

Backpropagation

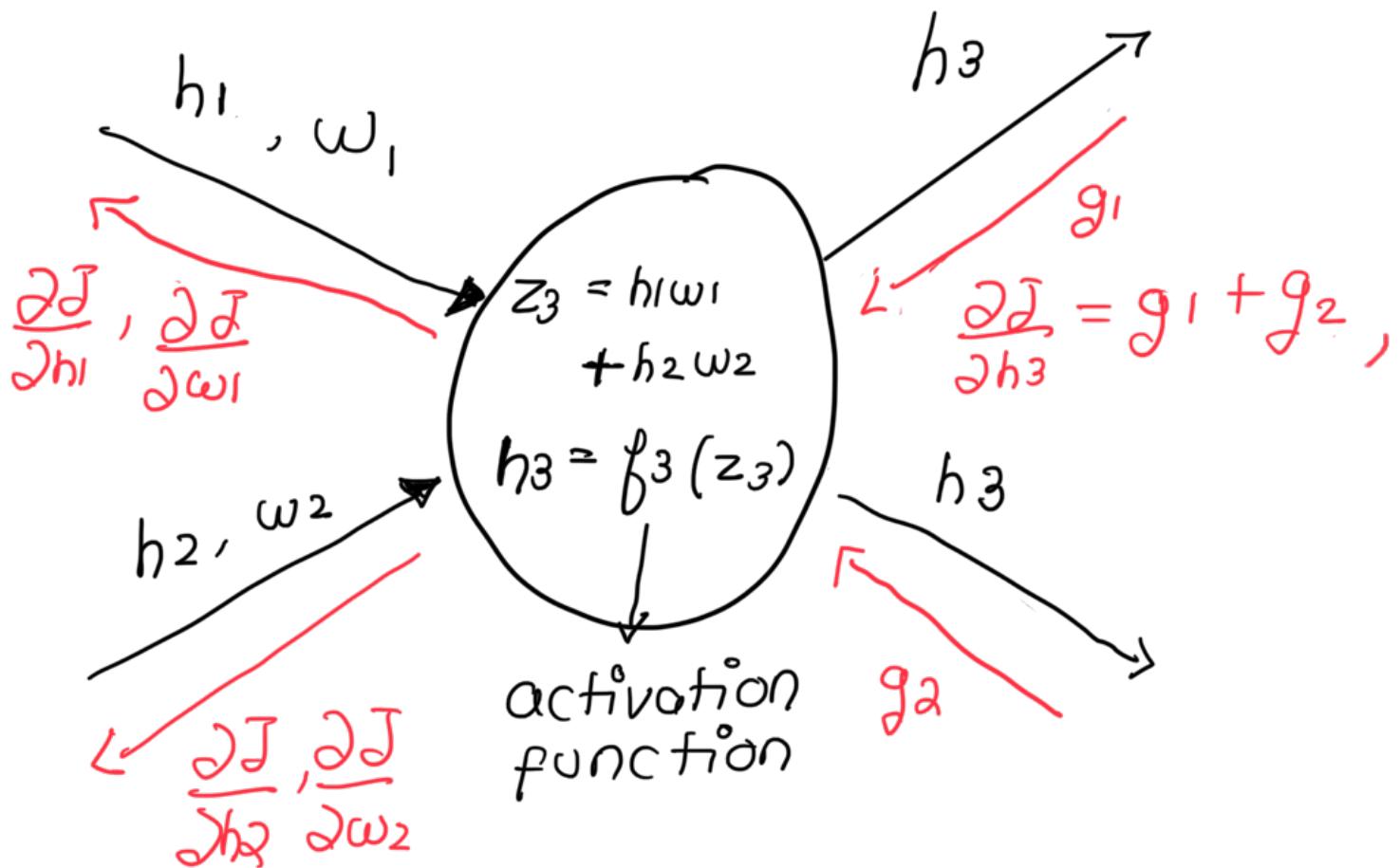
The goal of backpropagation is to compute $\frac{\partial J}{\partial \omega}$ for every weight ω in the network and perform the update

$$\omega = \omega - \ell \frac{\partial J}{\partial \omega}$$

so that on the next forward pass, the loss J will reduce.

Let's zoom in and observe what BP does at a particular neuron in the network:

.....



where

$$\frac{\partial J}{\partial h_1} = \frac{\partial J}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_1}$$

$$\frac{\partial J}{\partial w_1} = \frac{\partial J}{\partial h_3} \cdot \frac{\partial h_3}{\partial w_1}$$

$$\frac{\partial J}{\partial h_2} = \frac{\partial J}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_2}$$

$$\frac{\partial J}{\partial w_2} = \frac{\partial J}{\partial h_3} \cdot \frac{\partial h_3}{\partial w_2}$$

Also keep in mind that

$$\frac{\partial h^i}{\partial a} \quad \text{for some } i \text{ and variable } a$$

$$= \frac{\partial h^i}{\partial z^i} \cdot \frac{\partial z^i}{\partial a}$$

because of activations, but for the sake of keeping the chains shorter, I won't always expand

