Why Doesn't SGD Converge to the Minimum?

Due: Wednesday, October 21 at 11:59pm

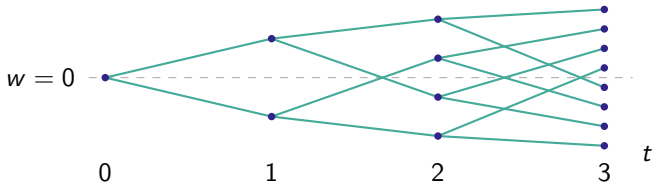
Instructions: Please submit your *handwritten* solutions as a single PDF on Gradescope.

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 $0 < \alpha < 2$. The average loss is $\mathcal{L}(w) = (w^2+1)/2$, uniquely minimized at $w = 0$.

The branches show the first three steps for $\alpha = 1/2$, starting at the minimizer.



1. Suppose SGD starts at $w^{(0)} = 0$. Show that $\mathbb{E}[w^{(t)}] = 0$ for every integer $t \geq 0$.
2. The squared distance from the minimizer is $(w^{(t)})^2$. Find $\lim_{t\to\infty} \mathbb{E}[(w^{(t)})^2]$.
3. Keep the learning rate α fixed. Does SGD converge to the minimizer?

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.*