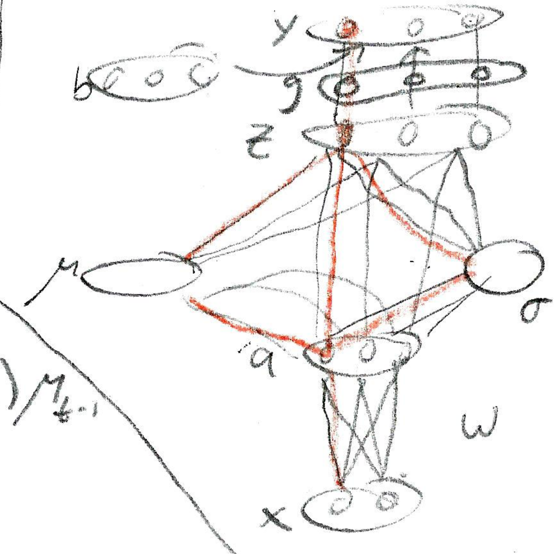
Layer Norm : $y = g \odot z + b$

Where: $z_o = \dfrac{a_o - \mu_t}{\sigma_t}$



- Averages
- Learnable

$D = n_{out}$

$m = \frac{1}{D}\sum_p a_p$

So: $\mu_t = \alpha_t m + (1-\alpha)\mu_{t-1}$

$$a_o = \sum_i w_{oi} x_i$$

$$\mu_t = \frac{\alpha_t}{D}\sum_p a_p + (1-\alpha_t)\mu_{t-1}$$

$$\sigma_t = \alpha_t \sqrt{\frac{1}{D-1}\sum_p (a_o - \mu_t)^2} + (1-\alpha_t)\mu_{t-1}$$

Goal: At time t, what are the gradients of the inputs and params: $\frac{\partial L}{\partial x_i}$, $\frac{\partial L}{\partial b_o}$, $\frac{\partial L}{\partial g_o}$, $\frac{\partial L}{\partial w_{oi}}$

Using chain rule, we can find the partials at every connection of the graph and combine.

$$\boxed{\frac{\partial y_o}{\partial g_o} = z_o} \qquad \boxed{\frac{\partial y_o}{\partial b_o} = 1} \qquad \boxed{\frac{\partial y_o}{\partial z_o} = g_o}$$

Prelim: Quotient Rule : $\frac{\partial}{\partial x}\left(\frac{g(x)}{h(x)}\right) = \frac{h(x)\frac{\partial}{\partial x}g(x) - g(x)\frac{\partial}{\partial x}h(x)}{h(x)^2}$

There is an alternate formulation which is more convenient for us here:

$$\frac{\partial}{\partial x}\left(\frac{g(x)}{h(x)}\right) = \frac{1}{h(x)}\cdot\frac{\partial}{\partial x}g(x) - \frac{g(x)}{h(x)^2}\frac{\partial}{\partial x}h(x)$$

The gradient flows from $z_o \to a_p$ in 3 ways... directly $z_o \to a_o$, and indirectly through μ_t and σ_t. This means when we compute $\frac{\partial z_o}{\partial x_i}$ or $\frac{\partial z_o}{\partial w_{pi}}$ that we need to sum over these pathways.

$$\frac{\partial z_o}{\partial a_p} = \frac{\partial}{\partial a_p}\left(\frac{a_o - \mu_t}{\sigma_t}\right)$$

Note: Use quotient rule with $g(x) = a_o - \mu_t \quad h(x) = \sigma_t$

$$= \underbrace{\frac{1}{\sigma_t}\frac{\partial}{\partial a_p}(a_o - \mu_t)}_{1_{o=p} - \frac{\partial \mu_t}{\partial a_p}} - \underbrace{\frac{a_o - \mu_t}{\sigma_t^2}}_{\frac{z_o}{\sigma_t}} \frac{\partial}{\partial a_p}\sigma_t = \frac{1}{\sigma_t}\left[1_{o=p} - \frac{\partial \mu_t}{\partial a_p} - z_o \frac{\partial \sigma_t}{\partial a_p}\right]$$

Indicator is 1 when $o=p$

So we need $\frac{\partial \mu_t}{\partial a_p}$ and $\frac{\partial \sigma_t}{\partial a_p}$. Lets derive those:

$$\boxed{\frac{\partial \mu_t}{\partial a_p} = \frac{\alpha_t}{D}}$$

$$\frac{\partial \sigma_t}{\partial a_p} = \alpha_t \frac{\partial}{\partial a_p}\left(\frac{1}{D-1}\sum_q (a_q - \mu_t)^2\right)^{1/2}$$

$$= \alpha_t \cdot \left(\frac{1}{2}\right) \cdot \underbrace{\left(\frac{1}{D-1}\sum_q (a_q - \mu_t)^2\right)^{-1/2}}_{\sigma_t^2} \cdot \frac{\partial}{\partial a_p}\left(\frac{1}{D-1}\sum_q (a_q - \mu_t)^2\right)$$

$$= \frac{\alpha_t}{2\sigma_t} \cdot \frac{1}{D-1} \cdot \sum_q \frac{\partial}{\partial a_p}(a_q - \mu_t)^2$$

$$= \frac{\alpha_t}{2\sigma_t (D-1)} \cdot \sum_q \left[2 \cdot (a_q - \mu_t) \cdot \frac{\partial}{\partial a_p}(a_q - \mu_t)\right]$$

