Blank title

The Blank Buccaneers

1 Introduction

Consider the problem

$$\min_{w\in\mathbb{R}^d} -\frac{1}{n}\sum_{i=1}\left(\frac{\hat{y}_i}{\Delta}\log\log\left(1+e^{w^\top x_i}\right)-\log\left(1+e^{w^\top x_i}\right)\right) =: \ell(w), \quad (1)$$

where $\Delta > 0$ is a fixed constant. To ease notation, let $y_i := \hat{y}_i/\Delta$. Let

$$\ell_i(w) := \phi(w^\top x_i), \quad \text{where} \quad \phi_i(\alpha) := \log\left(1+e^{\alpha}\right) - y_i \log\log\left(1+e^{\alpha}\right).$$

2 Bounding second derivative

Here we will show that the second derivative of $f(w)$ and the $f_i(w)$'s has an upper bound.

Lemma 2.1. Let

$$\boldsymbol{X} := [x_1,\ldots,x_n]$$
$$\boldsymbol{D}(a) := a \times \mathrm{diag}(y_1,\ldots,y_n) + \frac{1}{4}\boldsymbol{I}.$$

Numerically we can show (and probably prove with difficulty) that for $a = 0.17$

$$\|\nabla_w^2 \ell_i(w)\| \leq \max_{i=1,\ldots,n}\left\{\|x_i\|^2\left(a\times y_i+\frac{1}{4}\right)\right\} =: L_{\max} \quad (2)$$
$$\|\nabla_w^2 \ell(w)\| \leq\leq \|\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^\top\| =: L. \quad (3)$$

Failing that, we can prove the above for $a = 1.0$

Proof. First note that

$$\nabla_w^2 \ell_i(w) = x_i x_i^\top \phi_i''(w^\top x_i)$$

where

$$\begin{aligned}\phi_i''(\alpha) &= \frac{y_i e^{2\alpha} - y_i e^{\alpha}\log(1+e^{\alpha}) + e^{\alpha}\log^2(1+e^{\alpha})}{(1+e^{\alpha})^2\log^2(1+e^{\alpha})} \\ &= y_i \underbrace{\frac{e^{\alpha}}{(1+e^{\alpha})^2}\left(\frac{e^{\alpha} - \log(1+e^{\alpha}))}{\log^2(1+e^{\alpha})}\right)}_{I} + \underbrace{\frac{e^{\alpha}}{(1+e^{\alpha})^2}}_{II}.\end{aligned}$$

The second part is easier to bound with

$$II \;\leq\; \max_{\beta>0} \frac{\beta}{(1+\beta)^2} = \frac{1}{4}.$$

As for the I, it is much harder to bound. We can show through numeric experiments that

$$I \leq \max_{\beta\in\mathbb{R}} \frac{\beta}{(1+\beta)^2}\left(\frac{\beta - \log(1+\beta))}{\log^2(1+\beta)}\right) \leq 0.17.$$

Indeed, numerically the above has only one stationary point at $\beta \approx 1.64047$ which is a local maxima, and thus the global maxima. Thus in practice I suggest using the bounds

$$\begin{aligned}\|\nabla_w^2 \ell_i(w)\| &\leq\; y_i\|x_i\|^2 \times I + \|x_i\|^2 \times II && (4)\\ &\leq\; \max_{i=1,\dots,n}\left\{\|x_i\|^2\left(0.17 \times y_i + \frac{1}{4}\right)\right\} =: L_{\max} && (5)\end{aligned}$$

Furthermore, let

$$\begin{aligned}\boldsymbol{X} &:= [x_1, \dots, x_n] \\ \Phi(w) &:= \operatorname{diag}(\phi_1''(w^\top x_1), \dots, \phi_n''(w^\top x_n)) \\ \boldsymbol{D} &:= 0.17 \times \operatorname{diag}(y_1, \dots, y_n) + \frac{1}{4}\boldsymbol{I}.\end{aligned}$$

