Scientific Computing

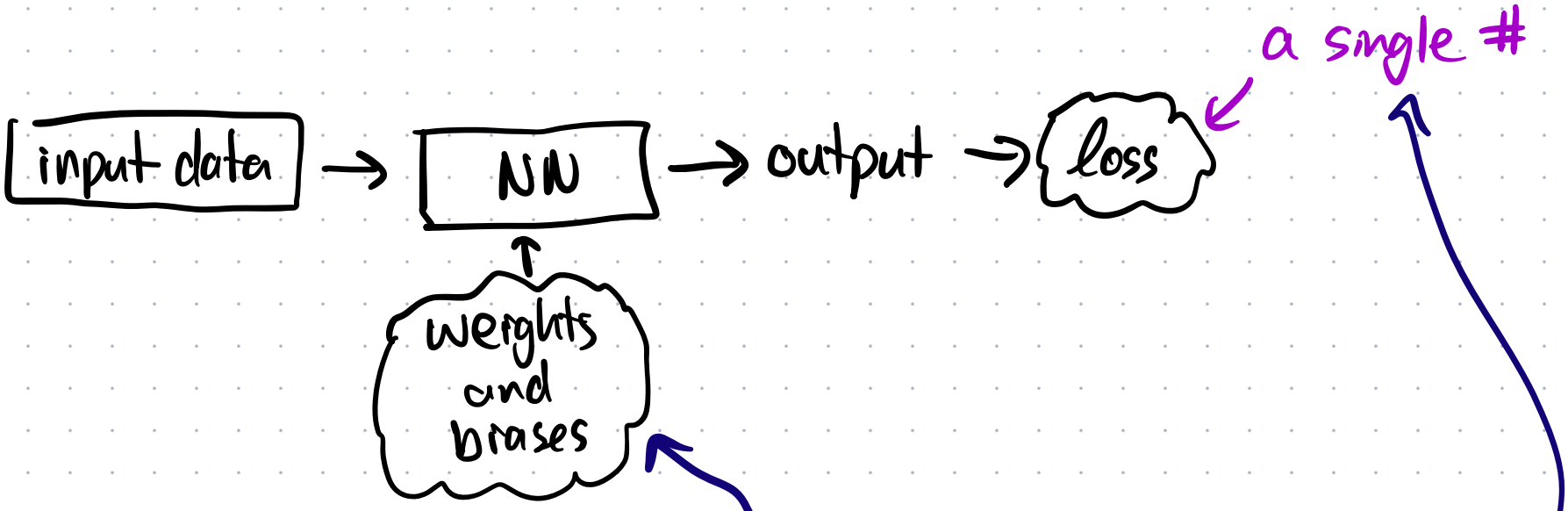
Wednesday, April 22

Announcements

* Homework 6 due next Monday, April 27

* Final Exam: Monday, 5/4,
  1pm-3pm
  Johnston Hall 417

Office Hours:
Mon, 9:30-10:30
Fri, 2:00-3:00
Cudahy 307



How can we find the inputs that minimize the output?

Gradient Descent!

Suppose the neural network has weights $w_1, w_2, \ldots, w_m$ and biases $b_1, b_2, \ldots, b_n$. Let ℓ be the mean loss over the whole batched input.

What is $-\nabla \underbrace{NN(w_1, w_2, \ldots w_m, b_1, b_2, \ldots, b_n)}$?

loss of output given these weights + biases + fixed input

$$\ell = NN(w_1, w_2, \ldots, w_m, b_1, b_2, \ldots b_n).$$

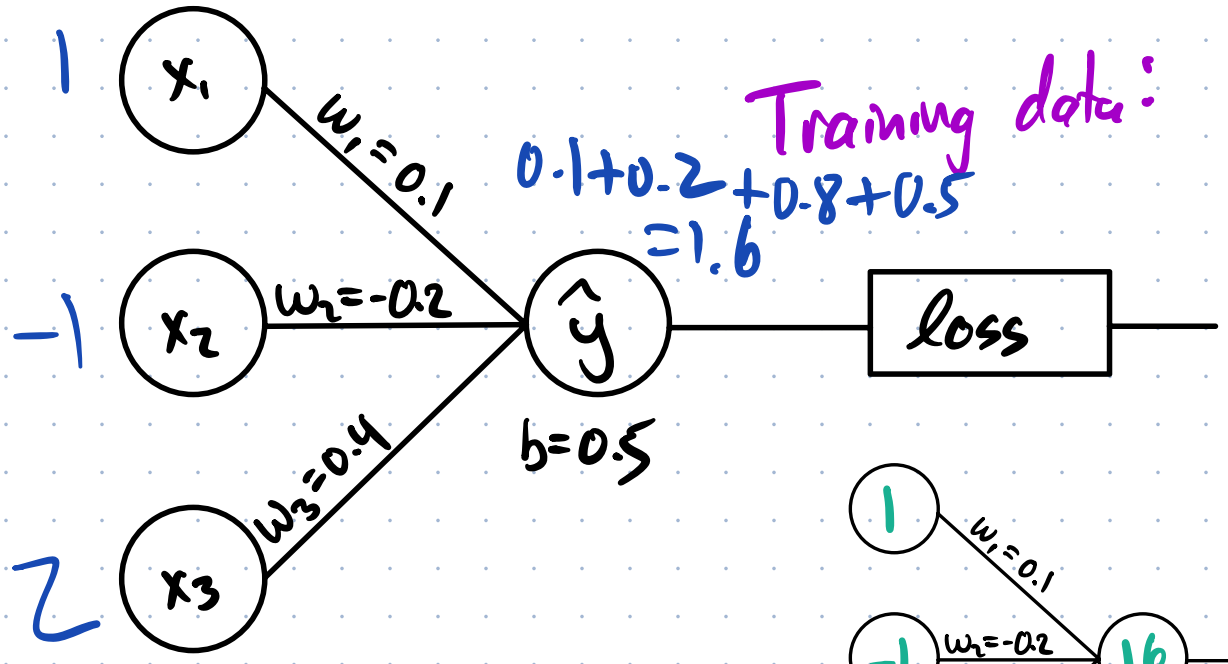
How will we compute

$$\nabla NN = \begin{bmatrix} \frac{\partial \ell}{\partial w_1} \\ \vdots \\ \frac{\partial \ell}{\partial w_m} \\ \frac{\partial \ell}{\partial b_1} \\ \vdots \\ \frac{\partial \ell}{\partial b_n} \end{bmatrix} ?$$

CALCULUS

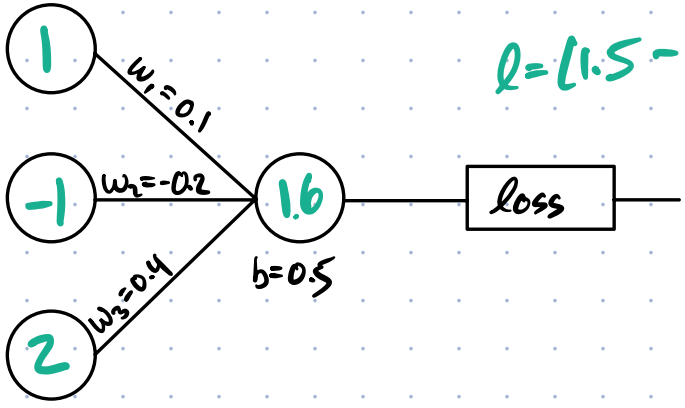
specifically
the CHAIN RULE

First example: No AF, no hidden layers, 3 input, 1 output, only one point of training data



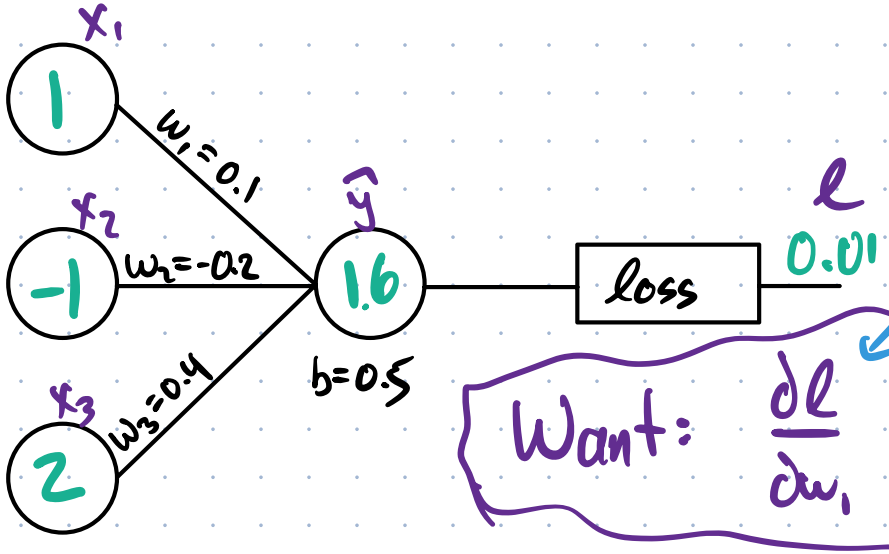
Training data: $(1, -1, 2) \rightarrow 1.5$

$0.1+0.2+0.8+0.5 = 1.6$



$\ell = (1.5-1.6)^2 = 0.01$

How do we adjust w_1, w_2, w_3, b to make ℓ go down for this training data?



Training data: $(1, -1, 2) \to 1.5$

when I change w_1 a little, how much does ℓ change?

Want: $\frac{\partial \ell}{\partial w_1}, \frac{\partial \ell}{\partial w_2}, \frac{\partial \ell}{\partial w_3}, \frac{\partial \ell}{\partial b}$

$\ell = (\hat{y} - 1.5)^2$.

So, $\frac{\partial \ell}{\partial \hat{y}} = 2(\hat{y} - 1.5)$.

0.2 $= 2 \cdot (0.1) = 0.2$

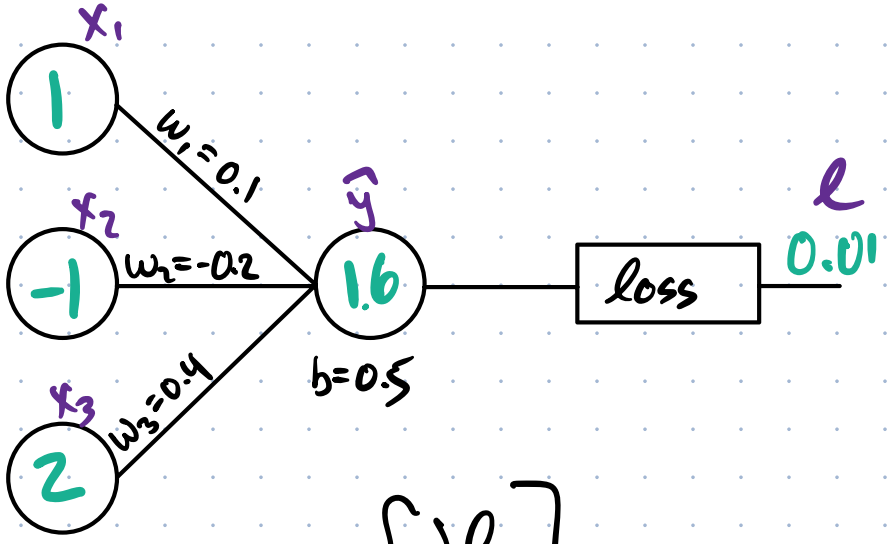
$\hat{y} = 1 \cdot w_1 - 1 \cdot w_2 + 2 \cdot w_3 + b$

So, $\frac{\partial \hat{y}}{\partial w_1} = 1$, $\frac{\partial \hat{y}}{\partial w_2} = -1$, $\frac{\partial \hat{y}}{\partial w_3} = 2$, $\frac{\partial \hat{y}}{\partial b} = 1$

$$\frac{\partial \ell}{\partial w_1} = \frac{\partial \ell}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} = 0.2 \cdot 1 = 0.2$$

If $\hat{y}$ changes by δ, this causes ℓ to change by $0.2 \cdot \delta$

$\frac{\partial \ell}{\partial w_2} = -0.2$ $\frac{\partial \ell}{\partial w_3} = 0.4$ $\frac{\partial \ell}{\partial b} = 0.2$



Training data: $(1, -1, 2) \to 1.5$

$$-\nabla NN = -\begin{bmatrix} \frac{\partial \ell}{\partial w_1} \\ \frac{\partial \ell}{\partial w_2} \\ \frac{\partial \ell}{\partial w_3} \\ \frac{\partial \ell}{\partial b} \end{bmatrix} = \begin{bmatrix} -0.2 \\ +0.2 \\ -0.4 \\ -0.2 \end{bmatrix}$$

