Batch Normalization : Back-propagation

$$X = \begin{matrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{matrix}$$

$$\mu = \frac{x_{11}+x_{21}}{2}, \quad \frac{x_{12}+x_{22}}{2}, \quad \frac{x_{13}+x_{23}}{2}$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_{i1} \quad \frac{1}{N}\sum_{i=1}^{N} x_{i2}, \quad$$

$$\sigma^2 = \frac{1}{2}\left((x_{11}-\mu_1)^2 + (x_{21}-\mu_1)^2\right), \quad \frac{1}{2}\left[(x_{12}-\mu_2)^2 + (x_{22}-\mu_2)^2\right],$$

$$\frac{1}{2}\left[(x_{13}-\mu_3)^2 + (x_{23}-\mu_3)^2\right]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_{i1}-\mu_1)^2, \quad \frac{1}{N}\sum_{i=1}^{N}(x_{i2}-\mu_2)^2 \ldots\ldots$$

$$\hat{x} = \frac{x_{11}-\mu_1}{\sqrt{\sigma_1^2+\epsilon}} \quad \frac{x_{12}-\mu_2}{\sqrt{\sigma_2^2+\epsilon}} \quad \frac{x_{13}-\mu_3}{\sqrt{\sigma_3^2+\epsilon}}$$

$$\frac{x_{21}-\mu_1}{\sqrt{\sigma_1^2+\epsilon}} \quad \frac{x_{22}-\mu_2}{\sqrt{\sigma_2^2+\epsilon}} \quad \frac{x_{23}-\mu_3}{\sqrt{\sigma_3^2+\epsilon}}.$$

$$y = \gamma_1\hat{x}_{11}+\beta_1 \quad \gamma_2\hat{x}_{12}+\beta_2 \quad \gamma_3\hat{x}_{13}+\beta_3$$

$$\gamma_1\hat{x}_{21}+\beta_1 \quad \gamma_2\hat{x}_{22}+\beta_2 \quad \gamma_3\hat{x}_{23}+\beta_3$$

$$\frac{dL}{dy} = \begin{matrix} \frac{dL}{dy_{11}} & \frac{dL}{dy_{12}} & \frac{dL}{dy_{13}} \\ \frac{dL}{dy_{21}} & \frac{dL}{dy_{22}} & \frac{dL}{dy_{23}} \end{matrix}$$

upstream × local gradient

$$\frac{dL}{dr_1} = \sum_{i,j} \frac{dL}{dy_{ij}} \times \frac{dy_{ij}}{dr_1} = \frac{dL}{dy_{11}} \cdot \hat{x}_{11} + \frac{dL}{dy_{21}} \hat{x}_{21} = \frac{dy}{dr}$$

$$\frac{dL}{dr} = \left(\frac{dL}{dy} \odot \hat{x}\right) \rightarrow \text{sum along axis} = 0$$

$$\frac{dL}{dr_j} = \sum_{i=1}^{N} \hat{x}_{ij} \cdot \frac{dL}{dy_{ij}}$$

$\odot$ elementwise multiplication

$$\frac{dL}{d\beta_1} = \sum_{i,j} \frac{dL}{dy_{ij}} \frac{dy_{ij}}{d\beta_1} = \frac{dL}{dy_{11}} + \frac{dL}{dy_{12}}$$

$$\therefore \frac{dL}{d\beta_j} = \sum_{i=1}^{N} \frac{dL}{dy_{ij}}$$

$$\frac{dL}{d\beta} = \frac{dL}{dy} \rightarrow \text{sum along axis} = 0$$

$$\frac{dL}{d\hat{x}_{11}} = \sum \frac{dL}{dy_{ij}} \frac{dy_{ij}}{d\hat{x}_{11}} = \frac{dL}{dy_{11}} \times r_1$$

