# CSCI 145 Problem 10 Solution: A Classifier That Refuses to Learn

## Solution

### 1. Gradient Descent at Zero

The hint rewrites the loss of one example as

$$\ell(z, y) = \log(1 + e^{-z}) + (1 - y)z.$$

With $n = 4$, write $z_i = b + w_1 x_{i1} + w_2 x_{i2}$ and $p_i = \sigma(z_i)$. Since $\partial \ell / \partial z = p - y$,

$$\nabla \mathcal{L} = \frac{1}{4} \sum_{i=1}^{4} (p_i - y_i) \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \end{pmatrix}.$$

At zero, every probability is $1/2$. In table order, the residuals are $(-1/2, 1/2, 1/2, -1/2)$, so

$$\begin{aligned} \sum_{i=1}^{4} (p_i - y_i) &= -1/2 + 1/2 + 1/2 - 1/2 = 0, \\ \sum_{i=1}^{4} (p_i - y_i) x_{i1} &= -1/2 + 1/2 - 1/2 + 1/2 = 0, \\ \sum_{i=1}^{4} (p_i - y_i) x_{i2} &= -1/2 - 1/2 + 1/2 + 1/2 = 0. \end{aligned}$$

Thus the gradient is zero. The full-batch update subtracts a learning rate times this zero vector, leaving every parameter at zero. The same calculation applies at every subsequent step.

**Why This Is a Global Minimum.** Differentiating this per-example loss gives

$$\frac{\partial \ell}{\partial z} = \sigma(z) - y, \qquad \frac{\partial^2 \ell}{\partial z^2} = \sigma(z)(1 - \sigma(z)) > 0.$$

The loss is convex in $z$. Each score is an affine function of $(b, w_1, w_2)$, so each composed loss, and hence their average $\mathcal{L}$, is convex in the parameters. A stationary point of a differentiable convex function is a global minimum. Thus gradient descent has already reached a global minimum, with value $\log 2$.

## 2. Better Accuracy, Worse Cross-Entropy

Choose $b = -1$ and $w_1 = w_2 = 1$, giving the score $x_1 + x_2 - 1$. The four scores are $(1, -1, -1, -3)$, so the predicted labels are $(1, 0, 0, 0)$. Only the fourth example is incorrect: accuracy is $3/4$. At zero, all four examples are predicted as class 0 under the stated tie rule, giving accuracy $1/2$.

The classifier with accuracy $3/4$ has higher cross-entropy:

$$\frac{3\log(1 + e^{-1}) + \log(1 + e^{3})}{4} \approx 0.9971 > \log 2 \approx 0.6931.$$

The fourth example receives a confident wrong prediction. Accuracy counts correct labels. Cross-entropy also measures the probability assigned to the observed label. Their preferred parameters can differ.

## 3. A Feature That Separates the Labels

Add the feature $u = x_1 x_2$. Its values are $(1, -1, -1, 1)$, positive exactly for the two examples with label 1. In the expanded model, choose score $u$: set the intercept and both original weights to zero and the new feature's weight to one. Then the probabilities are $(\sigma(1), \sigma(-1), \sigma(-1), \sigma(1))$, and all four labels are correct. The score is linear in the expanded features, even though $x_1 x_2$ is nonlinear in the original inputs.

The original stationary point reflects the combination of the available features and the chosen loss. Changing the features resolves the classification problem by making the labels linearly separable.