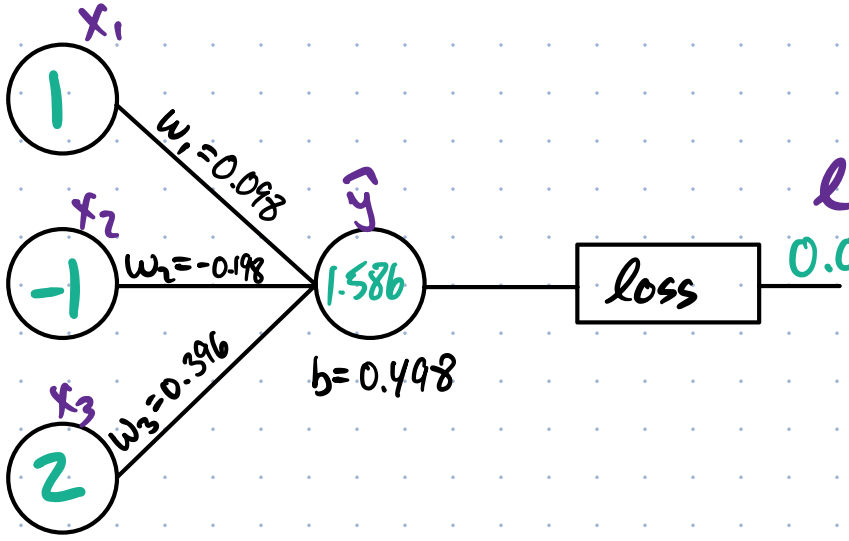
Now adjust each weight a small amount in this direction. Let's do $\frac{1}{100}$.

w_1: $0.1 \to -0.098$

w_2: $-0.2 \to -0.198$

w_3: $0.4 \to 0.396$

b: $0.5 \to 0.498$



Training data: $(1, -1, 2) \to 1.5$

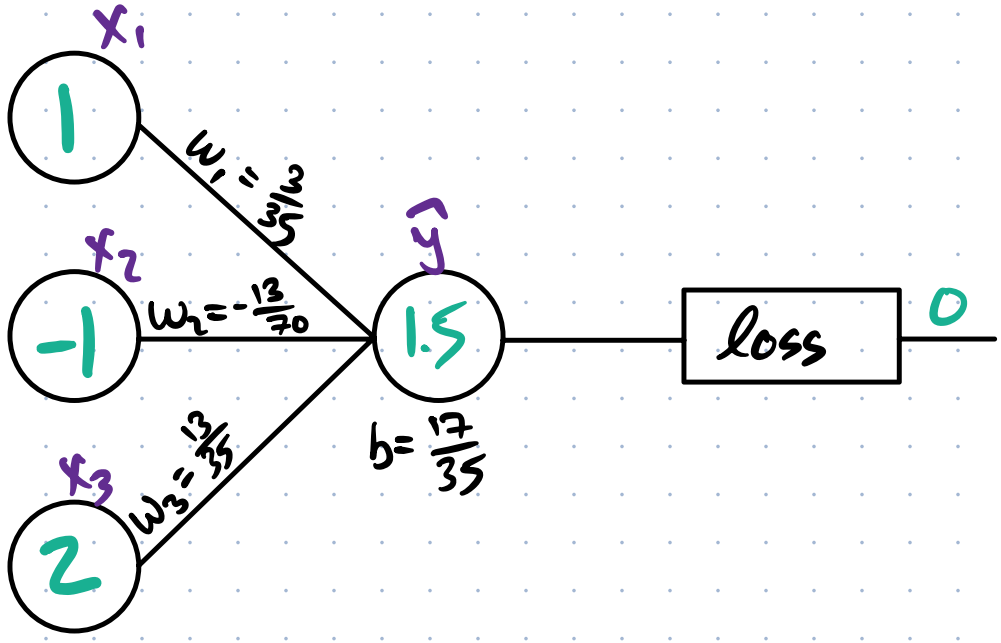
(prev. 0.01)

Now repeat. Compute $-\nabla NN$ with these altered weights, adjust by $\frac{1}{100}$ of it, and so on.

* Python demo.

Training data: $(1, -1, 2) \to 1.5$

After 100 or so repetitions:



Easy because only one piece of training data!

Example 2: Backpropagation on a Batch of data.

Training Data:

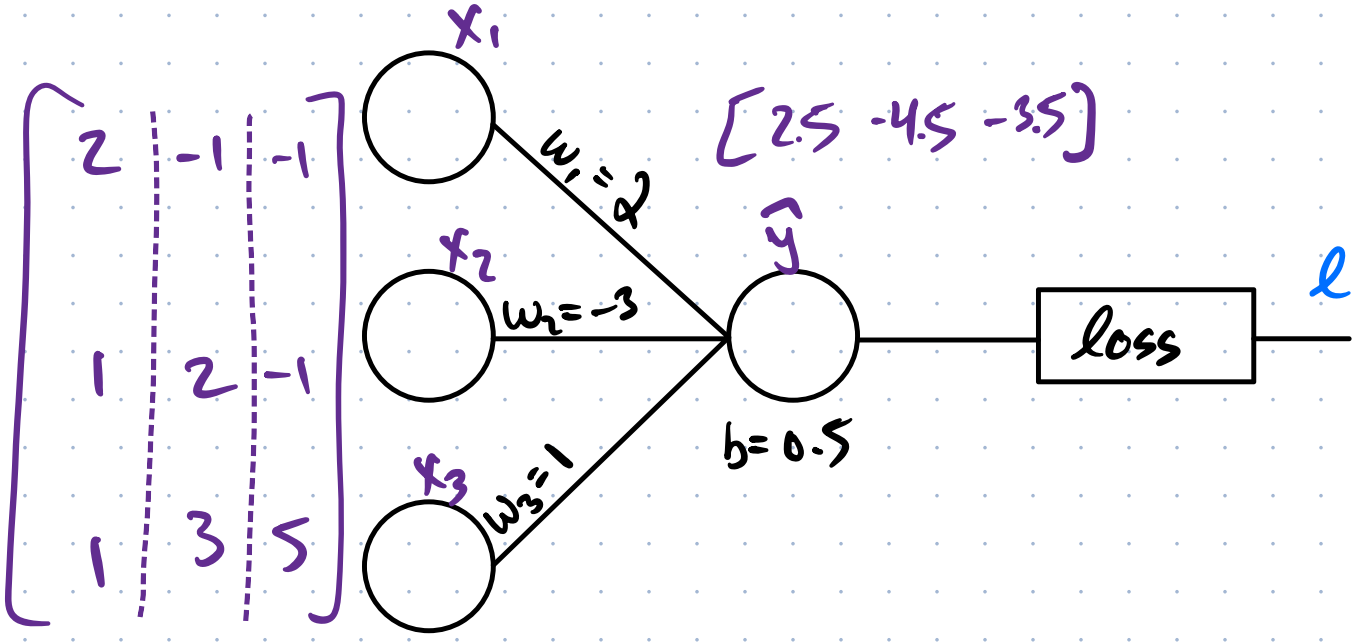
$(2, 1, 1) \to 3$

$(-1, 2, 3) \to -4$

$(-1, -1, -5) \to -5$

Still no AF

no hidden layers



$$\ell = \frac{1}{3}\left((2.5-3)^2 + (-4.5--4)^2 + (-3.5--5)^2\right)$$

$$= 0.916\ldots$$

(making the weights whole #s for now)

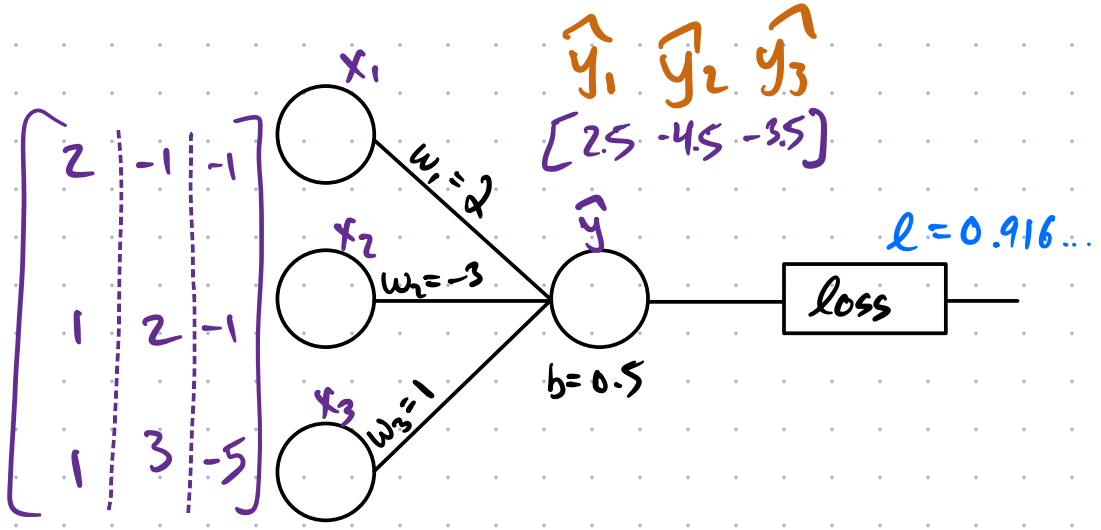
Example 2: Backpropagation on a Batch of data.

Training Data:

$(2, 1, 1) \to 3$

$(-1, 2, 3) \to -4$

$(-1, -1, -5) \to -5$



We fed forward, now we backpropagate the derivatives

$$\ell = \frac{1}{3}\left((\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2\right)$$

$$\frac{\partial \ell}{\partial \hat{y}_1} = \frac{2}{3}(\hat{y}_1 - y_1) \cdot (1) = \frac{2}{3} \cdot (-0.5) = -\frac{1}{3}$$

$$\frac{\partial \ell}{\partial \hat{y}_2} = \frac{2}{3}(\hat{y}_2 - y_2) = \frac{2}{3} \cdot (-0.5) = -\frac{1}{3}$$

$$\frac{\partial \ell}{\partial \hat{y}_3} = \frac{2}{3}(\hat{y}_3 - y_3) = \frac{2}{3} \cdot (1.5) = 1$$

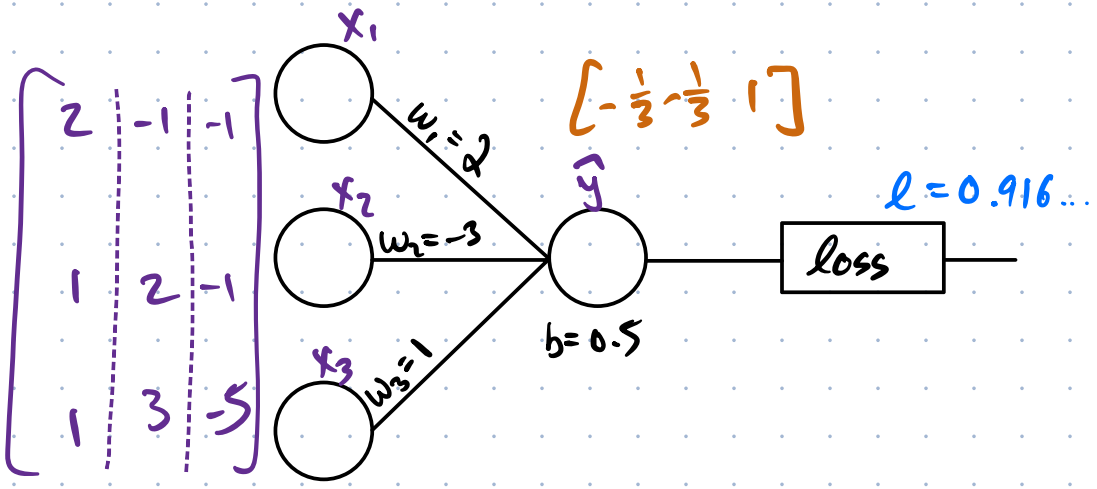
Example 2: Backpropagation on a Batch of data.

