CSCI 145 Problem 13 Solution: Can Layer Normalization Destroy All Gradients?

Solution

1. Two Coordinates

For $\mathbf{z} = (a, b)$, the mean, centered vector, and standard deviation are

$$\mu = \frac{a+b}{2}, \qquad \mathbf{z} - \mu\mathbf{1} = \left(\frac{a-b}{2}, \frac{b-a}{2}\right), \qquad \sigma = \frac{|a-b|}{2}.$$

Dividing the centered vector by σ gives

$$N(a, b) = \begin{cases} (1, -1) & \text{if } a > b, \\ (-1, 1) & \text{if } a < b. \end{cases}$$

Only the ordering of the two coordinates remains.

2. No Gradient Through the Layer

The output is constant throughout each of the open regions $a > b$ and $a < b$. At every point with $a \neq b$, sufficiently small changes leave that ordering unchanged. All partial derivatives are therefore zero:

$$J_N(a, b) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

For a differentiable loss that depends on $\mathbf{z}$ through $N(\mathbf{z})$, the chain rule gives

$$\nabla_{\mathbf{z}} \ell = J_N(\mathbf{z})^\top \nabla_{N(\mathbf{z})} \ell = \mathbf{0}.$$

No gradient passes through this normalization to its inputs. Parameters whose only influence on the loss is through these inputs receive zero gradients as well. The map is undefined when $a = b$, which the problem excludes.

3. Two Directions Disappear in Any Dimension

Adding a constant to every coordinate changes the mean by that constant and leaves both the centered vector and its standard deviation unchanged:

$$N(\mathbf{z} + c\mathbf{1}) = N(\mathbf{z}).$$

Differentiating with respect to c at zero gives

$$\begin{aligned}\frac{d}{dc}N(\mathbf{z}+c\mathbf{1})\bigg|_{c=0} &= J_N(\mathbf{z})\ \frac{d}{dc}(\mathbf{z}+c\mathbf{1})\bigg|_{c=0}\\ &= J_N(\mathbf{z})\mathbf{1} = \mathbf{0}.\end{aligned}$$

Thus $\mathbf{1}$ is in the null space of the Jacobian.

Multiplying every coordinate by $c > 0$ multiplies both the mean and the standard deviation by c. These factors cancel in the normalized output:

$$N(c\mathbf{z}) = \frac{c\mathbf{z} - c\mu\mathbf{1}}{c\sigma} = N(\mathbf{z}).$$

Differentiating with respect to c at one gives

$$\begin{aligned}\frac{d}{dc}N(c\mathbf{z})\bigg|_{c=1} &= J_N(\mathbf{z})\ \frac{d}{dc}(c\mathbf{z})\bigg|_{c=1}\\ &= J_N(\mathbf{z})\mathbf{z} = \mathbf{0}.\end{aligned}$$

Thus $\mathbf{z}$ is also in the null space.

Since $\sigma > 0$, the coordinates of $\mathbf{z}$ are not all equal, so $\mathbf{z}$ is not a multiple of $\mathbf{1}$. We therefore have two linearly independent vectors in the null space of the $d \times d$ Jacobian.

Choose an orthonormal basis of $\mathbb{R}^d$ whose first two vectors form a basis for $\operatorname{span}\{\mathbf{1}, \mathbf{z}\}$. The Jacobian maps both of these basis vectors to zero. Every output is therefore a linear combination of the images of the remaining $d-2$ basis vectors. Its column space has dimension at most $d-2$, giving

$$\operatorname{rank} J_N(\mathbf{z}) \leq d - 2.$$

Normalization removes changes in the common offset and the scale of the input. When $d = 2$, these account for every input direction.