Does Gradient Descent Implicitly Regularize?

Due: Wednesday, October 28 at 11:59pm

Instructions: Please submit your *handwritten* solutions as a single PDF on Gradescope.

Many weight vectors can make the same training predictions. For inputs $\mathbf{x}_i \in \mathbb{R}^d$ and labels $y_i \in \mathbb{R}$, train $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ by minimizing

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{n} \big(\mathbf{w}^\top \mathbf{x}_i - y_i\big)^2.$$

Starting at $\mathbf{w}^{(0)} = \mathbf{0}$, apply gradient descent with learning rate $\alpha > 0$:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \alpha \nabla \mathcal{L}(\mathbf{w}^{(t)}).$$

1. Each gradient determines an update direction. Show that $\nabla \mathcal{L}(\mathbf{w})$ lies in $\operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ for every $\mathbf{w} \in \mathbb{R}^d$.
2. Gradient descent starts at $\mathbf{w}^{(0)} = \mathbf{0}$. Prove that $\mathbf{w}^{(t)} \in \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ for every integer $t \geq 0$.
3. Different weights can make the same training predictions. Prove that every iterate has the smallest Euclidean norm among all weights with the same training predictions:

$$\mathbf{v}^\top \mathbf{x}_i = (\mathbf{w}^{(t)})^\top \mathbf{x}_i \quad \text{for } i = 1, \ldots, n \quad \Longrightarrow \quad \|\mathbf{w}^{(t)}\|_2 \leq \|\mathbf{v}\|_2.$$

Hint: Write $\mathbf{v} = \mathbf{w}^{(t)} + \mathbf{z}$ and use the matching predictions to show that $\mathbf{z}$ is orthogonal to the input span.

Takeaway: *For this linear model, gradient descent started at zero always finds weights with the smallest norm among those giving the same training predictions. If it reaches a perfect fit, that fit has the smallest weight norm. Gradient descent can favor small weights without an explicit penalty.*