Scientific Computing

Friday, April 24

Announcements

* Homework 6 due next Monday, April 27

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

On Monday I'll give you a list of topics.

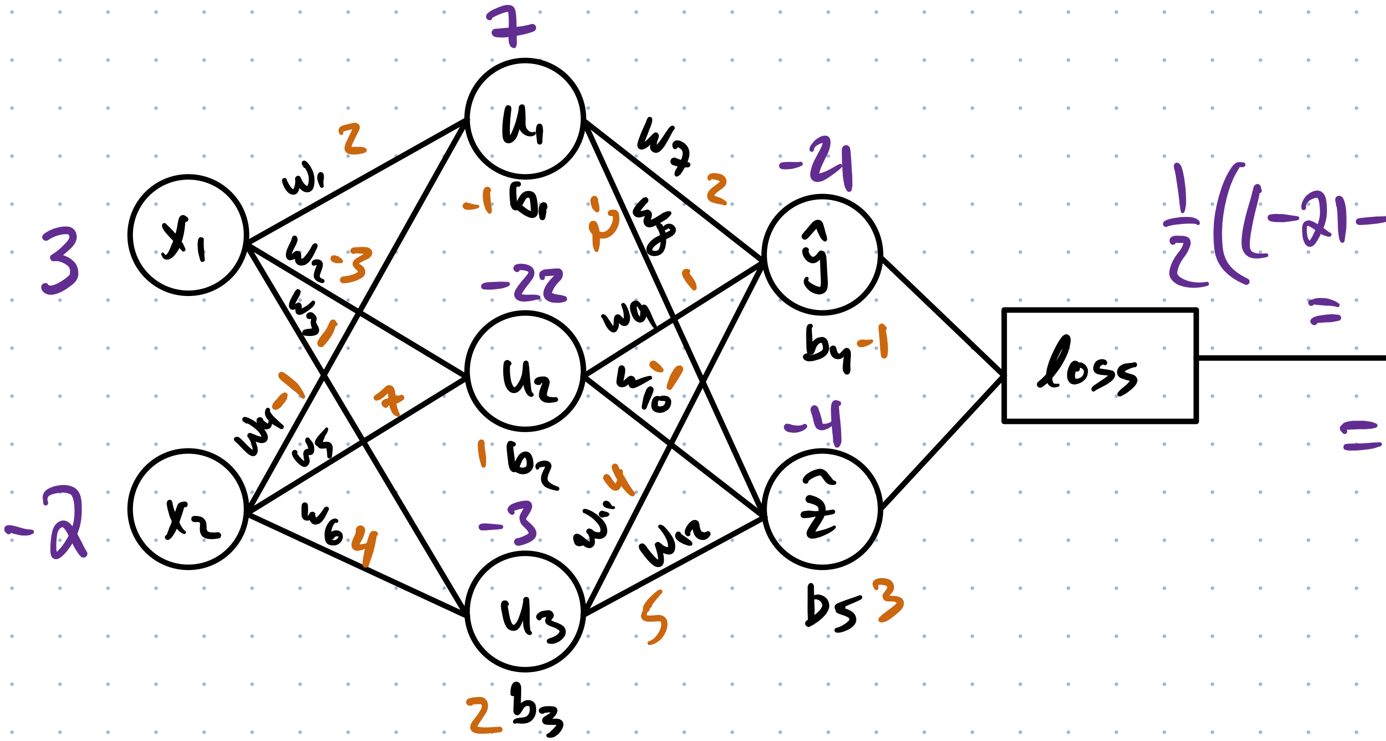
Office Hours canceled today. Email if Qs.

Office Hours:

