Homework 2

Ali Bemani **Ali.Bemani@hig.se**
Oscar Bautista Gonzalez **Oscar.Bautista.Gonzalez@hig.se**
Martin Hellkvist **Martin.Hellkvist@angstrom.uu.se**

Solutions for Homework 2 by Group 2 are in this document provided in the form of photocopies of handwritten notes.

Problem 2.1. a) 1/1

Q: Is $f(w) = \frac{1}{N} \sum_{i \in [N]} f_i(w) + \lambda \|w\|_2^2$ Lipschitz cont? If so, find the smallest B.

$f_i(w) = \log\left(1 + e^{-y_i x_i^T w}\right)$.

A: Solution: We will show that $\|\nabla f(w)\| \leq B$ for $\|w\| \leq D$

- $\nabla f(w) = \frac{1}{N} \sum_i \nabla f_i(w) + 2\lambda w$.
- $\nabla f_i(w) = \left(1 + e^{-y_i x_i^T w}\right)^{-1} e^{-y_i x_i^T w} \cdot (-y_i x_i)$

$$= \left(e^{y_i x_i^T w} + 1\right)(-y_i x_i) = -\frac{y_i}{1 + e^{y_i x_i^T w}} x_i$$

$$\Rightarrow \nabla f(w) = \frac{1}{N} \sum_{i \in [N]} -\frac{y_i}{1 + e^{y_i x_i^T w}} x_i + 2\lambda w.$$

$$\|\nabla f(w)\| = \left\| \frac{1}{N} \sum_i -\frac{y_i}{1 + e^{y_i x_i^T w}} x_i + 2\lambda w \right\|$$

triangular ineq. →

$$\leq \left\| \frac{1}{N} \sum_i \frac{-y_i}{1 + e^{y_i x_i^T w}} x_i \right\| + 2\lambda \|w\|$$

$$\leq \frac{1}{N} \sum_i \left\| -\frac{y_i}{1 + e^{y_i x_i^T w}} x_i \right\| + 2\lambda \|w\|$$

$$= \frac{1}{N} \sum_i \frac{|y_i| \|x_i\|}{1 + e^{y_i x_i^T w}} + 2\lambda \|w\|$$

{using that $-|y_i| \|x_i\| \|w\| \leq x_i^T w$ from Cauchy-Schwarz, (By $|y_i x_i^T w| \leq |y_i| \|x_i\| \|w\|$) we get}

$$\leq \frac{1}{N} \sum \frac{|y_i| \|x_i\|}{1 + e^{-|y_i| \|x_i\| \|w\|}} + 2\lambda \|w\|$$

{For $\|w\| \leq D$, we have}

$$\leq \underbrace{\frac{1}{N} \sum \frac{|y_i| \|x_i\|}{1 + e^{-|y_i| \|x_i\| D}} + 2\lambda D}_{\text{is the smallest } B \text{ such that } \|\nabla f(w)\| \leq B,\ \|w\| \leq D.} \leq \frac{1}{N} \sum_i |y_i| \|x_i\| + 2\lambda D$$

Problem 2.1 b)

Q: Is $f_i(w)$ smooth? If so, find a small L. What about f?

A: $f_i(w) = \log\left(1 + e^{-y_i x_i^T w}\right)$.

- Define $g_i(w) = -y_i x_i^T w = -b_i^T w$, $b_i = y_i x_i$.

$\rightarrow f_i(w) = \log\left(1 + e^{g_i(w)}\right)$.

- A twice differentiable function f is L-smooth iff $\nabla^2 f(x) \preceq LI$.

