Batch Norm computational graph:-

$$\underset{(N,D)}{X} \longrightarrow \boxed{\begin{matrix}\text{Batch Norm} \\ BN(\gamma,\beta)\end{matrix}} \longrightarrow \underset{(N,D)}{Y} \longrightarrow \boxed{\text{some transf.}} \longrightarrow \underset{(1,)}{L}$$

$$\underset{(D,)}{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \underset{(D,)}{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

$$\underset{(N,D)}{\hat{x}} = \frac{x-\mu}{\sqrt{\sigma^2+\epsilon}} \qquad \underset{(N,D)}{y} = \underset{(D,)}{\gamma}\hat{x}+\beta$$ where γ, β are learnable params of a BN layer

<u>problem</u>:- $\frac{\partial L}{\partial Y}$ is known/available, of shape (N, D)

calculate $\underset{(N,D)}{\frac{\partial L}{\partial x}}, \underset{(D,)}{\frac{\partial L}{\partial \gamma}}, \underset{(D,)}{\frac{\partial L}{\partial \beta}}$

$\frac{\partial L}{\partial \beta} = \frac{\partial L}{\partial Y}\frac{\partial Y}{\partial \beta} \rightarrow$ since $\frac{\partial Y}{\partial \beta} = 1$, $\frac{\partial L}{\partial \beta} = \sum_{i=1}^{N}\frac{\partial L}{\partial y_i}$, where y_i is the i-th row of Y

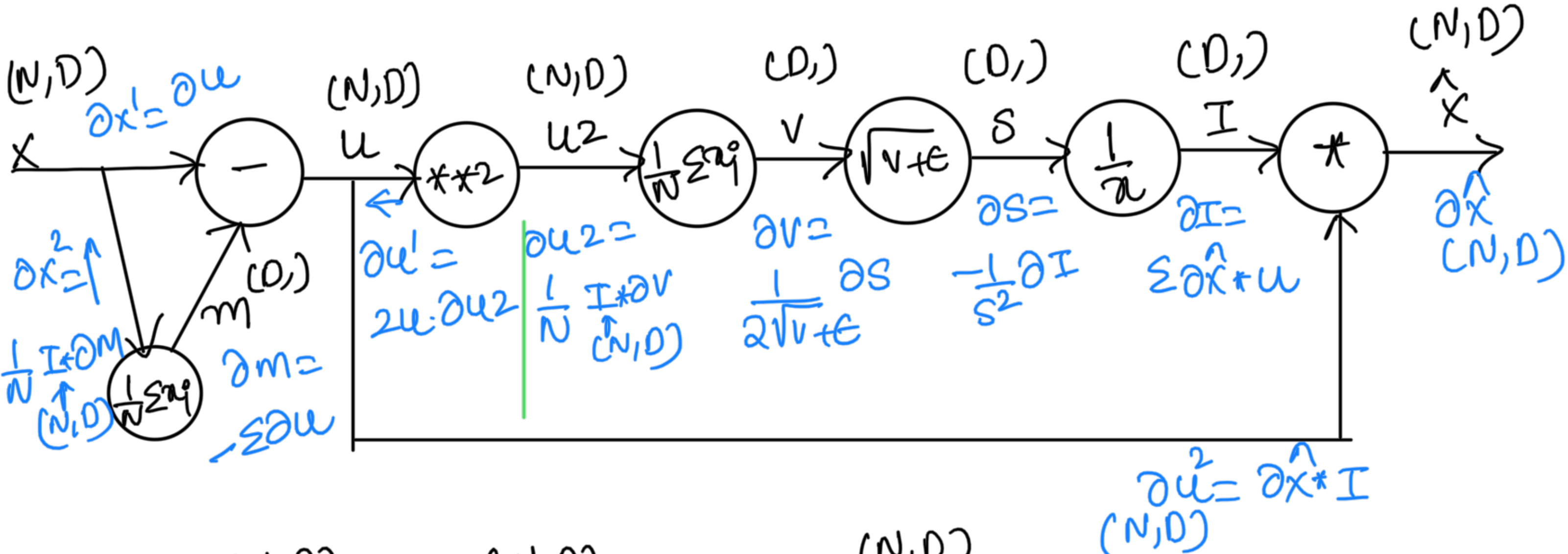
similarly, $\frac{\partial L}{\partial \gamma} = \frac{\partial L}{\partial Y}\frac{\partial Y}{\partial \gamma}$ and since $\frac{\partial Y}{\partial \gamma} = \hat{x}$,