Mon, 9:30-10:30
Fri, 2:00-3:00

Cudahy 307

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

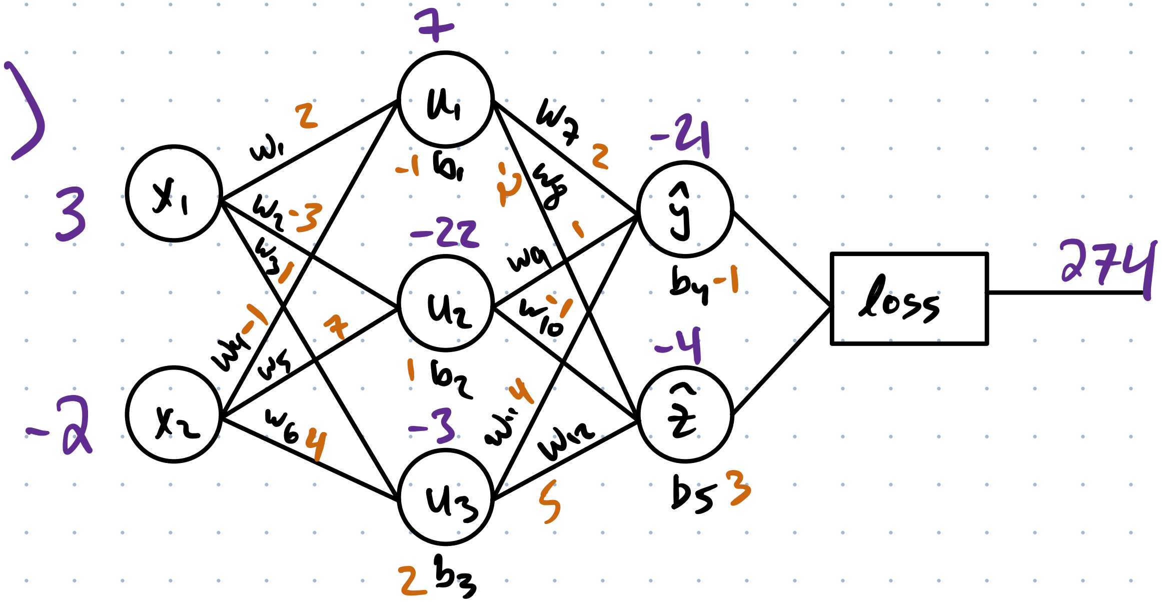


$$\frac{1}{2}\left((-21-1)^2+(-4-(4))^2\right)$$
$$=\frac{1}{2}\left((-22)^2+(-8)^2\right)$$
$$=\underline{274}$$

Backprop: Find:

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

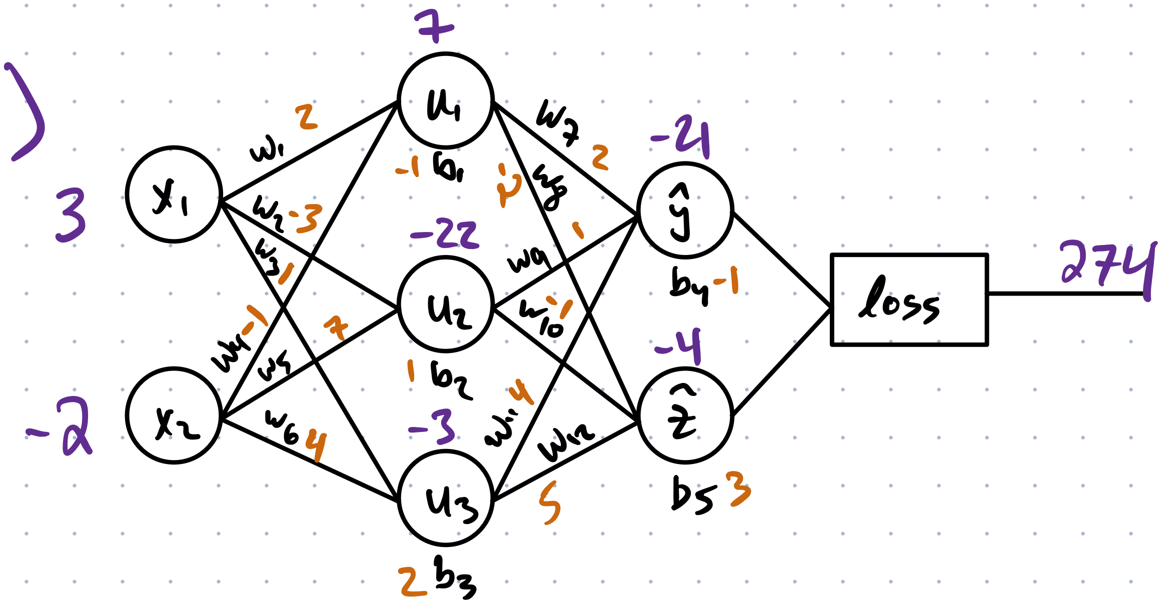
Sample: $(3,-2) \to (1,4)$



Work backward from the right.

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

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



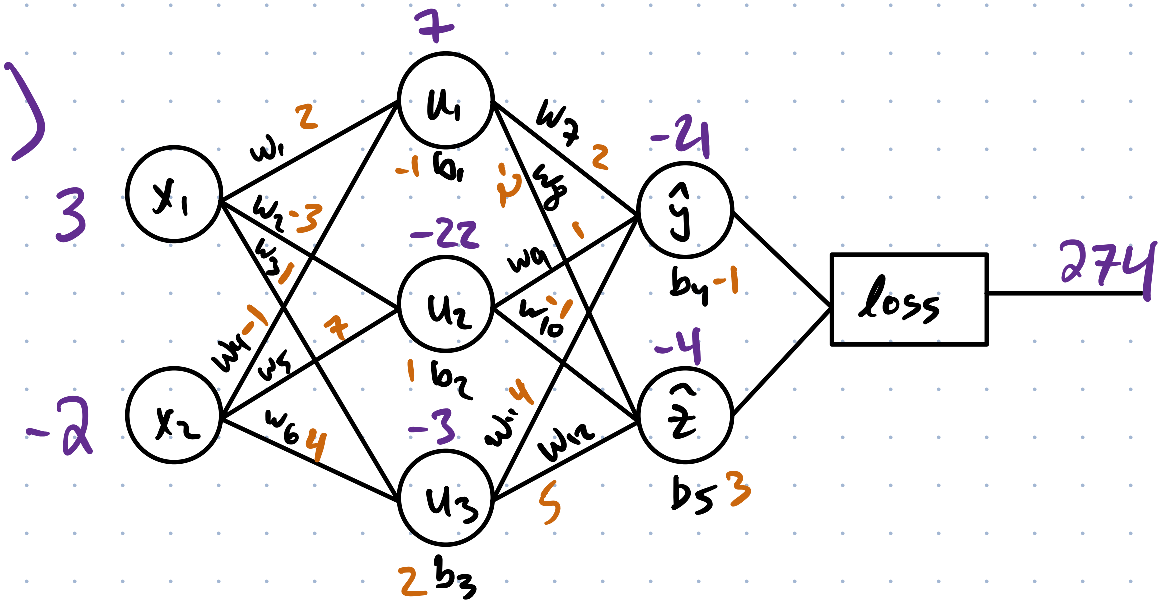
Work backward from the right.

