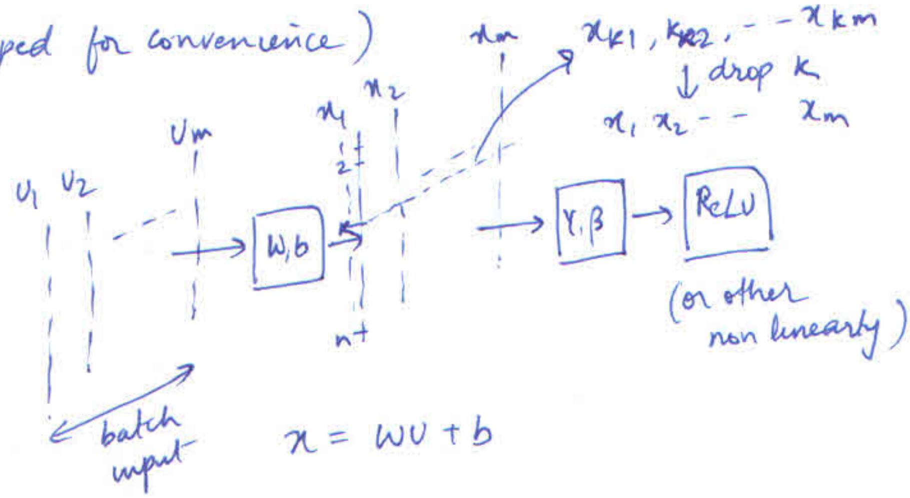
Derivation for Batch Normalization back Propagation Equations

Input: $x_{k1}\ x_{k2}\ --\ x_{km}$

m: minibatch size

K: element index (dropped for convenience)



$x = WU + b$

Parameters to be learnt: γ, β

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i$$ ← Sum over batch, not over output vector

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2$$

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$$ ← to avoid divide by 0 issues

effect of other layers + loss function

$$y_i = \gamma \hat{x}_i + \beta \quad , \quad y = f(y_1, y_2 -- y_m, \mu_B, \sigma_B^2)$$

need derivatives wrt x_i, ~~μ_B, σ_B^2~~, γ, β

μ_B & σ_B^2 don't depend on γ or β

$$\frac{\partial y}{\partial \lambda} = \frac{\partial f}{\partial y_1}\frac{\partial y_1}{\partial \lambda} + \frac{\partial f}{\partial y_2}\cdot\frac{\partial y_2}{\partial \lambda} -- \frac{\partial f}{\partial y_m}\cdot\frac{\partial y_m}{\partial \lambda} + \frac{\partial f}{\partial \mu_B}\underbrace{\frac{\partial \mu_B}{\partial \lambda}}_{0} + \frac{\partial f}{\partial \sigma_B^2}\underbrace{\frac{\partial \sigma_B^2}{\partial \lambda}}_{0}$$

should be γ (sorry)

$$= \sum_{i=1}^{m}\frac{\partial f}{\partial y_i}\gamma$$ ← back propagation from subsequent layers will give us these

Similarly,

$$\frac{\partial y}{\partial \beta} = \sum_{i=1}^{m}\frac{\partial f}{\partial y_i}$$

Now ~~lets~~ lets compute $\frac{\partial y}{\partial x_i}$

$$\frac{\partial y}{\partial x_i} = \sum_j \frac{\partial f}{\partial y_j}\frac{\partial y_j}{\partial x_i} + \frac{\partial f}{\partial \sigma_B^2}\frac{\partial \sigma_B^2}{\partial x_i} + \frac{\partial f}{\partial \mu_B}\cdot\frac{\partial \mu_B}{\partial x_i}$$

$$= \sum_j \frac{\partial f}{\partial y_j}\frac{\partial y_j}{\partial \hat{x}_i}\cdot\frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial f}{\partial \sigma_B^2}\cdot\frac{\partial \sigma_B^2}{\partial \hat{x}_i}\cdot\frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial f}{\partial \mu_B}\cdot\frac{\partial \mu_B}{\partial \hat{x}_i}\cdot\frac{\partial \hat{x}_i}{\partial x_i}$$

for $j \neq i$, these terms will become 0

$$+ \frac{\partial f}{\partial \sigma_B^2}\cdot\frac{\partial \sigma_B^2}{\partial x_i} + \frac{\partial f}{\partial \mu_B}\cdot\frac{\partial \mu_B}{\partial x_i}$$

