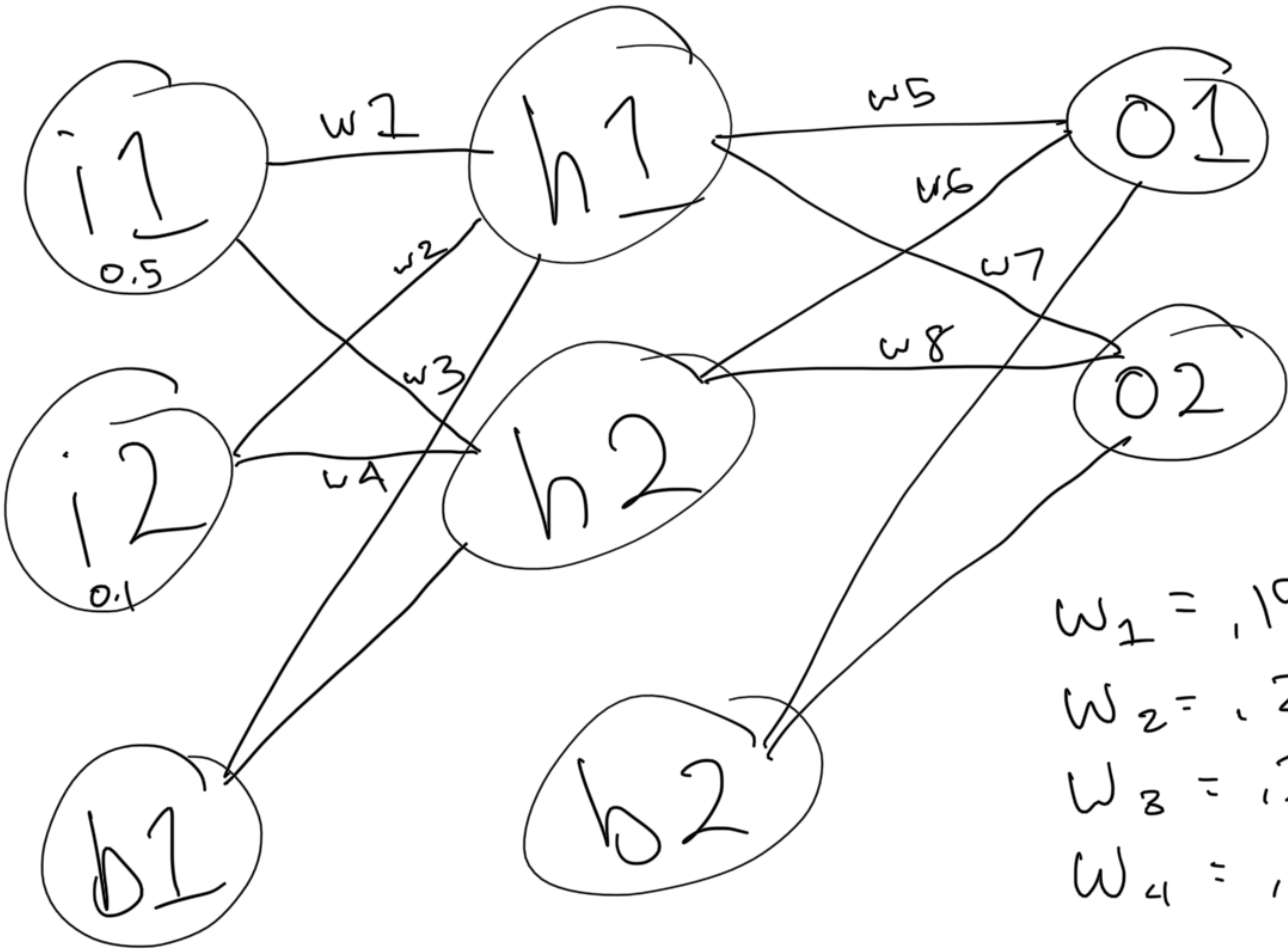


$w_1 = .15$

$w_2 = .2$

$w_3 = .25$

$w_4 = .3$

$b_1 = .35$

$b_2 = .6$

$$h1 = (i1 \times w1) + (i2 \times w2) + (b1 \times 1)$$

$$h2 = (i1 \times w3) + (i2 \times w4) + (b1 \times 1)$$

$$h1 = (0.5 \times 0.15) + (0.1 \times 0.2) + .35$$

$$h_2 = (0.5 \times .25) + (0.1 \times .3) + .35$$

$$\vdots$$

$$h_1 = 0.44$$

$$h_2 = 0.5$$

$$O_1 = (h_1 \times W_5) + (h_2 \times W_6) + b_2$$

$$O_2 = (h_1 \times w_7) + (h_2 \times w_8)$$

$$O_1 = 0.75, \quad O_2 = 0.77$$

Calculate loss using mean squared error...

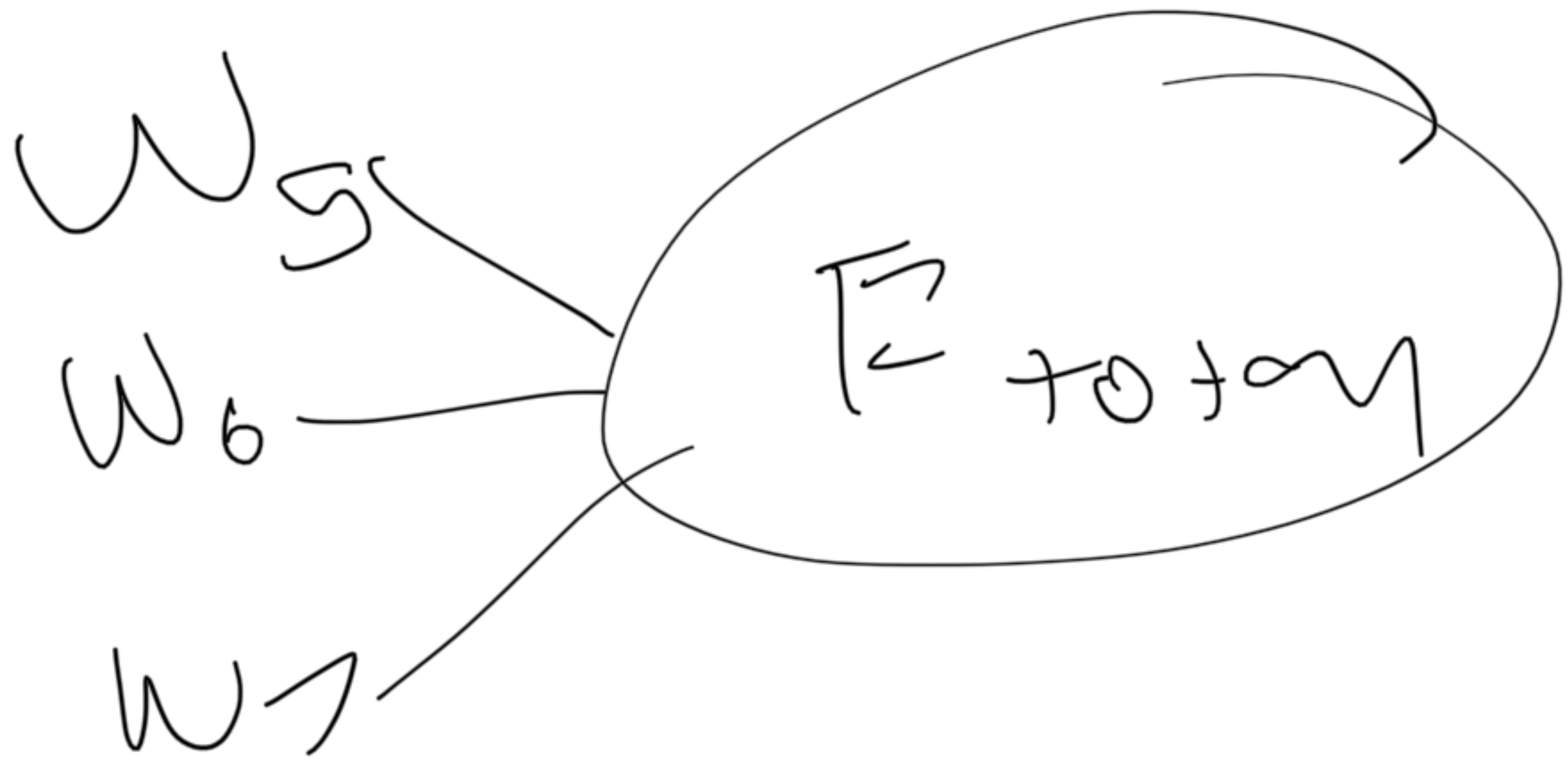
$$E_{total} = \sum \frac{1}{2} (\text{target} - \text{output})^2$$