- We will find $\nabla^2 f_i(w)$ and determine L:

$$\nabla f_i(w) = \frac{1}{1 + e^{g_i(w)}} \nabla\left(1 + e^{g_i(w)}\right) = \frac{1}{1 + e^{g_i(w)}} e^{g_i(w)} \nabla g_i(w)$$

$$= -b_i \frac{e^{g(w)}}{1 + e^{g(w)}}.$$

$$\nabla^2 f_i(w) = -b_i \left(\nabla \frac{e^{g_i(w)}}{1 + e^{g_i(w)}}\right)^T = -b_i \left(\frac{\left(\nabla(e^{g_i(w)})\right)(1 + e^{g_i(w)}) - e^{g(w)} \nabla(1 + e^{g_i(w)})}{(1 + e^{g_i(w)})^2}\right)^T$$

$$= -b_i \left(\frac{e^{g(w)}(-b_i)(1 + e^{g_i(w)}) - e^{g_i(w)} e^{g_i(w)}(-b_i)}{(1 + e^{g(w)})^2}\right)^T =$$

$$= b_i b_i^T \frac{e^{g_i(w)} + e^{2g_i(w)} - e^{2g_i(w)}}{(1 + e^{g_i(w)})^2}$$

$$= b_i b_i^T \frac{e^{g_i(w)}}{(1 + e^{g_i(w)})^2}.$$

Problem 2.1 b) cont'd

We have that $\nabla^2 f_i(w) = b_i b_i^T \frac{e^{g_i(w)}}{(1+e^{g_i(w)})^2}$,

which is bounded for bounded $b_i = y_i x_i$,

because $\frac{e^{g_i(w)}}{(1+e^{g_i(w)})}$ is lower bounded by zero, and upper bounded by $1/4$.

Now we find a small L: $\nabla^2 f_i(w)$ is symmetric,

$\rightarrow \nabla^2 f_i(w) \preceq LI \leftrightarrow x^T(LI - \nabla^2 f_i(w))x \geq 0 \quad \forall x, w$

$\leftrightarrow Lx^Tx - x^T b_i b_i^T x \frac{e^{g_i(w)}}{(1+e^{g_i(w)})^2} \geq 0$

$\rightarrow Lx^Tx \geq x^T b_i b_i^T x \frac{e^{g_i(w)}}{(1+e^{g_i(w)})^2} \quad \forall x, w$

$\rightarrow L \geq \frac{x^T b_i b_i^T x}{x^T x} \frac{e^{g_i(w)}}{(1+e^{g_i(w)})^2}, \quad \forall x, w, \quad x \neq 0$

in particular, we need $L \geq \max_{x \neq 0} \frac{x^T b_i b_i^T x}{x^T x} \frac{e^{g_i(w)}}{(1+e^{g_i(w)})^2}, \forall w$

$= \lambda_{max}(b_i b_i^T) \frac{e^{g_i(w)}}{(1+e^{g_i(w)})^2}, \forall w$

where $\lambda_{max}(b_i b_i^T) = \|b_i\|^2 = b_i^T b_i \geq 0$

$\rightarrow L \geq b_i^T b_i \frac{e^{g_i(w)}}{(1+e^{g_i(w)})^2} \quad \forall w$

where the right hand side is upper bounded by $\frac{1}{4} b_i^T b_i$

$\rightarrow$ Smallest L is $L = \frac{1}{4} b_i^T b_i = \frac{1}{4} y_i^2 x_i^T x_i$.

Answer: $f_i(w)$ is L-smooth with $L = \frac{1}{4} y_i^2 x_i^T x_i$.

Problem 2.1 b) cont'd 3/3

"What about f?"

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

which is bounded since $[N]$ is finite, and f_i's are bounded.