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

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

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

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



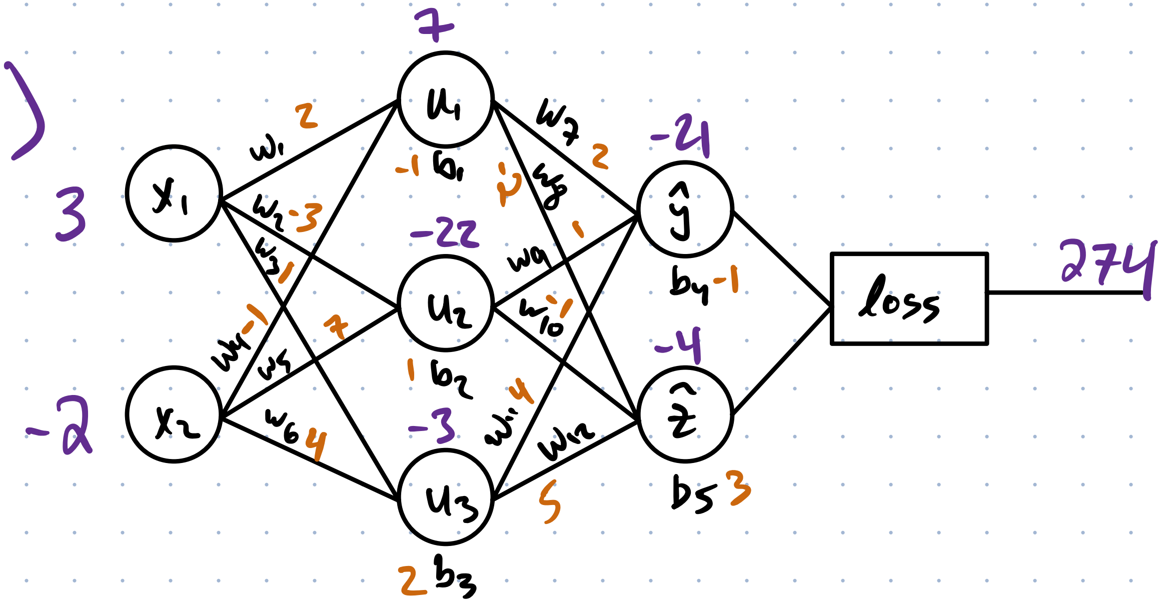
Work backward from the right.

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

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

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

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



Work backward from the right.

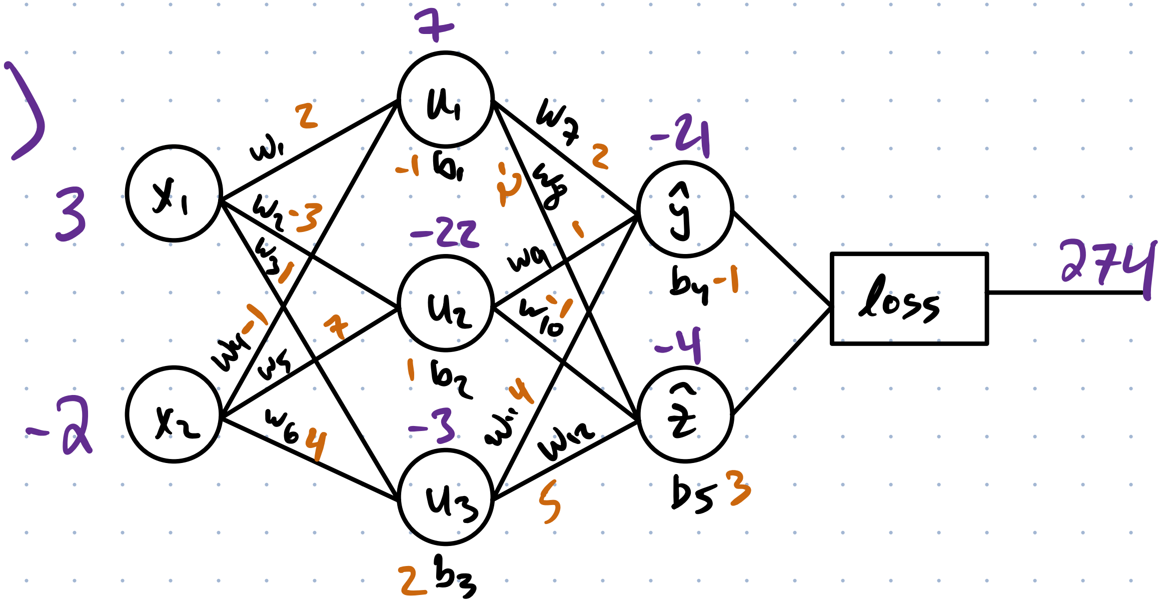
$$\frac{\partial l}{\partial \hat{y}} = -22 \qquad \frac{\partial l}{\partial \hat{z}} = -6$$

$$\frac{\partial l}{\partial w_7} = \frac{\partial l}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_7} \quad ?$$

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

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

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



Work backward from the right.