$$\therefore \frac{dL}{d\hat{x}} = \frac{dL}{dy} \odot r$$

x_{11} part of μ_1, σ_1^2, $\hat{x}_{11}$ → gradients sum

μ_1 — $\hat{x}_{11}, \hat{x}_{21}$

σ_1^2 — $\hat{x}_{11}, \hat{x}$

$$\frac{dL}{d\sigma_1^2} = \sum \frac{dL}{d\hat{x}_{ij}} \frac{d\hat{x}_{ij}}{d\sigma_1^2} = \frac{dL}{d\hat{x}_{11}} \cdot \frac{d\hat{x}_{11}}{d\sigma_1^2} + \frac{dL}{d\hat{x}_{21}} \cdot \frac{d\hat{x}_{21}}{d\sigma_1^2}$$

$$= \frac{dL}{d\hat{x}_{11}} \cdot \frac{d\hat{x}_{11}}{d\sigma_1^2} = \frac{d}{d\sigma_1^2} \frac{x_{11}-\mu_1}{(\sigma_1^2+\epsilon)^{0.5}} = -\frac{1}{2}(\sigma_1^2+\epsilon)^{-3/2}(x_{11}-\mu_1)$$

$$\therefore \boxed{\frac{dL}{d\sigma_1^2} = -\frac{1}{2}(\sigma_1^2+\epsilon)^{-3/2}\left[\frac{dL}{d\hat{x}_{11}}(x_{11}-\mu_1) + \frac{dL}{d\hat{x}_{21}}(x_{21}-\mu_1)\right]} \quad - A_1$$

$$= -\frac{1}{2}(\sigma_1^2+\epsilon)^{-1}\left[\frac{dL}{d\hat{x}_{11}}\frac{x_{11}-\mu_1}{(\sigma_1^2+\epsilon)^{1/2}} + \frac{dL}{d\hat{x}_{21}}\frac{(x_{21}-\mu_1)}{\sqrt{\sigma_1^2+\epsilon}}\right]$$

$$\frac{dL}{d\sigma_1^2} = -\frac{1}{2}(\sigma_1^2+\epsilon)^{-1}\left[\frac{dL}{d\hat{x}_{11}}\hat{x}_{11} + \frac{dL}{d\hat{x}_{21}}\hat{x}_{21}\right] \quad - A_1$$

$$\boxed{\frac{dL}{d\sigma_1^2} = -\frac{1}{2}(\sigma_1^2+\epsilon)^{-1}\left[\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}}\hat{x}_{i1}\right]}$$

$$\frac{dL}{d\mu_1} = \sum \frac{dL}{d\hat{x}_{ij}} \frac{d\hat{x}_{i1}}{d\mu_1} = \frac{dL}{d\hat{x}_{11}} \frac{d\hat{x}_{11}}{d\mu_1} + \frac{dL}{d\hat{x}_{21}} \frac{d\hat{x}_{21}}{d\mu_1}$$

$$\frac{d\hat{x}_{11}}{d\mu_1} = \frac{d}{d\mu_1}\left(\frac{x_{11}-\mu_1}{\sqrt{\sigma_1^2+\epsilon}}\right) = \frac{-1}{(\sigma_1^2+\epsilon)^{1/2}}$$

$$\boxed{\frac{dL}{d\mu_1} = \frac{-1}{(\sigma_1^2+\epsilon)^{0.5}}\left[\frac{dL}{d\hat{x}_{11}} + \frac{dL}{d\hat{x}_{21}}\right] - b_1}$$

$$\frac{d\sigma_1^2}{d\mu_1} = \frac{d}{d\mu_1}\left[\frac{x_{11}-\mu_1}{(\sigma_1^2+\epsilon)^{0.5}} + \frac{x_{21}-\mu_1}{(\sigma_1^2+\epsilon)^{0.5}}\right]$$

$$= \frac{d}{d\mu}\left(\frac{1}{2}\left[(x_{11}-\mu_1)^2 + (x_{21}-\mu_1)^2\right]\right)$$