Training Data:

$(2, 1, 1) \to 3$

$(-1, 2, 3) \to -4$

$(-1, -1, -5) \to -5$



We fed forward, now we backpropagate the derivatives

$\frac{\partial \ell}{\partial \hat{y}_1} = -\frac{1}{3}$ $\frac{\partial \ell}{\partial \hat{y}_2} = -\frac{1}{3}$ $\frac{\partial \ell}{\partial \hat{y}_3} = 1$

$\hat{y} = x_1 w_1 + x_2 w_2 + x_3 w_3 + b$

$\frac{\partial \ell}{\partial w_1} = \frac{\partial \ell}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial w_1} = -\frac{1}{3} \cdot 2 = -\frac{2}{3}$

OR $= \frac{\partial \ell}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial w_1} = -\frac{1}{3} \cdot -1 = \frac{1}{3}$

OR $= \frac{\partial \ell}{\partial \hat{y}_3} \cdot \frac{\partial \hat{y}_3}{\partial w_1} = 1 \cdot -1 = -1$

$\frac{\partial \ell}{\partial w_1}$ is different for each sample, which makes sense.

Example 2: Backpropagation on a Batch of data.

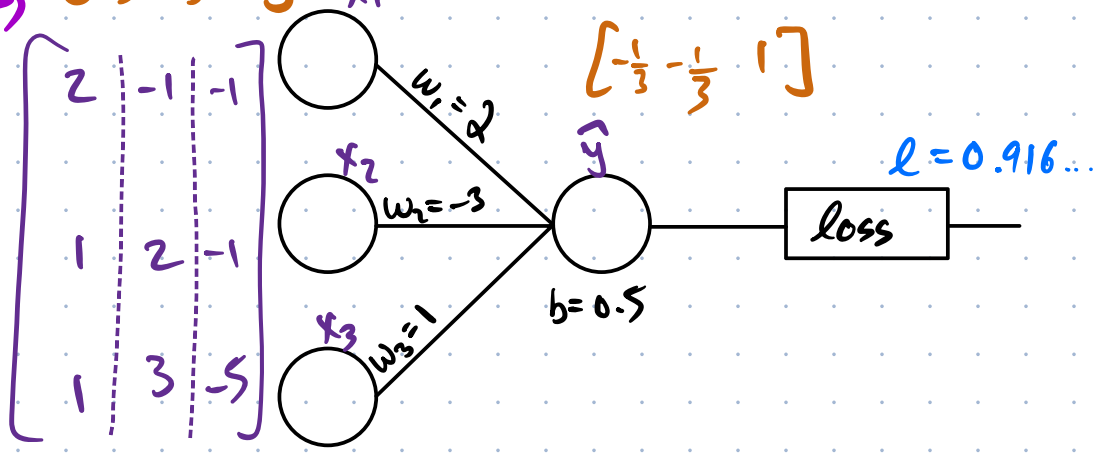
Training Data:

$(2,1,1) \to 3$

$(-1,2,3) \to -4$

$(-1,-1,-5) \to -5$

$\begin{bmatrix} 2 & -1 & -1 \\ 1 & 2 & -1 \\ 1 & 3 & -5 \end{bmatrix}$



We fed forward, now we backpropagate the derivatives

$\frac{\partial \ell}{\partial \hat{y}_1} = -\frac{1}{3} \quad \frac{\partial \ell}{\partial \hat{y}_2} = -\frac{1}{3} \quad \frac{\partial \ell}{\partial \hat{y}_3} = 1$

$\hat{y} = x_1 w_1 + x_2 w_2 + x_3 w_3 + b$

$\frac{\partial \ell}{\partial w_1} = \frac{\partial \ell}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial w_1} = -\frac{1}{3} \cdot 2 = -\frac{2}{3}$

OR $= \frac{\partial \ell}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial w_1} = -\frac{1}{3} \cdot -1 = \frac{1}{3}$

OR $= \frac{\partial \ell}{\partial \hat{y}_3} \cdot \frac{\partial \hat{y}_3}{\partial w_1} = 1 \cdot -1 = -1$

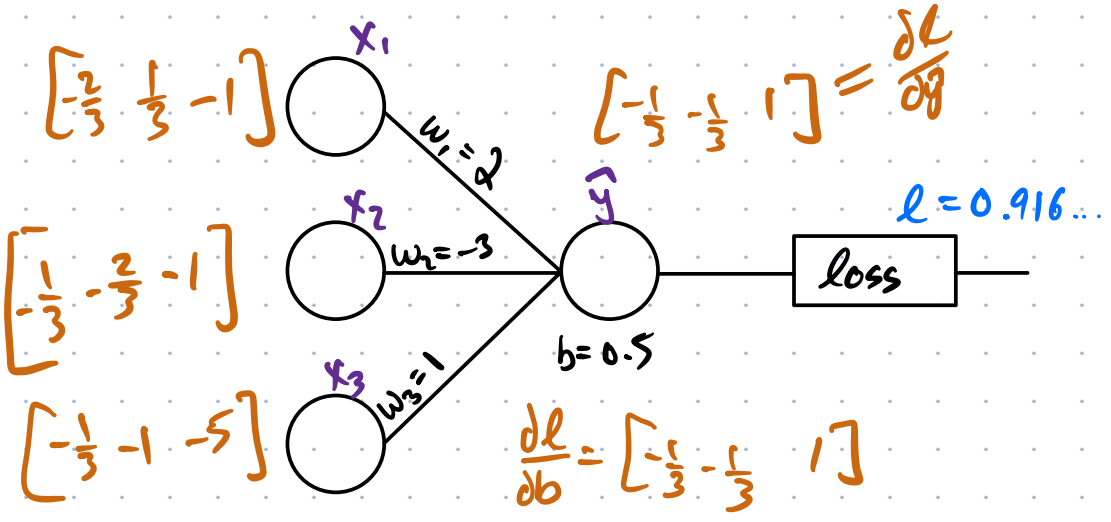
Example 2: Backpropagation on a Batch of data.

Training Data:

$(2, 1, 1) \to 3$

$(-1, 2, 3) \to -4$

$(-1, -1, -5) \to -5$



We fed forward, now we backpropagate the derivatives

$$\frac{\partial \ell}{\partial \hat{y}_1} = [-\tfrac{1}{3}\ -\tfrac{1}{3}\ 1]$$

$$\hat{y} = x_1 w_1 + x_2 w_2 + x_3 w_3 + b$$

$$\frac{\partial \ell}{\partial w_1} = [-\tfrac{2}{3}\ \tfrac{1}{3}\ -1]$$

$$\frac{\partial \ell}{\partial w_3} = [-\tfrac{1}{3}\ -1\ -5]$$

$$\frac{\partial \ell}{\partial w_2} = [-\tfrac{1}{3}\ -\tfrac{2}{3}\ -1]$$

$$\frac{\partial \ell}{\partial b} = [-\tfrac{1}{3}\ -\tfrac{1}{3}\ 1]$$

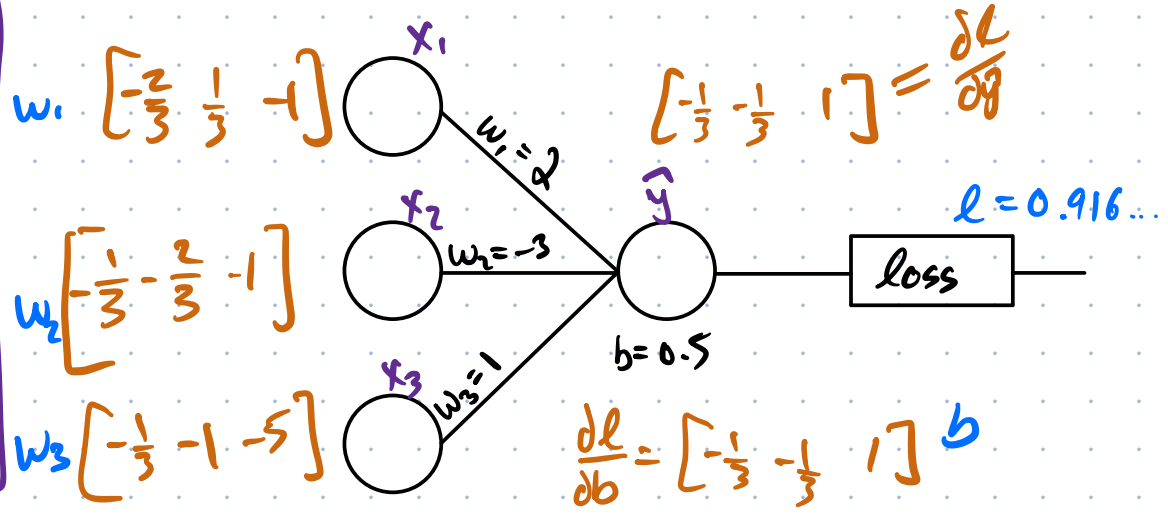
Example 2: Backpropagation on a Batch of data.

Training Data:

$(2, 1, 1) \to 3$

$(-1, 2, 3) \to -4$

$(-1, -1, -5) \to -5$



Now we add the gradients from each of the 3 samples together for each weight/bias.

$$\nabla = \begin{bmatrix} -4/3 \\ -2 \\ -19/3 \\ \frac{1}{3} \end{bmatrix} \begin{matrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{matrix}$$

Now adjust w_1, w_2, w_3, b by $-\nabla \cdot \frac{1}{100}$ and repeat.

Why? The loss already has $\frac{1}{3}$ factor, so this addition is really averaging the three gradients!

Example 2: Backpropagation on a Batch of data.

Training Data:

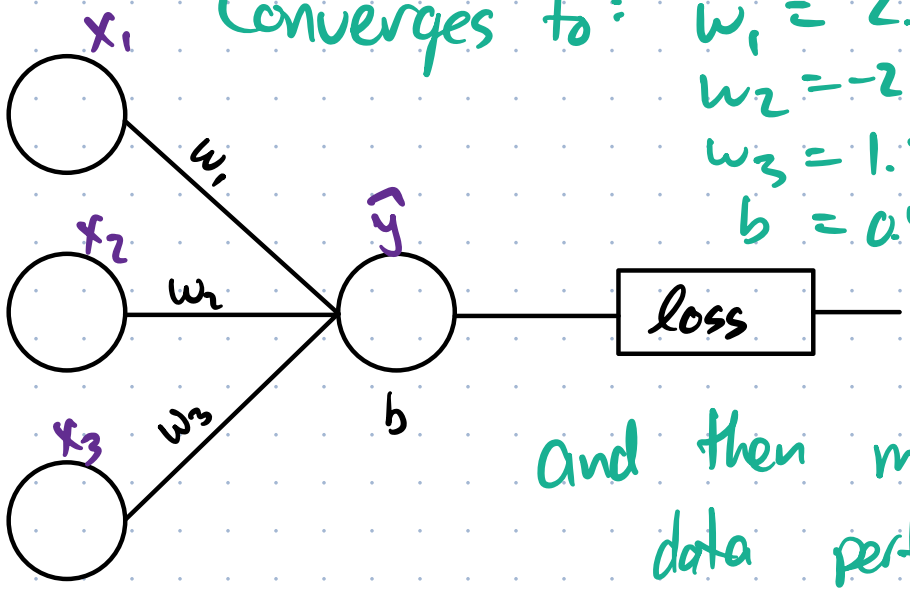
$(2, 1, 1) \to 3$

$(-1, 2, 3) \to -4$

$(-1, -1, -5) \to -5$

* Demo

Demo 3
batch of 6



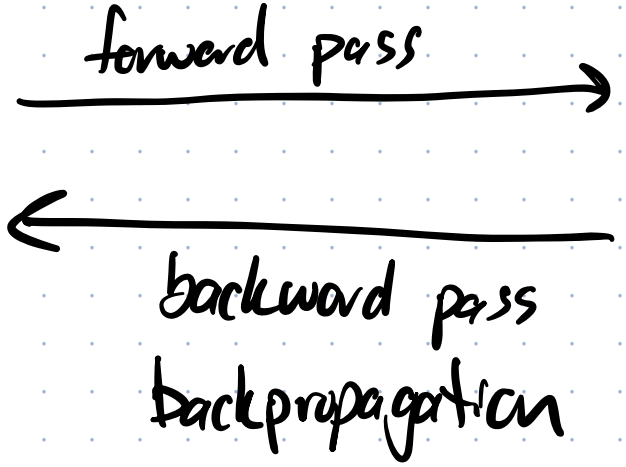
Converges to: $w_1 = 2.167\ldots$
$w_2 = -2.992\ldots$
$w_3 = 1.247\ldots$
$b = 0.4104\ldots$

and then matches the training data perfectly

* You feed forward training data
* Then you compute backward gradients
* You adjust the weights/biases accordingly
* Then you feed forward the same data again, or new data, over and over again.