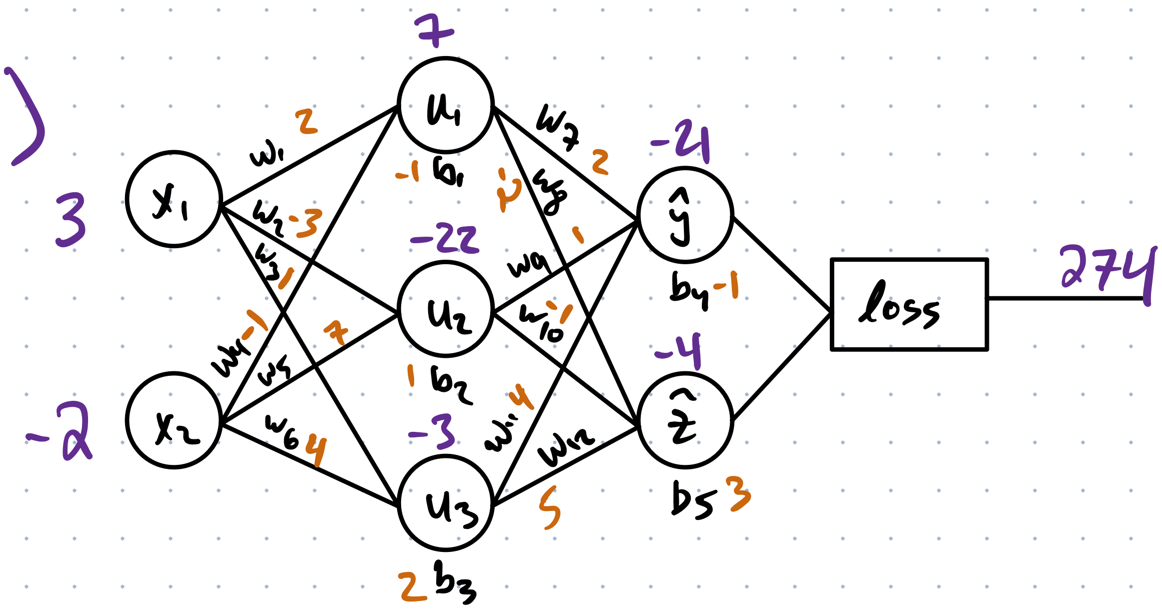
$$\frac{\partial l}{\partial \hat{y}} = -22 \qquad \frac{\partial l}{\partial \hat{z}} = -6$$

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

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

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

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



Work backward from the right.

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

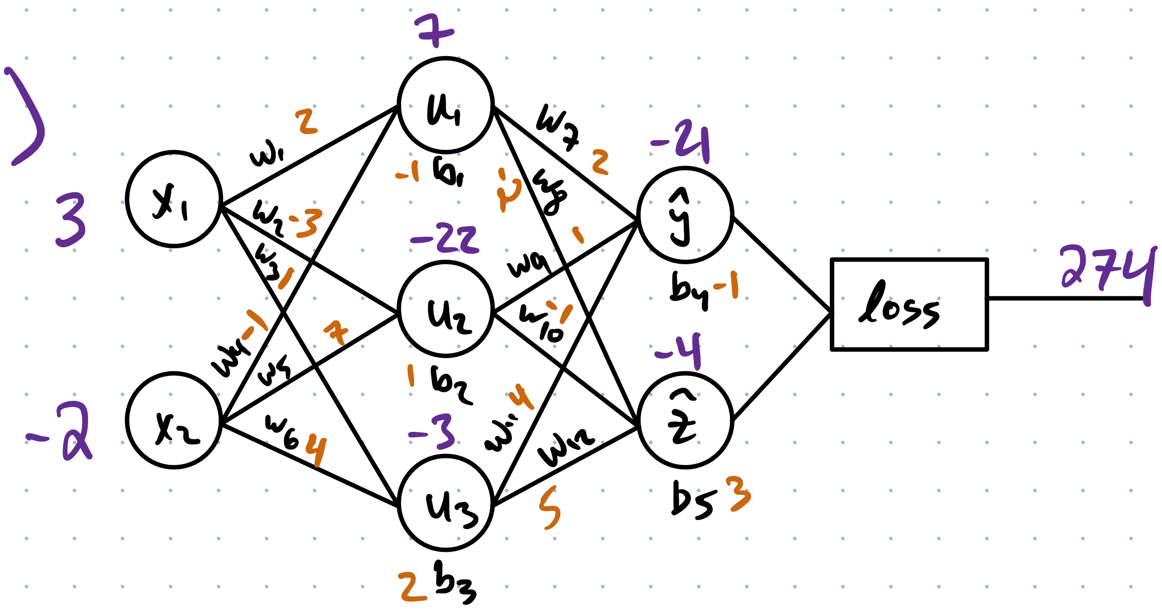
$$\frac{\partial l}{\partial w_7} = -154$$

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

$$\frac{\partial l}{\partial w_9} = (-22)(-22) = 484 \qquad \frac{\partial l}{\partial w_{10}} = (-6)(-22) = 132$$

$$\frac{\partial l}{\partial w_{11}} = (-22)(-3) = 66 \qquad \frac{\partial l}{\partial w_{12}} = (-6)(-3) = 18$$

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



Work backward from the right.

