problem 2.1)

a) $f(\omega_2) - f(\omega_1) = \frac{1}{N} \sum_i f_i(\omega_2) - f_i(\omega_1) + \lambda(\|\omega_2\|^2 - \|\omega_1\|^2)$

$$\leq \frac{1}{N} \sum_i f_i(\omega_2) - f_i(\omega_1) + \lambda \|\omega_2 - \omega_1\|^2$$

$$= \frac{1}{N} \sum_i \log \frac{1 + e^{-y_i \omega_2^T x_i}}{1 + e^{-y_i \omega_1^T x_i}} + \lambda \|\omega_2 - \omega_1\|^2$$

convexity of $\log(.)$

$$\leq \log \frac{1}{N} \sum_i \frac{1 + e^{-y_i \omega_2^T x_i}}{1 + e^{-y_i \omega_1^T x_i}} + \lambda \|\omega_2 - \omega_1\|^2$$

$$= \log \frac{1}{N} \sum_i \frac{e^{y_i \omega_1^T x_i} + e^{-y_i (\omega_2 - \omega_1)^T x_i}}{1 + e^{y_i \omega_1^T x_i}} + \lambda \|\omega_2 - \omega_1\|^2$$

$$= \log \frac{1}{N} \sum_i 1 + \frac{e^{-y_i (\omega_2 - \omega_1)^T x_i} - 1}{1 + e^{y_i \omega_1^T x_i}} + \lambda \|\omega_2 - \omega_1\|^2$$

$$\leq \log \frac{1}{N} \sum_i e^{-y_i (\omega_2 - \omega_1)^T x_i} + \lambda \|\omega_2 - \omega_1\|^2 = A$$

Now from Cauchy-schwarz we have:

$$|y_i (\omega_2 - \omega_1)^T x_i| \leq \underbrace{|y_i|}_{=1} \|\omega_2 - \omega_1\|_2 \underbrace{\|x_i\|_2}_{\leq C}$$

$$\leq C \|\omega_2 - \omega_1\|^2$$

so $A < \log \frac{1}{N} \sum_i e^{C \|\omega_2 - \omega_1\|^2} + \lambda \|\omega_2 - \omega_1\|^2$

$$= (\lambda + C) \|\omega_2 - \omega_1\|^2$$

similar steps hold for $f(\omega_1) - f(\omega_2)$. and

$f(\omega_1) - f(\omega_2) \leq (\lambda + C) \|\omega_2 - \omega_1\|^2$

so $|f(\omega_1) - f(\omega_2)| \leq (\lambda + C) \|\omega_2 - \omega_1\|^2$. so $f(.)$ is $(\lambda + C)$-Lipschitz

and $B = \lambda + C$

Scanned with CamScanner

2.1)

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

Since f_i is twice differentiable, smoothness is equivalent to $\nabla^2 f(w) \leq L I$ for some $L > 0$.

$\nabla^2 f_i = x_i x_i^T h''(y_i w^T x_i)$ where $h(z) = \log(1 + e^{-z})$ and $h'(z) = \frac{-e^{-z}}{1+e^{-z}}$ and $h''(z) = \frac{e^{z}}{(1+e^{z})^2}$

Now $h''(z) \leq 1$ so $\delta_{max}(\nabla^2 f_i) \leq \delta_{max}(x_i x_i^T)$

So $\nabla^2 f(w) \leq \underbrace{\delta_{max}(x_i x_i^T)}_{\text{A small suggested } L} I$ so it is L-smooth.

Similarly $f(.)$ is also smooth.

$\nabla f = \frac{1}{N} \sum_i \nabla f_i + 2 \lambda w$ $\nabla^2 f = \frac{1}{N} \sum_i \nabla^2 f_i + 2 \lambda I$

$\delta_{max}(\nabla^2 f) \leq \underbrace{\delta_{max}\left(\frac{1}{n} \sum x_i x_i^T\right) + 2\lambda}_{\triangleq \delta^*}$ $\Rightarrow \nabla^2 f(w) \leq \delta^* I$ so $f(w)$ is δ^*-smooth.

c) Being μ-strongly convex is equivalent to $\nabla^2 f(w) \geq \mu I$ for twice differentiable $f(.)$.