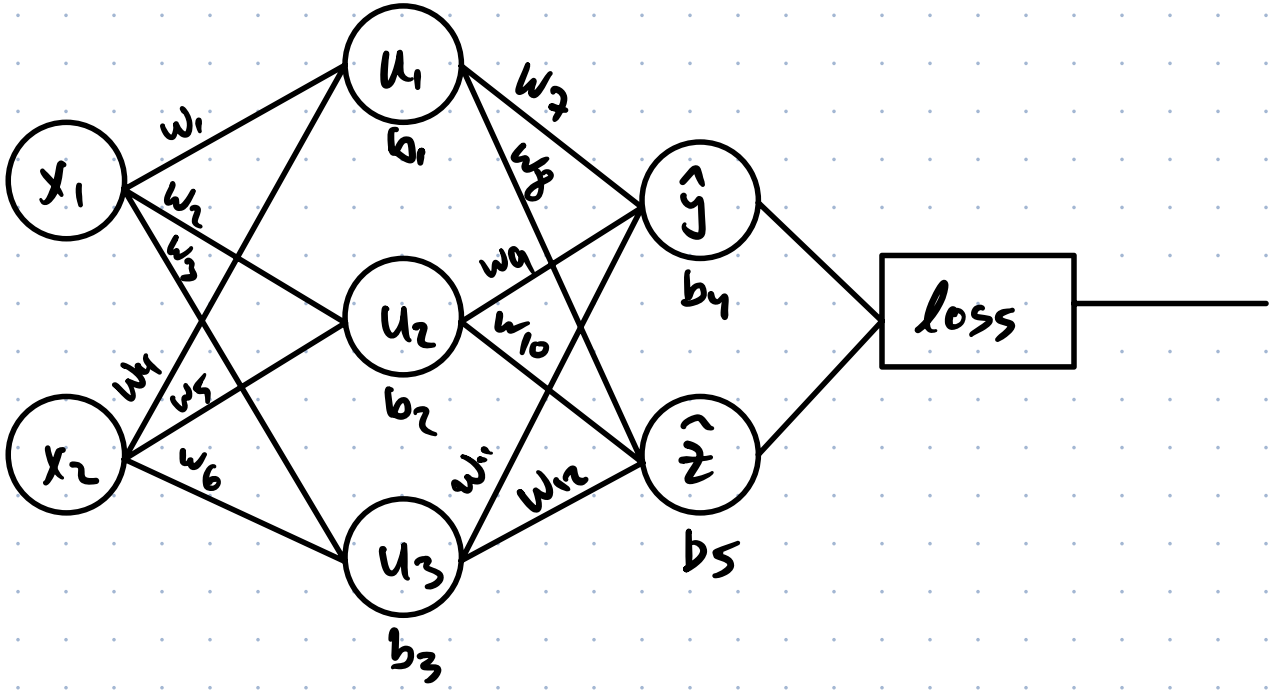
↳ many clever ways to do this beyond just $-\frac{2}{m} \cdot V$ "optimizers"

"Adam optimizer"



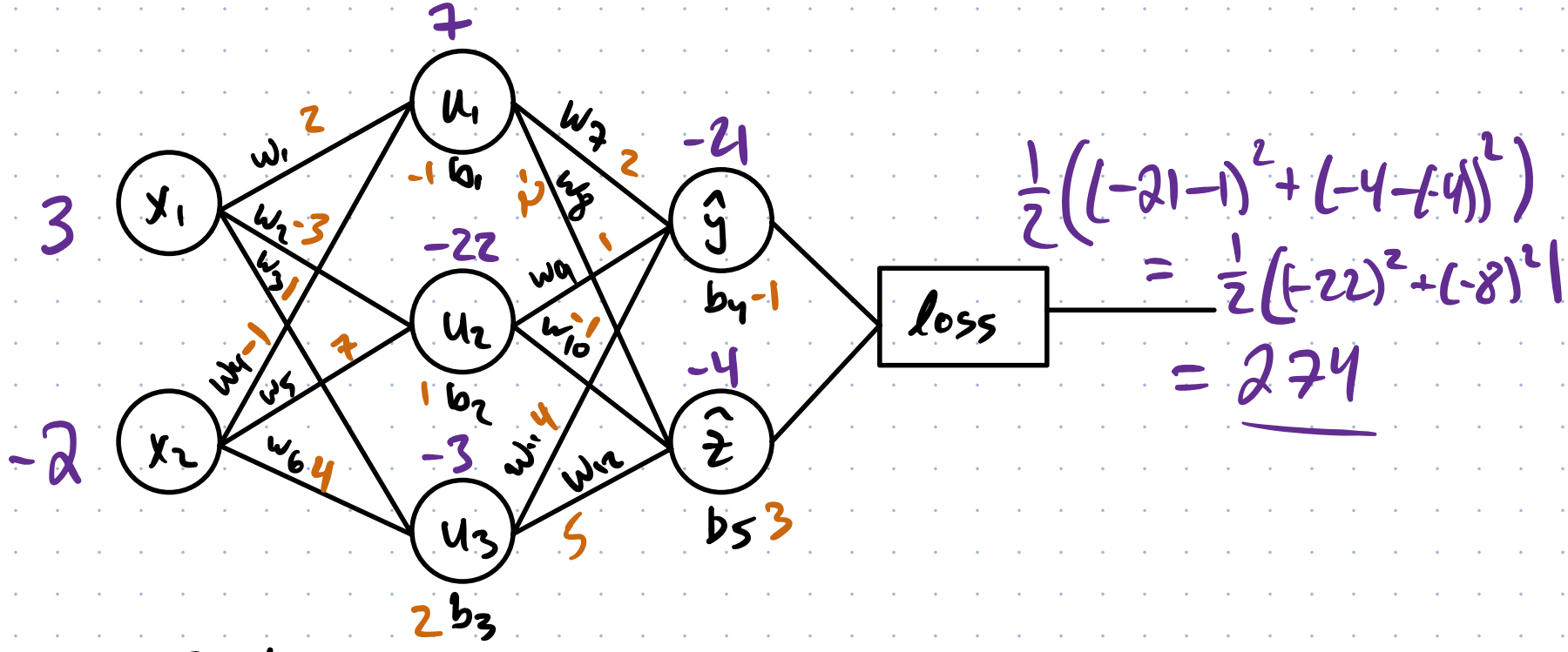
Example 3, with more layers,
still no activation functions
just 1 sample for simplicity

for more, just add them together to get the ∇ for weights and biases, as before



Sample: $(3, -2) \rightarrow (1, 4)$

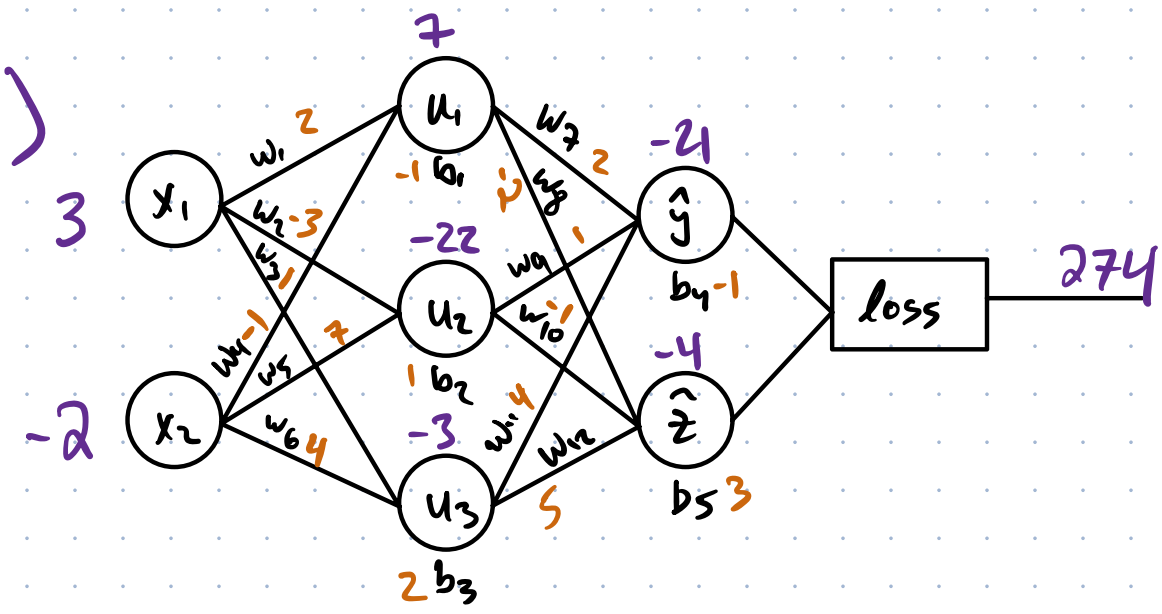
Sample: $(3, -2) \to (1, 4)$



Backprop: Find:

$$\frac{\partial \ell}{\partial w_1}, \frac{\partial \ell}{\partial w_2}, \ldots, \frac{\partial \ell}{\partial w_{12}}, \frac{\partial \ell}{\partial b_1}, \frac{\partial \ell}{\partial b_2}, \ldots, \frac{\partial \ell}{\partial b_5}$$

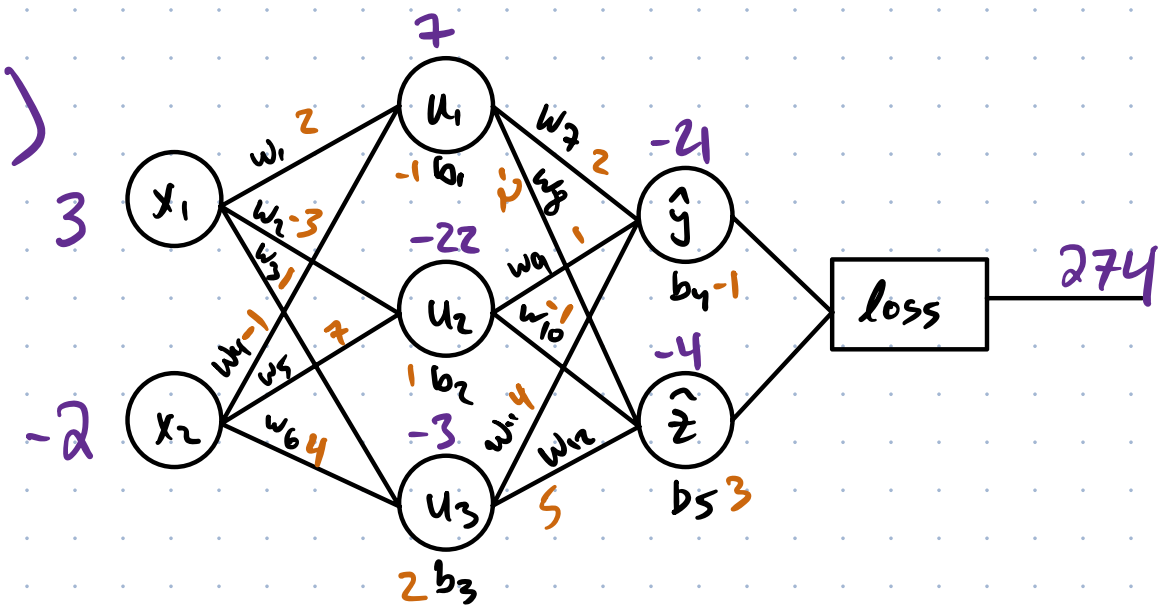
Sample: $(3, -2) \to (1, 4)$



Work backward from the right.

$\frac{\partial L}{\partial \hat{y}}$?

Sample: $(3, -2) \to (1, 4)$



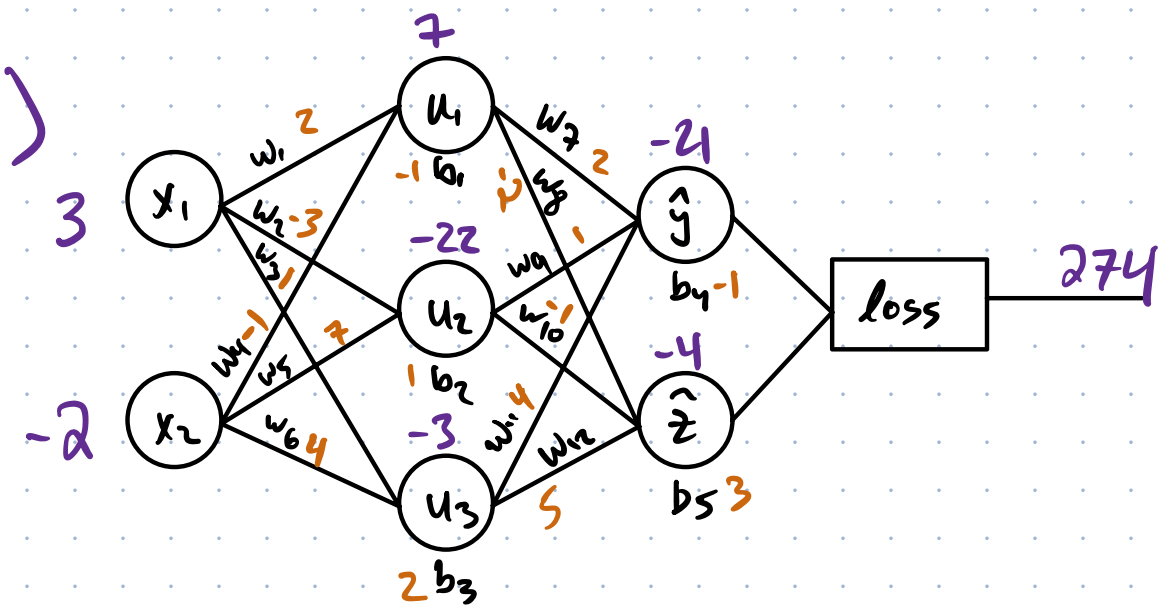
Work backward from the right.

