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

First, we consider the subgradient with respective to each variables:

$$\begin{aligned}
\partial_w f &= \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}'\left(b_i w + b_i \sum_{k=1}^{K} \sigma(a_i^\top x_k + y_k) z_k \right) b_i, \\
\partial_{z_k} f &= \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}'\left(b_i w + b_i \sum_{i=1}^{K} \sigma(a_i^\top x_k + y_k) z_k \right) b_i \sigma(a_i^\top x_k + y_k), \\
\partial_{y_k} f &= \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}'\left(b_i w + b_i \sum_{i=1}^{K} \sigma(a_i^\top x_k + y_k) z_k \right) b_i \sigma'(a_i^\top x_k + y_k) z_k, \\
\partial_{x_k} f &= \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}'\left(b_i w + b_i \sum_{i=1}^{K} \sigma(a_i^\top x_k + y_k) z_k \right) b_i \sigma'(a_i^\top x_k + y_k) z_k a_i,
\end{aligned} \tag{2.1}$$

where σ' and $\mathcal{L}'$ are both subgradients when the functions are not differentiable.

We may vectorize the computation. First, define the parameter vectors:

$$X = [x_1, x_2, \ldots, x_K] \in \mathbb{R}^{d \times K}, \quad Y = [y_1, y_2, \ldots, y_k]^\top \in \mathbb{R}^{K \times 1}, \quad Z = [z_1, z_2, \ldots, z_k]^\top \in \mathbb{R}^{K \times 1}, \tag{2.2}$$

and

$$W_i(x, y, z, w) = b_i w + b_i \sum_{k=1}^{K} \sigma(a_i^\top x_k + y_k) z_k = b_i w + b_i \sigma(a_i^\top X \oplus Y^\top) Z \in \mathbb{R}. \tag{2.3}$$

Here $\oplus, \odot$ denote pointwise operations. Let

$$\begin{aligned}
R_i &= a_i^\top X \oplus Y^\top \in \mathbb{R}^{1 \times K}, \\
R &= [R_1^\top, \ldots, R_n^\top]^\top \in \mathbb{R}^{n \times K}, \\
W &= [W_1, \ldots, W_n]^\top \in \mathbb{R}^{n \times 1},
\end{aligned} \tag{2.4}$$

then

$$
\begin{aligned}
R &= AX \oplus Y^\top \\
W &= b \odot (w + \sigma(R)Z),
\end{aligned}
\tag{2.5}
$$

and the target function and subgradients can be written as

$$
\begin{aligned}
f &= \frac{1}{n} 1^\top \mathcal{L}(W), \\
\partial_w f &= \frac{1}{n} b^\top \mathcal{L}'(W), \\
\partial_z f &= \frac{1}{n} \sigma(AX \oplus Y^\top)^\top (\mathcal{L}'(W) \odot b) = \frac{1}{n} \sigma(R)^\top (\mathcal{L}'(W) \odot b), \\
\partial_y f &= \frac{1}{n} (\sigma'(R)^\top (\mathcal{L}'(W) \odot b)) \odot Z, \\
\partial_x f &= \frac{1}{n} (A^\top (\mathcal{L}'(W) \odot b \odot \sigma'(R))) \odot Z^\top.
\end{aligned}
\tag{2.6}
$$

We use an iterative scheme to update the four variables.