Homework 1 - Extra Credit Proof

Haoli Yin, CS 6362

September 14, 2024

Show that for logistic regression, the Lipschitz constant for the gradient of the NLL loss gives the matrix norm of $X^T X$, with X being the data matrix.

Let $X \in \mathbb{R}^{n \times d}$ be the data matrix and $y \in \mathbb{R}^n$ be the vector of labels, such that $y_i \in \{0, 1\}$. The NLL of the logistic loss function is defined as:

$$L(w) = -[y^T \log(p(w)) + (1-y)^T \log(1 - p(w))]$$

where $p(w) = \sigma(Xw)$ and $\sigma(t) = 1/(1 + e^{-t})$.

The gradient of the loss function with respect to w is:

$$\nabla L(w) = X^T(p(w) - y) = X^T[\sigma(Xw) - y].$$

To show that $\nabla L(w)$ is Lipschitz continuous and determine the Lipschitz constant, we invoke the Mean Value Theorem (MVT) for vector-valued functions.

Consider two parameter vectors $w, w' \in \mathbb{R}^d$. The difference in gradients would be:

$$\nabla L(w) - L(w') = X^T[\sigma(Xw) - \sigma(Xw')].$$

By applying the MVT component-wise to the function σ, for each component $i \in \{i, \dots, n\}$, there exists θ_i between $x_i^T w$ and $x_i^T w'$ such that:

$$\sigma(x_i^T w) - \sigma(x_i^T w') = \sigma^{'}(\theta_i)(x_i^T w - x_i^T w')$$

where x_i^T is the i-th row of X, and $\sigma'(\theta_i) = \sigma(\theta_i)(1 - \sigma(\theta_i))$ is the derivative of σ evaluated at θ_i.

Let $D(\gamma)$ be a diagonal matrix with entries $\sigma'(\theta_i)$ and γ represents the vector of θ_i's. Then, we can rewrite:

$$\sigma(Xw) - \sigma(Xw') = \mathcal{D}(\gamma)X(w - w')$$

Substituting back into the gradient difference:

$$\nabla L(w) - \nabla L(w') = X^T \mathcal{D}(\gamma) X (w - w').$$

Taking norms on both sides and using the sub-multiplicative property of norms:

$$\nabla L(w) - \nabla L(w') = X^T \mathcal{D}(\gamma) X(w - w') \leq X^T \mathcal{D}(\gamma) X \cdot w - w'$$

Since $\sigma'(\theta_i) \leq \frac{1}{4}$ for all θ_i (because $\sigma(t)(1 - \sigma(t)) \leq \frac{1}{4}$ for all $t \in \mathbb{R}$), the norm of $\mathcal{D}(\gamma)$ satisfies $D(\gamma) \leq \frac{1}{4}$. Therefore:

$$|\nabla L(w) - \nabla L(w')| \leq \left| X^T \left(\frac{1}{4} I \right) X \right| \cdot |w - w'| = \frac{1}{4} |X^T X| \cdot |w - w'|.$$

Thus, the gradient $\nabla L(w)$ is Lipschitz continuous with Lipschitz constant $L = \frac{1}{4}|X^T X|$, where $|X^T X|$ denotes the matrix norm.