EP3260: Machine Learning Over Networks Homework 2

Stefanos Antaris*[1], Amaru Cuba Gyllensten†[1,2], Martin Isaksson‡[1,2,3], Sarit Khirirat§[1], and Klas Segeljakt¶[1,2]

[1]*KTH Royal Institute of Technology*
[2]*RISE AI*
[3]*Ericsson Research*

February, 2019

Contents

1 Homework assignment

1.1 Human Activity Recognition Using Smartphones

Problem 1.1.1. *Consider logistic ridge regression:*

$$\underset{\boldsymbol{w}}{\text{minimize}} f(\boldsymbol{w}) = \frac{1}{N} \sum_{i \in [N]} f_i(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_2^2 , \tag{1}$$

where

$$f_i(\boldsymbol{w}) = \log(1 + \exp\{-y_i \boldsymbol{w}^\intercal \boldsymbol{x}_i\}) \tag{2}$$

*antaris@kth.se
†amaru.cuba.gyllensten@ri.se
‡martisak@kth.se
§sarit@kth.se
¶klasseg@kth.se

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

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

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

Proof. **Statement 1.**
Lipschitz continuity (bounded gradients)

$$\|\boldsymbol{w}\|_2 \leq D \Rightarrow \left\|\nabla f(\boldsymbol{w})\right\|_2 \leq B$$
$$\|\boldsymbol{w}_1\|_2 , \|\boldsymbol{w}_2\|_2 \leq D \Rightarrow \left|f(\boldsymbol{w}_2) - f(\boldsymbol{w}_1)\right| \leq B\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2$$

Let $\sigma(a)$ be the Sigmoid function

$$\sigma(a) \coloneqq \frac{1}{1+\exp(-a)} = \frac{\exp(a)}{\exp(a)+1} \tag{3}$$

Since

$$\begin{aligned}\nabla f_i(\boldsymbol{w}) &= -y_i\boldsymbol{x}_i/(1+\exp(y_i\boldsymbol{w}^\intercal\boldsymbol{x}_i)),\\ &= -\sigma(y_i\boldsymbol{w}^\intercal\boldsymbol{x}_i)y_i\boldsymbol{x}_i\end{aligned}$$

$$\nabla f(\boldsymbol{w}) = 2\lambda\boldsymbol{w} - \frac{1}{N}\sum_{i\in[N]}\frac{1}{1+e^{y_i\boldsymbol{w}^\intercal\boldsymbol{x}_i}}y_i\boldsymbol{x}_i.$$

By the triangle inequality,

$$\|\nabla f(w)\|_2 \leq 2\lambda\|w\|_2 + \frac{1}{N}\sum_{i\in[N]}\frac{1}{|1+e^{y_i\boldsymbol{w}^\intercal\boldsymbol{x}_i}|}\|y_i\boldsymbol{x}_i\|_2.$$

Assume that $y_i \leq C < \infty$ and $\|y_ix_i\| \leq \tilde{C} < \infty$. Then,

$$\|\nabla f(w)\|_2 \leq 2\lambda D + \tilde{C}.$$

Using the dataset Human Activity Recognition Using Smartphones and finding a w that minimizes $f(\boldsymbol{w})$ we find that $\left\|f(\boldsymbol{w})\right\| \lessapprox 410$ ∎

Proof. **Statement 2.** The Hessian information of f_i is

$$\nabla^2 f_i(w) = \frac{\exp(y_iw^Tx_i)}{(1+\exp(y_iw^Tx_i))^2}x_ix_i^T \leq \frac{1}{4}x_ix_i^T$$

Therefore, the matrix norm of $\nabla^2 f_i(w)$ is

$$\|\nabla^2 f_i(w)\| \leq L,$$

where

$$L \leq \sigma_{\max}\left(\frac{1}{4}x_i x_i^T\right).$$

Here, $\sigma_{\max}(A)$ is the largest eigenvalue of a positive semidefinite matrix A.

Next, we can easily compute the Hessian information of f as follows:

$$\nabla^2 f(w) = \frac{1}{N}\sum_{i=1}^{N} \frac{\exp(y_i w^T x_i)}{(1+\exp(y_i w^T x_i))^2} x_i x_i^T + 2\lambda I \leq \frac{1}{4N}\sum_{i=1}^{N} x_i x_i^T + 2\lambda I.$$

From the definition of the matrix norm, and by the triangle inequality,

$$\|\nabla^2 f(w)\| \leq L$$

where

$$L \leq \sigma_{\max}\left(\frac{1}{4N}\sum_{i=1}^{N} x_i x_i^T + 2\lambda I\right).$$

■

Proof. **Statement 3.**

Define any $v \in \mathbb{R}^d$. Then, it is obvious that

$$v^\intercal x_i x_i^\intercal v = (x_i^\intercal v)^\intercal (x_i^\intercal v) = \|x_i^\intercal v\|^2 \geq 0,$$

for each $x_i \in \mathbb{R}^d$ and $i \in [N]$. This implies that $x_i x_i^\intercal$ is a positive semi-definite matrix. Therefore,

$$\frac{1}{N}\sum_{i=1}^{N} v^\intercal x_i x_i^\intercal v = \frac{1}{N}\sum_{i=1}^{N} \|x_i^\intercal v\|^2 \geq 0.$$