O_1 target

$$E_1 = \frac{1}{2} (.01 - O_1)$$

$$E_2 = \frac{1}{2}(.99 - O_2)$$

$E_1 =$



W_8

How much of a

change in w_5 changes E_{total}?

$$\frac{\partial E_{total}}{\partial w_5} \quad \text{or} \quad \frac{\Delta E_{total}}{\Delta w_5}$$

---

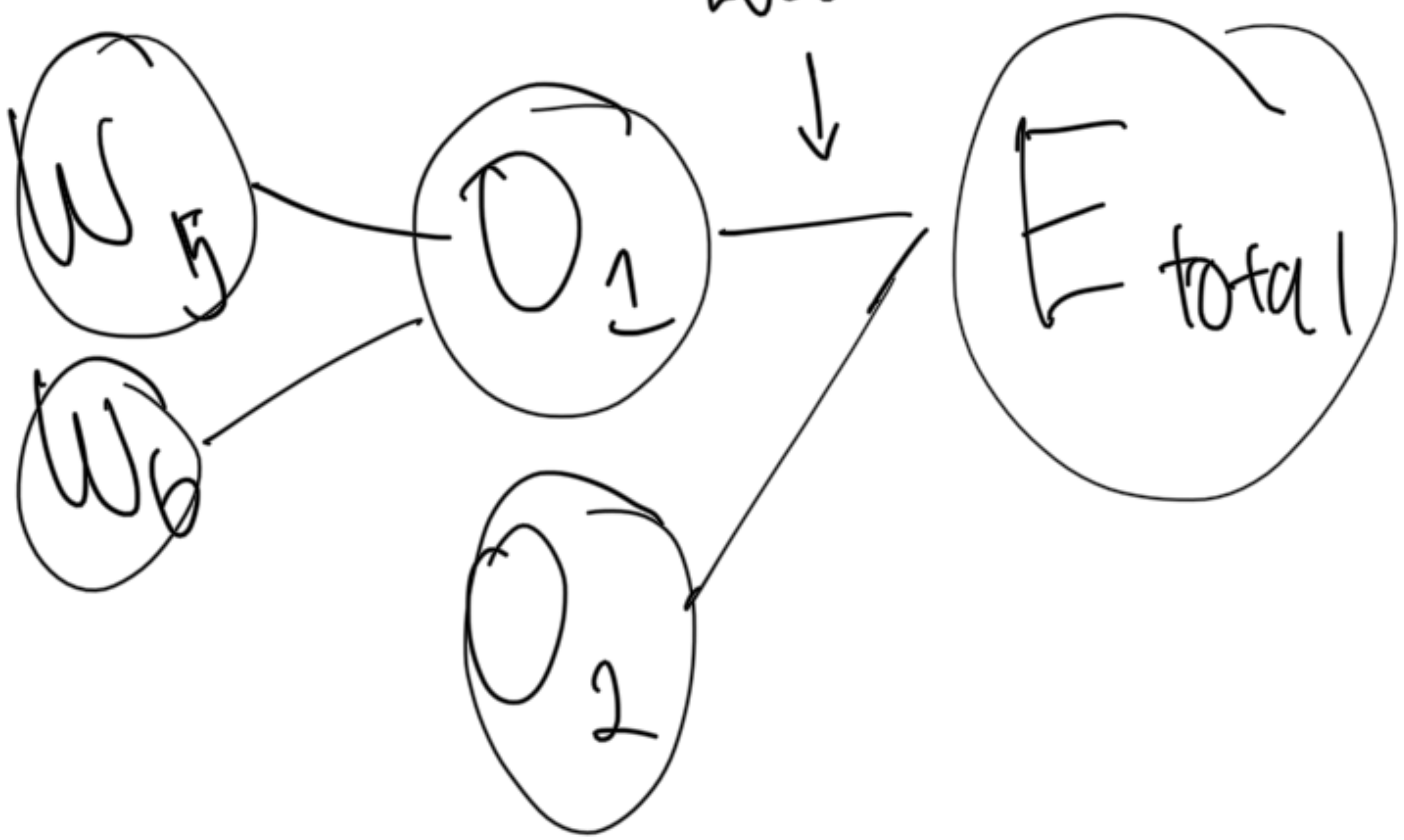
partial derivative of E_{total} w/ respect to w_5

partial derivative = gradient

partial derivative gradient

Since W_5 is connected to neurons before it, we have to consider them in $\frac{\partial E_{total}}{\partial W_5}$ b/c changing

Loss fn – mean squared error



2

If we wiggle w_5, how does that change our total error E_{total}?

$$\frac{\partial E_{total}}{\partial w_5} = \frac{\partial E_{total}}{\partial out_{o1}} \times \frac{\partial out_{o1}}{\partial net_{o1}} \times \frac{\partial net_{o1}}{\partial w_5}$$

O_1 can be thought of having two entities

O_1

$$E_{o1} = MSE(out_{o1})$$
$$= \frac{1}{2}(target_{o1} - out_{o1})^2$$

net_{o1} $\quad$ out_{o1} $\quad$ $E_{total} = E_{o1} + E_{o2}$

$\boxed{h_1 \times w_5 + h_2 \times w_6 + b_2}$ $\quad$ $\boxed{\text{sigmoid}(net_{o1})}$

o_1

net o_1 | out o_1

$$\frac{\partial E_{total}}{\partial w_5} = \frac{\partial E_{total}}{\partial out_{o1}} \times \frac{\partial out_{o1}}{\partial net_{o1}} \times \frac{\partial net_{o1}}{\partial w_5}$$

$$\frac{\partial E_{total}}{\partial out_{o1}} = \frac{d E_{total}}{d out_{o1}} \left(\frac{1}{2}\left(target_{o1} - out_{o1}\right)^2 + \frac{1}{2}\left(target_{o2} - out_{o2}\right)^2 \right)$$

$$= 2 \times \frac{1}{2}\left(target_{o1} - out_{o1}\right)^{2-1} + 0$$

← derivative of inner

$$= \left(target - out\right) \times -1$$

$$\left(\quad o_1 \qquad\qquad o_1 \right)$$

$$= -\text{target} + \text{out}$$

$$\frac{\partial \text{out}_{o1}}{\partial \text{net}_{o1}} \Rightarrow \text{out}_{o1} = \text{sigmoid}(\text{net}_{o1})$$

sigmoid

$$f(x) = \frac{1}{1+e^{-x}} \Rightarrow \frac{1}{1+e^{-\text{net}_{o1}}}$$

$$\Rightarrow \left(1+e^{-\text{net}_{o1}}\right)^{-1} \Rightarrow -1 * \left(1+e^{-\text{net}_{o1}}\right)^{-2} * \frac{d}{d\text{net}_{o1}}\left(1+e^{-\text{net}_{o1}}\right)$$

$$-2 \qquad\qquad -\text{net}_{o1}$$

$$\Rightarrow -\left(1+e^{-net_{o1}}\right)^{-} \times -e$$

$$\frac{\partial out_{o1}}{\partial net_1} = out_{o1}(1 - out_{o1})$$

$$\frac{\partial net_{o1}}{\partial w_5}$$

To decrease the error, subtract

∂E

$$\frac{\partial E_{total}}{\partial W_5} \text{ from } W_5$$

Goal: Lower E_{total}

If $$\frac{\partial E_{total}}{\partial W_5} = 0.0821$$

$$W_5^{+} = W_5 - \left(LR \times \frac{\partial E_{total}}{\partial W_5}\right) \qquad LR = .5$$

$$= 0.4 - (.5 \times 0.0821)$$

$$= 0.35$$

Let's run a forward pass w/ this updated weight & see what happens updated weight

