How Does Depth Change the Regularized Fit?

Due: Wednesday, October 14 at 11:59pm

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

Two linear layers can represent exactly the same functions as one, but use different parameters. Consider the models

$$f_w(x) = wx, \qquad f_{a,b}(x) = abx, \qquad w, a, b \in \mathbb{R}.$$

On the single training example $(x, y) = (1, 1)$, use squared-error loss and L_2 *regularization*, which penalizes the sum of squared weights:

$$\begin{aligned} \mathcal{L}_1(w) &= \frac{1}{2}(w-1)^2 + \frac{\lambda}{2}w^2, \\ \mathcal{L}_2(a,b) &= \frac{1}{2}(ab-1)^2 + \frac{\lambda}{2}(a^2+b^2), \end{aligned} \qquad \lambda > 0.$$

1. First consider the single linear layer. For every $\lambda > 0$, determine the globally optimal prediction $f_w(1)$.
2. Now consider the two-layer model. For every $\lambda > 0$, determine the globally optimal prediction $f_{a,b}(1)$.
3. Initialize both layer weights at zero. For which $\lambda > 0$ is this initialization stationary but not globally optimal?

Takeaway: *Both models can fit the same functions. Adding a layer changes the best fit when we penalize the weights. With two layers, gradient descent can also get stuck at zero even when better weights exist.*