Forward propagation

① $\text{sample\_mean} = E[X] = \frac{1}{N}\sum_{i=1}^{N} x_i$

② $\text{xsm} = x - \text{sample\_mean}$

③ $\text{sq\_xsm} = \text{xsm}^2$

④ $\text{sample\_var} = Var[X] = \frac{1}{N}\sum_{i=1}^{N}(x_i - E(x_i))^2 = \frac{1}{N}\sum_{i=1}^{N}(\text{xsm})^2 = \frac{1}{N}\sum_{i=1}^{N} \text{sq\_xsm}$

⑤ $\text{sve} = \text{sample\_var} + \text{eps}$

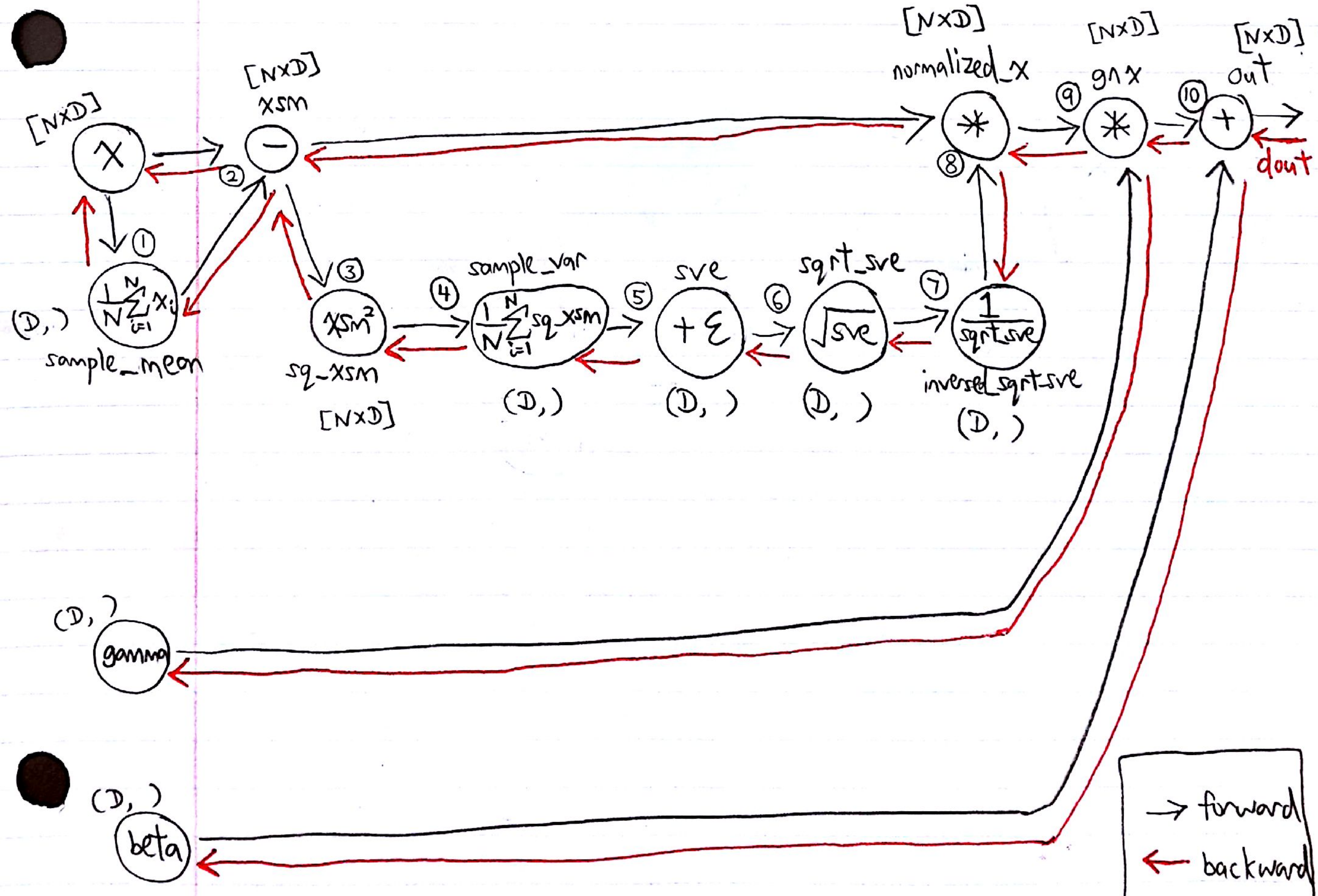
⑥ $\text{sqrt\_sve} = \sqrt{\text{sve}}$

⑦ $\text{inversed\_sqrt\_sve} = \frac{1}{\text{sqrt\_sve}}$

⑧ $\text{normalized\_x} = \text{xsm} * \text{inverse\_sqrt\_sve}$

⑨ $\text{gnx} = \text{gamma} * \text{normalized\_x}$

⑩ $\text{out} = \text{gnx} + \text{beta}$



Backward propagation

Recall Chain-Rule

$\boxed{\text{Eg.}}$ $f(x,y,z) = (x+y)z$

Let $q=(x+y)$ and $f=qz$, then $\underbrace{\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} * \frac{\partial q}{\partial x}}_{\text{Chain-Rule}}$, $\underbrace{\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} * \frac{\partial q}{\partial y}}_{\text{Chain-Rule}}$, $\frac{\partial f}{\partial q} = z$

$\boxed{\text{short-hand}}$: $dx = dq * \frac{\partial q}{\partial x} = dq * 1 = z$, $dy = dq * \frac{\partial q}{\partial y} = dq * 1 = z$, $dq = z$

---

(10) $\begin{cases} dgnx = 1 * dout \\ dbeta = 1 * dout \end{cases}$

(9) $\begin{cases} dgamma = normalized\_x * dgnx \\ dnormalized\_x = gamma * dgnx \end{cases}$

(8) $\begin{cases} dxsm_1 = inversed\_sqrt\_sve * dnormalized\_x \\ dinversed\_sqrt\_sve = xsm * dnormalized\_x \end{cases}$

(7) − $dsqrt\_sve = -\frac{1}{(sqrt\_sve)^2} * dinversed\_sqrt\_sve$

(6) − $dsve = \frac{1}{2} * \frac{1}{\sqrt{sve}} * dsqrt\_sve$

(5) $\begin{cases} dsample\_var = 1 * dsve \\ deps \leftarrow \text{hyperparameter, so we don't backpropagate it} \end{cases}$

(4) $dsq\_xsm = \frac{1}{N} * np.ones((N,D)) * dsample\_var$

(3) $dxsm_2 = 2 * xsm * dsq\_xsm$

$\boxed{\text{Note}}$: $dxsm = \underbrace{dxsm_1}_{\text{from normalized\_x}} + \underbrace{dxsm_2}_{\text{from sq\_xsm}}$ [which is true since node "xsm" inputs to two other nodes (i.e. node "normalized_x" and node "sq_xsm")]

(2) $\begin{cases} dx_1 = 1 * dxsm \\ dsample\_mean = -1 * dxsm \end{cases}$

(1) $dx_2 = \frac{1}{N} * np.ones((N,D)) * dsample\_mean$

$\boxed{\text{Note}}$ $dx = \underbrace{dx_1}_{\text{from xsm}} + \underbrace{dx_2}_{\text{from sample\_mean}}$