HW5: Question 3.1

Ian Dover

March 2023

Solve the following optimization problem using the proximal gradient descent:

$$f(\theta) = \min_{\theta \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \left[log(1 + e^{x_i \theta}) - y_i x_i \theta \right] + \frac{\lambda_2}{2} ||\theta||_2^2 + \lambda_1 ||\theta||_1$$

Where $\log(\cdot)$ is the natural logarithm; $x_i \in \mathbb{R}^{1 \times d}$ is sample i;
$\theta \in \mathbb{R}^d$, $y_i \in \{0, 1\}$ is the label for sample i and $\lambda > 0$.

1 Part 1

For proximal gradient descent, the function to minimize will be separated into a g(x) and h(x), with g(x) being differentiable and h(x) not:

$$g(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[log(1 + e^{x_i \theta}) - y_i x_i \theta \right] + \frac{\lambda_2}{2} ||\theta||_2^2$$

$$h(\theta) = \lambda_1 ||\theta||_1$$

Let us find the partial of $g(\theta)$ with respect to θ. This can also be considered the gradient (∇) with respect to θ:

$$\nabla_\theta = \frac{1}{m} \sum_{i=1}^{m} \left[\frac{x_i^T e^{x_i \theta}}{e^{x_i \theta} + 1} - (y_i x_i)^T \right] + \lambda_2 ||\theta||_2$$

Now we can drive the update statement for θ:

$$\theta_{k+1} = \theta_k - t \nabla_{\theta_k}$$

Given that $\theta_{k+1} \in \mathbf{R}^d$, we must perform soft-thresholding on each element of θ_{k+1}. The soft-thresholding is defined by:

$$[S_{\lambda t}(\theta_{k+1})]_i = \begin{cases} \theta_i - t\lambda_1 & \theta_i > t\lambda_1 \\ 0 & -t\lambda_1 \leq \theta_i \leq t\lambda_1 \\ \theta_i + t\lambda_1 & \theta_i < -t\lambda_1 \end{cases}$$