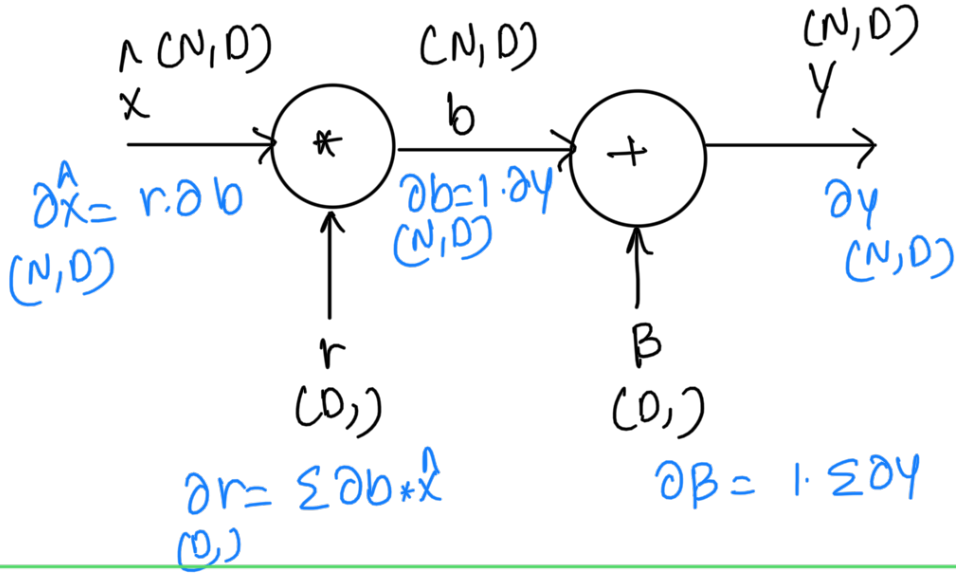
$\frac{\partial L}{\partial \gamma} = \sum_{i=1}^{N}\frac{\partial L}{\partial y_i} * \hat{x}_i$ where y_i & $\hat{x}_i$ are i-th rows of Y & $\hat{X}$, respectively.

$\frac{\partial L}{\partial \gamma}$ & $\frac{\partial L}{\partial \beta}$ are relatively straight forward to calculate.

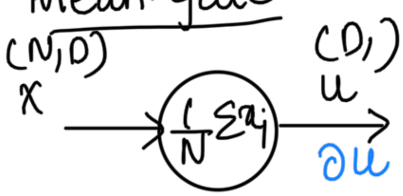
$\frac{\partial L}{\partial x}$ is easier to calculate using a computational graph.

a computational graph is shown below :-

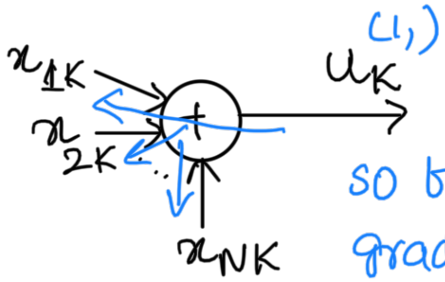




Mean-gate



mean-gate can be thought of as D sum-gates in parallel, each with N inputs corresponding to the N examples of a given feature.

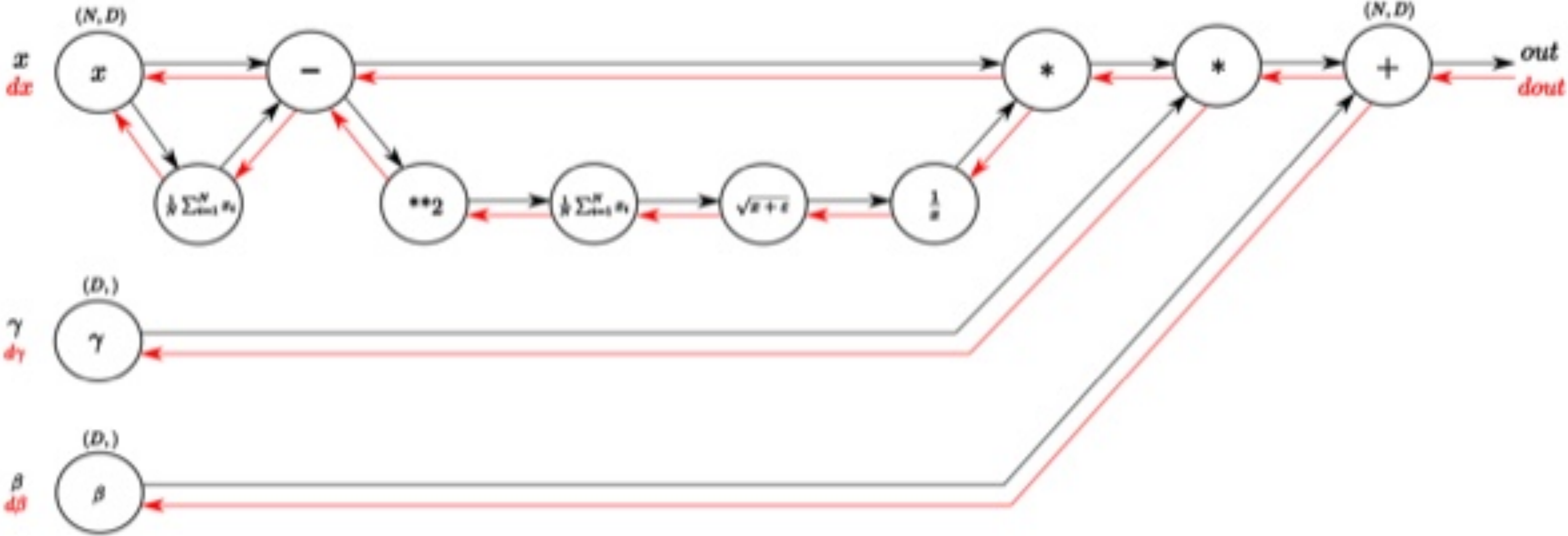


$K = \{1 \ldots D\}$, for D features

so for each feature u_K, the gradient is copied to the N samples

$\Rightarrow$ essentially copy u vector N times so it is of shape (N, D) and divide by N

$$\partial x = \frac{1}{N} \underset{(N,D)}{I} \, du$$



(N, D)
x
dx
x
−
*
*
(N, D)
+
out
dout
**2
(D,)
γ
dγ
γ
(D,)
β
dβ
β