Fundamentals of Machine Learning Over Networks

Group 6
HW 2

HW2 (a):

1) Define $z_i \triangleq y_i x_i \Rightarrow f_i(W) = \ln(1+e^{-W^T z_i})$

$$\nabla f_i(W) = \frac{-e^{-W^T z_i}}{1+e^{-W^T z_i}} z_i = \left(\frac{1}{1+e^{-W^T z_i}} - 1\right) z_i \Rightarrow \|\nabla f_i(W)\|_2 \leqslant \|z_i\|_2$$

$$\nabla f(W) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(W) + 2\lambda W \Rightarrow \|\nabla f(W)\|_2 \leqslant \frac{1}{N}\sum_{i=1}^{N}\|z_i\|_2 + 2\lambda\|W\|_2$$

Hence, if $\|W\|_2 \leqslant D \Rightarrow \|\nabla f(W)\|_2 \leqslant B$ for $\boxed{B = \frac{1}{N}\sum_{i=1}^{N}\|z_i\|_2 + 2\lambda D}$

So, it is Lipschitz

2), 3)

$$\nabla^2 f_i(W) = \frac{1}{1+e^{-W^T z_i}}\left(1 - \frac{1}{1+e^{-W^T z_i}}\right) z_i z_i^T$$

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

$$\lambda_{max}(\nabla^2 f_i(W)) = \frac{1}{1+e^{-W^T z_i}}\left(1 - \frac{1}{1+e^{-W^T z_i}}\right)\|z_i\|_2^2 \leqslant \frac{1}{4}\|z_i\|_2^2$$

$$\lambda_{min}(\nabla^2 f_i(W)) = 0$$

$$\lambda_{max}(\nabla^2 f(W)) \leqslant \frac{1}{N}\sum_{i=1}^{N}\lambda_{max}(\nabla^2 f_i(W)) + 2\lambda \leqslant \frac{1}{4N}\sum_{i=1}^{N}\|z_i\|_2^2 + 2\lambda$$

$$\lambda_{min}(\nabla^2 f(W)) \geqslant \frac{1}{N}\sum_{i=1}^{N}\lambda_{min}(\nabla^2 f_i(W)) + 2\lambda = 2\lambda$$

Hence, f is L-smooth with $\boxed{L = \frac{1}{4N}\sum_{i=1}^{N}\|z_i\|_2^2 + 2\lambda}$

f is μ-strongly convex with $\boxed{\mu = 2\lambda}$

HW2 (b):

$$\mathrm{Var}[g_k(W_k)\,|\,W_k] \triangleq \mathbb{E}[\|g_k(W_k)\|_2^2\,|\,W_k] - \|\mathbb{E}[g_k(W_k)\,|\,W_k]\|_2^2 \underset{\text{from (3)}}{\leqslant} M + M_V \|\nabla f(W_k)\|_2^2$$

$$\Longrightarrow \mathbb{E}[\|g_k(W_k)\|_2^2\,|\,W_k] \leqslant M + M_V \|\nabla f(W_k)\|_2^2 + \|\mathbb{E}[g_k(W_k)\,|\,W_k]\|_2^2$$

$$\overset{\text{from (2b)}}{\leqslant} M + (M_V + c_0^2)\|\nabla f(W_k)\|_2^2$$

Hence, $\boxed{\alpha = M}$, $\boxed{\beta = M_V + c_0^2}$

HW 2 (c) :

$$W_{k+1} = W_k - \alpha_k\, g_k(W_k)$$

f is L-smooth $\Longrightarrow$ $f(W_{k+1}) \leqslant f(W_k) + \nabla f(W_k)^T (W_{k+1} - W_k) + \frac{L}{2} \|W_{k+1} - W_k\|_2^2$

$$= f(W_k) - \alpha_k \nabla f(W_k)^T g_k(W_k) + \alpha_k^2 \frac{L}{2} \|g_k(W_k)\|_2^2$$

Hence,

$$\mathbb{E}[f(W_{k+1}) - f(W_k) | W_k] \leqslant - \alpha_k\ \mathbb{E}[\nabla f(W_k)^T g_k(W_k) | W_k] + \alpha_k^2 \frac{L}{2}\, \mathbb{E}[\|g_k(W_k)\|_2^2 | W_k]$$

from (2a) and HW2(b)
$(M_G \triangleq M_V + C_0^2)$

$$\leqslant - c\,\alpha_k \|\nabla f(W_k)\|_2^2 + \alpha_k^2 \frac{L}{2} \left(M + M_G \|\nabla f(W_k)\|_2^2\right)$$

$$= - \|\nabla f(W_k)\|_2^2\ \alpha_k \left(c - \frac{L M_G}{2} \alpha_k\right) + \alpha_k^2 \frac{LM}{2}$$

if $\forall k$ $\alpha_k \leqslant \frac{c}{LM_G}$ $\leqslant - \|\nabla f(W_k)\|_2^2\ \alpha_k \frac{c}{2} + \alpha_k^2 \frac{LM}{2}$

By summing from $k=0$ to $N-1$, we obtain that

$$\sum_{k=0}^{N-1} \mathbb{E}[f(W_{k+1}) - f(W_k) | W_k] \leqslant - \frac{c}{2} \sum_{k=0}^{N-1} \alpha_k \|\nabla f(W_k)\|_2^2 + \frac{LM}{2} \sum_{k=0}^{N-1} \alpha_k^2$$

$$\Rightarrow \mathbb{E}[f(W_N)] - f(W_0) \leqslant - \frac{c}{2} \sum_{k=0}^{N-1} \alpha_k \mathbb{E}[\|\nabla f(W_k)\|_2^2] + \frac{LM}{2} \sum_{k=0}^{N-1} \alpha_k^2$$

If f has an infimum greater than $-\infty$, the sequence $\mathbb{E}[f(W_N)]$ is always greater that f_{inf}

$$\underbrace{f_{inf} - f(W_0)}_{> -\infty} \leqslant - \frac{c}{2} \sum_{k=0}^{\infty} \alpha_k \mathbb{E}[\|\nabla f(W_k)\|_2^2] + \frac{LM}{2} \underbrace{\sum_{k=0}^{\infty} \alpha_k^2}_{< \infty \text{ becase of square summability}}$$

$$\Rightarrow \sum_{k=0}^{\infty} \alpha_k\, \mathbb{E}[\|\nabla f(W_k)\|_2^2] < \infty \quad \Longrightarrow \quad (11) \checkmark$$

If $\sum_{k=0}^{\infty} \alpha_k = \infty$, then

$$\lim_{N \to 0} \frac{\sum_{k=0}^{N} \alpha_k\, \mathbb{E}[\|\nabla f(W_k)\|_2^2]}{\sum_{k=0}^{N} \alpha_k} = \frac{\sum_{k=0}^{\infty} \alpha_k\, \mathbb{E}[\|\nabla f(W_k)\|_2^2]}{\sum_{k=0}^{\infty} \alpha_k} = 0 \quad \Longrightarrow \quad (12) \checkmark$$