Now, we've setup a procedure we can follow at every neuron that computes

$$h' = f'(z')$$
 where

z' is the activation and f'

$h^0 = \omega_1 h_1 + \dots + \omega_K h_K$ (the weighted sum of outputs from the previous layer)

1. If h' reaches the final output through multiple paths, it will accumulate multiple 'sub gradients' g^1, \dots, g_j . Add these up, this gives
$$g_1 + \dots + g_j = \frac{\partial J}{\partial h},$$

(You can think of each g_i as one of the summands when you compute $\frac{\partial J}{\partial h}$, using the chain rule)

2. Compute the partial derivative of h' w.r.t every input $\omega_1, h_1, \dots, \omega_K, h_K$ - $\frac{\partial J}{\partial \omega_i}$ and $\frac{\partial J}{\partial h_i}$.

(we also need the gradients
w.r.t h_1, \dots, h_K because
they are required to calculate
 $\frac{\partial J}{\partial w}$ for weights w in earlier
layers)

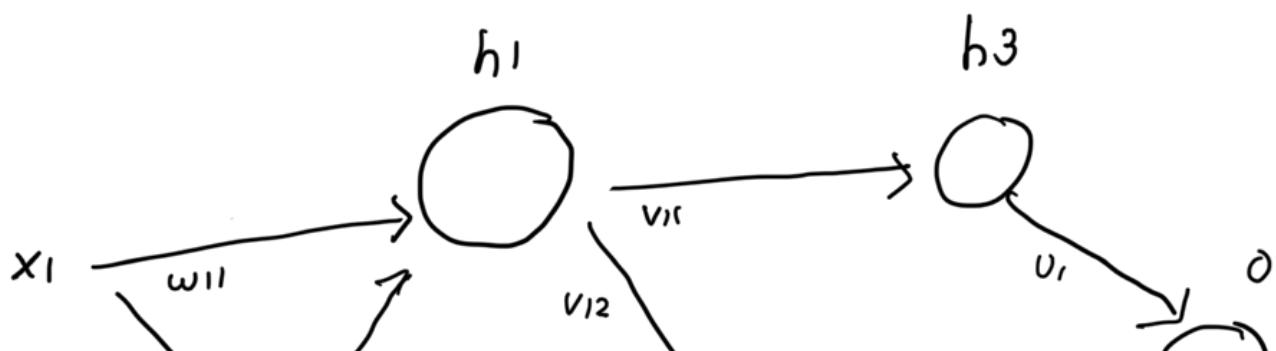
3. Applying the chain rule for
every input!

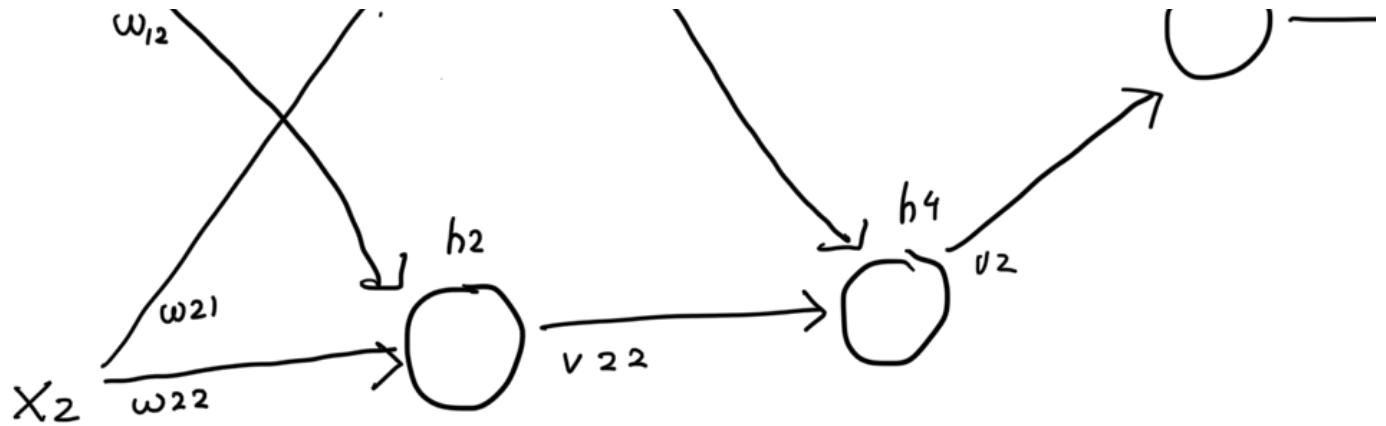
$$\text{Ex: } \frac{\partial J}{\partial h_i} = \underbrace{\frac{\partial J}{\partial h'}}_{\text{outgoing gradient}} \circ \underbrace{\frac{\partial h'}{\partial h_i}}_{\substack{\text{incoming,} \\ \text{precomputed} \\ \text{gradient}}} \underbrace{\frac{\partial h'}{\partial h_i}}_{\substack{\text{local} \\ \text{gradient} \\ \text{computed in} \\ \text{step 2}}}$$