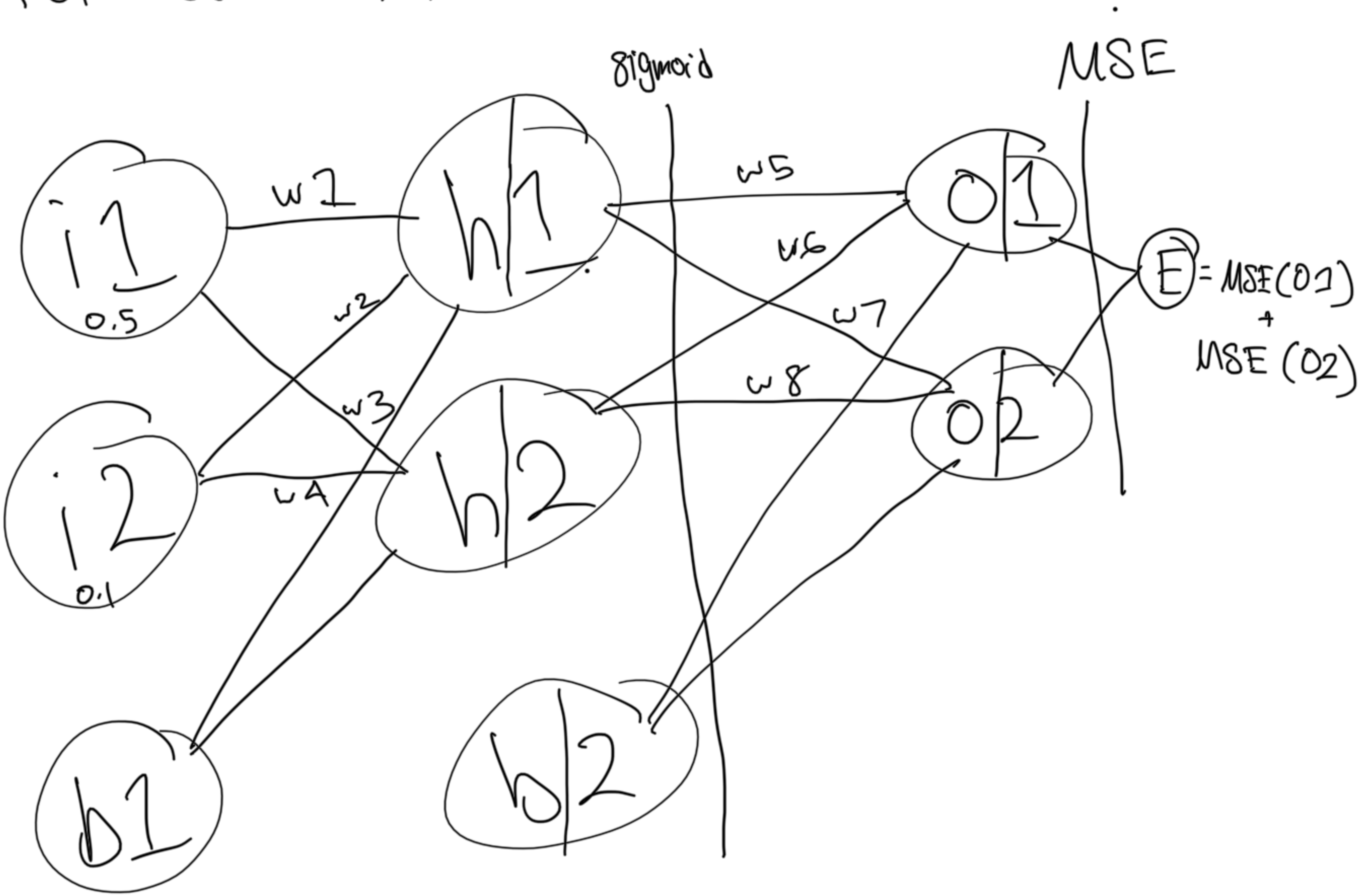
$a \approx$

OK apparently we need to start a new section, to avoid lag buildup. lol, wtf apple, this is the exact type of product bloat that Jobs tried to fight. But orgs w/ bad incentives push features for promotions instead of love & stability.

OK anyways, back to backprop :)

Repeat process for $\begin{matrix} w_5 \\ \vdots \\ w_8 \end{matrix}$

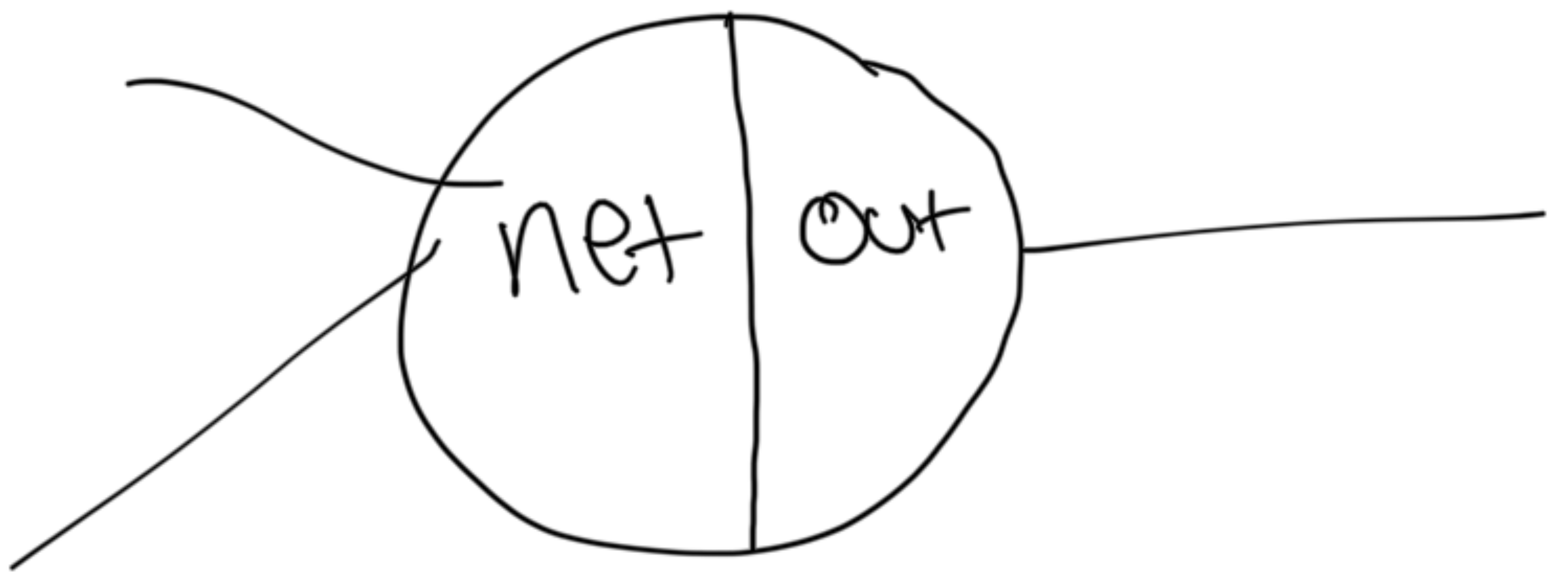
- Let's find out how ~~to~~ changing w_1 affects our total error



$$\frac{\partial E_{total}}{\partial w_1}$$

How does wiggling w_1 affect E_{total}?

For each "neuron" or unit after our input unit, we can think of it having two parts



The net is typically a linear combination of some units before it.

