1 Derivation of Batchnorm gradients

Let X of size (m, D) contain the data for a single batch of training. Let μ_B be the mean of that batch of size (1xD) and σ^2 of size (1xD) be the variance. $\bar{X}$ is the batch normalized output.

Notation: Data_k refers to the k^{th} row of Data matrix.

$\bar{X} = \frac{X-\mu_B}{\sqrt{\sigma^2+\epsilon}}$

$\text{y} = \gamma\bar{X} + \beta$

$$\frac{\partial y}{\partial \bar{X}} = \gamma$$

We know that:

$$\bar{X_i} = \frac{X_i - \mu_B}{\sqrt{\sigma^2+\epsilon}}$$

$$\frac{\partial \bar{X_i}}{\partial X_k} = \frac{\frac{\partial X_i}{\partial X_k} - \frac{\partial \mu_B}{\partial X_k}}{\sqrt{\sigma^2+\epsilon}} - \frac{X_i - \mu_B}{2\sqrt[3/2]{\sigma^2+\epsilon}}\frac{\partial \sigma^2}{\partial X_k} \tag{1}$$

$$\frac{\partial \mu_B}{\partial X_k} = \frac{1}{m} \tag{2}$$

We can express the variance in terms of mean using the following equation.

$$\sigma^2 = \sum_{k=0}^{m-1}\frac{X_k^2}{m} - \mu_B^2$$

$$\frac{\partial \sigma^2}{\partial X_k} = \frac{2X_k}{m} - 2\mu_B\frac{\partial \mu_B}{\partial X_k}$$

From (2), we get

$$\frac{\partial \sigma^2}{\partial X_k} = \frac{2(X_k - \mu_B)}{m}$$

Substituting this into (1) we have:

$$\frac{\partial \bar{X_i}}{\partial X_k} = \frac{\frac{\partial X_i}{\partial X_k} - \frac{1}{m}}{\sqrt{\sigma^2+\epsilon}} - \frac{(X_i - \mu_B)(X_k - \mu_B)}{m\sqrt[3/2]{\sigma^2+\epsilon}}$$

The above equation can be expressed in terms of $\bar{X}$ as

$$\frac{\partial \bar{X_i}}{\partial X_k} = \frac{m\frac{\partial X_i}{\partial X_k} - 1}{m\sqrt{\sigma^2+\epsilon}} - \frac{\bar{X_i}\bar{X_k}}{m\sqrt{\sigma^2+\epsilon}}$$

$$\frac{\partial \bar{X_i}}{\partial X_k} = \frac{m\frac{\partial X_i}{\partial X_k} - 1 - \bar{X_i}\bar{X_k}}{m\sqrt{\sigma^2+\epsilon}} \tag{3}$$

We need to find $\frac{\partial L}{\partial \bar{X_j}}$ where L is the loss. From chain rule, we have:

$$\frac{\partial L}{\partial X_j} = \frac{\partial L}{\partial \bar{X_0}}\frac{\partial \bar{X_0}}{\partial X_j} + \frac{\partial L}{\partial \bar{X_1}}\frac{\partial \bar{X_1}}{\partial X_j} + ... + \frac{\partial L}{\partial \bar{X_{m-1}}}\frac{\partial \bar{X_{m-1}}}{\partial X_j}$$

Which can be reduced as:

$$\frac{\partial L}{\partial X_j} = \sum_{k=0}^{m-1} \frac{\partial L}{\partial \bar{X_k}}\frac{\partial \bar{X_k}}{\partial X_j}$$

Substituting result from eq (3), we have:

$$\frac{\partial L}{\partial X_j} = \sum_{k=0}^{m-1} \frac{\partial L}{\partial \bar{X_k}}(\frac{m\frac{\partial X_k}{\partial X_j} - 1 - \bar{X_j}\bar{X_k}}{m\sqrt{\sigma^2+\epsilon}})$$

$$\frac{\partial L}{\partial X_j} = \frac{1}{m\sqrt{\sigma^2+\epsilon}} \sum_{k=0}^{m-1} (\frac{\partial L}{\partial \bar{X_k}}(m\frac{\partial X_k}{\partial X_j} - 1 - \bar{X_j}\bar{X_k}))$$

$$\frac{\partial L}{\partial X_j} = \frac{1}{m\sqrt{\sigma^2+\epsilon}}(m\sum_{k=0}^{m-1} \frac{\partial L}{\partial \bar{X_k}}\frac{\partial X_k}{\partial X_j} - \sum_{k=0}^{m-1} \frac{\partial L}{\partial \bar{X_k}} - \bar{X_j}\sum_{k=0}^{m-1} \frac{\partial L}{\partial \bar{X_k}}\bar{X_k})$$

We see that $\frac{\partial X_k}{\partial X_j}$ is 0 everywhere except when j=k. So that part of the summation reduces to summing all rows where j=k, which is just one row corresponding to $\frac{\partial L}{\partial \bar{X_j}}$. So the equation reduces to:

$$\frac{\partial L}{\partial X_j} = \frac{1}{m\sqrt{\sigma^2+\epsilon}}(m\frac{\partial L}{\partial \bar{X_j}} - \sum_{k=0}^{m-1} \frac{\partial L}{\partial \bar{X_k}} - \bar{X_j}\sum_{k=0}^{m-1} \frac{\partial L}{\partial \bar{X_k}}\bar{X_k}) \tag{4}$$

We receive the gradient of y from the function as dout. so

$$\frac{\partial L}{\partial y} = dout$$

$$\frac{\partial L}{\partial \bar{X}} = \gamma\frac{\partial L}{\partial y}$$

We use the matrix computed from above to substitute in equation (4) to get the loss w.r.t X for all its rows.