Notice how we are performing
the chain rule incrementally, one
chain 'link' at a time. This
is exactly why backprop is
incremental - it isn't repeatedly

efficiently calculating long chains for every variable because it doesn't need to. The compositional nature of neural networks means that 'sub' chains computed at later layers can be reused to compute the longer chains at earlier layers, meaning **HUGE** savings on computation overall, especially for large, wide, denser nets

Let's do an example:



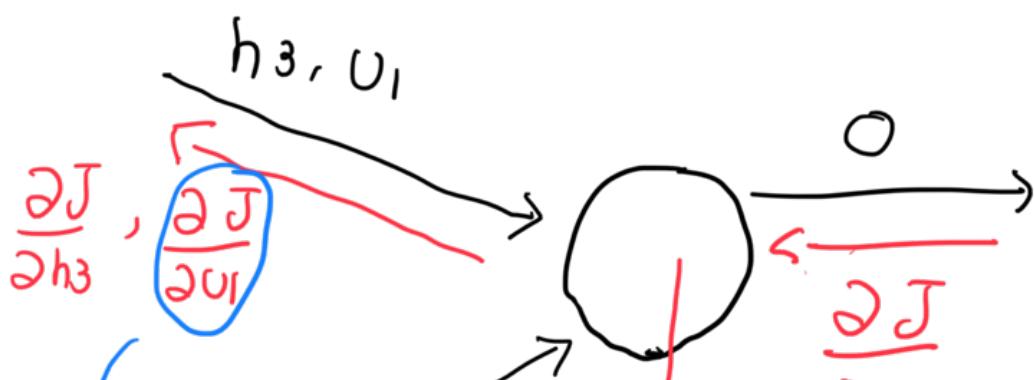


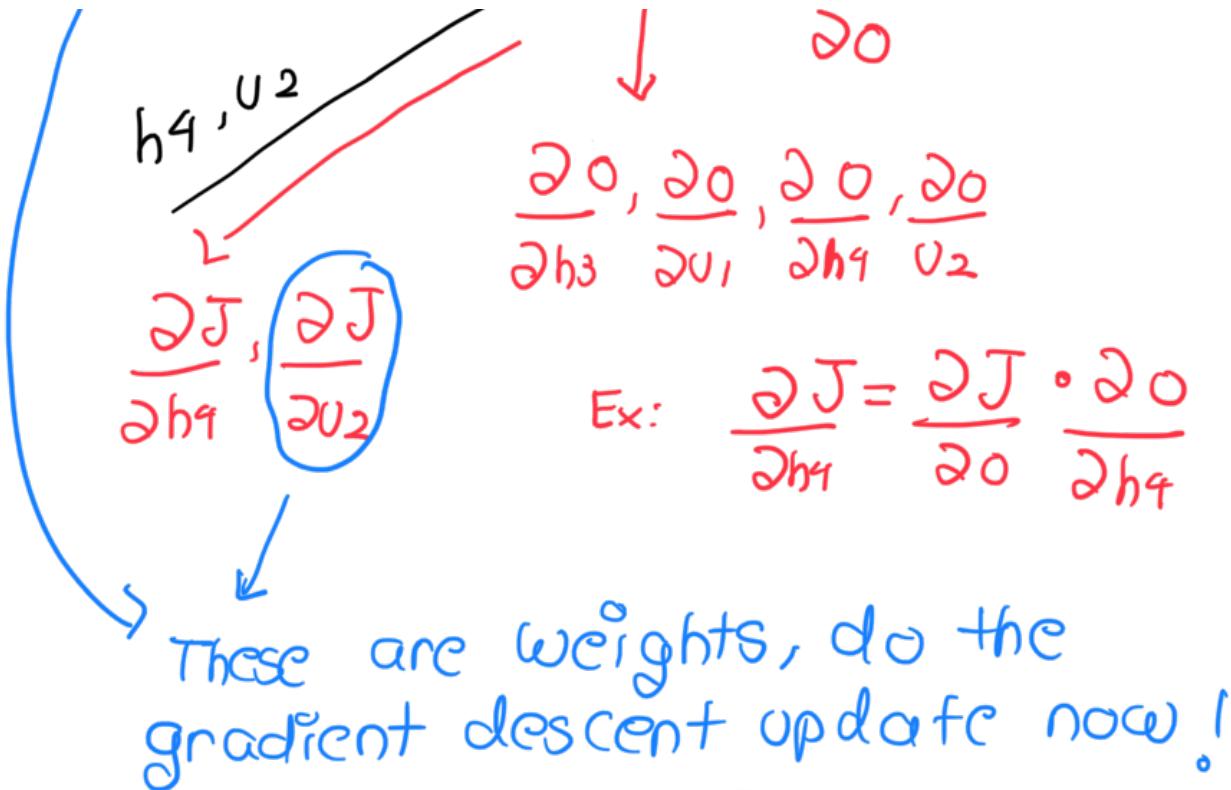
$h_i = f_i(z_i)$ where $z_i = \text{weighted sum of inputs}$

