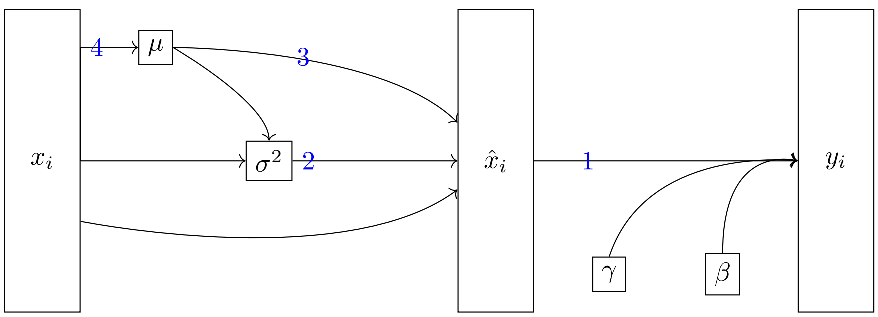
Backpropagation Derivation for Batch Normalization

Computational graph



Forward Pass

Given a mini-batch of inputs $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$:

1. Mean Calculation:

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

2. Variance Calculation:

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

3. Normalization: Each input is normalized as follows:

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

where ϵ is a small constant for numerical stability.

4. Scale and Shift: The final output is computed as:

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

where γ and β are learnable parameters.

Backpropagation

Let J be the loss function, and we want to compute the gradients with respect to the parameters and inputs.

1. Gradient with respect to Output: The gradient of the loss with respect to the output y_i:

$$\frac{\partial L}{\partial y_i}$$

2. Gradient with respect to Scale Parameter γ: The gradient with respect to γ:

$$\frac{\partial L}{\partial \gamma} = \sum_{i=1}^{N} \frac{\partial L}{\partial y_i} \hat{x}_i$$

3. Gradient with respect to Shift Parameter β: The gradient with respect to β:

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{N} \frac{\partial L}{\partial y_i}$$

Path 1: Gradient with respect to Normalized Input $\hat{x}_i$: The gradient with respect to the normalized input:

$$\begin{aligned}
\frac{\partial L}{\partial x_i} &= \frac{\partial L}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial x_i} \\
&= \gamma \cdot \frac{\partial L}{\partial y_i}
\end{aligned}$$

Path 2: Gradient with respect to Variance σ^2 : Using the chain rule, we have:

$$\begin{aligned}
\frac{\partial L}{\partial \sigma^2} &= \sum_{i=1}^{N} \frac{\partial L}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \sigma^2} \\
&= \gamma \cdot \sum_{i=1}^{N} \frac{\partial L}{\partial y_i} \cdot \frac{\hat{x}_i}{\partial \sigma^2} \\
&= \gamma \cdot \sum_{i=1}^{N} \frac{\partial L}{\partial y_i} \cdot \frac{-1}{2} \cdot \left(\frac{1}{\sqrt{\sigma^2 + \epsilon}} \right)^3 \cdot (x_i - \mu) \\
&= -\frac{\gamma}{2} \cdot \left(\frac{1}{\sqrt{\sigma^2 + \epsilon}} \right)^3 \sum_{i=1}^{N} \frac{\partial L}{\partial y_i} \cdot (x_i - \mu) \\
&= -\frac{\gamma}{2 \cdot (\sigma^2 + \epsilon)} \cdot \sum_{i=1}^{N} \frac{\partial L}{\partial y_i} \cdot \hat{x}_i
\end{aligned}$$

Given that,

$$
\begin{aligned}
\frac{\partial \sigma^2}{\partial x_i} &= -\frac{2}{n-1} \sum_{i=1}^{N} (x_i - \mu) \\
&= -\frac{2}{n-1} \cdot \sqrt{\sigma^2 + \epsilon} \cdot \sum_{i=1}^{N} \hat{x}_i
\end{aligned}
$$

Then,

$$
\begin{aligned}
\frac{\partial L}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial x_i} &= -\frac{\gamma}{(n-1) \cdot (\sigma^2 + \epsilon)} \cdot \sum_{j}^{N} \frac{\partial L}{\partial y_j} \cdot \hat{x}_j \cdot \hat{x}_i \\
&= -\frac{\gamma}{(n-1) \cdot (\sigma^2 + \epsilon)} \cdot \hat{x}_i \cdot \sum_{j}^{N} \frac{\partial L}{\partial y_j} \cdot \hat{x}_j
\end{aligned}
$$

Path 3: Gradient with respect to Mean μ : Similarly,

$$
\begin{aligned}
\frac{\partial L}{\partial \mu} &= \sum_{i=1}^{N} \frac{\partial L}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \mu} + \frac{\partial L}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial \mu} \\
&= \sum_{i=1}^{N} \frac{\partial L}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \mu} + 0 \\
&= \gamma \cdot \sum_{i=1}^{N} \frac{\partial L}{\partial y_i} \cdot \frac{\hat{x}_i}{\partial \mu} \\
&= -\frac{\gamma}{\sqrt{\sigma^2 + \epsilon)}} \cdot \sum_{i=1}^{N} \frac{\partial L}{\partial y_i}
\end{aligned}
$$

Note that,

$$
\frac{\partial L}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial \mu} = 0
$$

because

$$
\begin{aligned}
\frac{\partial \sigma^2}{\partial \mu} &= \frac{2}{n-1} \cdot \sum_{i=1}^{N} (x_i - \mu) \\
&= \frac{2}{n-1} \cdot \left(\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu) \right) \\
&= 0
\end{aligned}
$$

Given that,

$$\frac{\partial \mu}{\partial x_i} = \frac{1}{n}$$

Then,

$$\begin{aligned}
\frac{\partial L}{\partial \mu} \cdot \frac{\partial \mu}{\partial x_i} &= -\frac{\gamma}{\sqrt{\sigma^2 + \epsilon}} \cdot \sum_j^N \frac{\partial L}{\partial y_j} \cdot \frac{1}{n} \\
&= -\frac{\gamma}{n \cdot \sqrt{\sigma^2 + \epsilon}} \cdot \sum_j^N \frac{\partial L}{\partial y_j}
\end{aligned}$$

Finally: Gradient with respect to Input x_i:

Finally, the gradient with respect to the input is given by:

$$\begin{aligned}
\frac{\partial L}{\partial x_i} &= \sum_j^N \frac{\partial L}{\partial \hat{x_j}} \cdot \frac{\partial \hat{x}_j}{\partial x_i} + \frac{\partial L}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial x_i} + \frac{\partial L}{\partial \mu} \cdot \frac{\partial \mu}{\partial x_i} \\
&= \sum_j^N \frac{\partial L}{\partial \hat{x_j}} \cdot \delta_{ij} + \frac{\partial L}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial x_i} + \frac{\partial L}{\partial \mu} \cdot \frac{\partial \mu}{\partial x_i} \\
&= \frac{\gamma}{n\sqrt{\sigma^2 + \epsilon)}} \cdot \left(n \cdot \frac{\partial L}{\partial y_i} - \frac{n}{n-1} \hat{x}_i \sum_j^N \frac{\partial L}{\partial y_j} \hat{x}_j - \sum_j^N \frac{\partial L}{\partial y_j} \right)
\end{aligned}$$

This completes the derivation of the backpropagation equations for batch normalization.