MLoN Homework2

Group 1

Seems Ok. Maybe more clarification

1 Problem 1

Consider the logistic ridge regression loss function:

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

where $f_i(\boldsymbol{w}) = \log(1 + \exp\{-y_i \boldsymbol{w}^T \boldsymbol{x}_i\})$

1.1

To prove the Lipschitz continuity, if we have $\|\boldsymbol{w}\|_2 \le D$, $\lambda \ge 0$

$$\begin{aligned} \triangledown f(\boldsymbol{w}) &= 2\lambda \boldsymbol{w} - \frac{1}{N}\sum_{i\in[N]} \frac{y_i \boldsymbol{x}_i}{1 + e^{y_i \boldsymbol{w}^T \boldsymbol{x}_i}} \\ \|\triangledown f(\boldsymbol{w})\|_2 &\le 2\lambda \|\boldsymbol{w}\|_2 + \frac{1}{N}\sum_{i\in[N]} \frac{\|y_i \boldsymbol{x}_i\|_2}{|1 + e^{y_i \boldsymbol{w}^T \boldsymbol{x}_i}|} \\ &\le 2\lambda D + \frac{1}{N}\sum_{i\in[N]} \|y_i \boldsymbol{x}_i\|_2 \end{aligned} \tag{2}$$

Assume that $\|y_i \boldsymbol{x}_i\|_2 \le E$, then

$$\|\triangledown f(\boldsymbol{w})\|_2 \le 2\lambda D + E \tag{3}$$

So $f(\boldsymbol{w})$ is Lipschitz continuous.
To find a small B, we apply the whole human Activity Recognition Using Smartphones dataset, where $\boldsymbol{x}_i$ is normalized to range $[-1, 1]$ and y_i has range $[0, 5]$ (standing, sitting, laying, walking, walking_downstairs, walking_upstairs):

$$E = \frac{1}{N}\sum_{i\in[N]} \|y_i \boldsymbol{x}_i\|_2 = 36.078 \tag{4}$$

Then we have:

$$B = 2\lambda D + E = 2\lambda D + 36.078 \tag{5}$$

1.2

To prove the smoothness of $f_i(\boldsymbol{w})$, we only need to prove that $f_i(\boldsymbol{w})$ is differentiable (obviously) and $\bigtriangledown f_i(\boldsymbol{w})$ is Lipschitz continuous:

$$\begin{aligned}
\left\|\bigtriangledown^2 f_i(\boldsymbol{w})\right\|_2 &= \left\|\frac{(y_i\boldsymbol{x}_i)^2}{e^{y_i\boldsymbol{w}^T\boldsymbol{x}_i}+2+e^{-y_i\boldsymbol{w}^T\boldsymbol{x}_i}}\right\|_2 \\
&\leq \frac{\|y_i\boldsymbol{x}_i\|_2^2}{4} \\
&\leq \frac{25}{4}
\end{aligned} \tag{6}$$

Thus we can conclude that $\bigtriangledown f_i(\boldsymbol{w})$ is Lipschitz continuous, then $f_i(\boldsymbol{w})$ is L-smoothness and L is 6.25

As for $f(\boldsymbol{w})$, to apply the whole human Activity Recognition Using Smartphones dataset,

$$\begin{aligned}
\left\|\bigtriangledown^2 f(\boldsymbol{w})\right\|_2 &= \left\|2\lambda + \frac{1}{N}\sum_{i\in[N]}\frac{(y_i\boldsymbol{x}_i)^2}{e^{y_i\boldsymbol{w}^T\boldsymbol{x}_i}+2+e^{-y_i\boldsymbol{w}^T\boldsymbol{x}_i}}\right\|_2 \\
&\leq 2\lambda + \frac{1}{4N}\sum_{i\in[N]}\|y_i\boldsymbol{x}_i\|_2^2 \\
&\leq 2\lambda + 465.870
\end{aligned} \tag{7}$$

Thus $f(\boldsymbol{w})$ is L-smoothness with $L = 2\lambda + 465.870$

1.3

To prove the strongly-convexity of $f(\boldsymbol{w})$,

$$\bigtriangledown^2 f(\boldsymbol{w}) = 2\lambda + \frac{1}{N}\sum_{i\in[N]}\frac{(y_i\boldsymbol{x}_i)^2}{e^{y_i\boldsymbol{w}^T\boldsymbol{x}_i}+2+e^{-y_i\boldsymbol{w}^T\boldsymbol{x}_i}} \quad \geq 2\lambda \tag{8}$$

Thus $f(\boldsymbol{w})$ is μ-strongly convex with $\mu = 2\lambda$

Proof is Ok, but some more steps could be added as well for more `

2 Problem 2

Problem Let us assume that there exist scalars $c0 \geq c > 0$ such that for all $k \in N$

$$\bigtriangledown f(\mathbf{w}_k)^T \mathbb{E}_{\zeta_k}[g(\mathbf{w}_k;\zeta_k)] \geq c\| \bigtriangledown f(\mathbf{w}_k)\|_2^2 \tag{1a}$$

$$\|\mathbb{E}_{\zeta_k}[g(\mathbf{w}_k;\zeta_k)]\|_2 \leq c_0\| \bigtriangledown f(\mathbf{w}_k)\|_2 \tag{1b}$$