$\nabla^2 f = \frac{1}{N} \sum_i \nabla^2 f_i + 2\lambda I$ note that $\nabla^2 f_i = x_i x_i^T h''(y_i w^T x_i)$ and $h''(z) \geq 0$

so $\min_{\substack{u \\ \|u\|^2 = 1}} u^T \nabla^2 f \, u = \min_{\substack{u \\ \|u\|^2 = 1}} \frac{1}{N} \sum_i u^T x_i x_i^T u \, h''(y_i w^T x_i) + 2\lambda \|u\|^2$

$= \min_{\substack{u \\ \|u\|^2 = 1}} \frac{1}{N} \sum_i \underbrace{(u^T x_i)^2}_{\geq 0} \underbrace{h''(y_i w^T x_i)}_{\geq 0} + 2\lambda$

$\geq 2\lambda$

so $\nabla^2 f(w) \geq \underbrace{2\lambda}_{\to \text{ a large } \mu} I$ so $f(w)$ is 2λ-strongly convex

2.2)

① $\nabla f(\omega_k)^T E_{\zeta_k}[g(\omega_k;\zeta_k)] \geq c \| \nabla f(\omega_k)\|_2^2 \quad \exists c_0 \geq c > 0 \quad \forall k$

② $\| E_{\zeta_k}[g(\omega_k;\zeta_k)]\|_2 \leq c_0 \| \nabla f(\omega_k)\|_2$

③ $\mathrm{Var}_{\zeta_k}[g(\omega_k;\zeta_k)] \leq M + M_V \|\nabla f(\omega_k)\|^2 \quad \exists M, M_V \geq 0 \quad \forall k$

$$E_{\zeta_k}[\|g(\omega_k;\zeta_k)\|^2] = \mathrm{Var}_{\zeta_k}[g(\omega_k;\zeta_k)] + \| E[g(\omega_k;\zeta_k)]\|^2$$

$$\leq M + M_V \|\nabla f(\omega_k)\|_2^2 + c_0^2 \|\nabla f(\omega_k)\|_2^2$$

so $\alpha = M \quad \beta = M_V + c_0^2$

Homework Assignment #2

Problem 2.3

* $\mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w)) \leq -\mu \alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k,\xi_k)\|_2^2]$ (1)

$\leq -(\mu - \frac{1}{2}\alpha_k L M_G)\alpha_k \|\nabla F(w_k)\|_2^2] + \frac{1}{2}\alpha_k^2 K L M$ (2)

since $\alpha_k L M_G \leq \mu$ for all $k \in \mathbb{N}$ by assumption and $\mathbb{E}_{\xi_k}[\|g(w_k,\xi_k)\|_2^2] \leq \alpha + \beta \|\nabla f(w_k)\|_2^2$ from problem (2.2)

* <u>Step size requirements</u>: $\sum_{k=1}^{\infty} \alpha_k = \infty$ (a)
$\sum_{k=1}^{\infty} \alpha_k^2 < \infty$ (b) (3)

* summing both sides of the inequality in (2) for $k \in \{1, \cdots, K\}$ gives

$F_{inf} - \mathbb{E}[F(w_1)] \leq \mathbb{E}[F(w_{K+1})] - \mathbb{E}[F(w_1)]$ (4)

$\leq -\frac{1}{2}\mu \sum_{k=1}^{K} \alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2} L M \sum_{k=1}^{K} \alpha_k^2$ (5)

* dividing by $\mu/2$ and rearranging the terms, we obtain

$\sum_{k=1}^{K} \alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] \leq \frac{2(\mathbb{E}[F(w_1)] - F_{inf}}{\mu} + \frac{LM}{\mu} \sum_{k=1}^{\infty} \alpha_k^2$ (7)

* Let $A_K := \sum_{k=1}^{K} \alpha_k$

* 3(b) implies that the right-hand side of this inequality converges to a finite limit when K increases proving that: $\lim_{K \to \infty} \mathbb{E}[\sum_{k=1}^{\infty} \alpha_k \|\nabla F(w_k)\|_2^2] < \infty$

* 3(a) ensures that $A_K \to \infty$ as $K \to \infty$, proving that

$\mathbb{E}[\frac{1}{A_K} \sum_{k=1}^{\infty} \alpha_k \|\nabla F(w_k)\|_2^2] \xrightarrow{K \to \infty} 0$

References

L. Bottou, F.E. Curtis and J. Nocedal, "Optimization methods for large-scale machine learning," SIAM Review, 2018.