Homework 2 (Group 5)

HW2(a)

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

Solution: We know that f is Lipschitz continuous on $\|w\| \le D$ if

$$\|w\| \le D \Rightarrow \|\nabla f(w)\| \le B \,.$$

It is given that $f(w) = \frac{1}{N}\sum_{i\in[N]} f_i(w) + \lambda\|w\|^2$.

where $f_i(w) = \log(1+\exp\{-y_i w^T x_i\})$.

$$\Rightarrow \nabla f_i(w) = \frac{-y_i x_i}{1+\exp\{-y_i w^T x_i\}} \cdot \exp\{-y_i w^T x_i\}$$

Then we have that for any $\|w\| \le D$:

$$\nabla f(w) = \frac{1}{N}\sum_{i\in[N]} \nabla f_i(w) + 2\lambda w$$

$$\Rightarrow \|\nabla f(w)\| \le \frac{1}{N}\sum_{i\in[N]} \|\nabla f_i(w)\| + 2|\lambda|\cdot\|w\|$$

$$\le \frac{1}{N}\sum_{i\in[N]} \frac{\exp\{D\cdot|y_i|\cdot\|x_i\|\}}{1+\exp\{D\cdot|y_i|\cdot\|x_i\|\}} \cdot |y_i|\cdot\|x_i\| + 2\lambda D$$

Therefore, f is Lipschitz continuous, i.e. $\|w\| \le D \Rightarrow \|\nabla f(w)\| \le B$,

where B can be chosen as:

$$B = \frac{1}{N}\sum_{i\in[N]} \frac{\exp\{D\cdot|y_i|\cdot\|x_i\|\}}{1+\exp\{D\cdot|y_i|\cdot\|x_i\|\}} \cdot |y_i|\,\|x_i\| + 2\lambda D$$

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

Solution: Denote $z(w) = -y_i w^T x_i$, $g(z) = \frac{1}{1+e^{-z}}$.

Then it is obvious that:

$$\frac{\partial g(z)}{\partial z} = -g(z)(1-g(z))$$

Since $\nabla f_i(w) = -(1-g(z))y_i x$, it holds that:

$$\nabla^2 f_i(w) = \frac{\partial g(z)}{\partial z}\cdot\frac{\partial z(w)}{\partial w}\cdot y_i x^T = y_i^2 \cdot g(z)(1-g(z))\cdot x_i x_i^T$$

For binary classification problems, we often have $\max\{|y_i|\} = 1$.

Also since $x_i x_i^T \le \|x_i\|^2 \cdot I$ and $g(z)\cdot(1-g(z)) \le \frac{1}{4}$, we have:

$$\nabla^2 f_i(w) \le \frac{\|x_i\|^2}{4}\cdot I$$

Therefore, we know that $f_i(w)$ is $\frac{\|x_i\|^2}{4}$ - smooth.

Since $f(w) = \frac{1}{N}\sum f_i(w) + \lambda\|w\|^2$, we have that:

$$\nabla^2 f(w) = \frac{1}{N}\sum_{i\in[N]} \nabla^2 f_i(w) + 2\lambda \cdot I$$

$$\leq \left(\frac{1}{4N}\sum_{i\in[N]} \|x_i\|^2 + 2\lambda\right) I$$

Therefore, $f(w)$ is also smooth.

3) Is f strongly convex? If so, find a high μ.

Solution: For any x_i and y_i, we have:

$$g[z(w)] = \frac{1}{1+e^{-z}} \in (0,1)$$

$$\Rightarrow g(z)\cdot(1-g(z)) \geq 0$$

$$\Rightarrow \nabla^2 f_i(w) = y_i^2 g(z)(1-g(z))\cdot x_i x_i^T \geq 0$$

Thus, it holds that

$$\nabla^2 f(w) = \frac{1}{N}\sum f_i(w) + 2\lambda \cdot I \geq 2\lambda \cdot I$$

Therefore, $f(w)$ is 2λ — strongly convex.

Homework II (b) (Jan. 31st 2019)

$\nabla f(w_k)^T \mathbb{E}_{\xi_k}[g(w_k;\xi_k)] \geq c \|\nabla f(w_k)\|_2^2 \quad \cdots\cdots (2a)$ $\qquad c_0 \geq c > 0$

$\|\mathbb{E}_{\xi_k}[g(w_k;\xi_k)]\|_2 \leq c_0 \|\nabla f(w_k)\|_2 \quad \cdots\cdots (2b)$

$M \geq 0$ and $M_v \geq 0$. s.t for all $k \in \mathbb{N}$

$\mathrm{Var}_{\xi_k}[g(w_k;\xi_k)] \leq M + M_v \|\nabla f(w_k)\|_2^2 \quad \cdots\cdots (3)$

