leq \sum_{k=1}^{K}\Big(-\left(c - 0.5\alpha_k L M_G\right)\alpha_k \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2 + 0.5\alpha_k^2 LM\Big) && \text{(17b)}\\ &\leq -c\sum_{k=1}^{K}\alpha_k \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2 + 0.5LM_G\sum_{k=1}^{K}\alpha_k^2\left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2 + 0.5LM\sum_{k=1}^{K}\alpha_k^2 && \text{(17c)}\end{aligned}$$

Taking expectation over the above equation and then re-arrange the terms, we have

$$\mathbb{E}\left[\sum_{k\in[K]} \alpha_k \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2\right] \leq \frac{f(\boldsymbol{w}_1) - f_{\inf}}{c} + 0.5LM_G\mathbb{E}\left[\sum_{k=1}^{K}\alpha_k^2\left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2\right] + 0.5LM\mathbb{E}\left[\sum_{k=1}^{K}\alpha_k^2\right] \tag{18}$$

Note that a L-smooth function is also Lipschitz continuous, and thus we have

$$\left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2 \leq B \tag{19}$$

if $\left\|\nabla \boldsymbol{w}_k\right\|_2 \leq D$. Therefore, we have

$$\mathbb{E}\left[\sum_{k\in[K]} \alpha_k \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2\right] \leq \frac{f(\boldsymbol{w}_1) - f_{\inf}}{c} + 0.5LM_G B^2\mathbb{E}\left[\sum_{k=1}^{K}\alpha_k^2\right] + 0.5LM\mathbb{E}\left[\sum_{k=1}^{K}\alpha_k^2\right] \tag{20}$$

Since the square is summable, and thus, we have the three terms on the right side bounded. Therefore, the left side is also bounded, namely

$$\left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2 \leq B \tag{21}$$

if $\left\|\nabla \boldsymbol{w}_k\right\|_2 \leq D$. Therefore, we have

$$\mathbb{E}\left[\sum_{k\in[K]} \alpha_k \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2\right] \leq \infty. \tag{22}$$

Meanwhile, we have

$$\mathbb{E}\left[\frac{1}{\sum_{k\in[K]}\alpha_k}\sum_{k\in[K]} \alpha_k \left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2\right] \leq \frac{f(\boldsymbol{w}_1) - f_{\inf}}{c\sum_{k\in[K]}\alpha_k} + 0.5LM_G B^2\mathbb{E}\left[\frac{\sum_{k=1}^{K}\alpha_k^2}{\sum_{k\in[K]}\alpha_k}\right] + 0.5LM\mathbb{E}\left[\frac{\sum_{k=1}^{K}\alpha_k^2}{\sum_{k\in[K]}\alpha_k}\right] \tag{23}$$

Since α_k is not summable step-size, it is clear that the three terms on the right side approach zero. Therefore, we have

$$\mathbb{E}\left[\frac{1}{\sum_{k\in[K]}\alpha_k}\sum_{k\in[K]}\alpha_k\left\|\nabla f\left(\boldsymbol{w}_k\right)\right\|_2^2\right]\xrightarrow{K\to\infty}0. \tag{24}$$

$\square$