$o = f_o(z_o)$ where $z_o = v_1 h_3 + v_2 h^4$

o = predictions of the network

With o , we can compute our loss J since we have both the targets and predictions. And then as our first step in back prop, we compute $\frac{\partial J}{\partial o}$

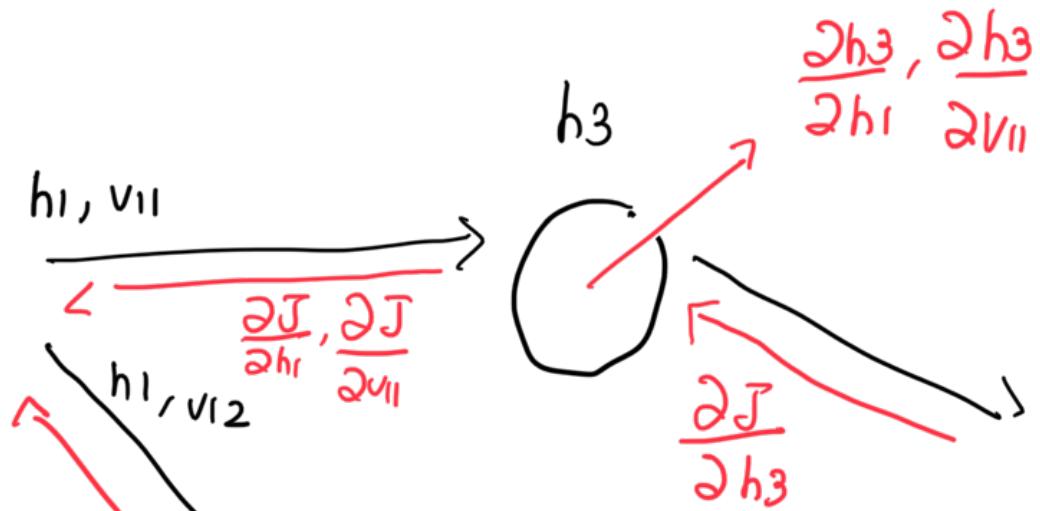


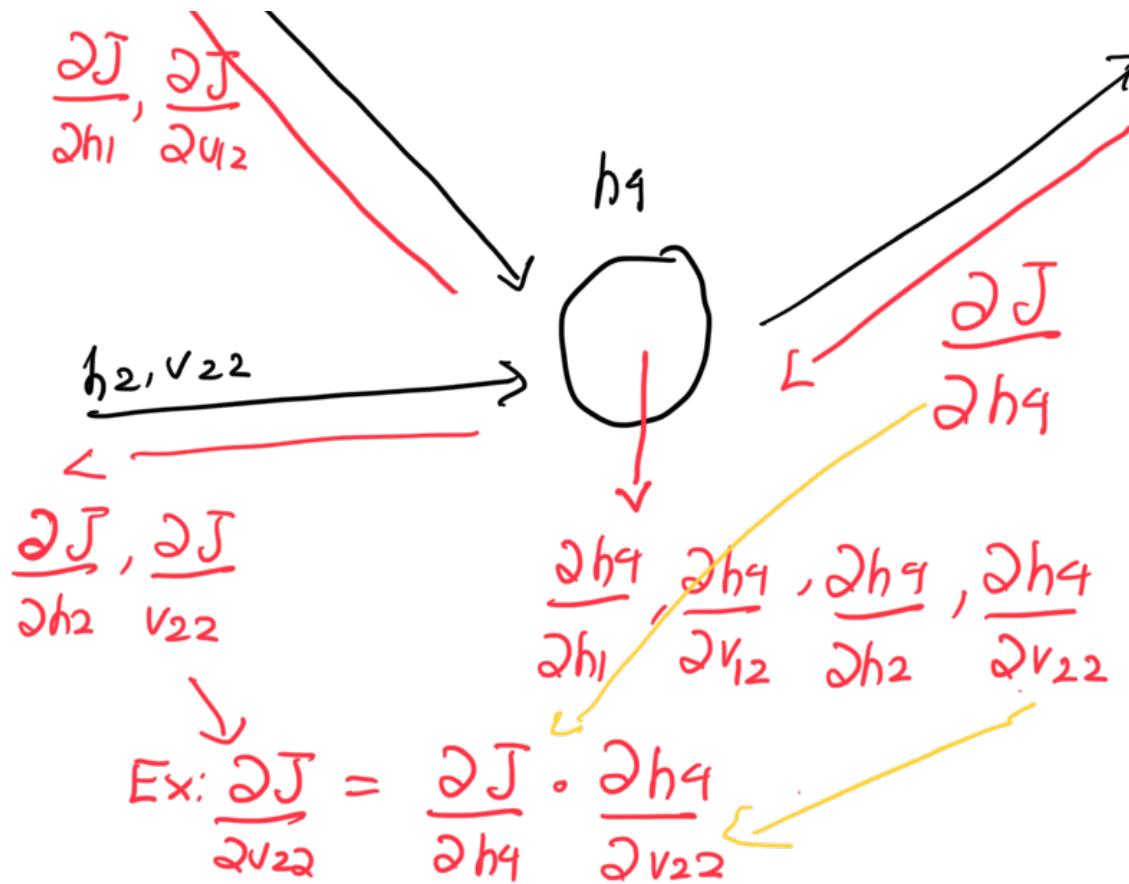


$$u_1 = u_1 - \ell \frac{\partial J}{\partial u_1}$$

$$u_2 = u_2 - \ell \frac{\partial J}{\partial u_2}$$

Now we move backward to the previous layer ...





We've calculated $\frac{\partial J}{\partial v_{11}}, \frac{\partial J}{\partial v_{12}}, \frac{\partial J}{\partial v_{22}}$.

Now we can do the gradient descent update for those

Last layer! (combining input arrows)



$$\left\langle \frac{\partial J}{\partial w_{11}}, \frac{\partial J}{\partial w_{21}} \right\rangle$$

$$\left\langle \frac{\partial h_1}{\partial w_{11}}, \frac{\partial h_1}{\partial w_{21}} \right\rangle$$

$$\frac{\partial J}{\partial h_1}$$

$$\left\langle \frac{\partial h_2}{\partial w_{12}}, \frac{\partial h_2}{\partial w_{22}} \right\rangle$$

$$\left\langle x_1, w_{12}, x_2, w_{22} \right\rangle$$

$$\left\langle \frac{\partial J}{\partial w_{12}}, \frac{\partial J}{\partial w_{22}} \right\rangle$$

$$\text{Ex: } \frac{\partial J}{\partial w_{22}} = \frac{\partial J}{\partial h_2} \circ \frac{\partial h_2}{\partial w_{22}}$$



Notice a few things:

- 1) There are 2 $\frac{\partial J}{\partial h_i}$'s coming into h_1 . These are the 'sub' gradients I was referring to earlier. The actual,

Official $\frac{\partial J}{\partial h_i}$ would be the

sum of these and what