(2) and (3) imply

$$\mathbb{E}_{\xi_k}[\|g(w_k;\xi_k)\|_2^2] \leq \alpha + \beta \|\nabla f(w_k)\|_2^2$$

$$\because \mathrm{Var}_{\xi_k}[g(w_k;\xi_k)] = \mathbb{E}_{\xi_k}[\|g(w_k;\xi_k)\|_2^2] - \left[\mathbb{E}_{\xi_k}[g(w_k;\xi_k)]\right]^2$$
$$\leq M + M_v \|\nabla f(w_k)\|_2^2$$

$$\therefore \mathbb{E}_{\xi_k}[\|g(w_k;\xi_k)\|_2^2] \leq M + M_v \|\nabla f(w_k)\|_2^2 + \left[\mathbb{E}_{\xi_k}[g(w_k;\xi_k)]\right]^2$$

from (2b) it holds that:

$$\mathbb{E}_{\xi_k}[\|g(w_k;\xi_k)\|_2^2 \leq M + M_v \|\nabla f(w_k)\|_2^2 + c_0^2 \|\nabla f(w_k)\|_2^2$$
$$\leq M + (M_v + c_0^2) \|\nabla f(w_k)\|_2^2$$

$\therefore \quad \alpha = M$

$\beta = M_v + c_0^2$

if unbiased gradient estimator : $c = c_0 = 1$

$\therefore \quad \alpha = M$

$\beta = M_v + 1$

HW2 (c)

With square summable but not summable step-size, we have for any $K \in \mathbb{N}$

$$\mathbb{E}\left[\sum_{k\in[K]} \alpha_k \|\nabla f(w_k)\|^2\right] < \infty,$$

and therefore

$$\mathbb{E}\left[\frac{1}{\sum \alpha_k} \sum_{k\in[K]} \alpha_k \|\nabla f(w_k)\|^2\right] \xrightarrow{K\to\infty} 0.$$

proof. Generic SG-algorithm on L-smooth function satisfies:

$$\mathbb{E}_{\xi_k}[f(w_{k+1})] - f(w_k) \le -(c - \tfrac{1}{2}\alpha_k L M_G)\alpha_k \|\nabla f(w_k)\|^2 + \tfrac{1}{2}\alpha_k^2 LM. \quad \text{—— } (*)$$

Recursively $\forall k \in [K]$, take total expectation from $(*)$;

$$f_{inf} - f(w_1) \le \mathbb{E}[f(w_{K+1})] - f(w_1)$$

$$\le -c\,\mathbb{E}\left[\sum_{k\in[K]} \alpha_k \|\nabla f(w_k)\|^2\right] + \tfrac{1}{2} L M_G \mathbb{E}\left[\sum_{k\in[K]} \alpha_k^2 \|\nabla f(w_k)\|^2\right] + \tfrac{LM}{2}\sum_{k\in[K]} \alpha_k^2$$

Then we have

$$\mathbb{E}\left[\sum_{k\in[K]} \alpha_k \|\nabla f(w_k)\|^2\right] \le \frac{f(w_1) - f_{inf}}{c} + \frac{LM_G}{2c}\mathbb{E}\left[\sum_{k\in[K]} \alpha_k^2 \|\nabla f(w_k)\|^2\right] + \frac{LM}{2c}\sum_k \alpha_k^2 \quad \text{—— } (**)$$

Since we assume f is L-smooth, f must be Lipschitz continuous.
Then we have $\|\nabla f(w_k)\|$ is bounded on the domain $\{w_k : \|w_k\| \le D\}$,
i.e. $\|\nabla f(w_k)\| \le B$.

Hence, it holds that

$$\mathbb{E}\left[\sum_{k\in[K]} \alpha_k^2 \|\nabla f(w_k)\|^2\right] \le B^2 \mathbb{E}\left[\sum_{k\in[K]} \alpha_k^2\right] < \infty$$

(by assumption, we have $\sum_{k\in[K]} \alpha_k^2 < \infty$)

Thus, we have $\mathbb{E}\left[\sum_{k\in[K]} \alpha_k \|\nabla f(w_k)\|^2\right] < \infty$ since each of the 3 terms on the right side of $(**)$ is bounded.

Dividing $(**)$ by $\sum_{k\in[K]} \alpha_k$, we have:

$$\mathbb{E}\left[\frac{1}{\sum \alpha_k}\sum \alpha_k \|\nabla f(w_k)\|^2\right] \le \frac{f(w_1) - f_{inf}}{c \cdot \sum \alpha_k} + \frac{LM_G B^2 + LM}{2c} \cdot \frac{\sum \alpha_k^2}{\sum \alpha_k}$$

Since $\lim_{K\to\infty} \sum_{k\in[K]} \alpha_k = \infty$ and $\lim_{K\to\infty} \sum_{k\in[K]} \alpha_k^2 < \infty$, it holds that

$$\frac{1}{\sum \alpha_k} \xrightarrow{K\to\infty} 0 \quad \text{and} \quad \frac{\sum \alpha_k^2}{\sum \alpha_k} \to 0.$$

Therefore, $\mathbb{E}\left[\frac{1}{\sum \alpha_k}\sum \alpha_k \|\nabla f(w_k)\|^2\right] \xrightarrow{K\to\infty} 0.$