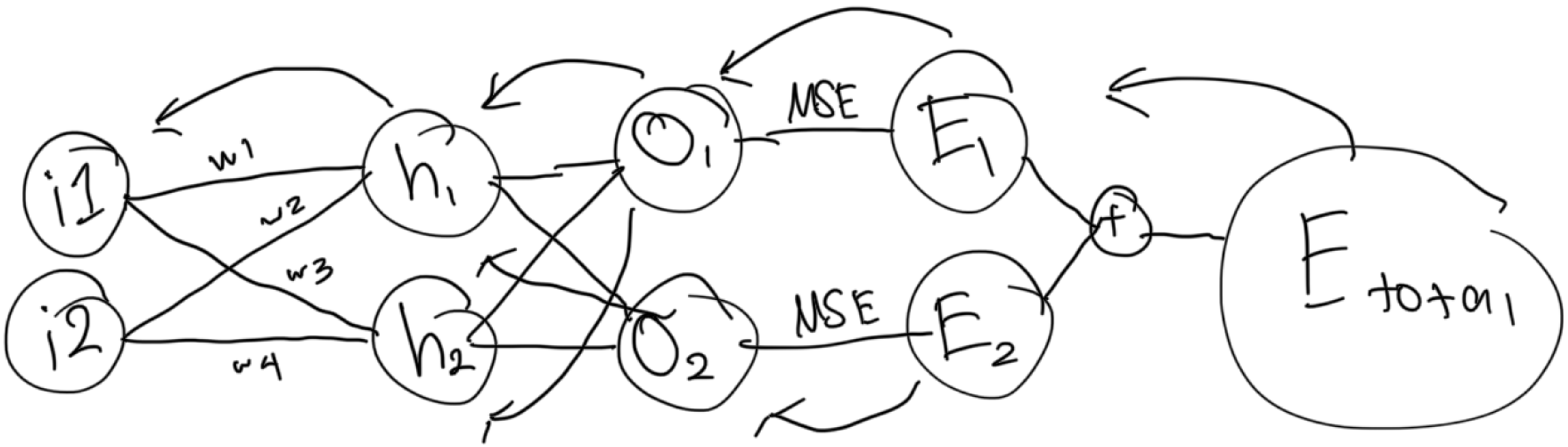
out is activated by a sigmoid fn

The $\underline{out}$ is derivative of a sigmoid fn.

OK so back to calculating

$$\frac{\partial E_{total}}{\partial w_1}$$

$$E_{total} = E_1 + E_2$$



$$\frac{\partial E_{total}}{\partial w_1} = \frac{\partial E_{total}}{\partial out_{h_1}} \times \frac{\partial out_{h_1}}{\partial net_{h_1}} \times \frac{\partial net_{h_1}}{\partial w_1}$$

∂w_1 ∂out_{h1}

How much does h1 affect Etotal?

② Take into account h1 activation

① How much does w1 affect h1?

b/c h1 affects E_2, we'll need to expand this out

$$\frac{\partial E_{total}}{\partial out_{h1}}$$

+	

Ok idk what's happening, this notes app is ok for light use, but shitty if you need anything more powerful.

$$\frac{\partial E_{total}}{\partial out_{h1}} = \frac{\partial E_{out_1}}{\partial out_{h_1}} + \frac{\partial E_{out_2}}{\partial out_{h_1}}$$

Ok now let's calculate what this should be

∂E_{o}

$$\boxed{\frac{\cdots}{\partial out_{h_1}}}$$

out_{h_1} $\leftarrow$ E_{o_1}

$$E_{o_1} = \frac{\partial net_{h_1}}{\partial w_1} \times \frac{\partial net_{o_1}}{\partial out_{h_1}} \times \frac{\partial out_{o_1}}{\partial net_{o_1}} \times \frac{\partial E_{o_1}}{\partial out_{o_1}}$$

$$\frac{\partial E_{total}}{\partial w_1} = \frac{\partial E_{total}}{\partial out_{h_1}} \times \frac{\partial out_{h_1}}{\partial net_{h_1}} \times \frac{\partial net_{h_1}}{\partial w_1}$$

Since w1 feeds

Since h1 gets activated, how does

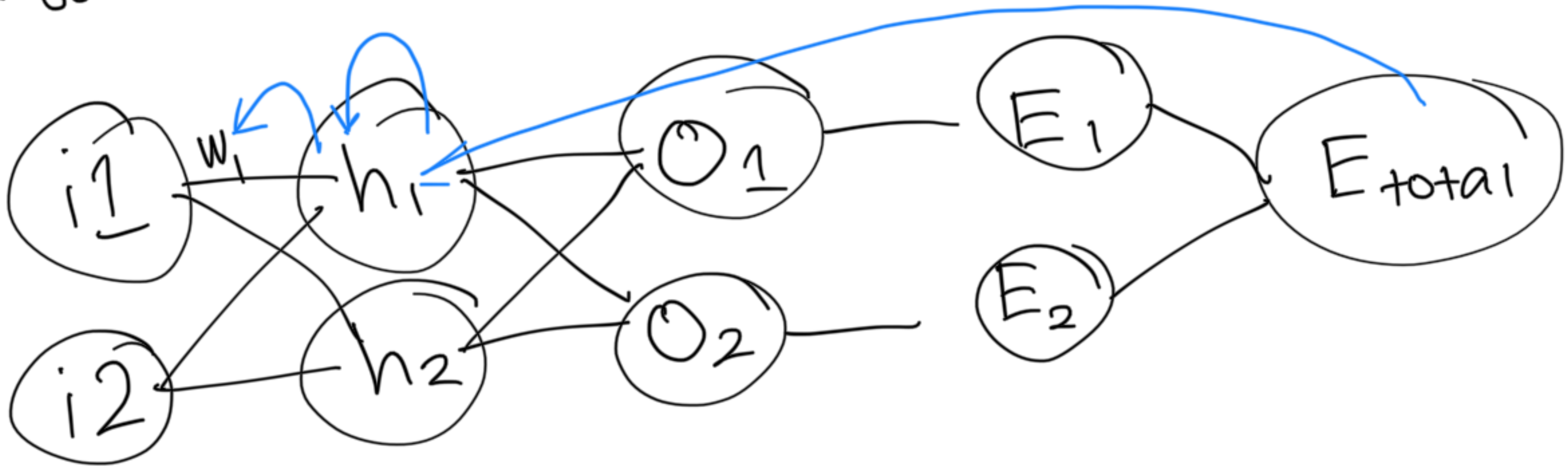
Finally, how does our input to h1 change as we

How does our total error change as we wiggle w1 ?

into h1, how does the output of h1 change our total error ?

the output of h1 change as the input to it changes ?

wiggle w1 ?



Chain Rule Analogy

To understand how $\frac{\partial E_{total}}{\partial w_1}$ is calculated visually, let's suppose the following

1. Wiggle the chain of w_1

   ↳ Observe how h1 changes

   connected on the

b/c they've connected
"same chain" thus $\dfrac{\partial net_{h_1}}{\partial w_1}$

(2) Because h1 really is net_{h_1} — out_{h_1},

where the sigmoid fn is

activating net_{h1} into out_{h2}

↳ Observe how wiggling net_{h_1}
will also wiggle out_{h1} b/c they're
connected on the same chain

thus ∴ $\dfrac{\partial out_{h_1}}{\partial net_{h_1}}$

(3) Now that out_{h_1} is getting wiggled,
we want to know how that will affect

we want to know how that will affect our total error E_{total}. Since the out_{h1} is connected to net_{o1} & net_{o2}, we can see how wiggling out_{h1} has ripple effects throughout our network. Nonetheless, for now

Observe how wiggling out_{h1} will affect our total error E_{total}

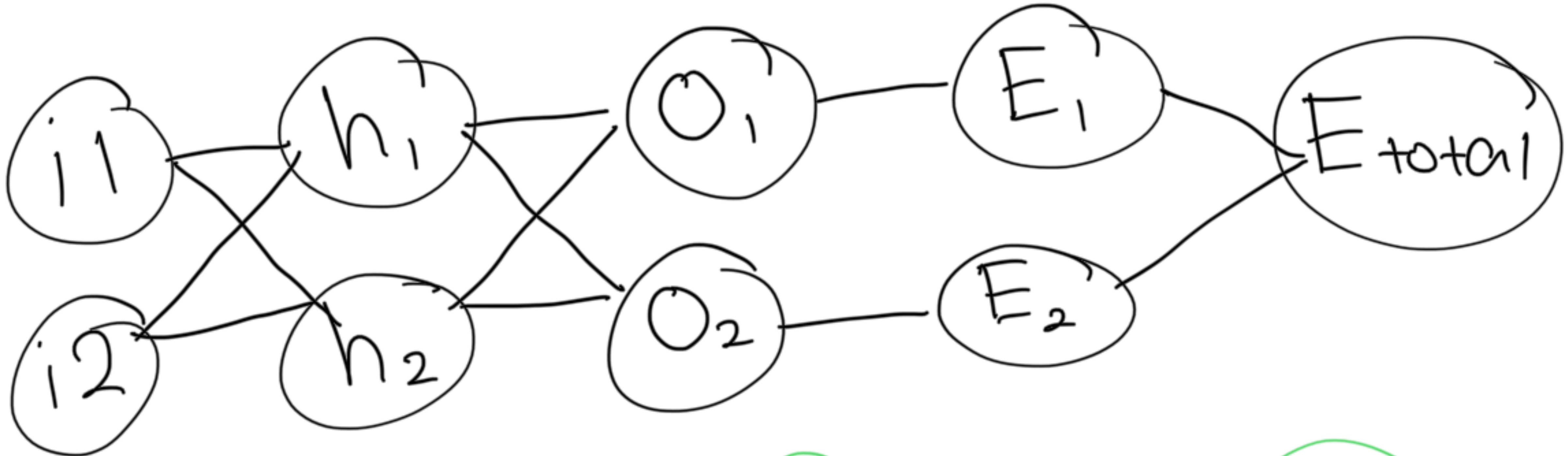
$$\frac{\partial E_{total}}{\partial out_{h1}}$$

