Batch normalization attempts to normalize inputs along their dimensions. It is defined as follows for a given input x in a batch of inputs X; hyper parameters γ (scaling), β (shifting); and producing an activation y:

$$y_i^{(m)} = \gamma_i \cdot \hat{x}_i^{(m)} + \beta_i$$

where $y_i^{(m)}$ is an activation for example m in dimension i. The normalized input $\hat{x}$ is defined as follows:

$$\hat{x}_i^{(m)} = \frac{(x_i^{(m)} - E[X])}{\sqrt{Var[X] + \epsilon}}$$

ϵ is simply a small value to avoid division by zero. We need to calcuate the back propagation logic for this by calculating derivatives of the loss function L with respect to γ, β, and each input. First, we can start with the more simple derivatives:

$$\frac{\partial L}{\partial \beta_i} = \sum_k \frac{\partial L}{\partial y_i^{(k)}} \frac{\partial}{\partial \beta_i} \left[\gamma_i \cdot \hat{x}_i^{(k)} + \beta_i\right] = \sum_k \frac{\partial L}{\partial y_i^{(k)}}$$

In vectorized form, this simply sums the upstream loss gradients along input dimensions, or given an activation matrix $Y \in \mathbb{R}^{NxD}$ where N is the number of examples and D is the number of dimensions:

$$\nabla_\beta L = J_{1N}\beta$$

Here J is the unit matrix (matrix of ones). γ is also simple:

$$\frac{\partial L}{\partial \gamma_i} = \sum_k \frac{\partial L}{\partial y_i^{(k)}} \frac{\partial}{\partial \gamma_i} \left[\gamma_i \cdot \hat{x}_i^{(k)} + \beta_i\right] = \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \hat{x}_i^{(k)}$$

In vectorized form, this is simply the sum of the elementwise multiplication of the upstream loss gradients with the normalized input:

$$\nabla_\gamma L = J_{1N}(\gamma \circ \nabla_Y L)$$

Deriving back propagations for the inputs is more complex. For a particular input in dimension i of example m, $x_i^{(m)}$, batch normalization produces an output $y_i^{(m)}$. However, it's important to note that all the dimension i inputs in the batch affect $y_i^{(m)}$. As an example, consider another example input in the batch, $x_i^{(l)}$ and assume it's extremely large or infinite. It's clear this will affect $y_i^{(m)}$ through the expected value and variance functions, probably increasing the loss. Thus we must backpropagate loss L from all outputs to each input:

$$\frac{\partial}{\partial x_i^{(m)}} L = \sum_k \frac{\partial}{\partial x_i^{(m)}} \left[\gamma_i \cdot \hat{x}_i^{(k)} + \beta_i\right]$$

We can first calculate the derivatives of the expectation and variance. In these derivations, we drop the previous notation and simply use x_i to denote

an element in a generic vector X. First we calculate the derivative of the expectation:

$$E[X] = \sum_k \frac{x_k}{n}$$

$$\frac{\partial}{\partial x_i} E[X] = \frac{\partial}{\partial x_i}\frac{x_1}{n} + \frac{\partial}{\partial x_i}\frac{x_2}{n} + ... + \frac{\partial}{\partial x_i}\frac{x_i}{n} + ...\frac{\partial}{\partial x_i}\frac{x_n}{n}$$
$$= 1/n$$

Now the variance:

$$Var[X] = \sum_k \frac{(x_k - E[X])^2}{n}$$

$$\frac{\partial}{\partial x_i} Var[X] = \frac{\partial}{\partial x_i}\left[\frac{(x_i - E[X])^2}{n}\right] + \frac{\partial}{\partial x_i}\left[\sum_{j\neq i}\frac{(x_j - E[X])^2}{n}\right]$$

$$= 2 * (x_i - E[X]) * \frac{1 - 1/n}{n} + \sum_{j\neq i}\frac{2 * (x_j - E[X]) * (-1/n)}{n}$$

$$= 2 * (n-1)/n^2 * (x_i - E[X]) - 2/n^2 * \sum_{j\neq i} x_j - E[X]$$

$$= 2\left[(n-1)/n^2 * (x_i - E[X]) - 1/n^2 * \left(\sum_{j\neq i} x_j - E[X]\right)\right]$$