$\frac{\partial \ell}{\partial \hat{y}}$?

$$\ell = \frac{1}{2}\left((\hat{y}-1)^2 + (\hat{z}-4)^2\right)$$

$$\frac{\partial \ell}{\partial \hat{y}} = \hat{y} - 1 = -22$$

Sample: $(3, -2) \to (1, 4)$



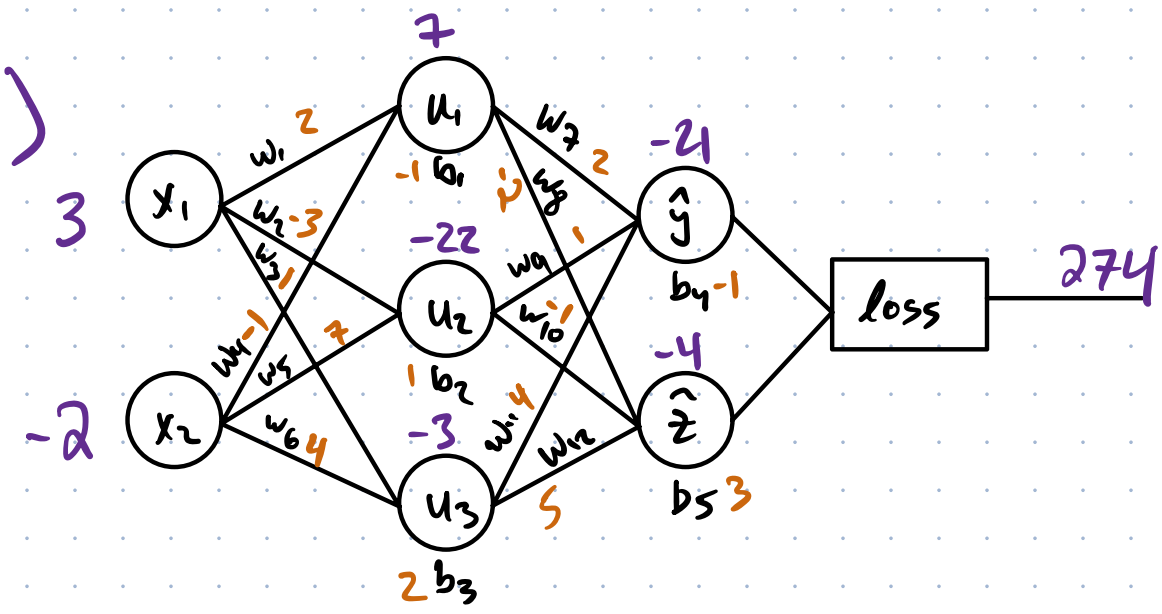
Work backward from the right.

$\frac{\partial \ell}{\partial \hat{y}} = -22$ $\frac{\partial \ell}{\partial \hat{z}} = ?$

$$\ell = \frac{1}{2}\left((\hat{y}-1)^2 + (\hat{z}-4)^2\right)$$

$$\frac{\partial \ell}{\partial \hat{z}} = \hat{z} - 4 = -2 - 4 = -6$$

Sample: $(3, -2) \to (1, 4)$



Work backward from the right.

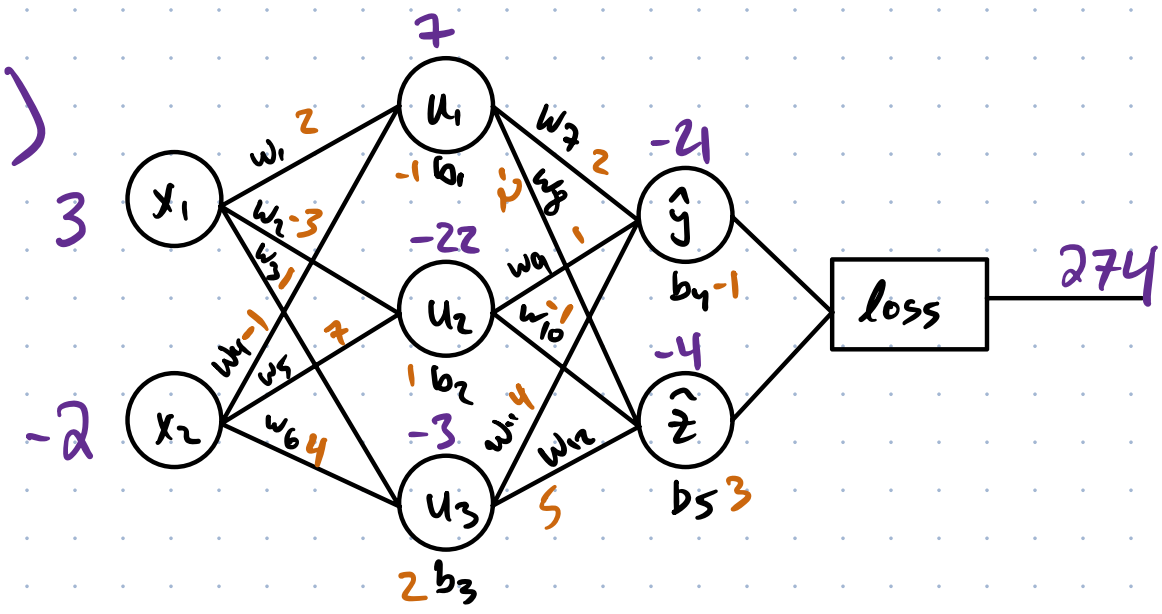
$$\frac{\partial \ell}{\partial \hat{y}} = -22 \qquad \frac{\partial \ell}{\partial \hat{z}} = -6$$

$$\frac{\partial \ell}{\partial w_7} = \frac{\partial \ell}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_7}$$ ↙ ?

$$\hat{y} = u_1 \cdot w_7 + u_2 \cdot w_9 + u_3 \cdot w_{11} + b_4$$

$$\Rightarrow \quad \frac{\partial \hat{y}}{\partial w_7} = u_1 = 7$$

Sample: $(3, -2) \to (1, 4)$



Work backward from the right.

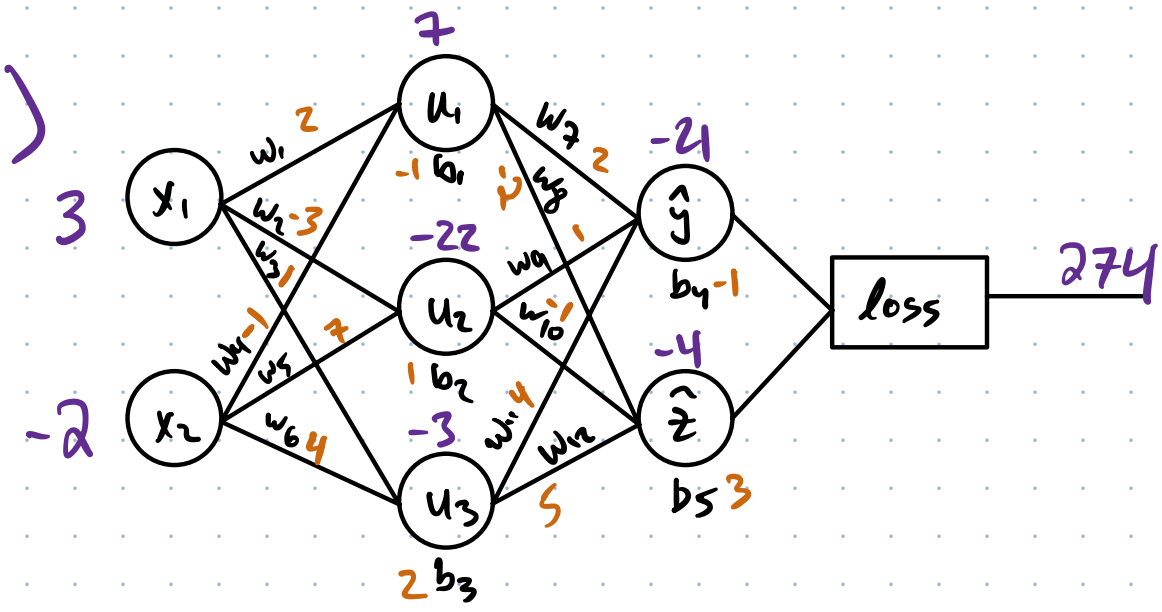
$\frac{\partial \ell}{\partial \hat{y}} = -22 \quad \frac{\partial \ell}{\partial \hat{z}} = -6$

$$\frac{\partial \ell}{\partial w_7} = \frac{\partial \ell}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_7} = -22 \cdot 7 = -154$$

$$\hat{y} = u_1 \cdot w_7 + u_2 \cdot w_9 + u_3 \cdot w_{11} + b_4$$

$$\Rightarrow \frac{\partial \hat{y}}{\partial w_7} = u_1 = 7$$

Sample: $(3, -2) \to (1, 4)$



Work backward from the right.

$\frac{\partial \ell}{\partial \hat{y}} = -22 \quad \frac{\partial \ell}{\partial \hat{z}} = -6$

$\frac{\partial \ell}{\partial w_7} = -154$

$\frac{\partial \ell}{\partial w_8} = (-6)(7) = -42$

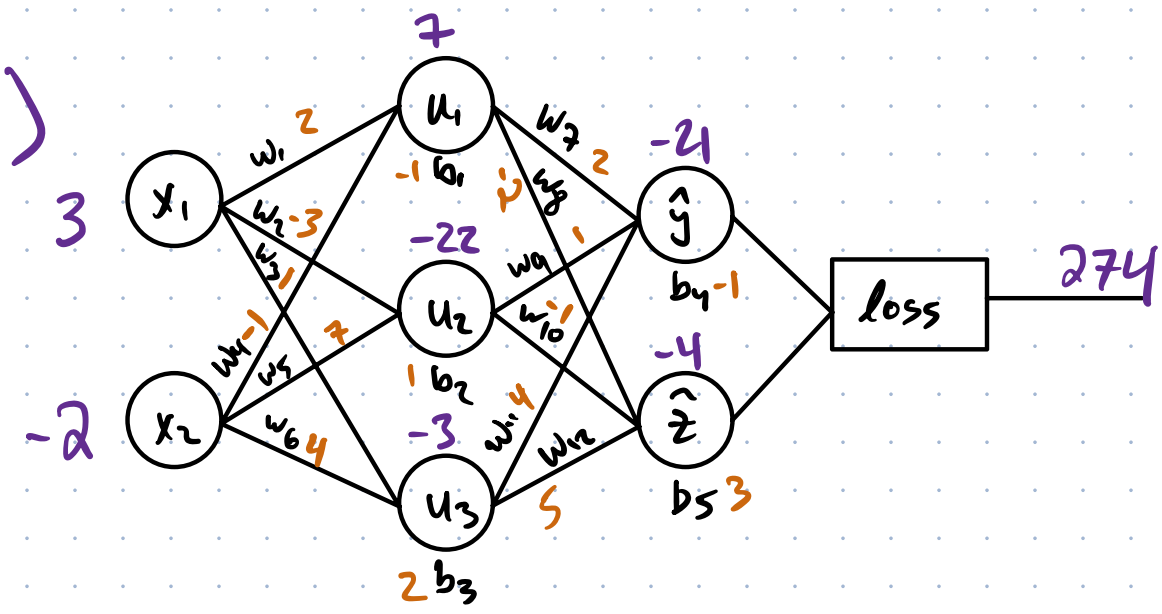
$\frac{\partial \ell}{\partial w_9} = (-22)(-22) = 484$

$\frac{\partial \ell}{\partial w_{10}} = (-6)(-22) = 132$

$\frac{\partial \ell}{\partial w_{11}} = (-22)(-3) = 66$

$\frac{\partial \ell}{\partial w_{12}} = (-6)(-3) = 18$

Sample: $(3, -2) \to (1, 4)$



Work backward from the right.