And so if we combine ①, ②, and ③, we ~~can~~ see that

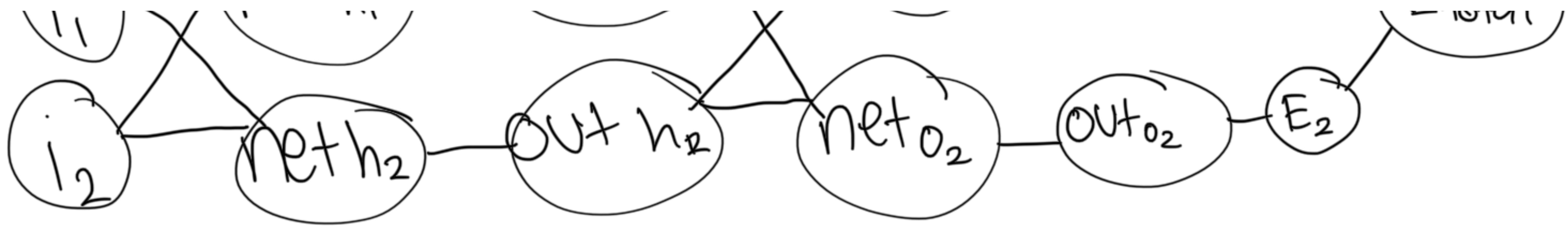
$$\partial E_{total} = \overset{①}{\partial net_{h1}} \times \overset{②}{\partial out_{h1}} \times \overset{③}{\partial E_{total}}$$

$$\frac{\quad}{\partial w_1} = \frac{\quad}{\partial w_1} \; \frac{\quad}{\partial net_{h_1}} \; \frac{\quad}{\partial out_{h_1}}$$

This is the chain rule!



Let's split h_1 into net_{h_1} and out_{h_1} to highlight that we have an activation fn transforming h_1 before it goes into O_1 and O_2



$$\frac{\partial E_{total}}{\partial w_1} = ① \times ② \times ③$$

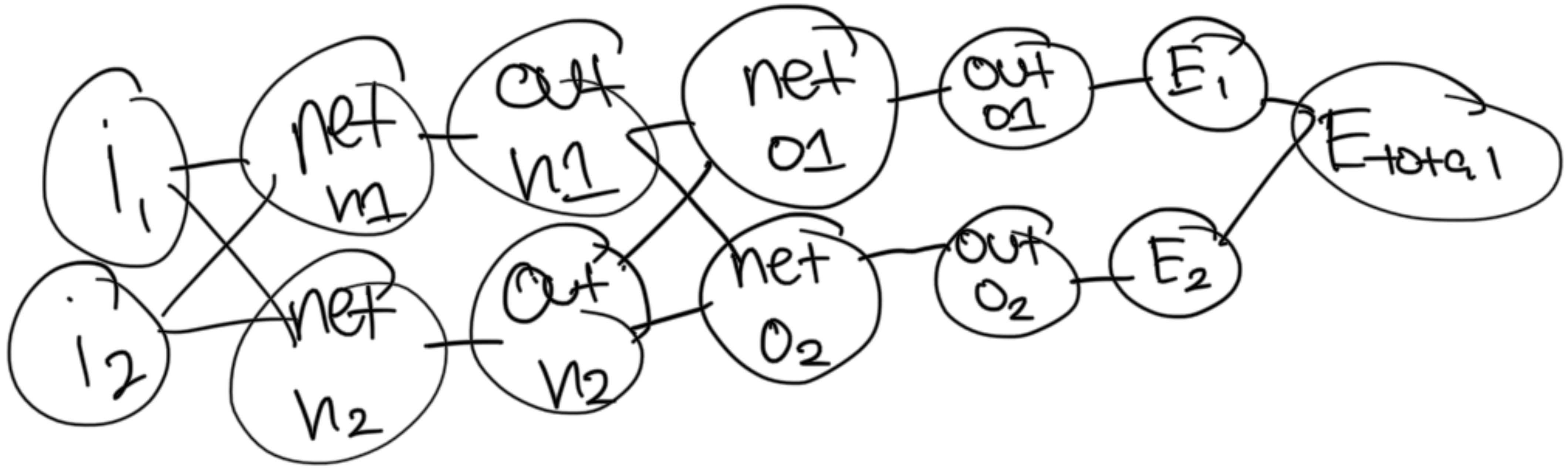
$$= \frac{\partial net_{h_1}}{\partial w_1} \times \frac{\partial out_{h_1}}{\partial net_{h_1}} \times \frac{\partial E_{total}}{\partial out_{h_1}}$$

- We can see that the "chain length" of ① and ② is quite short, this makes them easy to compute.

But what about ③, this chain is quite long, so let's break it down into smaller chains that we can compute

$$\frac{\partial E_{total}}{\partial out_{h1}}$$

How does E_{total} change as out_{h1} changes?



Going back to our chain analogy, if we wiggle out_{h1}, who else gets affected?

Since $E_{total} = E_1 + E_2$

$$\frac{\partial E_{total}}{\partial out_{h1}} = \frac{\partial E_1}{\partial out_{h1}} + \frac{\partial E_2}{\partial out_{h1}}$$

Let's focus on this for

∂E_2 shortly after

on ... now & then we'll calculate $\partial\, out_{h_1}$ ...

$$\frac{\partial E_1}{\partial Out_{h1}}$$

Observe that Out_{h1} has two chains connected to it. But, since we only care about how E_1 changes as we wiggle Out_{h1}, we don't care about the chain between Out_{h1} and net_{o2}.

$$\frac{\partial E_1}{\partial Out_{h1}}$$

So ignoring the chains not connected to $E1$, what we want to observe rn is just

$$Out_{h1} - net_{o1} - Out_{o1} - E_1$$

As we can see, by focusing on just observing how E_1 changes in respect to Out_{h1}, we can simplify our network by just inspecting a subset of what we saw before

$$\frac{\partial E_1}{\partial Out_{h1}} = \frac{\partial net_{o1}}{\partial Out_{h1}} \times \frac{\partial Out_{o1}}{\partial net_{o1}} \times \frac{\partial E_1}{\partial Out_{o1}}$$

Since everything is on the same chain, wiggling Out_{h1}, means we can observe how everything changes up to E_1. Originally, we wanted to know

$$\frac{\partial E_{total}}{\partial out_{h1}} = \frac{\partial E_1}{\partial out_{h1}} + \frac{\partial E_2}{\partial out_{h1}}$$

Now let's figure this out

$$\frac{\partial E_2}{\partial out_{h1}}$$

Our network of interest is

$$\frac{\partial E_2}{\partial out_{h1}} \quad \text{so} \quad out_{h1} - net_{o2} - out_{o2} - E_2$$

=

$$= \frac{\partial net_{o2}}{\partial out_{h1}} \times \frac{\partial out_{o2}}{\partial net_{o2}} \times \frac{\partial E_2}{\partial out_{o2}}$$

so now putting it all together

$$\frac{\partial E_{total}}{\partial out_{h1}} = \frac{\partial E_1}{\partial out_{h1}} + \frac{\partial E_2}{\partial out_{h1}}$$

