CSCI 145 Problem 11: Depth and Regularization

October 27, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on Wednesday, October 14**.

- Turn in a single PDF of *handwritten* solutions. Typed (including LaTeX) solutions will not be accepted.
- A complete, good-faith attempt earns submission credit. The grader may select any recent assigned problem for a brief conversation, so keep your work and be prepared to explain each solution and answer a related question. Scheduling details will be announced in class.

Problem 11: Depth and Regularization

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. For every $\lambda > 0$, find and compare the optimal predictions $f_w(1)$ and $f_{a,b}(1)$. Prove that each is globally optimal.
2. Compute both gradients. Does full-batch gradient descent move when all weights start at zero?
3. For which values of λ is the zero initialization stationary but not globally optimal in each model?

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.*