ultimately gets back propagated
/ used. Ex: when calculating

$$\frac{\partial J}{\partial w_{ii}} = \frac{\partial J}{\partial h_i} \cdot \frac{\partial h_i}{\partial w_{ii}}$$

I didn't relabel these sub
gradients because it keeps
things cleaner and easier
to follow. But remember,
neurons with more than 1
channel of output accumulate
/ add multiple subgradients
to compute a final gradient
that is shared with earlier
layers

2) we didn't calculate $\frac{\partial J}{\partial w}$ for x_1, x_2 . Because we don't have to! They would be useless - there are no earlier layers and it doesn't make sense to adjust inputs (we don't have control over that, duh)

That is one complete backward pass! We could compute $\frac{\partial J}{\partial w}$ for all weights w in the network. Next, we would repeat the cycle of doing a forward pass of predicting with the

updated weights, backpropagating
the gradients and so on till
convergence (weights stabilize)

To show that backprop is
correct and efficient,
consider what the formula
for $\frac{\partial J}{\partial w_{11}}$ would be.

w_1 influences o through 2
paths :

$$1) w_1 \rightarrow h_1 \rightarrow h_3 \rightarrow o$$

$$2) w_1 \rightarrow h_1 \rightarrow h_4 \rightarrow o$$

So we'll have 2 summands
in the formula for $\frac{\partial J}{\partial w_{11}}$:

$$\frac{\partial J}{\partial w_{11}} = \frac{\partial J}{\partial o} \frac{\partial o}{\partial h_3} \frac{\partial h_3}{\partial h_1} \frac{\partial h_1}{\partial w_{11}} +$$

$$\frac{\partial J}{\partial o} \frac{\partial o}{\partial h_4} \frac{\partial h_4}{\partial h_1} \frac{\partial h_1}{\partial w_{11}}$$