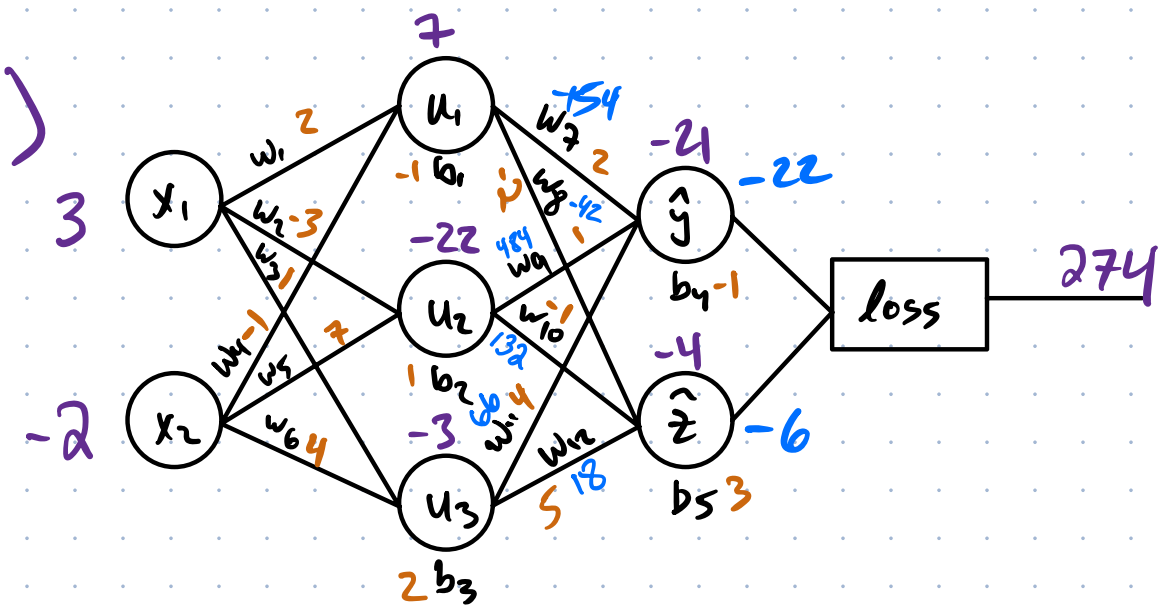
$\frac{\partial \ell}{\partial \hat{y}} = -22$ $\frac{\partial \ell}{\partial \hat{z}} = -6$

$\frac{\partial \ell}{\partial w_7} = -154$ $\frac{\partial \ell}{\partial w_8} = -42$

$\frac{\partial \ell}{\partial w_9} = 484$ $\frac{\partial \ell}{\partial w_{10}} = 132$

$\frac{\partial \ell}{\partial w_{11}} = 66$ $\frac{\partial \ell}{\partial w_{12}} = 18$

Sample: $(3, -2) \to (1, 4)$



Work backward from the right.

$\frac{\partial \ell}{\partial \hat{y}} = -22 \quad \frac{\partial \ell}{\partial \hat{z}} = -6$

$\frac{\partial \ell}{\partial w_7} = -154 \quad \frac{\partial \ell}{\partial w_8} = -42$

$\frac{\partial \ell}{\partial w_9} = 484 \quad \frac{\partial \ell}{\partial w_{10}} = 132$

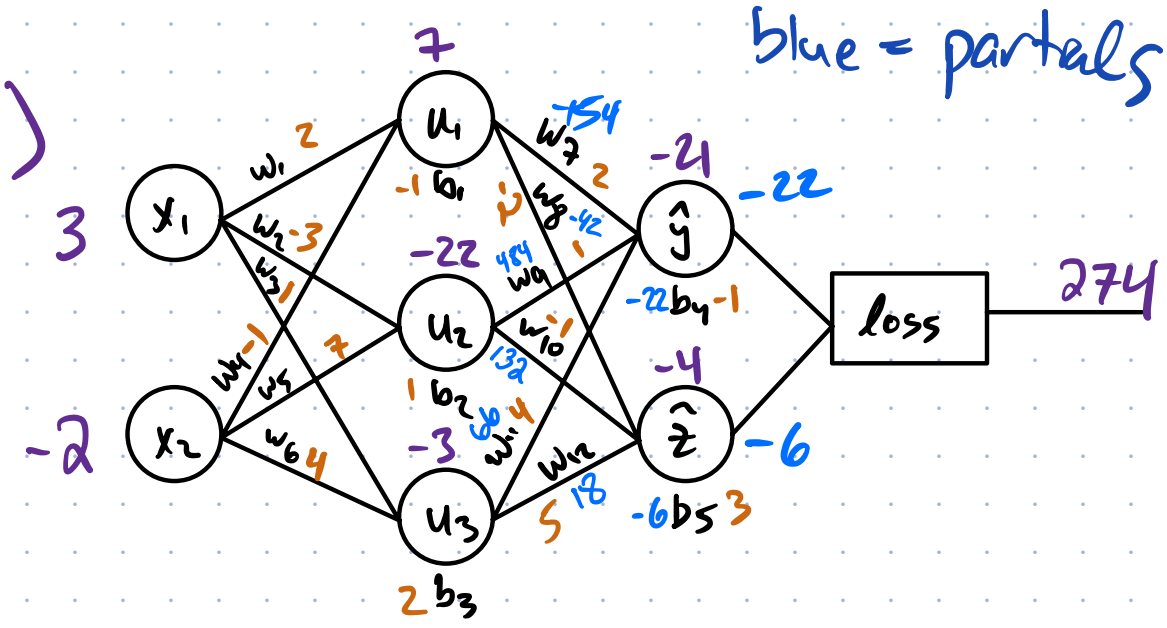
$\frac{\partial \ell}{\partial w_{11}} = 66 \quad \frac{\partial \ell}{\partial w_{12}} = 18$

$$\frac{\partial \ell}{\partial b_4} = \frac{\partial \ell}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} = (-22)(1) = -22$$

$$\hat{y} = u_1 w_7 + u_2 w_9 + u_3 w_{11} + b$$

Sample: $(3, -2) \to (1, 4)$

blue = partials



Work backward from the right.

$\frac{\partial \ell}{\partial \hat{y}} = -22$ $\frac{\partial \ell}{\partial \hat{z}} = -6$

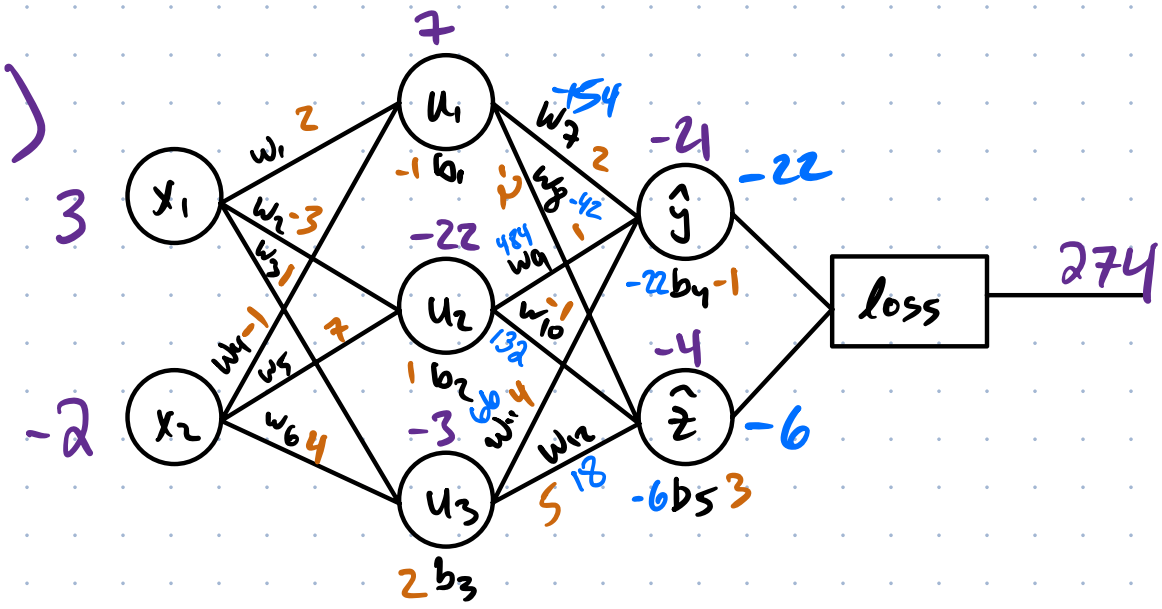
$\frac{\partial \ell}{\partial b_4} = -22$ $\frac{\partial \ell}{\partial b_5} = -6$

$\frac{\partial \ell}{\partial w_7} = -154$ $\frac{\partial \ell}{\partial w_8} = -42$

$\frac{\partial \ell}{\partial w_9} = 484$ $\frac{\partial \ell}{\partial w_{10}} = 132$

$\frac{\partial \ell}{\partial w_{11}} = 66$ $\frac{\partial \ell}{\partial w_{12}} = 18$

Sample: $(3, -2) \to (1, 4)$



Work backward from the right.

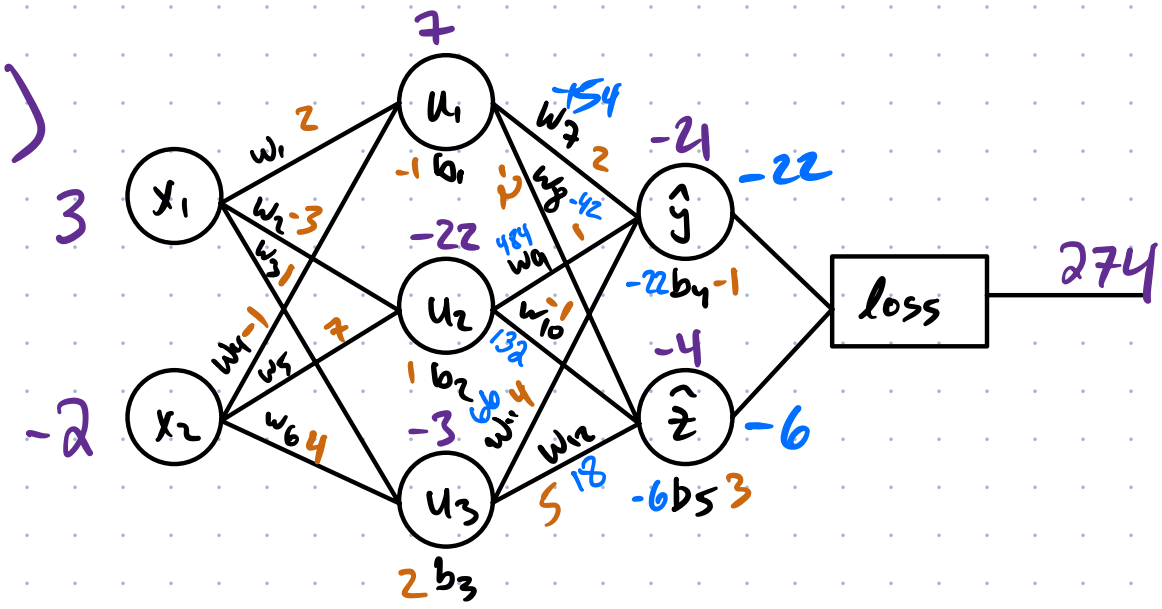
Next:

$$\frac{\partial \ell}{\partial w_1} = \cdots$$

$$u_1 = w_1 x_1 + w_4 x_2 + b_1$$

$$\text{So, } \frac{\partial u_1}{\partial w_1} = x_1 = 3$$

Sample: $(3, -2) \to (1, 4)$



Work backward from the right.