→ f is L-smooth, and we can find L as:

$$L x^T x - \frac{1}{N} x^T \left(\sum_i \nabla^2 f_i(w)\right) x + 2\lambda x^T x \geq 0 \qquad \forall x, w$$

$$\Leftrightarrow L \geq \frac{1}{N} \frac{x^T \left(\sum \nabla^2 f_i(w)\right) x}{x^T x} - 2\lambda \qquad \forall x \neq 0, w$$

to ensure for $\forall x \neq 0, w$ we choose L as:

$$L = \left(\sum_{i \in [N]} \max_{x_i \neq 0} \frac{1}{4N} \frac{x_i^T b_i b_i^T x_i}{x_i^T x_i} \right) - 2\lambda$$

$$= \frac{1}{4N} \sum_{i \in [N]} b_i^T b_i - 2\lambda$$

$$= \frac{1}{4N} \sum_{i \in [N]} y_i^2 x_i^T x_i$$

Answer: $f(w)$ is L-smooth with

$$L = \frac{1}{4N} \sum_{i \in [N]} y_i^2 x_i^T x_i - 2\lambda .$$

Problem 2.1 c) 1/1

Q: Is f strongly convex? If so, find a high μ.

A: f is twice differentiable, so it is strongly convex if $\nabla^2 f(w) \geq \mu I$, $\forall w$

We have $\nabla^2 f(w) = \frac{1}{N} \sum_i \nabla^2 f_i(w) + 2\lambda$,

so we need μ:

$$x^T \left(\nabla^2 f(w) - \mu I \right) x \geq 0, \quad \forall x, w$$

$$\leftrightarrow \quad x^T \nabla^2 f(w)\, x \geq \mu\, x^T x, \quad \forall x, w$$

$$\rightarrow \quad \mu \leq \frac{x^T \nabla^2 f(w)\, x}{x^T x} = \frac{1}{N} \sum_{i \in [N]} \underbrace{\frac{x^T \nabla^2 f_i(w)\, x}{x^T x}}_{\text{(A)}} + 2\lambda, \quad \forall x \neq 0, w$$

where (A) is lower bounded by zero.

Thus, we need $\mu \leq 2\lambda$, $\rightarrow$ choose $\mu = 2\lambda$.

Answer: $f(w)$ is μ-strongly convex, with μ

$$\mu = 2\lambda.$$

problem 2.2

$$\| E_{\xi_k}[g(\omega_k;\xi_k)] \|_2 \leq c_0 \| \nabla f(\omega_k) \|_2 \quad \text{assumption}$$

$$\Rightarrow \| E_{\xi_k}[g(\omega_k;\xi_k)] \|_2^2 \leq c_0^2 \| \nabla f(\omega_k) \|_2^2 \quad (1) \text{ to the } 2 \text{ power of}$$

$$\underset{\xi_k}{\text{var}}[g(\omega_k;\xi_k)] \leq M + M_V \| \nabla f(\omega_k) \|_2^2 \quad (2) \text{ assumption}$$

the variance is the mean square value:

$$\Rightarrow \underset{\xi_k}{E}\left\{ \left[g(\omega_k;\xi_k) - E\{g(\omega_k;\xi_k)\} \right]^2 \right\} = \underset{\xi_k}{\text{var}}[g(\omega_k;\xi_k)]$$

$$\underset{\xi_k}{\text{var}}[g(\omega_k;\xi_k)] = \underset{\xi_k}{E}\left\{ \| g(\omega_k;\xi_k) \|_2^2 \right\} + \underset{\xi_k}{E}\left\{ -2 g(\omega_k;\xi_k) \cdot E\{g(\omega_k;\xi_k)\} \right\}$$

$$+ \underset{\xi_k}{E}\left\{ \| E\{g(\omega_k;\xi_k)\} \|_2^2 \right\}$$

$$\underset{\xi_k}{\text{var}}[g(\omega_k;\xi_k)] = \underset{\xi_k}{E}\left[\| g(\omega_k;\xi_k) \|_2^2 \right] - 2 \| E_{\xi_k}[g(\omega_k;\xi_k)] \|_2^2$$

$$+ \| E_{\xi_k}[g(\omega_k;\xi_k)] \|_2^2$$

now we can write the assumption number 2 with this form:

