1 LayerNorm Backward Derivation for One Row

1.1 Forward Pass

For input $\mathbf{x} \in \mathbb{R}^{1\times N}$, LayerNorm computes:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{2}$$

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{3}$$

$$y_i = \gamma \hat{x}_i + \beta \tag{4}$$

For simplicity, let's assume $\gamma = 1$ and $\beta = 0$ (they can be added back easily), so:

$$y_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{5}$$

1.2 Jacobian Derivation

We need to compute $\frac{\partial y_i}{\partial x_j}$ for all $i, j \in \{1, 2, \ldots, N\}$.

Let $\sigma = \sqrt{\sigma^2 + \epsilon}$ for notation simplicity.

$$\frac{\partial y_i}{\partial x_j} = \frac{\partial}{\partial x_j}\left(\frac{x_i - \mu}{\sigma}\right) \tag{6}$$

Using the quotient rule and chain rule:

$$\frac{\partial y_i}{\partial x_j} = \frac{1}{\sigma}\frac{\partial (x_i - \mu)}{\partial x_j} - \frac{x_i - \mu}{\sigma^2}\frac{\partial \sigma}{\partial x_j} \tag{7}$$

1.2.1 Computing $\frac{\partial \mu}{\partial x_j}$

$$\frac{\partial \mu}{\partial x_j} = \frac{\partial}{\partial x_j}\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right) = \frac{1}{N} \tag{8}$$

1.2.2 Computing $\frac{\partial \sigma}{\partial x_j}$

$$\frac{\partial \sigma}{\partial x_j} = \frac{\partial}{\partial x_j}\sqrt{\sigma^2 + \epsilon} = \frac{1}{2\sigma}\frac{\partial \sigma^2}{\partial x_j} \tag{9}$$

$$\frac{\partial \sigma^2}{\partial x_j} = \frac{\partial}{\partial x_j}\left(\frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2\right) \tag{10}$$

$$= \frac{1}{N}\sum_{k=1}^{N} 2(x_k - \mu)\left(\delta_{kj} - \frac{1}{N}\right) \tag{11}$$

$$= \frac{2}{N}(x_j - \mu) - \frac{2}{N^2}\sum_{k=1}^{N}(x_k - \mu) \tag{12}$$

Since $\sum_{k=1}^{N}(x_k - \mu) = 0$:

$$\frac{\partial \sigma^2}{\partial x_j} = \frac{2}{N}(x_j - \mu) \tag{13}$$

Therefore:

$$\frac{\partial \sigma}{\partial x_j} = \frac{1}{2\sigma} * \frac{2}{N}(x_j - \mu) = \frac{x_j - \mu}{N\sigma} \tag{14}$$

1.2.3 Combining terms

$$J_{ij} = \frac{\partial y_i}{\partial x_j} = \frac{1}{\sigma}\left(\delta_{ij} - \frac{1}{N}\right) - \frac{x_i - \mu}{\sigma^2} * \frac{x_j - \mu}{N\sigma} \tag{15}$$

$$= \frac{1}{\sigma}\left(\delta_{ij} - \frac{1}{N}\right) - \frac{(x_i - \mu)(x_j - \mu)}{N\sigma^3} \tag{16}$$

$$= \frac{1}{\sigma}\left(\delta_{ij} - \frac{1}{N} - \frac{\hat{x}_i \hat{x}_j}{N}\right) \tag{17}$$

where $\hat{x}_i = \frac{x_i - \mu}{\sigma}$.

1.3 Jacobian Matrix

$$J_{(N\times N)} = \frac{\partial y}{\partial x} = \frac{1}{\sigma}\left(I - \frac{1}{N}\mathbf{1}^T\mathbf{1} - \frac{\hat{x}^T\hat{x}}{N}\right) \tag{18}$$

- $\mathbf{1}$ is a row vector with all elements equal to 1 and the same shape as x.

1.4 Backward Pass

Given $\frac{\partial L}{\partial \mathbf{y}} = \mathbf{dy} \in \mathbb{R}^{1\times N}$, we compute:

$$dx = dy \cdot J \tag{19}$$

$$= \frac{1}{\sigma}\left(dy - \frac{1}{N}\underbrace{(dy \cdot \mathbf{1}^T)}_{scalar1}\mathbf{1} - \frac{1}{N}\underbrace{(dy \cdot \hat{x}^T)}_{scalar2}\hat{x}\right) \tag{20}$$

$$= \frac{1}{\sigma}\left(dy - mean(dy)\mathbf{1} - mean(dy \odot \hat{x})\hat{x}\right) \tag{21}$$

$$\tag{22}$$

Let:

- $\text{mean}(\mathbf{dy}) = \frac{1}{N}\sum_{i=1}^{N} dy_i$
- $\text{mean}(\mathbf{dy} \odot \hat{\mathbf{x}}) = \frac{1}{N}\sum_{i=1}^{N} dy_i \hat{x}_i$

where $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 + \epsilon}$ and $\hat{\mathbf{x}} = \frac{\mathbf{x}-\mu}{\sigma}$.

1.5 Backward Pass add in w

$$dx = (w \odot dy) \cdot J \tag{23}$$

$$= \frac{1}{\sigma}\left((w \odot dy) - \frac{1}{N}\underbrace{((w \odot dy) \cdot \mathbf{1}^T)}_{scalar1}\mathbf{1} - \frac{1}{N}\underbrace{((w \odot dy) \cdot \hat{x}^T)}_{scalar2}\hat{x}\right) \tag{24}$$

$$= \frac{1}{\sigma}\left((w \odot dy) - mean((w \odot dy))\mathbf{1} - mean((w \odot dy) \odot \hat{x})\hat{x}\right) \tag{25}$$

$$\tag{26}$$