Next:

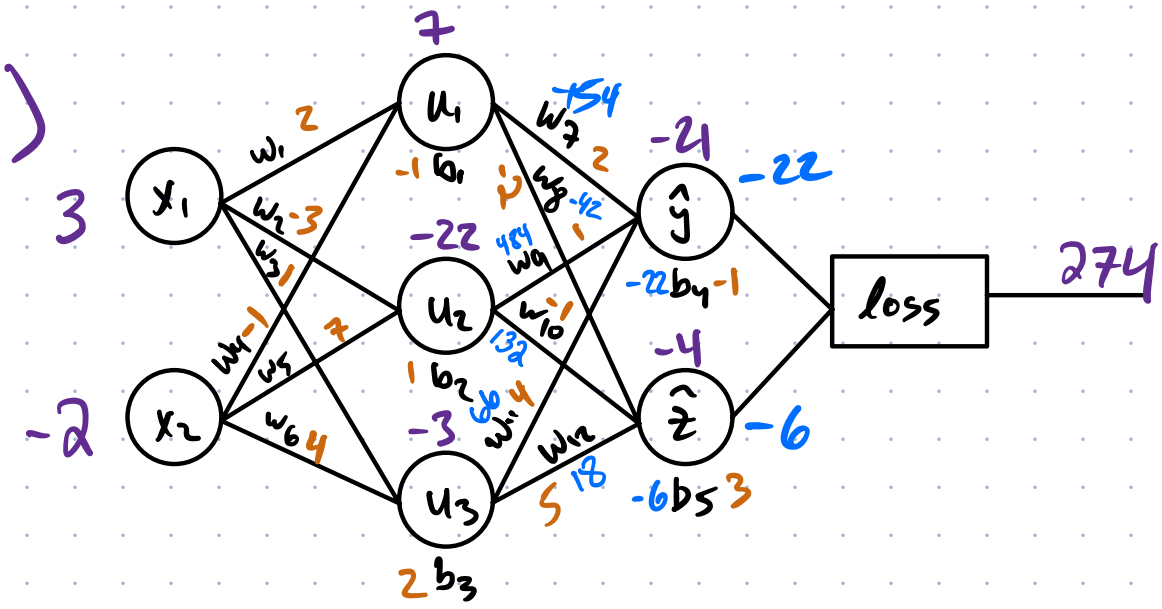
$$\frac{\partial \ell}{\partial w_1} = \frac{\partial \ell}{\partial u_1} \cdot \frac{\partial u_1}{\partial w_1}$$

not computed yet

$$u_1 = w_1 x_1 + w_2 x_2 + b_1$$

So, $$\frac{\partial u_1}{\partial w_1} = x_1 = 3$$

Sample: $(3, -2) \to (1, 4)$



Work backward from the right.

Next:

$$\frac{\partial \ell}{\partial w_1} = \frac{\partial \ell}{\partial u_1} \cdot \frac{\partial u_1}{\partial w_1}$$

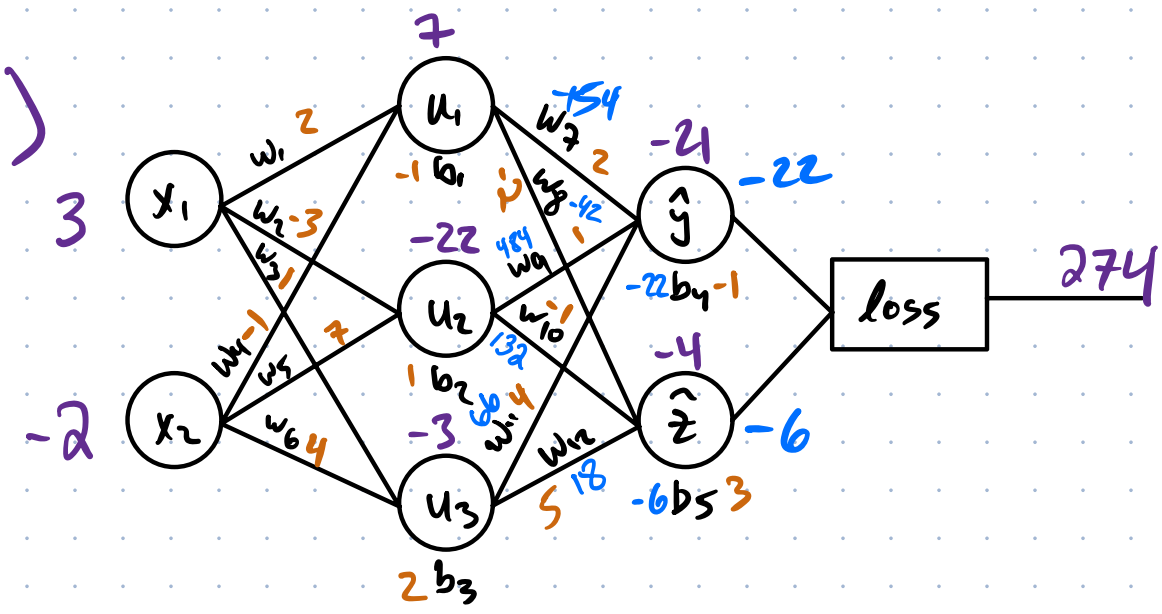
not computed yet

$$u_1 = w_1 x_1 + w_2 x_2 + b_1$$

So, $\frac{\partial u_1}{\partial w_1} = x_1 = 3$

Now we see that with hidden layers, we also need derivs of the neuron values to keep passing backward

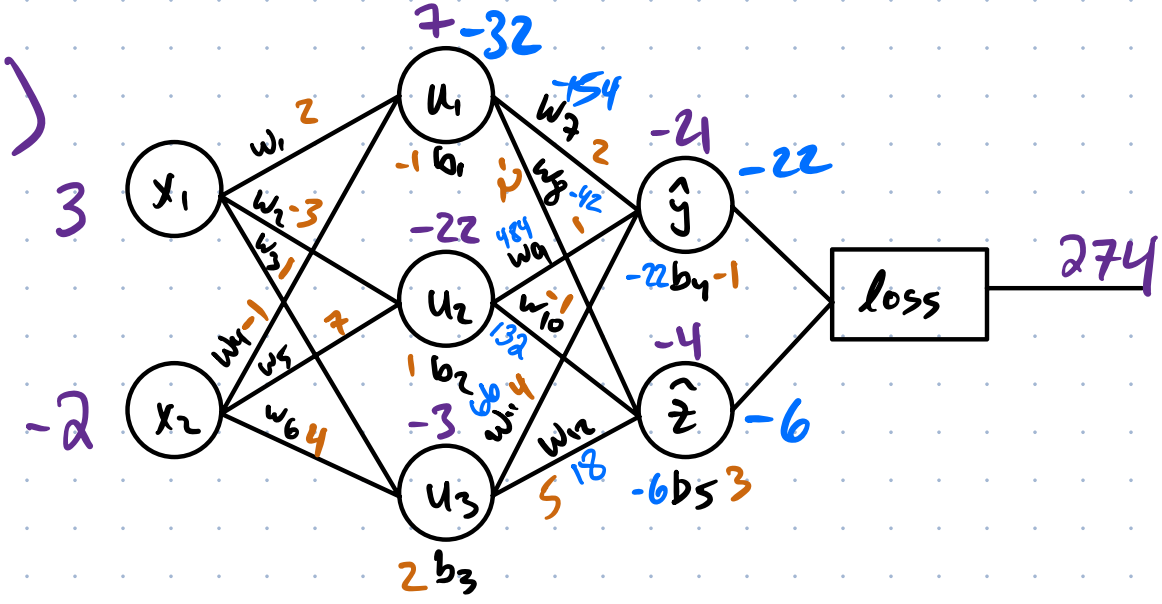
Sample: $(3, -2) \to (1, 4)$



Work backward from the right.

$$\frac{\partial \ell}{\partial w_1} = \underbrace{\frac{\partial \ell}{\partial u_1}}_{?} \cdot \frac{\partial u_1}{\partial w_1} = 3$$

Sample: $(3, -2) \to (1, 4)$



Work backward from the right.

$$\frac{\partial \ell}{\partial w_1} = \underbrace{\frac{\partial \ell}{\partial u_1}}_{?} \cdot \frac{\partial u_1}{\partial w_1} = 3$$

Faa di Bruno formula

u_1 affects the loss along two paths:

$$u_1 \xrightarrow{w_7} \hat{y} \to \ell$$

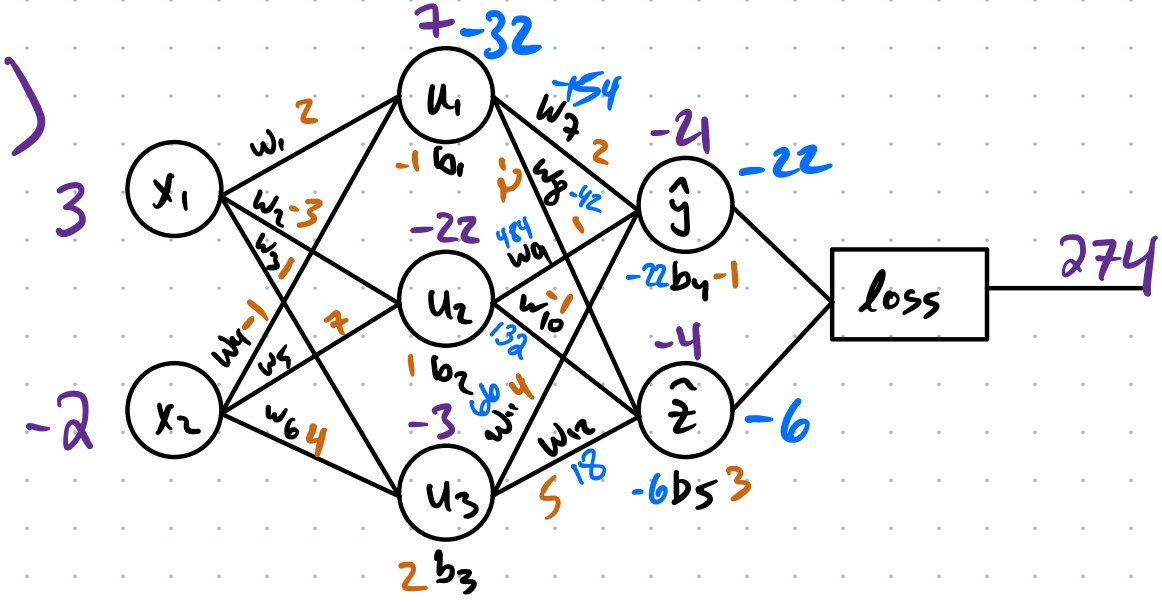
$$u_1 \xrightarrow{w_8} \hat{z} \to \ell$$

So we have the <u>multivariate chain</u> rule now:

$$\frac{\partial \ell}{\partial u_1} = \frac{\partial \ell}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial u_1} + \frac{\partial \ell}{\partial \hat{z}} \cdot \frac{\partial \hat{z}}{\partial u_1} = (-22)\cdot(2) + (-6)(-2)$$

$$= -32$$

Sample: $(3, -2) \to (1, 4)$



Work backward from the right.

$$\frac{\partial \ell}{\partial w_1} = \frac{\partial \ell}{\partial u_1} \cdot \frac{\partial u_1}{\partial w_1}$$

$\frac{\partial u_1}{\partial w_1} = 3$

$\frac{\partial \ell}{\partial u_1}$: ?

$$\frac{\partial \hat{y}}{\partial w_7} = u_1$$

$$\frac{\partial \hat{y}}{\partial u_1} = w_7$$

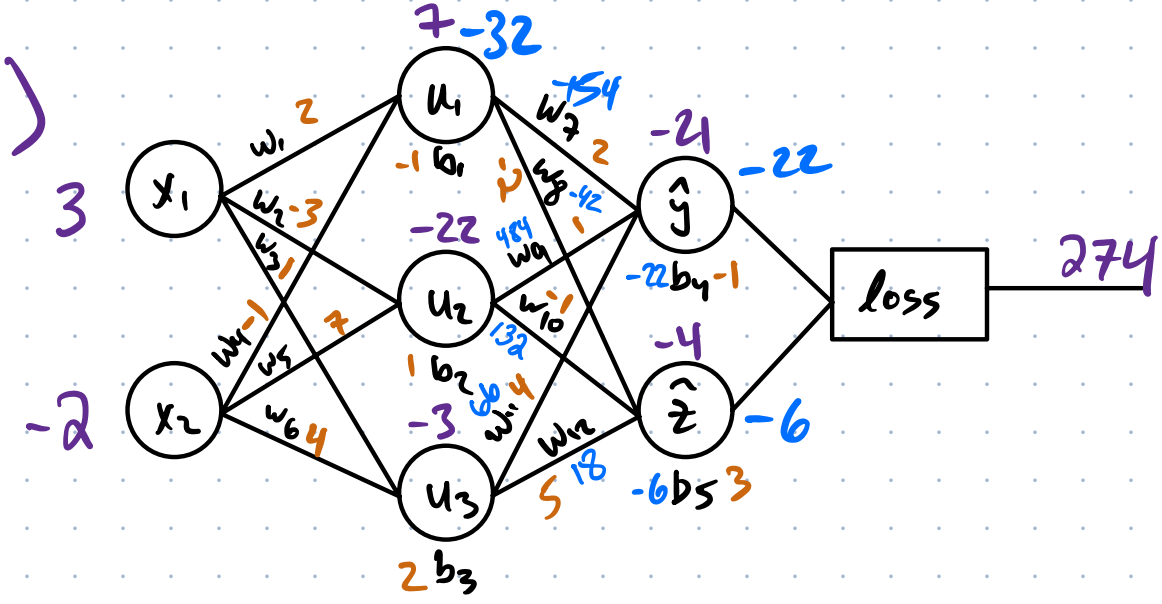
So we have the <u>multivariate chain</u> rule now:

$$\frac{\partial \ell}{\partial u_1} = \frac{\partial \ell}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial u_1} + \frac{\partial \ell}{\partial \hat{z}} \cdot \frac{\partial \hat{z}}{\partial u_1} = (-22)\cdot(2) + (-6)(-2) = -32$$

$$\frac{\partial \ell}{\partial u_1} = w_7 \cdot \frac{\partial \ell}{\partial \hat{y}} + w_8 \cdot \frac{\partial \ell}{\partial \hat{z}} = \begin{bmatrix} w_7 \\ w_8 \end{bmatrix} \cdot \begin{bmatrix} \partial \ell / \partial \hat{y} \\ \partial \ell / \partial \hat{z} \end{bmatrix}$$

Sample: $(3, -2) \to (1, 4)$

Work backward from the right.