Since $\exp(-x)/(1+\exp(-x))$ is always a positive number, and $(1/N)\sum_{i=1}^{N} x_i x_i^\intercal$ is the positive-semidefinite matrix, the Hessian information on the form

$$\nabla^2 f(w) = \frac{1}{N}\sum_{i=1}^{N} \frac{\exp(y_i w^\intercal x_i)}{(1+\exp(y_i w^\intercal x_i))^2} x_i x_i^\intercal + 2\lambda I$$

implies

$$\begin{aligned} v^T \nabla^2 f(w) v &= \frac{1}{N}\sum_{i=1}^{N} \frac{\exp(y_i w^\intercal x_i)}{(1+\exp(y_i w^\intercal x_i))^2} v^\intercal x_i x_i^T v + 2\lambda v^T v \\ &\geq v^\intercal (2\lambda \cdot I) v. \end{aligned}$$

Thus, $\nabla^2 f(w)$ has the smallest eigenvalue with 2λ, i.e. $\nabla^2 f(w) \geq \mu I$ with $\mu = 2\lambda$. This means that f is strongly convex with $\mu = 2\lambda$. ■

1.2 find α and β

Problem 1.2.1.

$$\mathop{\mathbb{E}}_{\zeta_k}\left[\left\|g(\boldsymbol{w}_k;\zeta_k)\right\|_2^2\right] \le \alpha + \beta\left\|\nabla f(\boldsymbol{w}_k)\right\|_2^2 \tag{4}$$

Proof. Notice that $\mathbb{E}_{\zeta_k} g(\boldsymbol{w}_k;\zeta_k) = \nabla f(w_k)$ and from the definition of the Euclidean norm,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\zeta_k}\left[\left\|g(\boldsymbol{w}_k;\zeta_k)\right\|_2^2\right] &= \mathop{\mathbb{E}}_{\zeta_k}\left[\left\|g(\boldsymbol{w}_k;\zeta_k) - \nabla f(w_k) + \nabla f(w_k)\right\|_2^2\right] \\
&= \mathop{\mathbb{E}}_{\zeta_k}\left\|g(\boldsymbol{w}_k;\zeta_k) - \nabla f(w_k)\right\|^2 - \|\nabla f(w_k)\|^2 \\
&\quad + 2\mathop{\mathbb{E}}_{\zeta_k}\left\langle g(\boldsymbol{w}_k;\zeta_k), \nabla f(w_k)\right\rangle \\
&\le \mathop{\mathbb{E}}_{\zeta_k}\left\|g(\boldsymbol{w}_k;\zeta_k) - \nabla f(w_k)\right\|^2 - \|\nabla f(w_k)\|^2 \\
&\quad + 2\left\|\mathop{\mathbb{E}}_{\zeta_k} g(\boldsymbol{w}_k;\zeta_k)\right\|\left\|\nabla f(w_k)\right\|.
\end{aligned}$$

Since

$$\mathop{\mathbb{E}}_{\zeta_k}\left\|g(\boldsymbol{w}_k;\zeta_k) - \nabla f(w_k)\right\|^2 - \|\nabla f(w_k)\|^2 = \mathrm{Var}_{\zeta_k} g(w_k;\zeta_k) \le M + M_V\|\nabla f(w_k)\|^2$$

$$\mathop{\mathbb{E}}_{\zeta_k}\left\|g(w_k;\zeta_k)\right\| \le c_0\left\|\nabla f(w_k)\right\|$$

we have:

$$\mathop{\mathbb{E}}_{\zeta_k}\left[\left\|g(\boldsymbol{w}_k;\zeta_k)\right\|_2^2\right] \le M + (M_V - 1 + 2c_0)\|\nabla f(w_k)\|^2.$$

Since $c_0 = 1$, we have: $\alpha = M$ and $\beta = M_V + 1$. ∎

1.3 Theorem 5

Theorem 1.3.1. *With square summable but not summable step-size, we have for any* $K \in \mathbb{N}$

$$\mathbb{E}\left[\sum_{k\in[K]} \alpha_k\left\|\nabla f(\boldsymbol{w}_k\right\|_2^2\right] < \infty \tag{5}$$

and therefore

$$\lim_{K\to\infty} \mathbb{E}\left[\frac{1}{\sum_{k\in[K]}\alpha_k}\sum_{k\in[K]} \alpha_k\left\|\nabla f(\boldsymbol{w}_k\right\|_2^2\right] = 0 \tag{6}$$

The expected gradient norm cannot stay bounded away from zero.

Proof. Since $\sum_{k\in[K]} \alpha_k^2 < \infty$ and $\sum_{k\in[K]} \alpha_k = \infty$, we have:

$$\mathbb{E}\left[\frac{1}{\sum_{k\in[K]}\alpha_k}\sum_{k\in[K]}\alpha_k\big\|\nabla f(\boldsymbol{w}_k\big\|_2^2\right] = \frac{1}{\sum_{k\in[K]}\alpha_k}\mathbb{E}\left[\sum_{k\in[K]}\alpha_k\big\|\nabla f(\boldsymbol{w}_k\big\|_2^2\right].$$

Since $\sum_{k\in[K]} \alpha_k = \infty$ and $\mathbb{E}\left[\sum_{k\in[K]}\alpha_k\big\|\nabla f(\boldsymbol{w}_k\big\|_2^2\right] = C < \infty$ for a finite constant C,

$$\begin{aligned}\lim_{K\to\infty}\mathbb{E}\left[\frac{1}{\sum_{k\in[K]}\alpha_k}\sum_{k\in[K]}\alpha_k\big\|\nabla f(\boldsymbol{w}_k\big\|_2^2\right] &= \lim_{K\to\infty}\frac{1}{\sum_{k\in[K]}\alpha_k}\mathbb{E}\left[\sum_{k\in[K]}\alpha_k\big\|\nabla f(\boldsymbol{w}_k\big\|_2^2\right]\\ &= C\lim_{K\to\infty}\frac{1}{\sum_{k\in[K]}\alpha_k} = 0.\end{aligned}$$

■