Furthermore, let us assume that there exist scalars $M \geq 0$ and $M_v \geq 0$ such that for all $k \in N$:

$$Var_{\zeta_k}[g(\mathbf{w}_k;\zeta_k)] \leq M + M_v\| \bigtriangledown f(\mathbf{w}_k)\|_2^2 \tag{2}$$

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

$$\mathbb{E}_{\zeta_k}[\|g(\mathbf{w}_k;\zeta_k)\|_2^2] \leq \alpha + \beta\| \bigtriangledown f(\mathbf{w}_k)\|_2^2$$

Proof Start with definition of variance:

$$\begin{aligned}
Var_{\zeta_k}[g(\mathbf{w}_k;\zeta_k)] &= \mathbb{E}_{\zeta_k}[\langle g(\mathbf{w}_k;\zeta_k), g(\mathbf{w}_k;\zeta_k)\rangle] - \langle \mathbb{E}_{\zeta_k}[g(\mathbf{w}_k;\zeta_k)], \mathbb{E}_{\zeta_k}[g(\mathbf{w}_k;\zeta_k)]\rangle \\
&(Second - moment\ boundary) \\
&\leq M + M_v\| \bigtriangledown f(\mathbf{w}_k)\|_2^2 \\
&\Rightarrow \\
\mathbb{E}_{\zeta_k}[\langle g(\mathbf{w}_k;\zeta_k), g(\mathbf{w}_k;\zeta_k)\rangle] &\leq M + M_v\| \bigtriangledown f(\mathbf{w}_k)\|_2^2 + \langle \mathbb{E}_{\zeta_k}[g(\mathbf{w}_k;\zeta_k)], \mathbb{E}_{\zeta_k}[g(\mathbf{w}_k;\zeta_k)]\rangle \\
&(First - moment\ boundary) \\
&\leq M + M_v\| \bigtriangledown f(\mathbf{w}_k)\|_2^2 + c_0^2\| \bigtriangledown f(\mathbf{w}_k)\|_2^2 \\
&= M + (M_v + c_0^2)\| \bigtriangledown f(\mathbf{w}_k)\|_2^2
\end{aligned}$$

Thus, proved with $\alpha = M, \beta = M_v + c_0^2$.

3 Problem 3

Seems OK

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

$$\mathbb{E}[\sum_{k\in[K]} \alpha_k \| \bigtriangledown f(\mathbf{w}_k)\|_2^2] < \infty \tag{9}$$

and therefore:

$$\lim_{k\to\infty} \mathbb{E}[\frac{\sum_{k\in[K]} \alpha_k \| \bigtriangledown f(\mathbf{w}_k)\|_2^2]}{\sum_{k\in[K]} \alpha_k} = 0 \tag{10}$$

Proof Using the inequality (4) and (5) in the slice of lecture 3, we have:

$$\begin{aligned}
&\mathbb{E}[f(\omega_{k+1})] - f(\omega_k) \\
&\leq -(c - \frac{1}{2}\alpha_k LM_G)\alpha_k \| \bigtriangledown f(\mathbf{w}_k)\|_2^2 + \frac{1}{2}\alpha_k^2 LM \\
&\leq -(c - \frac{c}{2LM_G} LM_G)\alpha_k \| \bigtriangledown f(\mathbf{w}_k)\|_2^2 + \frac{1}{2}\alpha_k^2 LM \\
&\leq -\frac{c}{2}\alpha_k \| \bigtriangledown f(\mathbf{w}_k)\|_2^2 + \frac{1}{2}\alpha_k^2 LM
\end{aligned} \tag{11}$$

Recursively $\forall k \in [K]$, take total expectation for both side and do summation, we have:

$$\begin{aligned}
&\mathbb{E}[f(\omega_{k+1})] - \mathbb{E}[f(\omega_1)] \\
&\leq -\frac{c}{2}\mathbb{E}[\sum_{k\in[K]} \alpha_k \| \bigtriangledown f(\mathbf{w}_k)\|_2^2] + \frac{1}{2}LM \sum_{k\in[K]} \alpha_k^2
\end{aligned} \tag{12}$$

Rearrange (12), we have:

$$\begin{aligned}
&\mathbb{E}[\sum_{k\in[K]} \alpha_k \| \bigtriangledown f(\mathbf{w}_k)\|_2^2] \\
&\leq -\frac{2\mathbb{E}[f(\omega_{k+1})] - \mathbb{E}[f(\omega_1)]}{c} + \frac{LM\sum_{k\in[K]} \alpha_k^2}{c}
\end{aligned} \tag{13}$$

The step size is square summable $\sum_{k\in[K]} \alpha_k^2 < \infty$, the right hand size is smaller than ∞, so we have:

$$\mathbb{E}[\sum_{k\in[K]} \alpha_k \| \bigtriangledown f(\mathbf{w}_k)\|_2^2] < \infty \tag{14}$$

The step size is not summable $\sum_{k\in[K]} \alpha_k = \infty$, so we have:

$$\lim_{k\to\infty} \mathbb{E}[\frac{\sum_{k\in[K]} \alpha_k \| \bigtriangledown f(\mathbf{w}_k)\|_2^2]}{\sum_{k\in[K]} \alpha_k} = 0 \tag{15}$$