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

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

also, $\dfrac{\partial \hat{x}_i}{\partial x_i} = \dfrac{1}{\sqrt{\sigma_B^2+\epsilon}}$, $\dfrac{\partial \sigma_B^2}{\partial x_i} = \dfrac{1}{m} 2(x_i - \mu_B)$ derivatives for $j \neq i = 0$

$$\frac{\partial \mu_B}{\partial x_i} = \frac{1}{m}$$

$\therefore$ summation over j reduces to 1 term as $\dfrac{\partial y_j}{\partial \hat{x}_i} = 0$ for $j \neq i$

$$\frac{\partial y}{\partial x_i} = \cancel{\sum} \frac{\partial f}{\partial y_i} \gamma \cdot \frac{1}{\sqrt{\sigma_B^2+\epsilon}} + \frac{\partial f}{\partial \sigma_B^2} \cdot \frac{2(x_i - \mu_B)}{m} + \frac{\partial f}{\partial \mu_B} \cdot \frac{1}{m} \quad - (1)$$

we need $\dfrac{\partial f}{\partial \sigma_B^2}$, $\dfrac{\partial f}{\partial \mu_B}$. Notice σ_B^2 is a func of μ_B but μ_B doesn't depend on σ_B^2

$$\frac{\partial f}{\partial \sigma_B^2} = \sum_i \frac{\partial f}{\partial y_i} \cdot \frac{\partial y_i}{\partial \sigma_B^2} + \frac{\partial f}{\partial \mu_B} \cdot \underbrace{\frac{\partial \mu_B}{\partial \sigma_B^2}}_{\to 0}$$

$$= \sum_i \frac{\partial f}{\partial y_i} \frac{\partial y_i}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \sigma_B^2}$$

$$= \sum_i \frac{\partial f}{\partial y_i} \gamma \cdot - \frac{\hat{x}_i}{2(\sigma_B^2+\epsilon)}$$

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

$$\frac{\partial \hat{x}_i}{\partial \sigma_B^2} = -\frac{1}{2}(x_i - \mu_B)(\sigma_B^2+\epsilon)^{-3/2}$$

$$= \frac{-1}{2} \frac{(x_i - \mu_B)}{(\sqrt{\sigma_B^2+\epsilon})(\sigma_B^2+\epsilon)}$$

$$= -\frac{1}{2} \frac{\hat{x}_i}{\sigma_B^2+\epsilon}$$

$$\frac{\partial f}{\partial \mu_B} = \sum_i \frac{\partial f}{\partial y_i} \frac{\partial y_i}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \mu_B} + \frac{\partial f}{\partial \sigma_B^2} \cdot \frac{\partial \sigma_B^2}{\partial \mu_B}$$

$$= \sum_i \frac{\partial f}{\partial y_i} \gamma \cdot \frac{-1}{\sqrt{\sigma_B^2+\epsilon}} + \frac{\partial f}{\partial \sigma_B^2} \cdot \frac{2}{m} \underbrace{\sum_i (x_i - \mu_B)}_{\to 0}$$ by definition of μ_B

$$= \sum_i \frac{\partial f}{\partial y_i} \gamma \frac{-1}{\sqrt{\sigma_B^2+\epsilon}}$$

plugging into (1),

$$\frac{\partial y}{\partial x_i} = \cancel{\sum} \frac{\partial f}{\partial y_i} \gamma \frac{1}{\sqrt{\sigma_B^2+\epsilon}} - \frac{(x_i - \mu_B)}{m(\sigma_B^2+\epsilon)} \sum_{j=1}^{m} \frac{\partial f}{\partial y_j} \gamma \hat{x}_j - \frac{1}{m\sqrt{\sigma_B^2+\epsilon}} \sum_{j=1}^{m} \frac{\partial f}{\partial y_j} \gamma$$

$$\boxed{\frac{\partial y}{\partial x_i} = \frac{\partial f}{\partial y_i} \gamma \frac{1}{\sqrt{\sigma_B^2+\epsilon}} - \frac{\hat{x}_i}{\sqrt{\sigma_B^2+\epsilon}} \sum_{j=1}^{m} \frac{\partial f}{\partial y_j} \gamma \hat{x}_j - \frac{1}{m\sqrt{\sigma_B^2+\epsilon}} \sum_{j=1}^{m} \frac{\partial f}{\partial y_j} \gamma}$$