Management Sciences Topics: Convex Optimization
Final Project

1 Problem setup

We need to solve the optimization problem of a one-hidden-layer neural network

$$\min_{x_k \in \mathbb{R}^d, y_k \in \mathbb{R}, z \in \mathbb{R}^K, w \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(b_i w + b_i \sum_{k=1}^{K} \sigma(a_i^\top x_k + y_k) z_k \right), \tag{1.1}$$

where K is the number of neurons, $a_i \in \mathbb{R}^d$ is a data point, $b_i \in \{-1, 1\}$ is the class label of a_i, $\sigma(z) = \max(z, 0)$ or $\frac{\exp(z)}{1+\exp(z)}$, $\mathcal{L}(z) = \max(1 - z, 0)$ or $\log(1 + \exp(-z))$.

2 Stochastic subgradient method

We first consider the subgradient with respect to each variables.

First, define the variables

$$X = [x_1, x_2, \ldots, x_K] \in \mathbb{R}^{d \times K}, \quad Y = [y_1, y_2, \ldots, y_k]^\top \in \mathbb{R}^{K \times 1}, \quad Z = [z_1, z_2, \ldots, z_k]^\top \in \mathbb{R}^{K \times 1}. \tag{2.1}$$

Then a forward pass through the network can be written as

$$\begin{aligned} A_1 &= AX \oplus Y^\top, \\ A_2 &= \sigma(A_1) Z \oplus w \\ f &= \frac{1}{n} 1^\top L(b \odot A_2). \end{aligned} \tag{2.2}$$

By chain rule, the subgradients of f with respect to each variable are

$$\begin{aligned} \frac{\partial f}{\partial A_2} &= \frac{1}{n} L'(b \odot A_2) \odot b \\ \frac{\partial f}{\partial w} &= \frac{\partial f}{\partial A_2} \frac{\partial A_2}{\partial w} = \text{rowsum}(\frac{\partial f}{\partial A_2} \odot 1) = \left(\frac{\partial f}{\partial A_2} \right)^\top 1, \\ \frac{\partial f}{\partial Z} &= \frac{\partial f}{\partial A_2} \frac{\partial A_2}{\partial Z} = \sigma(A_1)^\top \frac{\partial f}{\partial A_2}, \\ \frac{\partial f}{\partial A_1} &= \frac{\partial f}{\partial A_2} \frac{\partial A_2}{\partial A_1} = \frac{\partial f}{\partial A_2} Z^\top \odot \sigma'(A_1), \\ \frac{\partial f}{\partial Y} &= \frac{\partial f}{\partial A_1} \frac{\partial A_1}{\partial Y} = \left(\frac{\partial f}{\partial A_1} \right)^\top 1, \\ \frac{\partial f}{\partial X} &= \frac{\partial f}{\partial A_1} \frac{\partial A_1}{\partial X} = A^\top \frac{\partial f}{\partial A_1}. \end{aligned} \tag{2.3}$$

The stochastic subgradients can be chosen to be the subgradient when input is a minibatch of the whole dataset, i.e.

$$G(x, \xi_i) = \partial_x (x; A_{\xi_i}, b_{\xi_i}) \tag{2.4}$$

for each variable x, where ξ_i is a uniformly sample index set for each i.

3 Accelerated proximal gradient method

In order to use APG for this problem, we need to choose

$$\sigma(z) = \frac{\exp(z)}{1+\exp(z)}, \quad \mathcal{L}(z) = \log(1+\exp(-z)) \tag{3.1}$$

to guarantee the objective function is smooth. Notice

$$\mathcal{L}'(z) = -\frac{1}{1+\exp(z)}, \tag{3.2}$$

and

$$|\mathcal{L}''(z)| = \left|\frac{e^z}{(1+e^z)^2}\right| \leq \frac{1}{4}, \tag{3.3}$$

by Lagrange mean value theorem, we know the Lipschitz constant for $\mathcal{L}'$ is $L = \frac{1}{4}$.

Now let's consider the Lipschitz constant for each derivatives. For ∂_{A_2},

$$\|\partial_{A_2} \tag{3.4}$$

For $\frac{\partial f}{\partial w}$,

$$\begin{aligned} |\partial_w(w_1) - \partial_w(w_2)| &= \frac{1}{n}\sum_{i=1}^{n} b_i(\mathcal{L}'(b_i(\sigma(A_1)Z)_i + b_i w_1) - \mathcal{L}'(b_i(\sigma(A_1)Z)_i + b_i w_2) \\ &\leq \frac{1}{n}\sum_{i=1}^{n} b_i(\frac{1}{4}b_i|w_1 - w_2|) = \frac{1}{4n}\sum_{i=1}^{n}|w_1 - w_2| \\ &= \frac{1}{4}|w_1 - w_2|. \end{aligned} \tag{3.5}$$

Then the Lipschitz constant for ∂_w is $L_w = \frac{1}{4}$.