CSCI 145 Problem 12: Why Doesn't SGD Converge to the Minimum?

October 27, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on Wednesday, October 21**.

- 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 12: Why Doesn't SGD Converge to the Minimum?

Consider two training examples with losses

$$\ell_1(w) = \frac{1}{2}(w-1)^2, \qquad \ell_2(w) = \frac{1}{2}(w+1)^2.$$

At each iteration, stochastic gradient descent (SGD) chooses one example uniformly at random, independently of earlier choices, and takes a gradient step with fixed learning rate $0 < \alpha < 2$.

1. Find the minimizer of the average loss $\mathcal{L}(w) = (\ell_1(w) + \ell_2(w))/2$. If SGD reaches this minimizer, will it stay there? Explain.
2. Suppose SGD starts at $w^{(0)} = 0$. Show that $\mathbb{E}[w^{(t)}] = 0$ for every t.
3. Find

   $$\lim_{t\to\infty} \mathbb{E}\big[(w^{(t)})^2\big].$$

   Does SGD converge to the minimizer? Explain what happens to the limiting expected squared distance as the learning rate approaches zero.

Takeaway: *The average gradient can be zero while individual examples still push the weight in different directions. With a fixed learning rate, SGD keeps fluctuating around the minimum. A smaller learning rate reduces these fluctuations.*