$$= -\frac{2}{N}\left[(x_{11}-\mu_1) + (x_{21}-\mu_1)\right]$$

$$\boxed{\frac{d\sigma_1^2}{d\mu_1} = -\frac{2}{N}\sum_{i=1}^{N}(x_{i1}-\mu_1)}$$

$$\therefore \frac{dL}{d\mu_1} = b_1 + \frac{dL}{d\sigma_1^2}\cdot\frac{d\sigma_1^2}{d\mu_1} + \frac{-1}{2(\sigma_1^2+\epsilon)}\left[\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}}\,d\hat{x}_{i1}\right]\cdot\frac{-2}{N}\sum_{i=1}^{N}(x_{i1}-\mu_1$$

$$= \frac{-1}{(\sigma_1^2+\epsilon)^{0.5}}\left[\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}}\right] + \left[\frac{-1}{2(\sigma_1^2+\epsilon)}\left[\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}}\hat{x}_{i1}\right]\cdot\frac{-2}{N}\sum_{i=1}^{N}(x_{i1}-\right.$$

$$\frac{dL}{d\mu_1} = \frac{-1}{(\sigma_1^2+\epsilon)^{0.5}}\left(\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}}\right) + \frac{1}{N}\left(\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}}\hat{x}_{i1}\right)\cdot\sum_{i=1}^{N}(x_{i1}-\mu_1)$$

$$\boxed{\frac{dL}{d\mu_1} = \frac{-1}{(\sigma_1^2+\epsilon)^{0.5}}\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}} + \frac{1}{N}\left(\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}}\hat{x}_{i1}\right)\sum_{i=1}^{N}(x_{i1}-\mu_1) - B_1}$$

Consider $\sum_{i=1}^{N}(x_{i1}-\mu_1) = N.\mu_1 - N\mu_1 = 0$

$$\therefore \boxed{\frac{dL}{d\mu_1} = \frac{-1}{(\sigma_1^2+\epsilon)^{0.5}}\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}} - B_2}$$

$$\frac{dL}{dx_{11}} = \sum_i \sum_j \frac{dL}{d\hat{x}_{ij}} \frac{d\hat{x}_{ij}}{dx_{11}} = \frac{dL}{d\hat{x}_{11}} \cdot \frac{d\hat{x}_{11}}{dx_{11}}$$

$$\boxed{\frac{dL}{dx_{11}} = \frac{dL}{d\hat{x}_{11}} \cdot \frac{1}{(\sigma_1^2+\epsilon)^{0.5}}}$$

$$\boxed{\frac{d\mu_1}{dx_{11}} = \frac{1}{N}}$$

$$\frac{d\sigma_1^2}{dx_{11}} = \frac{2}{N}\sum_{i=1}^{N}(x_{i1}-\mu_1) + \frac{-2}{N}\sum_{i=1}^{N}(x_{i1}-\mu_1)\cdot\frac{d\mu_1}{dx_{11}} = \frac{-2}{N}\sum_{i=1}^{N}(x_{i1}-\mu_1)\frac{d\mu_1}{dx_{11}}$$

($\sum_{i=1}^{N}(x_{i1}-\mu_1) = 0$)

$$\frac{dL}{dx_{11}} = \frac{dL}{d\hat{x}_{11}}\cdot\frac{1}{(\sigma_1^2+\epsilon)^{0.5}} + \frac{dL}{d\mu_1}\frac{d\mu_1}{dx_{11}} + \frac{dL}{d\sigma_1^2}\times\frac{d\sigma_1^2}{dx_{11}}$$

$$= \frac{dL}{d\hat{x}_{11}}\cdot\frac{1}{(\sigma_1^2+\epsilon)^{0.5}} + \frac{-1}{(\sigma_1^2+\epsilon)^{0.5}}\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}}\cdot\frac{1}{N} + \frac{2}{N}\sum_{i=1}^{N}$$

