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

Solution

1. SGD Leaves the Minimum

The average loss and its derivative are

$$\mathcal{L}(w) = \frac{1}{4}(w-1)^2 + \frac{1}{4}(w+1)^2 = \frac{1}{2}w^2 + \frac{1}{2}, \qquad \mathcal{L}'(w) = w.$$

The unique minimizer is $w = 0$. At zero, the two example derivatives are -1 and 1, so an SGD step starting at the minimizer moves to α or $-\alpha$.

2. The Mean Stays at Zero

Let $z^{(t)}$ be 1 when SGD chooses the first example and -1 when it chooses the second. Each draw is independent of the earlier draws, with mean zero and squared value one. The update is

$$w^{(t+1)} = w^{(t)} - \alpha\big(w^{(t)} - z^{(t)}\big) = (1-\alpha)w^{(t)} + \alpha z^{(t)}.$$

Taking expectations gives

$$\mathbb{E}[w^{(t+1)}] = (1-\alpha)\mathbb{E}[w^{(t)}].$$

Since $w^{(0)} = 0$, induction gives $\mathbb{E}[w^{(t)}] = 0$ for every t. The expectation averages the weights over random runs. Individual runs fluctuate around zero.

3. Persistent Fluctuations

Squaring the update and taking expectations gives

$$\begin{aligned}\mathbb{E}\big[(w^{(t+1)})^2\big] &= (1-\alpha)^2\mathbb{E}\big[(w^{(t)})^2\big] + 2\alpha(1-\alpha)\mathbb{E}[w^{(t)}z^{(t)}] + \alpha^2 \\ &= (1-\alpha)^2\mathbb{E}\big[(w^{(t)})^2\big] + \alpha^2.\end{aligned}$$

The cross term is zero because $w^{(t)}$ depends only on earlier draws, so it is independent of the mean-zero draw $z^{(t)}$.

Each update multiplies the previous expected squared weight by $(1-\alpha)^2$ and adds α^2 from the new random draw. Starting from $w^{(0)} = 0$, the first three iterations give

$$\begin{aligned}
\mathbb{E}\big[(w^{(1)})^2\big] &= (1-\alpha)^2 \cdot 0 + \alpha^2 = \alpha^2, \\
\mathbb{E}\big[(w^{(2)})^2\big] &= (1-\alpha)^2\alpha^2 + \alpha^2 = \alpha^2\big(1 + (1-\alpha)^2\big), \\
\mathbb{E}\big[(w^{(3)})^2\big] &= (1-\alpha)^2\alpha^2\big(1 + (1-\alpha)^2\big) + \alpha^2 \\
&= \alpha^2\big(1 + (1-\alpha)^2 + (1-\alpha)^4\big).
\end{aligned}$$

After t updates, the newest draw contributes α^2. The preceding draw's contribution has been multiplied by $(1-\alpha)^2$ once, and the first draw's contribution has been multiplied by it $t-1$ times. Thus

$$\begin{aligned}
\mathbb{E}\big[(w^{(t)})^2\big] &= \alpha^2\big(1 + (1-\alpha)^2 + \cdots + (1-\alpha)^{2(t-1)}\big) \\
&= \alpha^2 \sum_{k=0}^{t-1} (1-\alpha)^{2k}.
\end{aligned}$$

We can verify this formula by induction: multiplying the sum by $(1-\alpha)^2$ raises each exponent by two, and adding α^2 supplies the new term with exponent zero.

To evaluate the sum, multiply it by $(1-\alpha)^2$ and subtract. All terms except the first and last cancel:

$$\begin{aligned}
\big(1 - (1-\alpha)^2\big) \sum_{k=0}^{t-1} (1-\alpha)^{2k} &= \sum_{k=0}^{t-1} (1-\alpha)^{2k} - \sum_{k=1}^{t} (1-\alpha)^{2k} \\
&= 1 - (1-\alpha)^{2t}.
\end{aligned}$$

The factor on the left is

$$1 - (1-\alpha)^2 = 1 - (1 - 2\alpha + \alpha^2) = \alpha(2-\alpha) > 0,$$

where positivity follows from $0 < \alpha < 2$. Dividing by this factor and multiplying by α^2 gives

$$\begin{aligned}
\mathbb{E}\big[(w^{(t)})^2\big] &= \alpha^2 \frac{1 - (1-\alpha)^{2t}}{\alpha(2-\alpha)} \\
&= \frac{\alpha}{2-\alpha}\big(1 - (1-\alpha)^{2t}\big).
\end{aligned}$$

Since $0 < \alpha < 2$, we have $|1-\alpha| < 1$. Therefore

$$\lim_{t\to\infty} \mathbb{E}\big[(w^{(t)})^2\big] = \frac{\alpha}{2-\alpha} > 0.$$

The expected squared distance from the minimizer stays positive. In fact, the weights cannot converge to zero along any sequence of draws. The update always satisfies

$$\big|w^{(t+1)} - (1-\alpha)w^{(t)}\big| = \alpha,$$

whereas the left side would tend to zero if the weights converged to zero.

As α approaches zero, the limiting expected squared distance $\alpha/(2-\alpha)$ approaches zero. A smaller fixed learning rate reduces the fluctuations, but they persist for every fixed positive learning rate.