CSCI 145 Problem 11 Solution: Depth and Regularization

Solution

Both models represent every function $x \mapsto zx$: the first uses $w = z$, and the second can use $a = z$, $b = 1$. Their predictions on the training input are w and ab, respectively.

1. Optimal Predictions

One Linear Layer. Differentiate with respect to w and set the derivative to zero:

$$\mathcal{L}_1'(w) = (w - 1) + \lambda w = (1 + \lambda)w - 1 = 0.$$

Since $\mathcal{L}_1''(w) = 1 + \lambda > 0$ everywhere, the loss is strictly convex. Thus this stationary point is the unique global minimum:

$$w^\star = f_{w^\star}(1) = \frac{1}{1 + \lambda}.$$

Two Linear Layers. Fix the prediction $z = ab$. Every pair with this product has the same error, so we first find the pair with the smallest regularization penalty. Since

$$0 \le (|a| - |b|)^2 = a^2 + b^2 - 2|ab|,$$

we have $a^2 + b^2 \ge 2|z|$, with equality exactly when $|a| = |b|$. Every z has such a factorization: take $a = \sqrt{|z|}$ and choose $b = \sqrt{|z|}$ for $z \ge 0$ or $b = -\sqrt{|z|}$ for $z < 0$. Thus the smallest penalty for prediction z is $(\lambda/2) \cdot 2|z| = \lambda|z|$. The smallest loss for that prediction is

$$q(z) = \min_{a,b:\, ab=z} \mathcal{L}_2(a, b) = \frac{1}{2}(z - 1)^2 + \lambda|z|.$$

For $z < 0$, the squared-error term alone exceeds $1/2 = q(0)$, so a negative prediction cannot be optimal. We only need to consider $z \ge 0$, where $|z| = z$. Differentiate the resulting quadratic and set its derivative to zero:

$$q'(z) = (z - 1) + \lambda = 0 \quad \Longrightarrow \quad z = 1 - \lambda.$$

If $0 < \lambda < 1$, this point is positive and is the unique global minimum, since the quadratic has second derivative 1. If $\lambda \ge 1$, the derivative is positive for every $z > 0$, so the minimum is at the boundary $z = 0$.

Hence the optimal prediction is

$$f_{a^*,b^*}(1) = a^* b^* = \max(0, 1 - \lambda).$$

For $0 < \lambda < 1$, it is attained by $a^* = b^* = \sqrt{1-\lambda}$ or by their simultaneous negatives. For $\lambda \geq 1$, the unique optimal weights are $a^* = b^* = 0$.

The two optimal predictions differ for every $\lambda > 0$. When $0 < \lambda < 1$,

$$\frac{1}{1+\lambda} - (1-\lambda) = \frac{\lambda^2}{1+\lambda} > 0.$$

When $\lambda \geq 1$, the first model's prediction is positive and the second's is zero. Penalizing the individual layer weights has changed a quadratic penalty on the prediction into an absolute-value penalty.

2. Gradient Descent from Zero

The gradients are

$$\mathcal{L}_1'(w) = (1+\lambda)w - 1, \qquad \nabla \mathcal{L}_2(a,b) = \begin{pmatrix} b(ab-1) + \lambda a \\ a(ab-1) + \lambda b \end{pmatrix}.$$

At $w = 0$, the first gradient is -1, so any positive learning rate α moves the first weight to α. Its initialization is never stationary or globally optimal for finite $\lambda > 0$.

At $a = b = 0$, both components of the second gradient vanish. Full-batch gradient descent therefore leaves both weights at zero at every step, regardless of the learning rate.

3. Stationary but Not Globally Optimal

Our calculation above shows that these weights are globally optimal exactly when $\lambda \geq 1$. Thus the two-layer model is stationary but not globally optimal at zero exactly when $0 < \lambda < 1$.