$$\frac{\partial \sigma_t}{\partial a_p} = \frac{\alpha_t}{\cancel{2}\sigma_t(D-1)} \cdot \cancel{2} \cdot \sum_q (a_q - \mu_t)\left[\underbrace{\frac{\partial a_q}{\partial a_p}}_{1_{p=q}} - \underbrace{\frac{\partial \mu_t}{\partial a_p}}_{\frac{\alpha_t}{D}}\right]$$

$$= \frac{\alpha_t}{\sigma_t(D-1)}\left[(a_p - \mu_t) - \frac{\alpha_t}{D}\sum_q (a_q - \mu_t)\right]$$

$$= \frac{\alpha_t}{D-1} \cdot z_p - \frac{\alpha_t^2}{\sigma_t D(D-1)}\left[\sum_q a_q - \underbrace{\sum_q \mu_t}_{D\mu_t}\right]$$

$$= \frac{\alpha_t}{D-1} z_p - \frac{\alpha_t^2}{\sigma_t(D-1)}\mu_t - \frac{\alpha_t^2}{\sigma_t(D-1)}\left(\frac{1}{D}\sum_q a_q\right)$$

$$\boxed{\frac{\partial \sigma_t}{\partial a_p} = \frac{\alpha_t}{D-1} \cdot \left[z_p - \frac{\alpha_t}{\sigma_t}\left(\mu_t - \underbrace{\frac{1}{D}\sum_q a_q}_{\mu_t}\right)\right]}$$

Difference of online mean vs. this timestep's mean

Back to computing $\frac{\partial z_o}{\partial a_p}$ ...

$$\frac{\partial z_o}{\partial a_p} = \frac{1}{\sigma_t}\left[1_{o=p} - \frac{\partial \mu_t}{\partial a_p} - z_o \frac{\partial \sigma_t}{\partial a_p}\right]$$

$$= \frac{1}{\sigma_t}\left[1_{o=p} - \frac{\alpha_t}{D} - z_o\left(\frac{\alpha_t}{D-1}\right)\left[z_p - \alpha_t \cdot \frac{\mu_t - m}{\sigma_t}\right]\right]$$

$$\boxed{\frac{\partial z_o}{\partial a_p} = \frac{1}{\sigma_t}\left[1_{o=p} - \alpha_t\left[\frac{1}{D} + \frac{z_o z_p}{D-1} - \alpha_t z_o\left(\frac{\mu_t - m}{\sigma_t}\right)\right]\right]}$$

$$\boxed{\frac{\partial a_p}{\partial x_i} = w_{pi}} \qquad \boxed{\frac{\partial a_p}{\partial w_{qi}} = 1_{p=q} \cdot x_i}$$

The formulas we really care about are $\frac{\partial L}{\partial x_i}$ and $\frac{\partial L}{\partial w_{qi}}$:

$$\frac{\partial L}{\partial x_i} = \sum_o \frac{\partial L}{\partial y_o}\cdot\frac{\partial y_o}{\partial x_i} = \sum_o \frac{\partial L}{\partial y_o}\frac{\partial y_o}{\partial z_o}\frac{\partial z_o}{\partial x_i} = \sum_o \nabla_{y_o}\cdot g_o\cdot\frac{\partial z_o}{\partial x_i}$$

$$\frac{\partial L}{\partial w_{qi}} = \dots = \sum_o \nabla_{y_o}\cdot g_o\cdot\frac{\partial z_o}{\partial w_{qi}}$$

need these

$$\frac{\partial z_o}{\partial x_i} = \sum_p \frac{\partial z_o}{\partial a_p}\underbrace{\frac{\partial a_p}{\partial x_i}}_{w_{pi}} = \sum_p w_{pi}\frac{\partial z_o}{\partial a_p}$$

$$\frac{\partial L}{\partial x_i} \overset{\text{rearrange}}{=} \sum_p w_{pi}\cdot\underbrace{\left(\sum_o \nabla_{y_o} g_o \frac{\partial z_o}{\partial a_p}\right)}_{\nabla a_p}$$

$$\boxed{\nabla x_i = \sum_p \nabla a_p \, w_{pi}}$$

$$\frac{\partial L}{\partial w_{qi}} = \sum_p \nabla a_p \frac{\partial a_p}{\partial w_{qi}} = \sum_p \nabla a_p \left(1_{p=q}\, x_i\right)$$

$$\boxed{\nabla w_{qi} = \nabla a_q \cdot x_i}$$

Backprop Affine w/o bias!

Conclusion:

- Split LayerNorm into $\underbrace{a \to y}_{\text{norm}}$ and $\underbrace{x \to a}_{\text{linear}}$.

- Compute ∇g, ∇b, ∇a in the norm layer.

- Compute ∇w, ∇x in the linear layer using backprop eqns and no bias.

$$\nabla a_p = \sum_o \frac{\partial L}{\partial z_o} \frac{\partial z_o}{\partial a_p} = \sum_o \underbrace{\frac{\partial L}{\partial y_o}}_{\nabla y_o} \cdot \underbrace{\frac{\partial y_o}{\partial z_o}}_{g_o} \cdot \frac{\partial z_o}{\partial a_p}$$

$$\boxed{\nabla a_p = \sum_o \nabla y_o \, g_o \cdot \left(\frac{\partial z_o}{\partial a_p} \right)}$$

$$\boxed{\nabla g_o = \nabla y_o \cdot z_o}$$

$$\boxed{\nabla b_o = \nabla y_o}$$

6)