

Homework 2

Ali Bemani
Oscar Bautista Gonzalez
Martin Hellkvist

Ali.Bemani@hig.se
Oscar.Bautista.Gonzalez@hig.se
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 .

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

A: Solution: We will show that $\|\nabla f(w)\| \leq B$ for $\|w\| \leq 1$

$$\nabla f(w) = \frac{1}{N} \sum_i \nabla f_i(w) + 2\lambda w.$$

$$\begin{aligned} \nabla f_i(w) &= (1 + e^{-y_i x_i^T w})^{-1} e^{-y_i x_i^T w} (-y_i x_i) \\ &= (e^{y_i x_i^T w} + 1) (-y_i x_i) = -\frac{y_i}{1 + e^{y_i x_i^T w}} x_i \end{aligned}$$

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

$$\begin{aligned} \|\nabla f(w)\| &= \left\| \frac{1}{N} \sum_i -\frac{y_i}{1 + e^{y_i x_i^T w}} x_i + 2\lambda w \right\| \\ \text{triangular ineq.} \rightarrow &\leq \left\| \frac{1}{N} \sum_i -\frac{y_i}{1 + e^{y_i x_i^T w}} x_i \right\| + 2\lambda \|w\| \\ &\leq \frac{1}{N} \sum_i \left\| -\frac{y_i}{1 + e^{y_i x_i^T w}} x_i \right\| + 2\lambda \|w\| \end{aligned}$$

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

{ 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 }

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

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

$$\leq \frac{1}{N} \sum_i \underbrace{\frac{|y_i| \|x_i\|}{1 + e^{-|y_i| \|x_i\| D}}} + 2\lambda D \leq \frac{1}{N} \sum_i |y_i| \|x_i\| + 2\lambda D$$

is the smallest B such that $\|\nabla f(w)\| \leq B$,
 $\|w\| \leq D$.

Problem 2.1 b) 1/3

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

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

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

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

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

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

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

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

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

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

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

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

Problem 2.1 b) cont'd ^{2/3}

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 $\frac{e^{g_i(w)}}{(1 + e^{g_i(w)})^2}$ is lower bounded by zero, and upper bounded by $\frac{1}{4}$.

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

$$\rightarrow \nabla^2 f_i(w) \preceq L I \Leftrightarrow x^T (L I - \nabla^2 f_i(w)) x \geq 0, \forall x, w$$

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

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

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

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

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

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

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

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

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

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

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

~~we~~ $\rightarrow f$ is L -smooth, and we can find L as:

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

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

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

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

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 $\nabla^2 f(w) \succ \mu I$, $\forall w$

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

so we need μ :

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

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

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

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

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

$$= \|\mathbb{E}_{\xi_k} [g(\omega_k; \xi_k)]\|_2^2 \leq c_0^2 \|\nabla f(\omega_k)\|_2^2 \quad \textcircled{1} \text{ to the } 2 \text{ power of}$$

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

the variance is the mean square value:

$$\begin{aligned} \Rightarrow \mathbb{E}_{\xi_k} \left\{ \left[g(\omega_k; \xi_k) - \mathbb{E}\{g(\omega_k; \xi_k)\} \right]^2 \right\} &= \text{var}_{\xi_k} [g(\omega_k; \xi_k)] \\ \text{var}_{\xi_k} [g(\omega_k; \xi_k)] &= \mathbb{E}_{\xi_k} \left\{ \|g(\omega_k; \xi_k)\|_2^2 \right\} + \mathbb{E}_{\xi_k} \left\{ -2g(\omega_k; \xi_k) \cdot \mathbb{E}\{g(\omega_k; \xi_k)\} \right\} \\ &\quad + \mathbb{E}_{\xi_k} \left\{ \|\mathbb{E}\{g(\omega_k; \xi_k)\}\|_2^2 \right\} \\ \text{var}_{\xi_k} [g(\omega_k; \xi_k)] &= \mathbb{E}_{\xi_k} \left\{ \|g(\omega_k; \xi_k)\|_2^2 \right\} - 2 \|\mathbb{E}_{\xi_k} [g(\omega_k; \xi_k)]\|_2^2 \\ &\quad + \|\mathbb{E}_{\xi_k} [g(\omega_k; \xi_k)]\|_2^2 \end{aligned}$$

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

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

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

$$\mathbb{E}_{\delta_k} [\|g(\omega_k; \delta_k)\|_2^2] \leq M + \underbrace{(M_r + C_0^2)}_{\alpha} \|\nabla f(\omega_k)\|_2^2$$

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

Problem 2.3

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

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

and therefore

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

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

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

which is equivalent to

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

$$\text{where } A_k = c - \frac{1}{2} \alpha_k L M G$$

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

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

Then,

$$\begin{aligned} E[f(w_{k+1})] &\leq E[f(w_k)] + \sum_{i=1}^k -c\alpha_i E[\|\nabla f(w_i)\|_2^2] \\ &\quad + \frac{1}{2} \alpha^2 L M_G E[\|\nabla f(w_i)\|_2^2] + \frac{1}{2} \alpha^2 L M \Rightarrow \end{aligned}$$

$$\begin{aligned} \Rightarrow c \sum_{i=1}^k \alpha_i E[\|\nabla f(w_i)\|_2^2] &\leq E[f(w_k)] - E[f(w_{k+1})] \\ &\quad + \sum_{i=1}^k \frac{1}{2} \alpha_i^2 L M_G E[\|\nabla f(w_i)\|_2^2] \\ &\quad + \frac{1}{2} \alpha^2 L M \end{aligned}$$

with α square summable and not summable step size

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

Finally

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