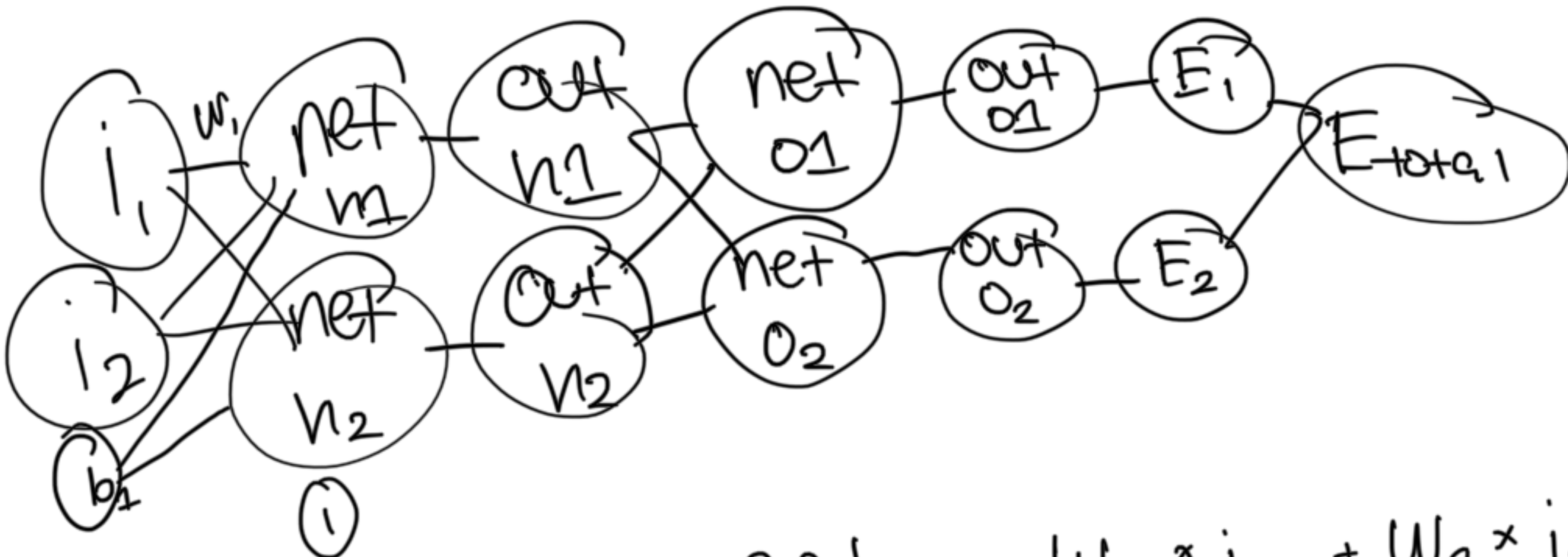
$$= \frac{\partial net_{o1}}{\partial out_{h1}} \times \frac{\partial out_{o1}}{\partial net_{o1}} \times \frac{\partial E_1}{\partial out_{o1}}$$

$$+ \frac{\partial net_{o2}}{\partial out_{h1}} \times \frac{\partial out_{o2}}{\partial net_{o2}} \times \frac{\partial E_2}{\partial out_{o2}}$$

Now that we've calculated 3, let's put it all together

$$\frac{\partial E_{total}}{\partial w_1} = \underset{1}{\frac{\partial net_{h1}}{\partial w_1}} \times \underset{2}{\frac{\partial out_{h1}}{\partial net_{h1}}} \times \underset{3}{\frac{\partial E_{total}}{\partial out_{h1}}}$$

Let's now actually calculate 1, 2, 3



$\partial\, net_h$

$$net_{h1} = w_1 \times i_1 + w_2 \times i_2 + b_1$$

$$\frac{\partial\, net_{h_1}}{\partial w_1} \quad \text{So} \quad net_{h_1} \ldots$$

$$\frac{d}{d w_1}\left(net_{h_1}\right) = i_1 \quad (1) \checkmark$$

Now that (1) is calculated, let's focus on (2)

$$\frac{\partial\, out_{h_1}}{\partial\, net_{h_1}} \quad \text{so} \quad out_{h_1} = \text{sigmoid}\left(net_{h_1}\right)$$

$$\frac{d}{d\, net_{h_1}}\left(out_{h_1}\right) = out_{h_1} \times \left(1 - out_{h_1}\right)$$

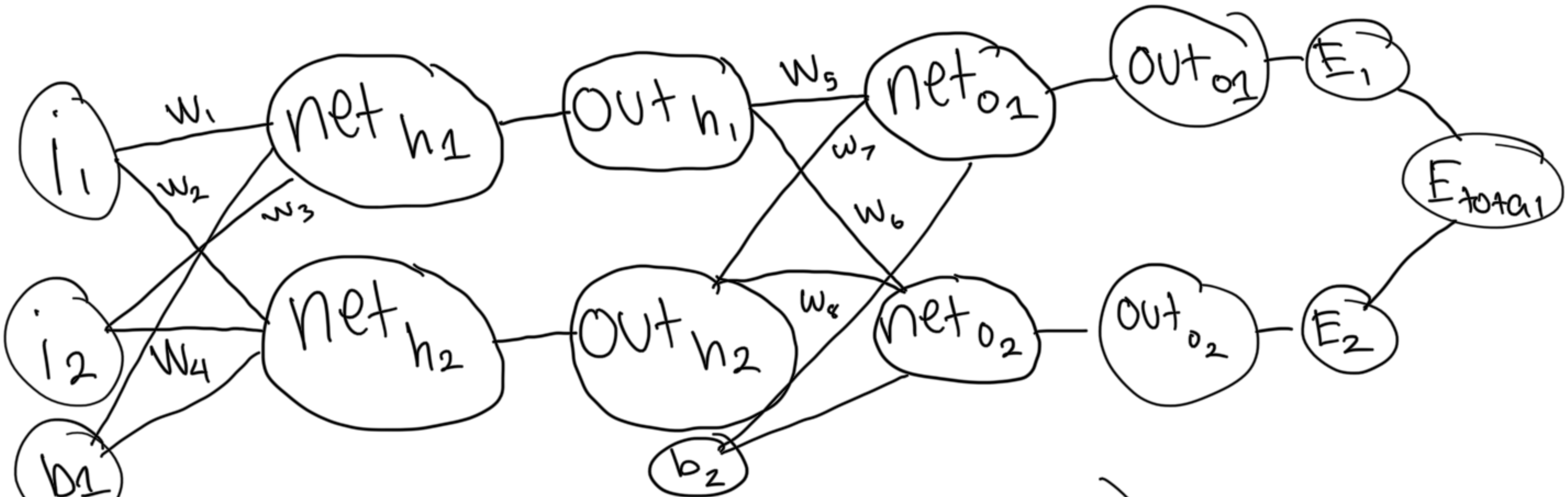
(2) ✓

$$(3) \quad \frac{\partial E_{total}}{\partial\, out_{h_1}} = \overset{(1)}{\frac{\partial\, net_{o_1}}{\partial\, out_{h_1}}} \times \overset{(2)}{\frac{\partial\, out_{o_1}}{\partial\, net_{o_1}}} \times \overset{(3)}{\frac{\partial E_1}{\partial\, out_{o_1}}}$$

$$+ \overset{(4)}{\frac{\partial\, net_{o_2}}{\ldots}} \times \overset{(5)}{\frac{\partial\, out_{o_2}}{\partial\, net \ldots}} \times \overset{(6)}{\frac{\partial E_2}{\partial\, out_{o_2}}}$$

$\partial out_{h_1} \quad \partial net_{o_2} \quad \partial \; o_2$

$$\frac{\partial net_{o_1}}{\partial out_{h_1}}$$ Let's draw our network out to get a clearer picture



$$net_{o_1} = (w_5 * out_{h_1}) + (w_7 * out_{h_2}) + b_2$$

$$\frac{d}{} \left(net_{o_1} \right) = w_5 \quad 1 \; \checkmark$$

$d\, out_{h1}$

(2) $\frac{\partial out_{o1}}{\partial net_{o1}} = (out_{o1}) \times (1 - out_{o1})$ (2) ✓