$$= 2\left[(n-1)/n^2 * (x_i - E[X]) - 1/n^2 * \sum_{j\neq i} x_j + 1/n^2 \sum_{j\neq i} E[X]\right]$$

$$= 2\left[(n-1)/n^2 * (x_i - E[X]) - 1/n^2 * \sum_{j\neq i} x_j + (n-1)/n^2 * E[X]\right]$$

$$= 2\left[x_i * (n-1)/n^2 - E[X] * (n-1)/n^2 - 1/n^2 * \sum_{j\neq i} x_j + E[X] * (n-1)/n^2\right]$$

$$= \frac{2}{n^2}\left[x_i * (n-1) - \sum_{j\neq i} x_j\right]$$

$$= \frac{2}{n^2}\sum_k x_i - x_k$$

The last line comes about as the multiplication can be thought of as a sum of x_i over 1 to $n-1$. This matches the number of terms in the latter summation.

Then note that when $k = i$ in the summation, the result is zero. So we can simply sum over all k rather than using the previous restriction of $\neq i$.

For $\frac{\partial}{\partial x}$, we must use the quotient rule (also note the β term can be ignored as it's not a function of x):

$$\left[\frac{f(x)}{g(x)}\right]' = \frac{g(x)f'(x) - f(x)g'(x)}{[g(x)]^2}$$

In this case, for a particular input example m along dimension i, we need to calculate how the loss has changed for each activation $y_i^{(k)}$:

$$f(X_i)^{(k)} = x_i^{(k)} - E[X_i]$$

$$\frac{\partial}{\partial x_i^{(m)}} f(X_i)^{(k)} = \frac{\partial}{\partial x_i^{(m)}} x_i^{(k)} - 1/n$$

$$g(X_i)^{(k)} = \sqrt{Var[X_i] + \epsilon}$$

$$\frac{\partial}{\partial x_i^{(m)}} g(X_i)^{(k)} = \frac{1}{2 * \sqrt{Var[X] + \epsilon}} \cdot \frac{2}{n^2} \sum_t x_i^{(m)} - x_i^{(t)}$$

$$= \frac{1}{n^2\sqrt{Var[X_i] + \epsilon}} \sum_t x_i^{(m)} - x_i^{(t)}$$

$$= \frac{1}{n^2\sqrt{Var[X_i] + \epsilon}} (n \cdot x_i^{(m)} - \sum_t x_i^{(t)})$$

$$= \frac{1}{n^2\sqrt{Var[X_i] + \epsilon}} n(x_i^{(m)} - (\sum_t x_i^{(t)})/n)$$

$$= \frac{(x_i^{(m)} - E[X_i])}{n\sqrt{Var[X_i] + \epsilon}} = \hat{x}_i^{(m)}/n$$

Note the numerator derivative differs for the $m = k$ and $m \neq k$ cases, while the denominator is identical in both.

$$\frac{\partial}{\partial x_i^{(m)}} L = \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \frac{\partial}{\partial x_i^{(m)}} \left[\gamma_i \cdot \hat{x}_i^{(k)} + \beta_i\right]$$

$$\frac{\partial}{\partial x_i^{(m)}} L = \gamma_i \cdot \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \frac{\partial}{\partial x_i^{(m)}} \hat{x}_i^{(k)} + \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \frac{\partial}{\partial x_i^{(m)}} \beta_i = \gamma_i \cdot \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \frac{\partial}{\partial x_i^{(m)}} \hat{x}_i^{(k)}$$