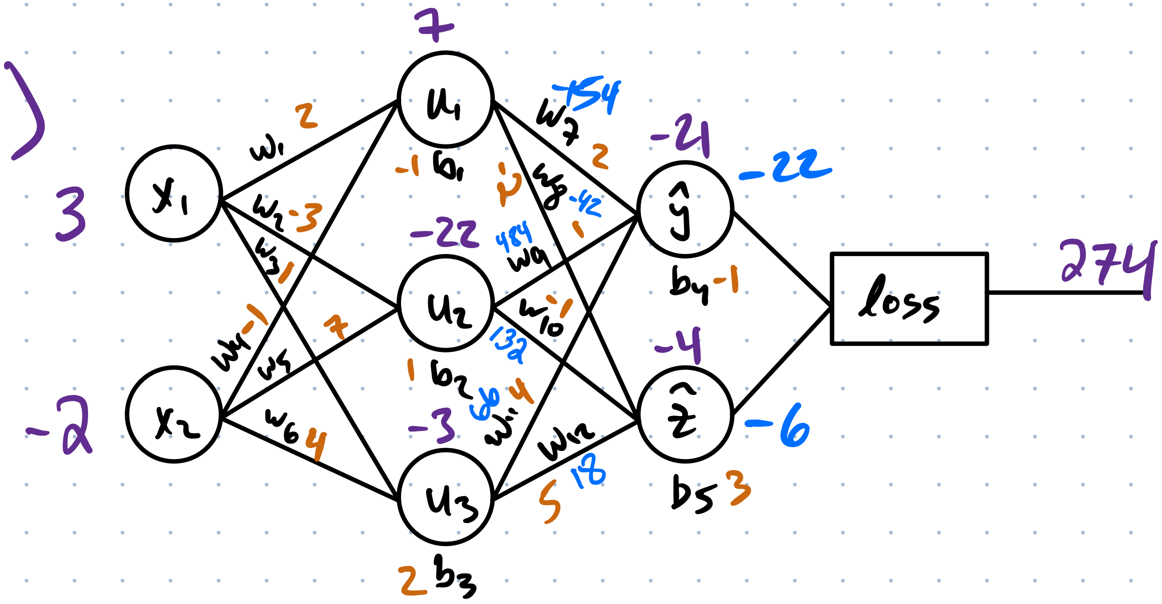
$$\frac{\partial l}{\partial \hat{y}} = -22 \qquad \frac{\partial l}{\partial \hat{z}} = -6$$

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

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

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

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



Work backward from the right.

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

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

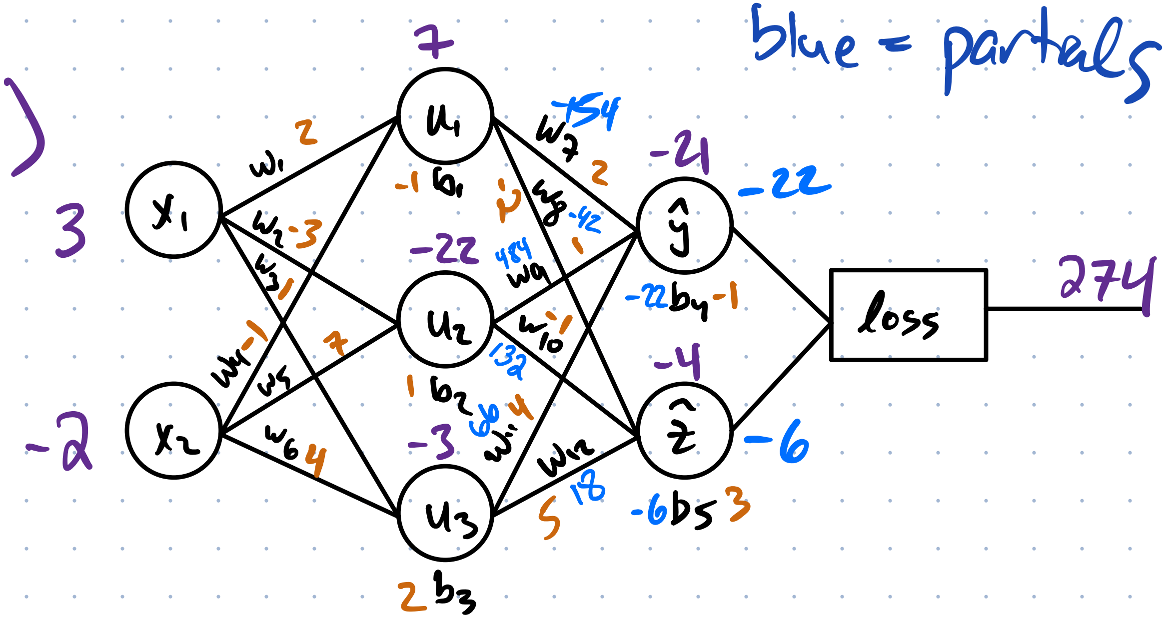
$$\frac{\partial l}{\partial w_9} = 484 \qquad \frac{\partial l}{\partial w_{10}} = 132$$

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

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

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

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



Work backward from the right.