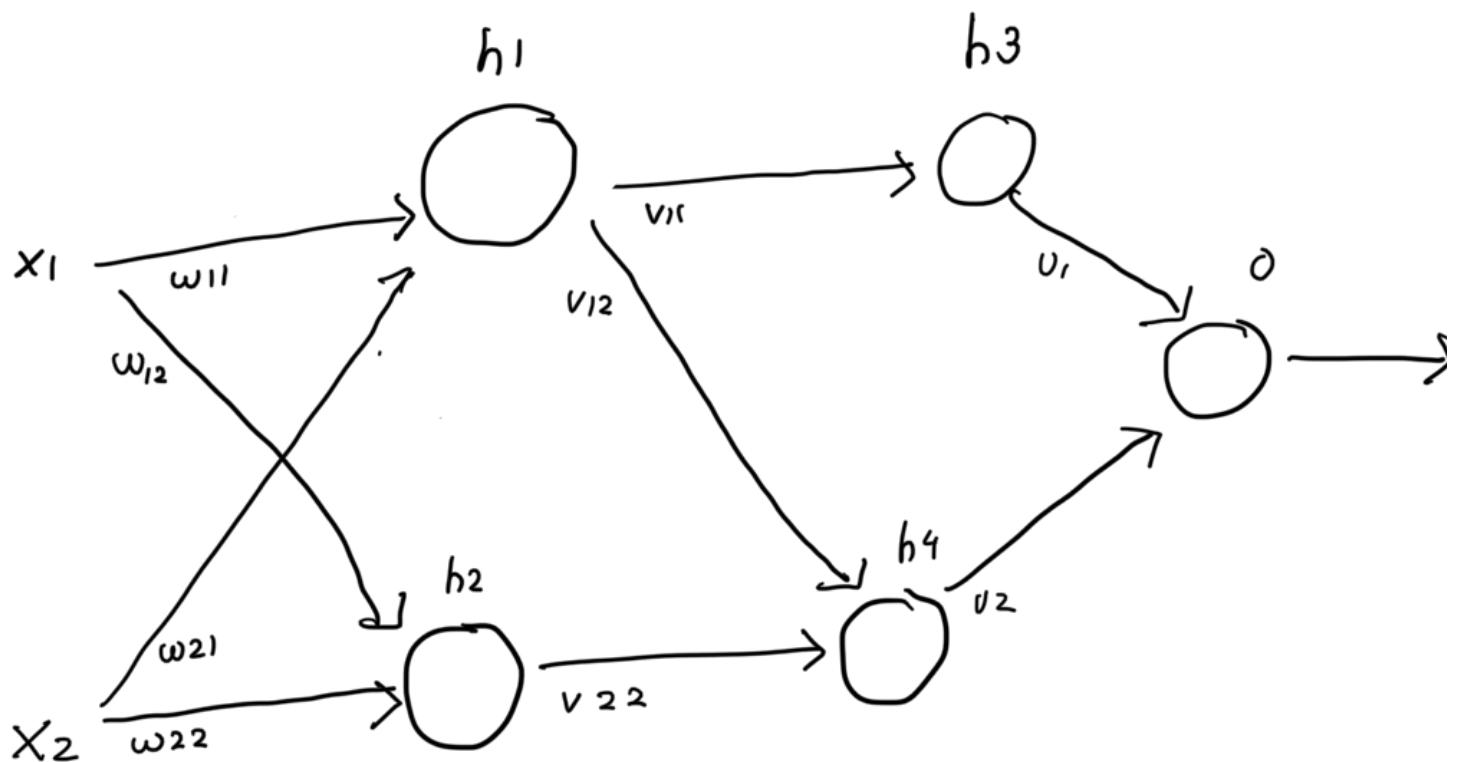
Now let's see if we get this from our backprop calculation.
Continuously unavelling the chain:

$$\begin{aligned}\frac{\partial J}{\partial w_{11}} &= \frac{\partial J}{\partial h_1} \cdot \frac{\partial h_1}{\partial w_{11}} \\ &\quad \parallel \\ &= \left(\underbrace{\frac{\partial J}{\partial h_3} \frac{\partial h_3}{\partial h_1} + \frac{\partial J}{\partial h_4} \frac{\partial h_4}{\partial h_1}}_{\left(\frac{\partial J}{\partial o} \frac{\partial o}{\partial h_3} + \frac{\partial J}{\partial o} \frac{\partial o}{\partial h_4} \right) \frac{\partial h_1}{\partial h_1}} \right) \frac{\partial h_1}{\partial w_{11}} \\ &= \left(\underbrace{\frac{\partial J}{\partial o} \frac{\partial o}{\partial h_3} \frac{\partial h_3}{\partial h_1}}_{\frac{\partial J}{\partial o} \frac{\partial o}{\partial h_3} \frac{\partial h_3}{\partial h_1}} + \underbrace{\frac{\partial J}{\partial o} \frac{\partial o}{\partial h_4} \frac{\partial h_4}{\partial h_1}}_{\frac{\partial J}{\partial o} \frac{\partial o}{\partial h_4} \frac{\partial h_4}{\partial h_1}} \right) \frac{\partial h_1}{\partial w_{11}} \\ &= \underbrace{\frac{\partial J}{\partial o} \frac{\partial o}{\partial h_3} \frac{\partial h_3}{\partial h_1}}_{\frac{\partial J}{\partial o} \frac{\partial o}{\partial h_3} \frac{\partial h_3}{\partial h_1}} + \underbrace{\frac{\partial J}{\partial o} \frac{\partial o}{\partial h_4} \frac{\partial h_4}{\partial h_1}}_{\frac{\partial J}{\partial o} \frac{\partial o}{\partial h_4} \frac{\partial h_4}{\partial h_1}} \frac{\partial h_1}{\partial w_{11}}\end{aligned}$$

20 ans JHS 2011 20 ENT JH1 2011
and voila !