(3) $\frac{\partial E_1}{\partial out_{o1}}$ so $E_1 = MSE(out_{o1})$

$= \frac{1}{2}(target - out_{o1})^2$

$\frac{d}{d\, out_{o1}}(E_1) = (target_{o1} - out_{o1}) \times (-1)$ ✓

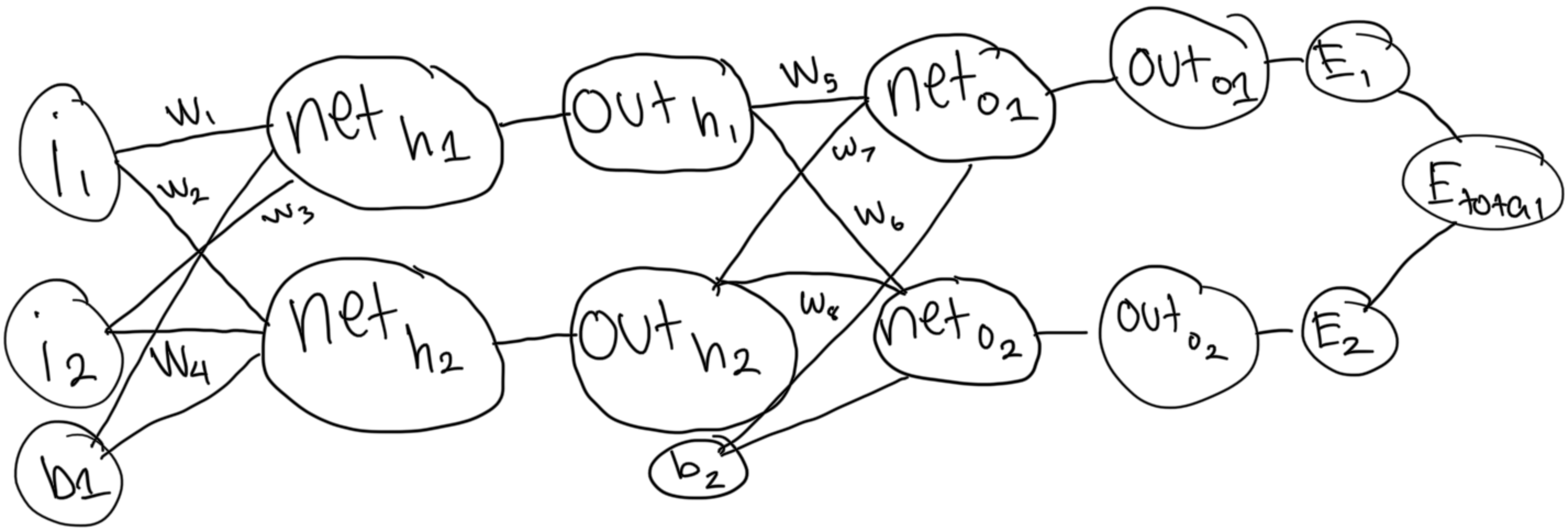
(4) $\frac{\partial net_{o2}}{\partial out_{h1}} = W_6$

(5) $\frac{\partial out_{o2}}{\partial net_{o2}} = (out_{o2}) \times (1 - out_{o2})$

6 $$\frac{\partial E_2}{\partial out_{o_2}} = \left(target_{o_2} - out_{o_2}\right) \times \left(-1\right)$$

$$\frac{\partial E_{total}}{\partial W_1}$$ We can now calculate this gradient!

OK now let's see how $\frac{\partial E_{total}}{\partial w_2}$



$$\frac{\partial E_{total}}{\partial W_2} = \frac{\partial net_{h2}}{\partial W_2} \times \frac{\partial out_{h2}}{\partial net_{h_2}} \times \frac{\partial E_{total}}{\partial out_{h_2}}$$

$$= i1 \times (out_{h_2}) \times (1 - out_{h2}) \times \frac{\partial E_{total}}{\partial out_{h_2}}$$

let's calculate this and see what happens

$$E_{total} = E_1 + E_2$$

$$\frac{\partial E_{total}}{\partial out_{h_2}} = \frac{\partial E_1}{out_{h_2}} + \frac{\partial E_2}{out_{h_2}}$$

$$= \frac{\partial net_{o_1}}{\partial out_{h_2}} \times \frac{\partial out_{o_1}}{\partial net_{o_1}} \times \frac{\partial E_1}{\partial out_{o_1}}$$

$$+$$

$$\frac{\partial net_{o_2}}{\partial out_{h_2}} \times \frac{\partial out_{o_2}}{\partial net_{o_2}} \times \frac{\partial E_2}{\partial out_{o_2}}$$

$$= W_7 \times (out_{o_1}) \times (1 - out_{o_1}) \times -(target_{o_1} - out_{o_1})$$

$$+$$

$$W \times (out_{o_2}) \times (1 - out_{o_2}) \times -(target_{o_2} - out_{o_2})$$

cool now let's find

$$\frac{\partial E_{total}}{\partial w_3} = \frac{\partial net_{h_1}}{\partial w_3} \times \underbrace{\frac{out_{h_1}}{\partial net_{h_1}} \times \frac{\partial E_{total}}{\partial out_{h_1}}}_{\text{we've already calculated this}}$$

$$= i_2$$

$$\frac{\partial E_{total}}{\partial w_4} = \frac{\partial net_{h_2}}{\partial w_4} \times \frac{\partial out_{h_2}}{\partial net_{h_2}} \times \frac{\partial E_{total}}{\partial out_{h_2}}$$

$$\frac{\partial E_{total}}{\partial w_5} = \frac{\partial net_{o_1}}{\partial w_5} \times \frac{\partial out_{o_1}}{\partial net_{o_1}} \times \frac{\partial E_{total}}{\partial out_{o_1}}$$

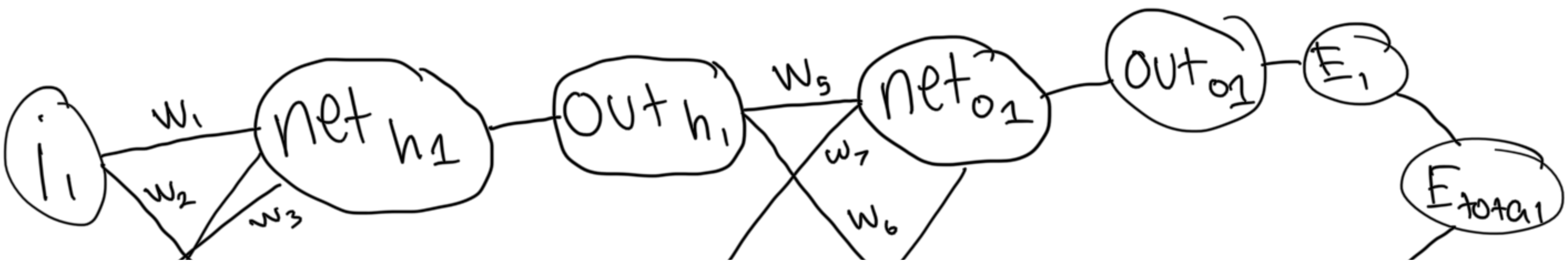
$$\frac{\partial E_{total}}{\partial W_6}$$

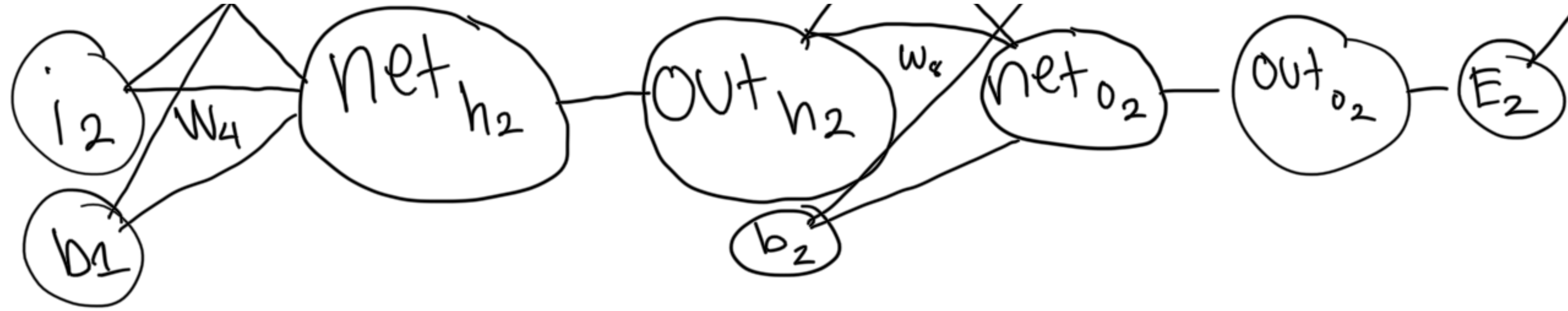
r

What about our bias parameters b_1 and b_2?