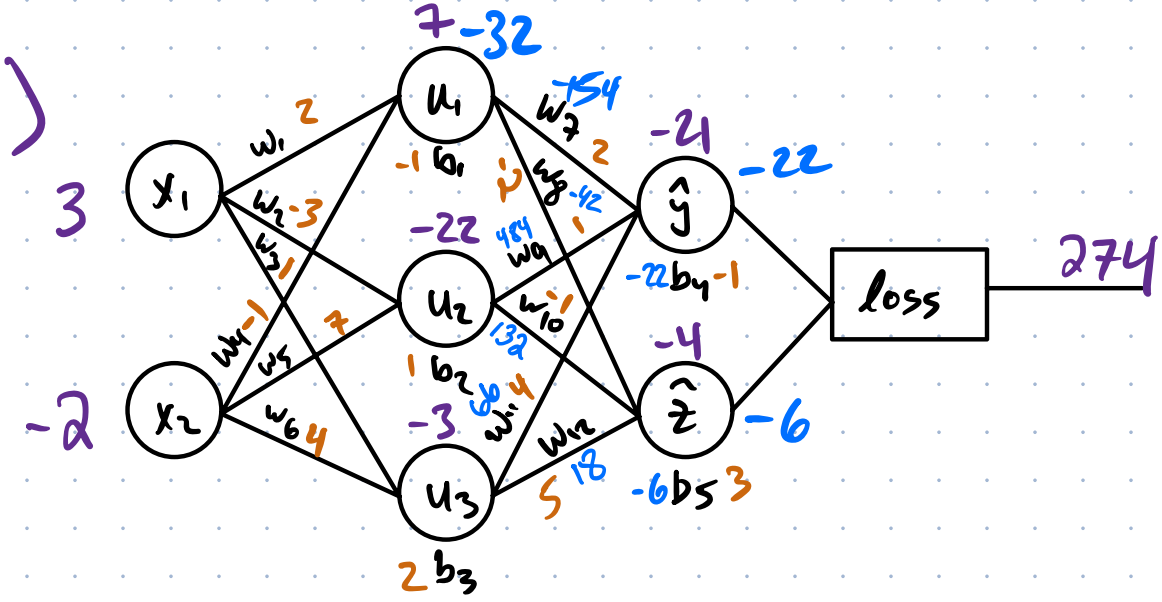
$$\frac{\partial \ell}{\partial u_1} = w_7 \cdot \frac{\partial \ell}{\partial \hat{y}} + w_8 \cdot \frac{\partial \ell}{\partial \hat{z}} = \begin{bmatrix} w_7 \\ w_8 \end{bmatrix} \cdot \begin{bmatrix} \partial \ell / \partial \hat{y} \\ \partial \ell / \partial \hat{z} \end{bmatrix}$$

$$\frac{\partial \ell}{\partial u_2} = \begin{bmatrix} w_9 \\ w_{10} \end{bmatrix} \begin{bmatrix} \partial \ell / \partial \hat{y} \\ \partial \ell / \partial \hat{z} \end{bmatrix}$$

$$\frac{\partial \ell}{\partial u_3} = \begin{bmatrix} w_{11} \\ w_{12} \end{bmatrix} \begin{bmatrix} \partial \ell / \partial \hat{y} \\ \partial \ell / \partial \hat{z} \end{bmatrix}$$

Sample: $(3, -2) \to (1, 4)$

Work backward from the right.



$$\frac{\partial \ell}{\partial u_1} = w_7 \cdot \frac{\partial \ell}{\partial \hat{y}} + w_8 \cdot \frac{\partial \ell}{\partial \hat{z}} = \begin{bmatrix} w_7 \\ w_8 \end{bmatrix} \cdot \begin{bmatrix} \partial \ell / \partial \hat{y} \\ \partial \ell / \partial \hat{z} \end{bmatrix}$$

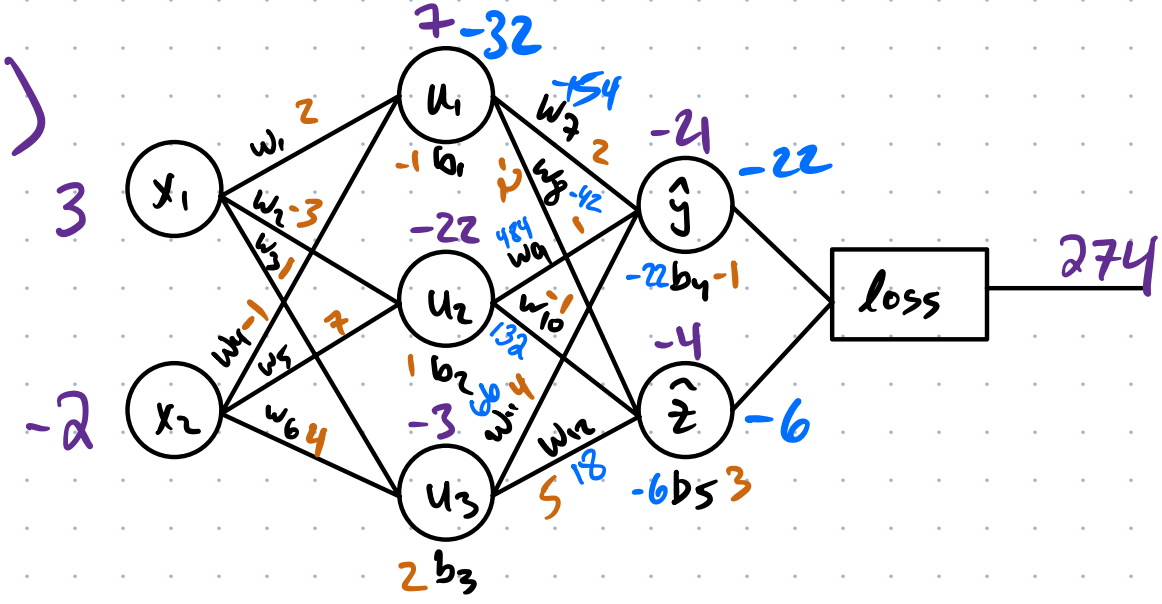
$$\frac{\partial \ell}{\partial u_2} = \begin{bmatrix} w_9 \\ w_{10} \end{bmatrix} \begin{bmatrix} \partial \ell / \partial \hat{y} \\ \partial \ell / \partial \hat{z} \end{bmatrix}$$

$$\frac{\partial \ell}{\partial u_3} = \begin{bmatrix} w_{11} \\ w_{12} \end{bmatrix} \begin{bmatrix} \partial \ell / \partial \hat{y} \\ \partial \ell / \partial \hat{z} \end{bmatrix}$$

$$\begin{bmatrix} \partial \ell / \partial u_1 \\ \partial \ell / \partial u_2 \\ \partial \ell / \partial u_3 \end{bmatrix} = \begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix} \begin{bmatrix} \partial \ell / \partial \hat{y} \\ \partial \ell / \partial \hat{z} \end{bmatrix}$$

Sample: $(3, -2) \to (1, 4)$

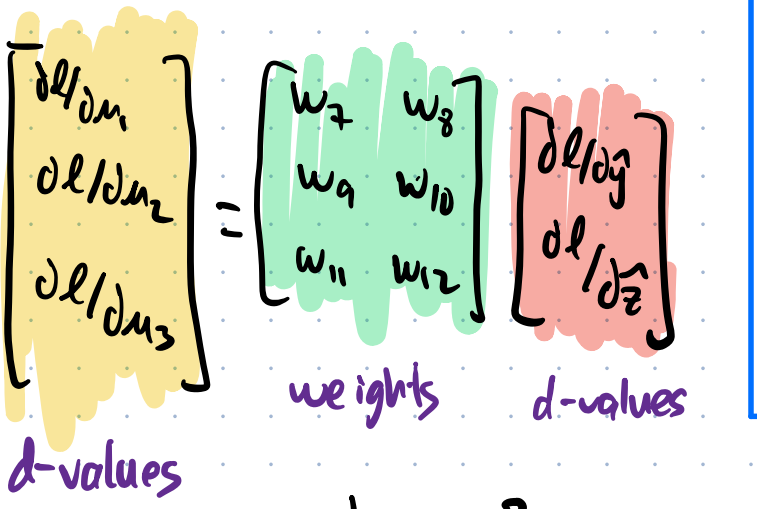
Work backward from the right.

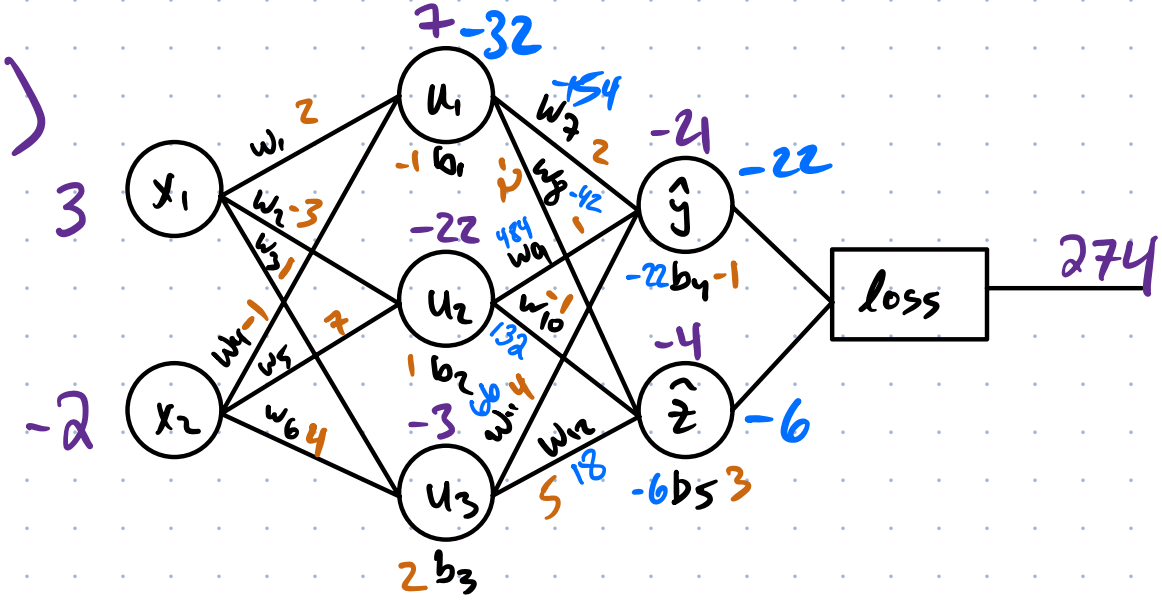


$$\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix} = \begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix} \begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix}$$

The vector of derivs. of the neuron values comes from a matrix operation between the weights to the next layer and the vector of derivs of the next layers neurons

Sample: $(3, -2) \rightarrow (1, 4)$





biases?

$b_1 \rightarrow u_1 \xrightarrow{w_7} \hat{y} \rightarrow \text{loss}$

$b_1 \rightarrow u_1 \xrightarrow{w_8} \hat{z} \rightarrow \text{loss}$

but don't have to do this much work!

because b_1 directly only affects u_1

and now we know $\frac{\partial l}{\partial u_1}$