$$+\left(\frac{-1}{2}\right)(\sigma_1^2+\epsilon)^{-1}\left(\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}}\hat{x}_{i1}\right)\cdot\frac{-2}{N}\sum_{i=1}^{N}(x_{i1}-\mu_1)\frac{d\mu_1}{dx_{11}}\left(\frac{1}{N}\right)$$

$$= \frac{dL}{dx_1}\frac{dL}{d\hat{x}_{11}}\frac{1}{(\sigma_1^2+\epsilon)^{0.5}} - \frac{1}{N(\sigma_1^2+\epsilon)^{0.5}}\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}} + \frac{1}{N^2(\sigma_1^2+\epsilon)^{0.5}}\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{11}}\hat{x}_{i1}$$

$$\cdot\sum_{i=1}^{N}\hat{x}_{i1}$$

$$= \frac{1}{N(\sigma_i^2+\epsilon)^{0.5}}\left[N \cdot \frac{dL}{d\hat{x}_{i1}} - \sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}} + \frac{1}{N}\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}}\hat{x}_{i1}\sum_{i=1}^{N}\hat{x}_{i1}\right]$$

$$\frac{d\sigma_i^2}{dx_{i1}} = \underbrace{\frac{2}{N}\sum_{i=1}^{N}(x_{i1}-\mu_1)}_{0} + \frac{-2}{N}\sum_{i=1}^{N}(x_{i1}-\mu_1)\cdot\frac{d\mu_1}{dx_{i1}}, \quad \frac{d\mu_1}{dx_{i1}} = \frac{1}{N}$$

$$\frac{d\sigma_i^2}{dx_{i1}} = 0 + \frac{-2}{N^2}\sum_{i=1}^{N}(x_{i1}-\mu_1) = 0.$$

$$\therefore \frac{dL}{dx_{i1}} = \frac{dL}{d\hat{x}_{i1}}\cdot\frac{1}{(\sigma_i^2+\epsilon)^{0.5}} + \frac{1}{N}\Bigg[$$

$$\frac{d\sigma_i^2}{dx_{i1}} = \frac{d}{dx_{i1}}\frac{1}{N}\sum_{i=1}^{N}(x_{i1}-\mu_1)^2$$

we don't take $\frac{d\sigma^2}{d\mu_1}\times\frac{d\mu_1}{dx_{i1}}\cdot\frac{dx_{i1}}{d\mu}$ since we already computed $\frac{dL}{dx_{i1}}$ wrt μ

$$= \frac{2}{N}(x_{i1}-\mu_1) + \left(\frac{-2}{N}\sum_{i=1}^{N}(x_{i1}-\mu_1)\frac{d\mu_1}{dx_{i1}}\right)$$

$$= \frac{2}{N}(x_{i1}-\mu_1) + \frac{-2}{N^2}\underbrace{\left(\sum_{i=1}^{N}x_{i1}-\mu_1\right)}_{=0} \quad \because \frac{d\mu_1}{dx_{i1}} = 1/N$$

$$\boxed{\frac{d\sigma_i^2}{dx_{i1}} = \frac{2}{N}(x_{i1}-\mu_1)}$$

$$\therefore \frac{dL}{dx_{i1}} = \frac{dL}{d\hat{x}_{i1}}\cdot\frac{1}{(\sigma_i^2+\epsilon)^{0.5}} + \frac{-1}{(\sigma_i^2+\epsilon)^{0.5}}\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}}\cdot\frac{1}{N}$$

$$\frac{-1}{2(\sigma_i^2+\epsilon)}\left(\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}}\hat{x}_{i1}\right)\cdot\frac{2}{N}(x_{i1}-\mu_1)$$

$$\boxed{\frac{dL}{dx_{i1}} = \frac{1}{N(\sigma_i^2+\epsilon)^{0.5}}\left[\frac{dL}{d\hat{x}_{i1}} - \sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}} - \left(\sum_{i=1}^{N}\frac{dL}{d\hat{x}_{i1}}\hat{x}_{i1}\right)\cdot\hat{x}_{i1}\right]} \quad - C$$