# 1 Deriving the BatchNorm Gradients

The goal is to derive the gradient across the whole batch normalization process for a batch B of size m.
Assume we already have gradient w.r.t logits: $\frac{\partial L}{\partial y_i}$
Going backwards, we will find $\frac{\partial L}{\partial \hat{x}_i}, \frac{\partial L}{\partial \sigma_B^2}, \frac{\partial L}{\partial \mu_B}$, and finally $\frac{\partial L}{\partial x_i}$.

**Finding gradient for normalized unshifted logits:**

$$\frac{\partial L}{\partial \hat{x}_i} = \frac{\partial L}{\partial y_i}\frac{\partial y_i}{\partial \hat{x}_i} = \frac{\partial L}{\partial y_i}\gamma$$

**Finding gradient for variance across batch:**

$$\frac{\partial L}{\partial \sigma_b^2} = \sum_j \frac{\partial L}{\partial \hat{x}_j}\frac{\partial \hat{x}_j}{\partial \sigma_B^2} \Rightarrow \sum_j \frac{\partial \hat{x}_j}{\partial \sigma_B^2} = \frac{\partial}{\partial \sigma_B^2}(\sum_j (x_j - \mu_B)(\sigma_b^2 + \epsilon)^{-\frac{1}{2}})$$
$$= -\frac{1}{2}\sum_j (x_j - \mu_B)(\sigma_B^2 + \epsilon)^{-\frac{3}{2}} \Rightarrow \frac{\partial L}{\partial \sigma_b^2} = -\gamma\frac{1}{2}(\sigma_B^2 + \epsilon)^{-\frac{3}{2}}\sum_j (x_j - \mu_B)\frac{\partial L}{\partial y_j}$$

**Finding gradient for mean across batch:**

$$\frac{\partial L}{\partial \mu_b} = \sum_j \frac{\partial L}{\partial \hat{x}_j}\frac{\partial \hat{x}_j}{\partial \mu_B} + \frac{\partial L}{\partial \sigma_B^2}\frac{\partial \sigma_B^2}{\partial \mu_B}$$

$$\sum_j \frac{\partial L}{\partial \hat{x}_j}\frac{\partial \hat{x}_j}{\partial \mu_B} = -\gamma(\sigma_B + \epsilon)^{-\frac{1}{2}}\sum_j \frac{\partial L}{\partial y_j}$$

$$\frac{\partial \sigma_B^2}{\partial \mu_B} = \frac{\partial}{\partial \mu_B}(\frac{1}{m-1}\sum_j (x_j - \mu_B)^2) = \frac{2}{m-1}\sum_j (x_j - \mu_B) = \frac{2}{m-1}(\sum_j x_j - \sum_j \mu_B)$$
$$= \frac{2}{m-1}(m\mu_B - m\mu_B) = 0 \Rightarrow \frac{\partial L}{\partial \sigma_B^2}\frac{\partial \sigma_B^2}{\partial \mu_B} = 0 \Rightarrow \frac{\partial L}{\partial \mu_B} = -\gamma(\sigma_B + \epsilon)^{-\frac{1}{2}}\sum_j \frac{\partial L}{\partial y_j}$$

**Finding total gradient for BatchNorm input:**

$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial \hat{x}_i}\frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial L}{\partial \mu_B}\frac{\partial \mu_B}{\partial x_i} + \frac{\partial L}{\partial \sigma_b^2}\frac{\partial \sigma_B^2}{\partial x_i}$$
$$\frac{\partial \hat{x}_i}{\partial x_i} = \frac{\partial}{\partial x_i}((x_i - \mu_B)(\sigma_B^2 + \epsilon)^{-\frac{1}{2}}) = (\sigma_B^2 + \epsilon)^{-\frac{1}{2}} \Rightarrow \frac{\partial L}{\partial \hat{x}_i}\frac{\partial \hat{x}_i}{\partial x_i} = \gamma\frac{\partial L}{\partial y_i}(\sigma_B^2 + \epsilon)^{-\frac{1}{2}}$$
$$\frac{\partial \mu_B}{\partial x_i} = \frac{1}{m} \Rightarrow \frac{\partial L}{\partial \mu_B}\frac{\partial \mu_B}{\partial x_i} = -\frac{\gamma}{m}(\sigma_B^2 + \epsilon)^{\frac{1}{2}}\sum_j \frac{\partial L}{\partial y_j}$$
$$\frac{\partial \sigma_B^2}{\partial x_i} = \frac{2}{m-1}(x_i - \mu_B) \Rightarrow \frac{\partial L}{\partial \sigma_B^2}\frac{\sigma_B^2}{\partial x_i} = -\frac{\gamma}{m-1}(x_i - \mu_B)(\sigma_B^2 + \epsilon)^{-\frac{3}{2}}\sum_j (x_j - \mu_B)\frac{\partial L}{\partial y_j}$$
$$\frac{\partial L}{\partial x_i} = \gamma\frac{\partial L}{\partial y_i}(\sigma_B^2 + \epsilon)^{-\frac{1}{2}} - \frac{\gamma}{m}(\sigma_B^2 + \epsilon)^{\frac{1}{2}}\sum_j \frac{\partial L}{\partial y_j} - \frac{\gamma}{m-1}(x_i - \mu_B)(\sigma_B^2 + \epsilon)^{-\frac{3}{2}}\sum_j (x_j - \mu_B)\frac{\partial L}{\partial y_j}$$
$$= \gamma(\sigma_B^2 + \epsilon)^{-\frac{1}{2}}(\frac{\partial L}{\partial y_i} - \frac{1}{m}\sum_j \frac{\partial L}{\partial y_j} - \frac{\hat{x}_i}{m-1}\sum_j (x_j - \mu_B)\frac{\partial L}{\partial y_i})$$