Project 1: Report

Justin Baker, Eric Brown, Trent DeGiovanni,
Edward Gu, Rebecca Hardenbrook

October 14, 2021

Consider training the following regularized logisti regression model

$$\min_x F(x) := f(x) + \lambda R(x)$$

where

$$f(x) = \frac{1}{2n}\sum_{i=1}^{n} \log(1 + \exp(-b_i a_i^T x)),$$

with n being the sample size and $a_i \in \mathbb{R}^d$ ($d = 50$) is the training data, $b_i \in \{-1, 1\}$ be the label of a_i. Here, we consider two different regularization functions i.e. ℓ_1-regularization ($R(x) = ||x||_1$) and ℓ_2-regularization ($R(x) = ||x||_2^2$).
Please use the code in the zip file to genreate 1000 data-label pairs $\{a_i, b_i\}_{i=1}^{1000}$.

1. Derive $prox_{\lambda||x||_1}$ and $prox_{\lambda||x||_2}$.

Solution:

By definition

$$prox_{\lambda h(x)} = \text{argmin}_v \{h(v) + \frac{1}{2\lambda}||x - v||_2^2\}$$

Let $h(x) = ||x||_1$

$$prox_{\lambda||x||_1} = \text{argmin}_v \{||v||_1 + \frac{1}{2\lambda}||x - v||_2^2\}$$

With insight we anticipate that the optimum of mixed $\ell_1 - \ell_2$ norms is given by the soft-threshold or shrinkage operator.

For this problem we can use an extension of the optimality coniditons to subdifferentiable functions.

$$0 \in \partial_v F = \partial_v[||v||_1 + \frac{1}{2\lambda}||v - x||_2^2]$$
$$0 \in \partial_v F = \partial_v||v||_1 + \frac{1}{2\lambda}\partial_v||v - x||_2^2$$
$$0 \in \partial_v F = \partial_v||v||_1 + \frac{1}{2\lambda}\nabla||v - x||_2^2$$
$$0 \in \lambda\partial_v||v||_1 + v - x$$

Now we consider the subdifferential for ℓ_1 component wise.

$$\partial_v ||v||_1 = \begin{cases} \text{sign}(v_i) & \text{for } v_i \neq 0 \\ [-1, 1] & \text{for } v_i = 0 \end{cases}$$

Analyzing both cases we have the following.

$$\begin{cases} 0 = v_i^* - x + \lambda \text{sign}(v_i^*) & v_i \neq 0 \\ 0 \in x + \lambda[-1, 1] & v_i = 0 \end{cases}$$

Solving for the minimizer v^* in terms of x.

$$\begin{cases} v_i^* = x - \lambda \text{sign}(v_i^*) & v_i \neq 0 \\ x \in \lambda[-1, 1] & v_i = 0 \end{cases}$$

From the first condition we see that if $v_i^* \leq 0$ then $x \leq 0$ (notice that $\lambda > 0$).

$$0 > v^* = x + \lambda$$

Similarly for $v^* > 0$ the $x > 0$.

$$0 < v^* = x - \lambda$$

Now using the fact that x and v^* have similar signs we may write the solution for v^* exclusively in terms of x.

$$v_i = \begin{cases} 0 & x \in [-\lambda, \lambda] \\ x - \lambda \text{sign}(x) & \text{otherwise} \end{cases}$$

This is exactly the shrinkage operator we anticipated to find.

Now consider $prox_{\lambda ||x||_2} x$

Again by definition

$$prox_{\lambda h(x)} = \text{argmin}_v \{h(v) + \frac{1}{2\lambda} ||x - v||_2^2\}$$

Let $h(x) = ||x||_2$

$$prox_{\lambda ||x||_2} = \text{argmin}_v \{||v||_2 + \frac{1}{2\lambda} ||x - v||_2^2\}$$

In this instance the function is differentiable everywhere.

$$0 = \nabla[||v||_2 + \frac{1}{2\lambda} ||v - x||_2^2]$$

$$
\begin{aligned}
0 &= \nabla ||v||_2 + \frac{1}{2\lambda}\nabla ||v - x||_2^2 \\
0 &= \frac{1}{2}v + \frac{1}{\lambda}v - x \\
(\frac{1}{2} + \lambda^{-1})v &= \lambda^{-1}x \\
v &= \frac{2}{2+\lambda}x
\end{aligned}
$$

Thus the optimal value si given by $v^* = \frac{2}{2+\lambda}x$.

2. For $\lambda = 0.001$, numerically solve the problem $\min_x F(x)$ using subgradient method, proximal gradient method, accelerated proximal gradient method with heavy-ball momentum and Nesterov's acceleration. Plot $F(x^k) - F(x^*)$ over the iteration k for each method, where x^* is in the code that used to generate the training data.

3. Test different λ, e.g. $0.005, 0.01, 0.05, 0.1$ and see how x^k changes after you run enough number of iterations.

4. Can you propose any approach to further accelerate the training process?