We have that

$$\begin{aligned}\|\nabla_w^2 \ell(w)\| &= \|\boldsymbol{X}\Phi(w)\boldsymbol{X}^\top\| && (6)\\ &\leq \|\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^\top\| =: L, && (7)\end{aligned}$$

which follows since $\Phi(w) \preceq \boldsymbol{D}$. To compute the right hand side of (7), I recommend a few steps of the power method. This way, you need only implement a function that computes the matrix vector product $v \mapsto \boldsymbol{X}\Phi(w)\boldsymbol{X}^\top v$, without ever forming the matrix $\boldsymbol{X}\Phi(w)\boldsymbol{X}^\top$.

For a more algebraic proof, we have the following looser bound on I.

Lemma 2.2.
$$I \leq 1.0$$

Proof. First note that

$$\begin{aligned} I &\leq \frac{e^{2\alpha}}{(1+e^\alpha)^2}\frac{1}{\log^2(1+e^\alpha)} \\ &\leq \max_{\beta\geq 0} \frac{\beta^2}{(1+\beta)^2}\frac{1}{\log^2(1+\beta)} \leq 1, \end{aligned}$$

where it remains to prove that the above is less than 1. The proof is as follows. First, we show that the function

$$g(\beta) := \frac{\beta^2}{(1+\beta)^2}\frac{1}{\log^2(1+\beta)}$$

has only on stationary point at $\beta \to \infty$. At this limit we can show that

$$\lim_{\beta\to\infty} g(\beta) = 0.$$

The other candidate for the global maxima is at the boundary of our constraint set, which is $\beta = 0$, for which we can show that

$$\lim_{\beta\to 0} g(\beta) = 1.$$

Consequently the maxima is attained at $\beta = 0$. □

□

3 Lower bounds

Consider instead the more general model given by

$$\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}\left(f(w^\top x_i) - y_i \log\left(f(w^\top x_i)\right)\right) =: \ell(w). \tag{8}$$

Let $u_i = f(w^\top x_i) > 0$ and note that $u_i \mapsto u_i - y_i \log(u_i)$ is convex for $y_i \geq 0$. Thus we can obtain a lower bound:

$$f(w^\top x_i) - y_i \log(f(w^\top x_i)) \geq \min_{u>0}[u - y_i \log(u)] \geq y_i - y_i \log(y_i) \quad (9)$$

where we define $y_i \log(y_i) = 0$ for $y_i = 0$. Above we solved the minimization over u explicitly by setting the derivative with respect to u equal to zero and finding that the minimum is attained when $u = y_i$.

Thus a lower bound on the full batch objective function is

$$\ell(w) \geq \frac{1}{n} \sum_{i=1}^{n} f(w^\top x_i) - y_i \log(f(w^\top x_i) \geq \frac{1}{n} \sum_{i=1}^{n} y_i - y_i \log(y_i)$$

4 Algorithms

4.1 SAGA with optimal step size

The following version of SAGA is taken from [GGS19]. This implementation makes use of both large step size set using the smoothness constants, and the structure of a generalized linear model. That is, the stochastic gradients of (8) are always combinations of the features vectors $x_1, \ldots, x_n$. That is,

$$\nabla \ell(w) = \frac{1}{n} \sum_{i=1}^{n} x_i f'(\mathbf{w}_k^\top x_i) \left(1 - \frac{y_i}{f(\mathbf{w}_k^\top x_i)}\right) =: \frac{1}{n} \sum x_i z_i$$

where $z_i = f'(\mathbf{w}_k^\top x_i) \left(1 - \frac{y_i}{f(\mathbf{w}_k^\top x_i)}\right)$. Because of this structure, we need only store and update the values of the vector $\mathbf{z} = [z_1, \ldots, z_n] \in \mathbb{R}^n$.