Now let's do an example
with actual numbers and
functions

Reminder about the network:



$$h_i^o = f_i^o(z_i^o)$$

$z_i^o = \text{weighted sum of inputs}$

Define

$$h_1 = f_1(x) = x$$

$$h_2 = f_2(x) = \frac{1}{2}x^2$$

$$h_3 = f_3(x) = 2x$$

$$h_4 = f_4(x) = e^x$$

$$O = f_0(x) = \sin(x)$$

And assume all weights are initialized to 0, i.e

$$w_{11} = w_{12} = w_{21} = w_{22} = v_{11} = v_{12} = v_{22} = v_1 = v_2 = 0$$

Let's do one cycle of a forward and backward pass with the training instance

$$(x_1, x_2, y) = (0, 1, 0)$$

Step 1: forward pass

Calculate:

$$L = n \cdot n + n \cdot 1 = 7$$

$$h_1 = \frac{1}{2}(0.0 + 0.1)^2 = \frac{1}{2} \cdot 0 = 0$$

$$h_2 = \frac{1}{2}(0.0 + 0.0) = 0$$

$$h_3 = e^{(0.0 + 0.0)} = e^0 = 1$$

$$o = \sin(0.0 + 0.1) = \sin(0) = 0$$

Step 2 : Compute loss

Let's define

$$J \text{ as } \frac{1}{2}(o - t)^2$$

where t is the target
and o is the output of
the network

Calculate

$$\frac{\partial J}{\partial o} = o - t$$

and plug in t, o

$$t - o = o - o = 0$$

So $\frac{\partial J}{\partial o} = 0$. At this point, you could realize that since all the network chains include $\frac{\partial J}{\partial o}$, all partials collapse to 0, meaning all weights stay the same!

Logically, this makes sense - if the network is producing the right output, why should any weights update?

Doing the calculations for completeness of understanding backprop

Class 2 • Back propagate

STCP \circ Layer 3

(Remember to plug in values you've already computed, including from the forward pass)

O

Calculate :

$$\frac{\partial O}{\partial h_3} = \frac{\partial f_0}{\partial z_0} \cdot \frac{\partial z_0}{\partial h_3} = \cos(z_0) U_1 \\ = 0 \cdot (\cos(0)) \\ = 0 \cdot 1 = 0$$

$$\frac{\partial O}{\partial h_4} = \frac{\partial f_0}{\partial z_0} \cdot \frac{\partial z_0}{\partial h_4} = \cos(z_0) U_2 \\ = 0 \cdot (\cos(0)) \\ = 0 \cdot 1 = 0$$

$$\frac{\partial O}{\partial U_1} = \frac{\partial f_0}{\partial z_0} \cdot \frac{\partial z_0}{\partial U_1} = \cos(z_0) h_3 \\ = (\cos(0)) 0 \\ = 1 \cdot 0 = 0$$

$$\frac{\partial O}{\partial U_2} = \frac{\partial f_0}{\partial z_0} \cdot \frac{\partial z_0}{\partial U_2} = \cos(z_0) h_4 \\ = (\cos(0)) 1 = 1$$

$$\frac{\partial v_2}{\partial z_0} \quad \alpha v_2$$

$$\frac{\partial J}{\partial h_3} = \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial h_3} = 0 \cdot 0 = 0$$

$$\frac{\partial J}{\partial h_4} = \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial h_4} = 0 \cdot 0 = 0$$

$$\frac{\partial J}{\partial v_1} = \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial v_1} = 0 \cdot 0 = 0$$

$$\frac{\partial J}{\partial v_2} = \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial v_2} = 0 \cdot 1 = 0$$

Layer 2

h_3

Calculate:

$$\frac{\partial h_3}{\partial h_1} = \frac{\partial f_3}{\partial z_3} \cdot \frac{\partial z_3}{\partial h_1} = 2 \cdot v_{11} = 2 \cdot 0 = 0$$

$$\frac{\partial h_3}{\partial v_{11}} = \frac{\partial f_3}{\partial z_3} \cdot \frac{\partial z_3}{\partial v_{11}} = 2 \cdot h_1 = 2 \cdot 0 = 0$$

$$\frac{\partial J}{\partial h_1} = \frac{\partial J}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_1} = 0 \cdot 0 = 0$$

$$\frac{\partial J}{\partial v_{11}} = \frac{\partial J}{\partial h_3} \cdot \frac{\partial h_3}{\partial v_{11}} = 0 \cdot 0 = 0$$

h_4

Calculate:

