Normalization:

One important thing about BN algorithm is that it normalize the input right **before** activation:

$$
\begin{aligned}
y^l &\leftarrow \mathcal{BN}(W^{l-1,l}u^{l-1}) \\
u^l &= g(y^l)
\end{aligned}
$$

$g(u)$ is activation function. It can be *sigmoid* ,*ReLu* etc. $W^{l-1,l}$ is the weights linking layers $\{l-1, l\}$, u^l is the nonlinear activation of layer l. To summarize the normalization in B.N.:

$$
\begin{aligned}
x^l &= W^{l-1,l}u^{l-1} \\
\hat{x}^l &= \frac{x^l - \mu_B^l}{\sqrt{\sigma_B^{2,l}}} \\
y^l &= \gamma^l \hat{x}^l + \beta^l \\
u^l &= g(y^l)
\end{aligned}
$$

Back-Propagation:

Cost function $\mathcal{C}(\gamma, \beta, W, \hat{\theta}, \psi, \phi)$ is a function of hyper parameter $\hat{\theta}$, training parameter$\{\gamma, \beta, W\}$, input ψ and label ϕ. The whole business is centered at minimizing this beast. Define growth rate at layer l: $\Delta_s^l \equiv \dfrac{\partial C}{\partial y_s^l}$. Here y_s^l is the linear combination of sample s that waits to be activated. Using chain rule, BP for BN can be derived as below. It differs with the baseline BP slightly.

$$
\begin{aligned}
\frac{\partial C}{\partial W^{l-1,l}} &= \sum_s \frac{\partial C}{\partial y_s^l} \frac{\partial y_s^l}{\partial W^{l-1,l}} \equiv \sum_s \Delta_s^l \frac{\partial y_s^l}{\partial W^{l-1,l}} \\
\frac{\partial y_s^l}{\partial W^{l-1,l}} &= \frac{\gamma^l}{\sqrt{\sigma_B^2}^l} \left[u_s^{l-1} - \left\langle u^{l-1} \right\rangle - \hat{x}_s^l \left\langle u^{l-1} \hat{x}^l \right\rangle \right] \\
\Delta_s^l &= \frac{\gamma^{l+1}}{\sqrt{\sigma_B^2}^{l+1}} W^{l,l+1} g'(y_s^l) \left[\Delta_s^{l+1} - \left\langle \Delta^{l+1} \right\rangle - \hat{x}_s^{l+1} \left\langle \Delta^{l+1} \hat{x}^{l+1} \right\rangle \right] \\
\frac{\partial C}{\partial \beta^l} &= \sum_s \frac{\partial C}{\partial y_s^l} \\
\frac{\partial C}{\partial \gamma^l} &= \sum_s \frac{\partial C}{\partial y_s^l} \hat{x}_s^l
\end{aligned}
$$

$\langle \cdots \rangle$ is the sample average. s is the label of sample in a mini-batch. l labels layer. $g'(x)$ is the derivative of activation function. At the boundary:

$$
\Delta_s^L = \frac{\partial C}{\partial u_s^L} g'(y_s^L)
$$

L labels the last layer in network.