Algorithm 1: `SAGA`: step size as function of mini-batch

Input: Input:

1 $\mathbf{w}_0 \in \mathbb{R}^d$, batch size $b \in [n]$, and smoothness $L > 0$ and $L_{\max}$.

2 **Initiate:** full batch gradient $\bar{g} = \nabla \ell(\mathbf{w}_0)$, batch smoothness

$$L(b) := L \times \frac{n}{b}\frac{b-1}{n-1} + L_{\max} \times \frac{1}{b}\frac{n-b}{n-1}$$

step size $\gamma = \frac{1}{4}\frac{1}{L(b)}$ and scalar derivative table

$$\mathbf{z} = \left[f'(\mathbf{w}_0^\top x_1)\left(1 - \frac{y_1}{f(\mathbf{w}_0^\top x_1)}\right), \ldots, f'(\mathbf{w}_0^\top x_n)\left(1 - \frac{y_n}{f(\mathbf{w}_0^\top x_n)}\right)\right] \in \mathbb{R}^n$$

3 . **for** $k = 1$ ***to*** $K - 1$ **do**

4   Sample batch $B \subset n$ with $|B| = b$

5   **for** $i \in B$ **do**

6         $\hat{z}_i = f'(\mathbf{w}_k^\top x_i)\left(1 - \frac{y_i}{f(\mathbf{w}_k^\top x_i)}\right)$

7     $g_k = \bar{g} + \frac{1}{b}\sum_{i \in B}^n x_i\,(\hat{z}_i - z_i)$

8     $\mathbf{w}_{k+1} = \mathbf{w}_k - \gamma g_k$

9   **for** $i \in B$ **do**

10         $\bar{g} = \bar{g} - \frac{1}{n}z_i + \frac{1}{n}\hat{z}_i$;

11         $z_i = \hat{z}_i$

Output: $\mathbf{w}_K$

In particular, for a batch of data $B \subset [n]$, and batch gradient $\nabla \ell_B(w)$, the SAGA gradient estimator given by

$$g_k = \frac{1}{n}\sum x_i z_i + \nabla \ell_B(w) - \frac{1}{b}\sum_{i \in B} x_i z_i$$

see line 7 in Algorithm 1.

5 Templates

Theorem 5.1. I like my theorems like this, with boldface vectors $\mathbf{x}$ and matrices $\boldsymbol{A}$

It's also nice to use clever referencing like **??** or **??**.

6 Population GLM

In NeMoS we also allow to fit jointly a population of neurons. In this case, the counts are y_{ij}, where i indexes time and j neurons, and we have one set of weights per neuron:

$$\ell_i(w_1, \dots, w_n) = \sum_j \phi(x_i w_j).$$

Robert: Does this mean the full batch objective is given by

$$\min_w \frac{1}{n} \sum_{i=1}^{n} \ell_i(w_1, \dots, w_n) \;=\; \sum_j \frac{1}{n} \sum_{i=1}^{n} \phi(x_i w_j)$$

The Hessian over the vector $w = [w_1, \dots, w_n]$ is therefore block diagonal,

$$\nabla_w^2 \ell_i(w) = \begin{bmatrix} \nabla_{w_1}^2 \ell_i(w_1) & 0 & \cdots & 0 \\ 0 & \nabla_{w_2}^2 \ell_i(w_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \nabla_{w_n}^2 \ell_i(w_n). \end{bmatrix}$$

In this case,

$$\begin{aligned} \|\nabla_w^2 \ell_i(w)\| &= \max_j \|\nabla_{w_j}^2 \ell_i(w_j)\| \\ &\le \max_j \left(y_{ij} \|x_i\|^2 \times I + \|x_i\|^2 \times II \right) \\ &\le \|x_i\|^2 \left(0.17 \times \max_j (y_{ij}) + \frac{1}{4} \right) \end{aligned}$$

References

[GGS19] N. Gazagnadou, R. M. Gower, and J. Salmon. “Optimal mini-batch and step sizes for SAGA”. In: *International Conference on Machine Learning*. 2019.