$$\begin{aligned} \frac{\partial J}{\partial h_1} &= \frac{\partial J}{\partial h_4} \cdot \frac{\partial h_4}{\partial h_1} = \frac{\partial J}{\partial h_4} \cdot \frac{\partial f_4}{\partial z_4} \cdot \frac{\partial z_4}{\partial h_1} \\ &= 0 \cdot e^{z_4} \cdot v_{12} = 0 \cdot e^0 \cdot 0 = 0 \end{aligned}$$

$$\frac{\partial J}{\partial h_2} = \frac{\partial J}{\partial h_4} \cdot \frac{\partial h_4}{\partial h_2} = \frac{\partial J}{\partial h_4} \cdot \frac{\partial f_4}{\partial z_4} \cdot \frac{\partial z_4}{\partial h_2}$$

$$= 0^\circ C^{z^q} \cdot V_{22} = 0^\circ C^0 \cdot 0 = 0$$

$$\begin{aligned}\frac{\partial J}{\partial V_{12}} &= \frac{\partial J}{\partial h^q} \cdot \frac{\partial h^q}{\partial V_{12}} = \frac{\partial J}{\partial h^q} \cdot \frac{\partial f^q}{\partial z^q} \cdot \frac{\partial z^q}{\partial V_{12}} \\ &= 0^\circ C^{z^q} \cdot h_1 = 0^\circ C^0 \cdot 0 = 0\end{aligned}$$

$$\begin{aligned}\frac{\partial J}{\partial V_{22}} &= \frac{\partial J}{\partial h^q} \cdot \frac{\partial h^q}{\partial V_{22}} = \frac{\partial J}{\partial h^q} \cdot \frac{\partial f^q}{\partial z^q} \cdot \frac{\partial z^q}{\partial V_{22}} \\ &= 0^\circ C^{z^q} \cdot h_2 = 0^\circ C^0 \cdot 0 = 0\end{aligned}$$

Add up the 2 $\frac{\partial J}{\partial h_i}$'s you calculated and treat this

as your new $\frac{\partial J}{\partial h_1}$

$$\frac{\partial J}{\partial h_1} = 0 + 0 \\ = 0$$

Layer 3

h_1

Calculate

$$\frac{\partial J}{\partial w_{11}} = \frac{\partial J}{\partial h_1} \cdot \frac{\partial h_1}{\partial w_{11}} = \frac{\partial J}{\partial h_1} \cdot \frac{\partial f_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_{11}} \\ = 0 \cdot 1 \cdot x_1 = 0 \cdot 1 \cdot 0 = 0$$

$$\frac{\partial J}{\partial w_{12}} = \frac{\partial J}{\partial h_1} \cdot \frac{\partial h_1}{\partial w_{12}} = \frac{\partial J}{\partial h_1} \cdot \frac{\partial f_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_{12}} = 0 \cdot 1 \cdot x_2 \\ = 0 \cdot 1 \cdot 1 = 0$$

h_2

Calculate

$$\frac{\partial J}{\partial w_{21}} = \frac{\partial J}{\partial h_2} \cdot \frac{\partial h_2}{\partial w_{21}} = \frac{\partial J}{\partial h_2} \cdot \frac{\partial f_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_{21}}$$
$$= 0 \cdot z_2 \cdot x_1$$
$$= 0 \cdot 0 \cdot 0 = 0$$

$$\frac{\partial J}{\partial w_{22}} = \frac{\partial J}{\partial h_2} \cdot \frac{\partial h_2}{\partial w_{22}} = \frac{\partial J}{\partial h_2} \cdot \frac{\partial f_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_{22}}$$
$$= 0 \cdot z_2 \cdot x_2$$
$$= 0 \cdot 0 \cdot 1 = 0$$

Finally, do all the
weight updates using

$$\omega' = \omega - \ell \frac{\partial J}{\partial \omega}$$

What are the new values

~ ~ ~

of

$$\omega_{11} = 0 - 0.001 \frac{\partial J}{\partial \omega_{11}} = 0 - 0.001 \cdot 0 = 0$$

$$\omega_{12} = 0 - 0.001 \frac{\partial J}{\partial \omega_{12}} = 0 - 0.001 \cdot 0 = 0$$

$$\omega_{21} = 0 - 0.001 \frac{\partial J}{\partial \omega_{21}} = 0 - 0.001 \cdot 0 = 0$$

$$\omega_{22} = 0 - 0.001 \frac{\partial J}{\partial \omega_{22}} = 0 - 0.001 \cdot 0 = 0$$

$$v_{11} = 0 - 0.001 \frac{\partial J}{\partial v_{11}} = 0 - 0.001 \cdot 0 = 0$$

$$v_{12} = 0 - 0.001 \frac{\partial J}{\partial v_{12}} = 0 - 0.001 \cdot 0 = 0$$

$$v_{22} = 0 - 0.001 \frac{\partial J}{\partial v_{22}} = 0 - 0.001 \cdot 0 = 0$$

$$v_1 = 0 - 0.001 \frac{\partial J}{\partial v_1} = 0 - 0.001 \cdot 0 = 0$$

→ ~ ~ - ~ ~ ~ ~ ~ = ~