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

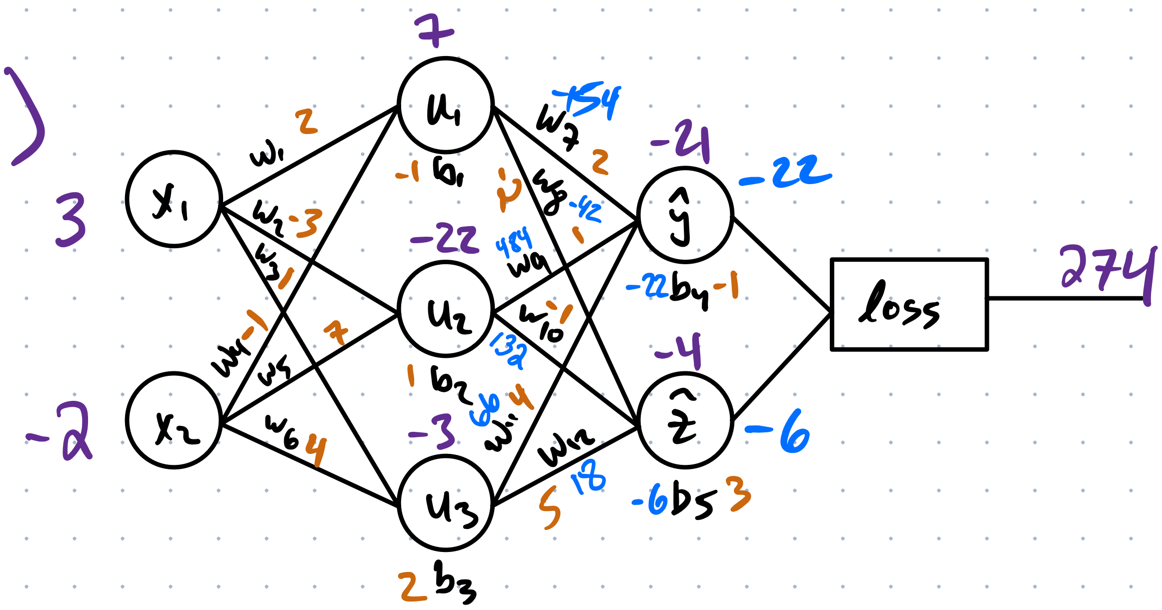
$\frac{\partial l}{\partial b_4} = -22$ $\frac{\partial l}{\partial b_5} = -6$

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

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

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

Sample: (3, -2) → (1, 4)



Work backward from the right.

Next:

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

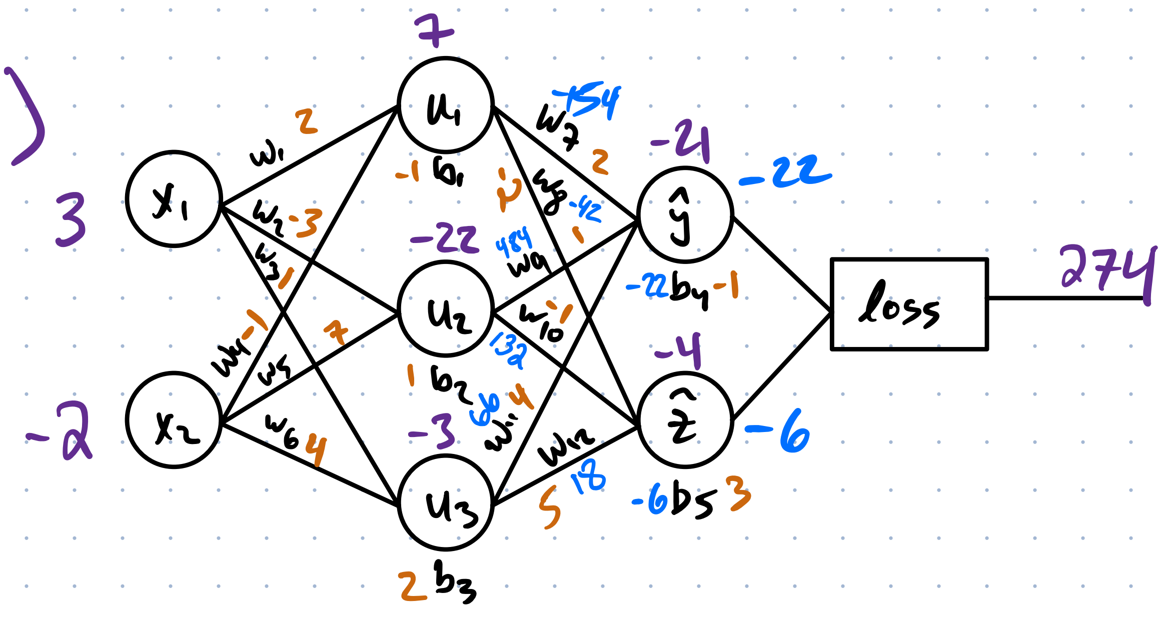
$$u_1 = w_1 x_1 + w_4 x_2 + b_1$$

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

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

Work backward from the right.