$$\underset{\xi_k}{E}\left[\| g(\omega_k;\xi_k) \|_2^2 \right] - \| E_{\xi_k}(g(\omega_k;\xi_k)) \|_2^2 \leq M + M_V \| \nabla f(\omega_k) \|_2^2 \quad (3)$$

now we can combine the equation number 1 with number 3 :

$$\mathbb{E}_{\delta_k}\left[\|g(\omega_k;\delta_{ik})\|_2^2\right] \lesssim \underbrace{M}_{\alpha} + \underbrace{(M_v + c_0^2)}_{\Lambda} \|\nabla f(\omega_{ik})\|_2^2$$

$$\Rightarrow \mathbb{E}_{\delta}\left[\|g(\omega_k;\delta_k)\|_2^2\right] \lesssim \alpha + \Lambda \|\nabla f(\omega_k)\|_2^2$$

Problem 2.3

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

$$(1) \quad E\left[\sum_{k \in \mathbb{K}} \alpha_k \| \nabla f(w_k) \|_2^2\right] < \infty$$

and therefore

$$E\left[\frac{1}{\sum_{k \in [K]} \alpha_k} \sum_{k \in [K]} \alpha_k \| \nabla f(w_k) \|_2^2\right] \xrightarrow{K \to \infty} 0$$

We can prove (1) by assuming that the function f is L-smooth and using the condition that the SG algorithm satisfy under the condition of smoothness

$$E[f(w_{k+1})] - f(w_k) \leq -\left(c - \frac{1}{2}\alpha_k L M_G\right)\alpha_k \| \nabla f(w_k) \|_2^2 + \frac{1}{2}\alpha_k^2 L M$$

which is equivalent to

$$E[f(w_{k+1})] - E[f(w_k)] \leq -A_k \alpha_k E\left[\| \nabla f(w_k) \|_2^2\right] + \frac{1}{2}\alpha_k^2 L M$$

where $A_k = c - \frac{1}{2}\alpha_k L M_G$

Specifying $E[f(w_k)]$ we have

$$E[f(w_{k+1})] \leq E[f(w_1)] - \sum_{i=1}^{k} A_i \alpha_i E\left[\| \nabla f(w_i) \|_2^2\right] + \frac{1}{2}\alpha_i L M$$

Then,

$$E\left[f(w_{k+1})\right] \leq E\left[f(w_1)\right] + \sum_{i=2}^{k} -c\,\alpha_i E\left[\|\nabla f(w_i)\|_2^2\right] + \frac{1}{2}\alpha_i^2 L M_G E\left[\|\nabla f(w_i)\|_2^2\right] + \frac{1}{2}\alpha_i^2 L M \Rightarrow$$

$$\Rightarrow c\sum_{i=1}^{k} \alpha_i E\left[\|\nabla f(w_i)\|_2^2\right] \leq E\left[f(w_1)\right] - E\left[f(w_{k+1})\right] + \sum_{i=1}^{k} \frac{1}{2}\alpha_i^2 L M_G E\left[\|\nabla f(w_i)\|_2^2\right] + \frac{1}{2}\alpha_i^2 L M$$

with square summable and not summable step size

$$c\sum_{i=1}^{k} \alpha_i E\left[\|\nabla f(w_i)\|_2^2\right] < \infty$$

Finally

$$E\left[\sum_{k\in[k]} \alpha_i \|\nabla f(w_i)\|_2^2\right] < \infty$$

Defining

$$\sum_{k\in[k]} a_k = \sum_{i=1}^{k} \frac{1}{2}\alpha_i^2 L M_G E\left[\|\nabla f(w_i)\|_2^2\right] + \frac{1}{2}\alpha_i^2 L M$$

Therefore if $k \to \infty$ and as before $\sum \alpha_i^2 < \infty$

$$E\left[\frac{1}{\sum_{k\in[k]} a_k} \sum_{k\in[k]} \alpha_k \|\nabla f(w_k)\|_2^2\right] \longrightarrow 0$$