Fundamentals of Machine Learning Over Networks

Group 6
Descriptions of CA 4

$$f_i(w) \triangleq \ln\left[1+e^{-y_i x_i^T w}\right] \Longrightarrow \nabla f_i(w) = \left[\frac{1}{1+e^{-y_i x_i^T w}} - 1\right] y_i x_i$$

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

$$g_j(w) \triangleq \frac{1}{M'} \sum_{i=1}^{M} \nabla f_{i+(j-1)M'}(w) \quad , \quad \bar{f}_j(w) \triangleq \frac{1}{M'} \sum_{i=1}^{M} f_{i+(j-1)M'}(w) \quad , \quad j=1,\dots,N$$

$\hookrightarrow$ length of the block of dataset

$$G_j^{(k)}(w) \triangleq g_j^{(k)}(w) + Z^{(k)} \quad , \quad Z^{(k)} \text{ are iid } \sim (1-p)\,\delta_m(w) + p\,\frac{1}{(2\pi R)^{\frac{m}{2}}} e^{-\frac{\|w\|^2}{2R}}$$

$$\mathbb{E}\left[G_j^{(k)}(w_k) \mid w_k\right] = g_j^{(k)}(w_k) \Rightarrow G_j \text{ is unbiased} \Rightarrow \boxed{c = c_0 = 1}$$

$$f(w) = \frac{1}{N} \sum_{j=1}^{N} \bar{f}_j(w) + \lambda \|x\|^2 \quad , \quad G^{(k)}(w) = \frac{1}{M'} \sum_{i=1}^{N} g_j(w) \Rightarrow \boxed{\mu \geqslant 2\lambda,\ L \leqslant 2\lambda + \frac{1}{4}\sum y_i^2 \|x_i\|^2}$$

$$\mathbb{E}\left[\|G^{(k)}(w_k)\|^2 \mid w_k\right] = \|g^{(k)}(w_k)\|^2 + p\,\frac{mR}{N} \Rightarrow \boxed{M = \frac{m}{N} pR,\ M_G = 1}$$

Lec 3, Thm. 2 $\Rightarrow \alpha_k = \frac{\beta}{\gamma + k} \Rightarrow \mathbb{E}\left[f(w_k) - f^*\right] \leqslant \frac{\nu' M}{\gamma + k}$ for some β, γ, ν

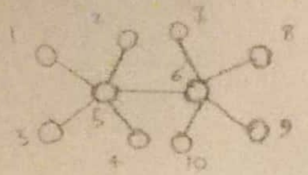
iterations for error less than ε is $\boxed{O\left(\frac{pR}{N}\frac{1}{\varepsilon}\right)}$

In order to make it robust, we propose the following:

If pR is very large: Lec 3, Thm. 4 $\Rightarrow \bar{w}_k \triangleq \frac{1}{k}\sum_{i=1}^{k} w_i \ , \ w_{k+1} = \bar{w}_k - g(\bar{w}_k)$

$$\Rightarrow \#\text{iterations} = \boxed{O\left(\sqrt{\frac{pR}{N}}\,\frac{1}{\varepsilon^2}\right)}$$

* We can also use other variance reduction techniques like SVRG



$*_2 = \frac{1}{5}$, $*_1 = \frac{4}{5}$

$$A \triangleq \begin{bmatrix} *_1 & 0 & 0 & 0 & *_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 & *_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & *_1 & 0 & *_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & *_1 & *_1 & 0 & 0 & 0 & 0 & 0 \\ *_1 & *_1 & *_1 & *_1 & 0 & *_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & *_1 & 0 & *_1 & *_2 & *_2 & *_2 \\ 0 & 0 & 0 & 0 & 0 & *_2 & *_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & *_2 & 0 & *_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & *_2 & 0 & 0 & *_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & *_2 & 0 & 0 & 0 & *_1 \end{bmatrix}$$

$\sigma_2(A) \approx 0.94$

iterations $O\left(N^2 \frac{pR}{N} \frac{1}{\varepsilon \ln \frac{1}{\sigma_2}}\right) = O\left(N pR \frac{1}{\varepsilon \ln \frac{1}{\sigma_2}}\right)$

In order to make it robust, when pR is large, we can do as following:

$\overline{W}_i^{(k)} = \frac{1}{k} \sum_{q=1}^{N} W_i^{(q)}$, $W_i^{(k+1)} = \overline{W}_i^{(k)} - g(\overline{W}_i^{(k)}) \Rightarrow$ #iterations $= O\left(N\sqrt{NpR}\, \frac{1}{\varepsilon \ln \frac{1}{\sigma_2}}\right)$

Since all of the nodes should converge to the optimal point, we should consider the worst node in each case. To this end, one solution is to consider the variance of the noise in each node.

$$\mathbb{E}[(Az)(Az)^T] = \sigma_e^2 AA^T \Rightarrow \mathbb{E}[W^{(k)} | W^{(k-1)}] = AW^{(k-1)} + \sigma_e^2 AA^T$$

where the equality holds since we assume that the noise is iid at the nodes.

Therefore, we should minimize the maximum value on the diag of $A_S A_S^T$, where S is the set of noisy nodes.

By simulating all 6 possible case for our proposed A, we obtained that the best set happens when nodes 5, 6, and another arbitrary node are noiseless.

Intuitively, it is reasonable to make nodes 5 and 6 noiseless; otherwise they propagate noise to the other nodes.