Next:



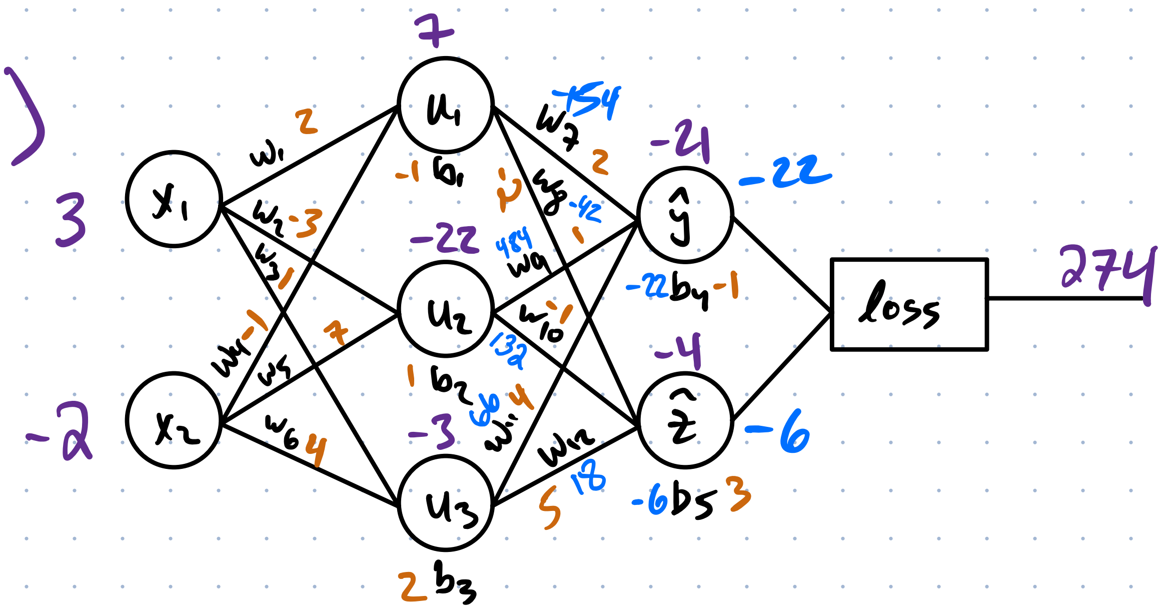
$$\frac{\partial \ell}{\partial w_1} = \frac{\partial \ell}{\partial u_1} \cdot \frac{\partial u_1}{\partial w_1}$$

not computed yet

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

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

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



Work backward from the right.

<u>Next:</u>

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

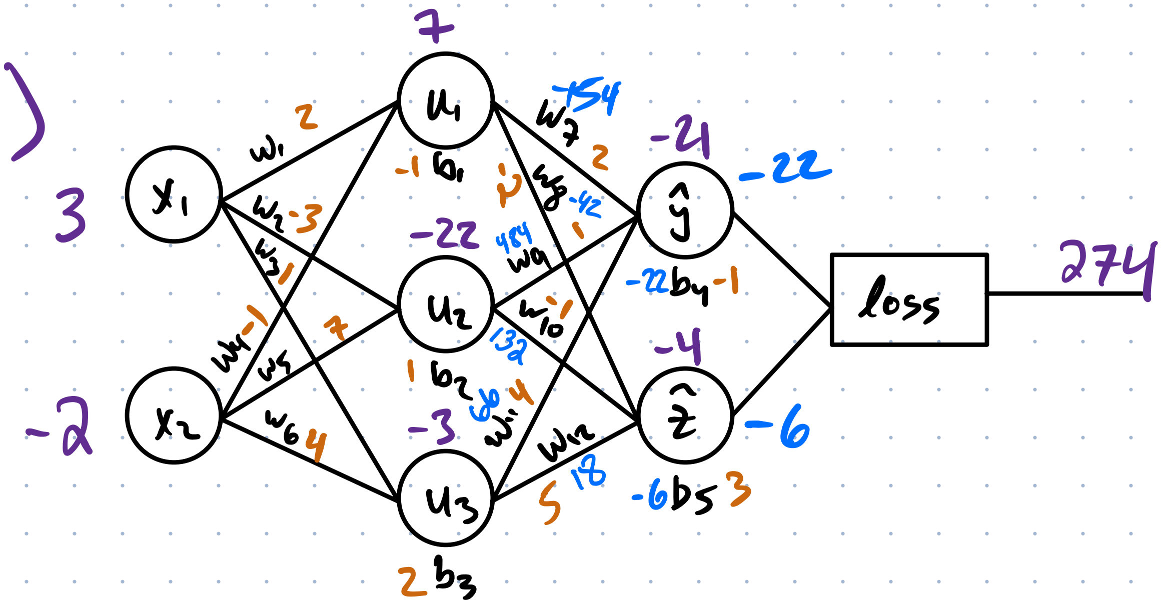
not computed yet

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

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

Now we see that with hidden layers, we also need derivs of the <u>neuron values</u> to keep passing backward.

