CSCI 145 Problem 14 Solution: Does Gradient Descent Implicitly Regularize?

Solution

1. Iterates Stay in the Input Span

Let $S = \text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Differentiating the loss gives

$$\nabla \mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i)\mathbf{x}_i.$$

Thus every gradient is a linear combination of the training inputs and belongs to S. The initial iterate $\mathbf{w}^{(0)} = \mathbf{0}$ also belongs to S. If $\mathbf{w}^{(t)} \in S$, then subtracting $\alpha \nabla \mathcal{L}(\mathbf{w}^{(t)}) \in S$ leaves $\mathbf{w}^{(t+1)} \in S$. Induction proves

$$\mathbf{w}^{(t)} \in S \qquad \text{for every } t.$$

This argument holds for every positive learning rate, even if the iterates do not converge.

2. Comparing Norms

Fix an iteration t and suppose $\mathbf{v}$ makes the same predictions as $\mathbf{w}^{(t)}$. Put $\mathbf{z} = \mathbf{v} - \mathbf{w}^{(t)}$. For each training example,

$$\mathbf{z}^\top \mathbf{x}_i = \mathbf{v}^\top \mathbf{x}_i - (\mathbf{w}^{(t)})^\top \mathbf{x}_i = 0.$$

Therefore $\mathbf{z}$ is perpendicular to every vector in S, including $\mathbf{w}^{(t)}$. By the Pythagorean theorem,

$$\|\mathbf{v}\|_2^2 = \|\mathbf{w}^{(t)} + \mathbf{z}\|_2^2 = \|\mathbf{w}^{(t)}\|_2^2 + \|\mathbf{z}\|_2^2 \geq \|\mathbf{w}^{(t)}\|_2^2.$$

Taking square roots proves the requested inequality. Equality holds only when $\mathbf{z} = \mathbf{0}$, so the smallest-norm vector with these predictions is in fact unique.

The comparison fixes the training predictions at the current iteration. Those predictions may still have positive training error, and the weight norm may grow from one iteration to the next.

Zero initialization and the directions available to gradient descent produce this preference for the smallest-norm fit.