(continuing with only the summand):

$$\frac{\partial}{\partial x_i^{(m)}} \hat{x}_i^{(k)} = \frac{\partial L}{\partial y_i^{(k)}} \cdot \left[\frac{\sqrt{Var[X_i] + \epsilon} \cdot (\frac{\partial}{\partial x_i^{(m)}} x_i^{(k)} - 1/n) - (x_i^{(k)} - E[X^{(i)}]) \cdot \hat{x}_i^{(m)}/n}{Var[X_i] + \epsilon}\right]$$

$$= \frac{\partial L}{\partial y_i^{(k)}} \cdot \left[\frac{\frac{\partial}{\partial x_i^{(m)}} x_i^{(k)} - 1/n}{\sqrt{Var[X_i] + \epsilon}} - \frac{(x_i^{(k)} - E[X^{(i)}]) \cdot \hat{x}_i^{(m)}}{n \cdot Var[X_i] + \epsilon} \right]$$

$$= \frac{\partial L}{\partial y_i^{(k)}} \cdot \left[\frac{n \cdot \frac{\partial}{\partial x_i^{(m)}} x_i^{(k)} - 1}{n \cdot \sqrt{Var[X_i] + \epsilon}} - \frac{\hat{x}_i^{(k)} \cdot \hat{x}_i^{(m)}}{n \cdot \sqrt{Var[X_i] + \epsilon}} \right]$$

$$= \frac{\partial L}{\partial y_i^{(k)}} \cdot \left[\frac{n \cdot \frac{\partial}{\partial x_i^{(m)}} x_i^{(k)} - 1 - \hat{x}_i^{(k)} \cdot \hat{x}_i^{(m)}}{n \cdot \sqrt{Var[X_i] + \epsilon}} \right]$$

(going back to the full expression):

$$\frac{\partial}{\partial x_i^{(m)}} L = \gamma_i \cdot \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \frac{n \cdot \frac{\partial}{\partial x_i^{(m)}} x_i^{(k)} - 1 - \hat{x}_i^{(k)} \cdot \hat{x}_i^{(m)}}{n \cdot \sqrt{Var[X_i] + \epsilon}}$$

$$= \gamma_i \cdot \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \frac{n \cdot \frac{\partial}{\partial x_i^{(m)}} x_i^{(k)}}{n \cdot \sqrt{Var[X_i] + \epsilon}} - \gamma_i \cdot \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \frac{1}{n \cdot \sqrt{Var[X_i] + \epsilon}} - \gamma_i \cdot \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \frac{\hat{x}_i^{(k)} \cdot \hat{x}_i^{(m)}}{n \cdot \sqrt{Var[X_i] + \epsilon}}$$

$$= \gamma_i \cdot \frac{\partial L}{\partial y_i^{(m)}} \cdot \frac{1}{\sqrt{Var[X_i] + \epsilon}} - \gamma_i \cdot \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \frac{1}{n \cdot \sqrt{Var[X_i] + \epsilon}} - \gamma_i \cdot \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \frac{\hat{x}_i^{(k)} \cdot \hat{x}_i^{(m)}}{n \cdot \sqrt{Var[X_i] + \epsilon}}$$

$$= \gamma_i \cdot \left[\frac{\partial L}{\partial y_i^{(m)}} \cdot \frac{1}{\sqrt{Var[X_i] + \epsilon}} - \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \frac{\hat{x}_i^{(k)} \cdot \hat{x}_i^{(m)} + 1}{n\sqrt{Var[X_i] + \epsilon}} \right]$$

$$= \gamma_i \frac{1}{\sqrt{Var[X_i] + \epsilon}} \left[\frac{\partial L}{\partial y_i^{(m)}} - \frac{1}{n} \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \left(\hat{x}_i^{(k)} \cdot \hat{x}_i^{(m)} + 1 \right) \right]$$

$$= \gamma_i \frac{1}{\sqrt{Var[X_i] + \epsilon}} \left[\frac{\partial L}{\partial y_i^{(m)}} - \frac{1}{n} \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \hat{x}_i^{(k)} \cdot \hat{x}_i^{(m)} - \frac{1}{n} \sum_k \frac{\partial L}{\partial y_i^{(k)}} \right]$$

$$= \gamma_i \frac{1}{\sqrt{Var[X_i] + \epsilon}} \left[\frac{\partial L}{\partial y_i^{(m)}} - \frac{\hat{x}_i^{(m)}}{n} \sum_k \frac{\partial L}{\partial y_i^{(k)}} \cdot \hat{x}_i^{(k)} - \frac{1}{n} \sum_k \frac{\partial L}{\partial y_i^{(k)}} \right]$$