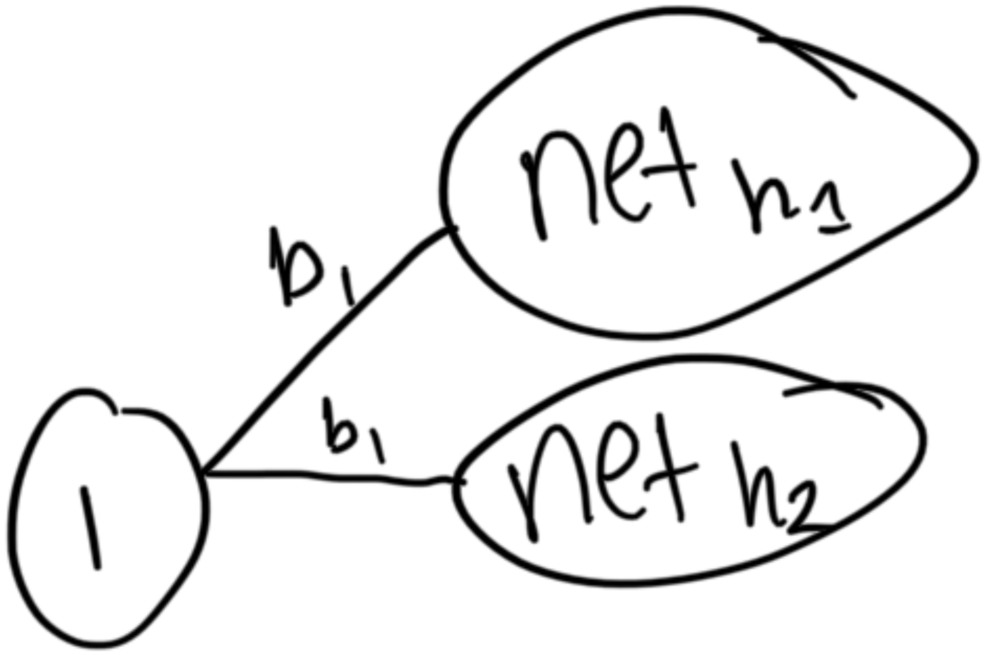
Let's see how they affect our total error over time.

$$\frac{\partial E_{total}}{\partial b_1}$$





To be more precise...

$b_1 =$ 

so $\frac{\partial E_{total}}{\partial b_1}$

If we shake b_1, what happens?

$\frac{\partial E_{total}}{\partial b_1} \Rightarrow \quad E_{total} = E_1 + E_2$

1 2

$\partial E_{total} = \frac{\partial E_1}{} + \frac{\partial E_2}{}$

$$\frac{\partial E_{total}}{\partial b_1} \quad \partial b_1 \quad \partial b_1$$

Focus

(1)

$$\frac{\partial E_1}{\partial b_1} = \frac{net_{h1}}{\partial b_1} \times \frac{net_{h2}}{\partial b_1} \times \frac{\partial out_{h1}}{\partial net_{h1}} \times \frac{\partial out_{h2}}{\partial net_{h2}} \times \frac{\partial E_1}{\partial out_{h1}} \times \frac{\partial E_1}{\partial out_{h2}}$$

$$= 1 \times 1 \times \frac{\partial out_{h1}}{\partial net_{h1}} \times \frac{\partial out_{h2}}{\partial net_{h2}} \times \frac{\partial E_1}{\partial out_{h1}} \times \frac{\partial E_1}{\partial out_{h2}}$$

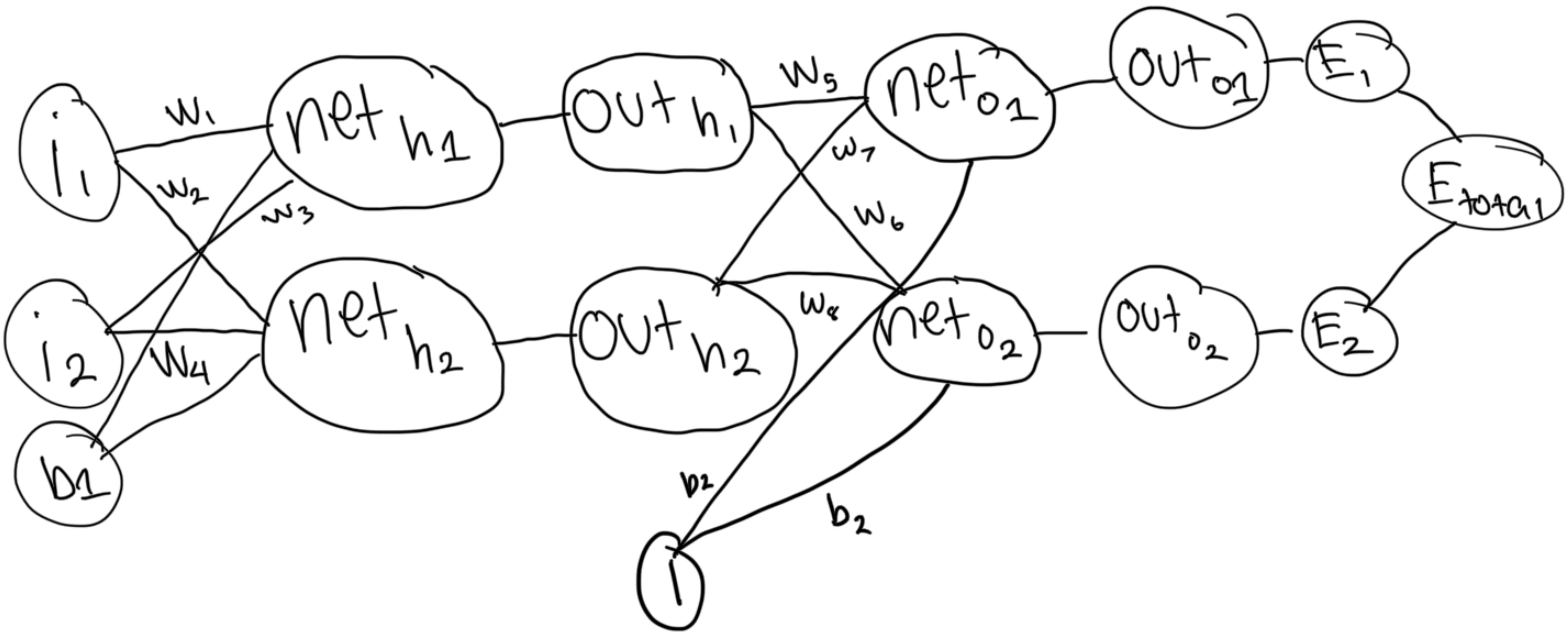
$$\ldots) \times (out \ldots) \times (1 - out_{h2})$$

$$= (out_{h_1})(1 - out_{h_1}) \times (out_{h_2}) \times (1 - out_{h_2})$$

$$\times \frac{\partial E_1}{\partial out_{h_1}} \times \frac{\partial E_1}{\partial out_{h_2}}$$

r

ok cool now $\dfrac{\partial E_{total}}{\partial b_2}$



$$\frac{\partial E_{total}}{\partial b} = \frac{\partial E_1}{\partial b} + \frac{\partial E_2}{\partial b}$$

∂b_2 ∂b_2 ∂b_2

$\uparrow$

Focus

$$\frac{\partial E_1}{\partial b_2} = \frac{\partial net_{o_1}}{\partial b_2} * \frac{\partial out_{o_1}}{\partial net_{o_1}} * \frac{\partial E_1}{\partial out_{o_1}}$$

$$= 1 * (out_{o_1}) * (1 - out_{o_1}) * -(target_1 - out_{o_1})$$