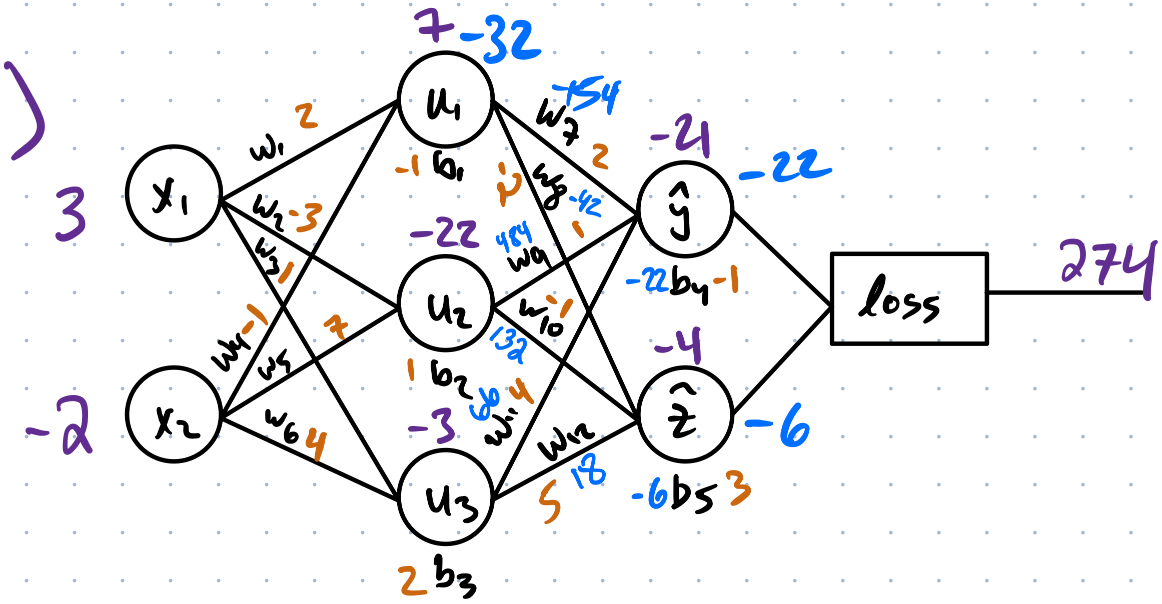
Sample: (3, −2) → (1, 4)



Work backward from the right.

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

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:

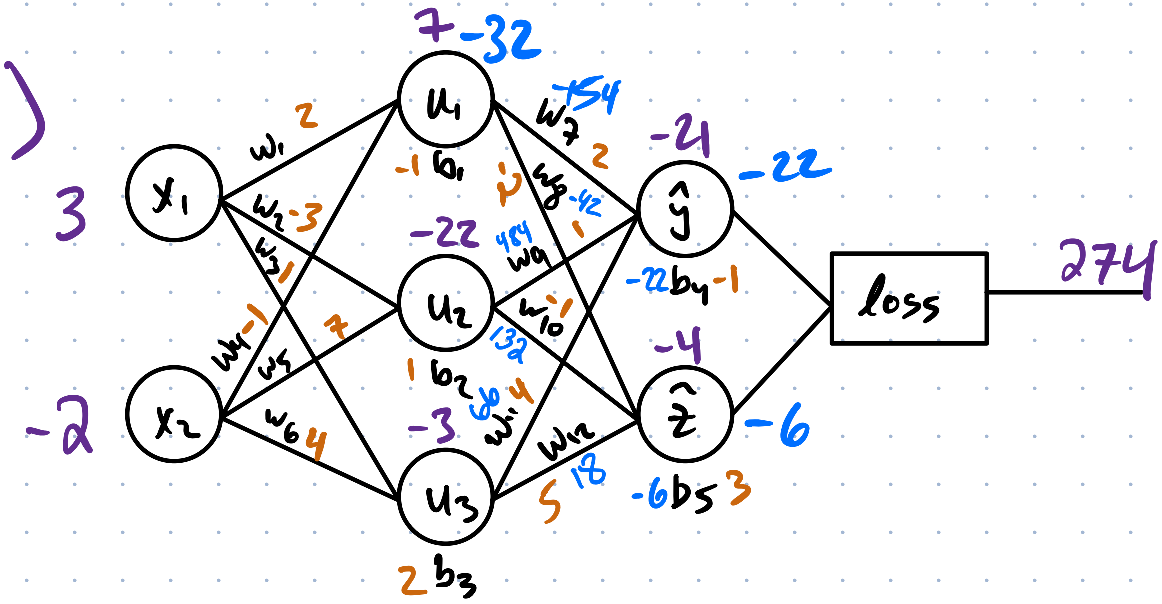
$$u_1 \xrightarrow[w_7]{} \hat{y} \to \ell$$

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

So we have the multivariate chain 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.

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

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

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

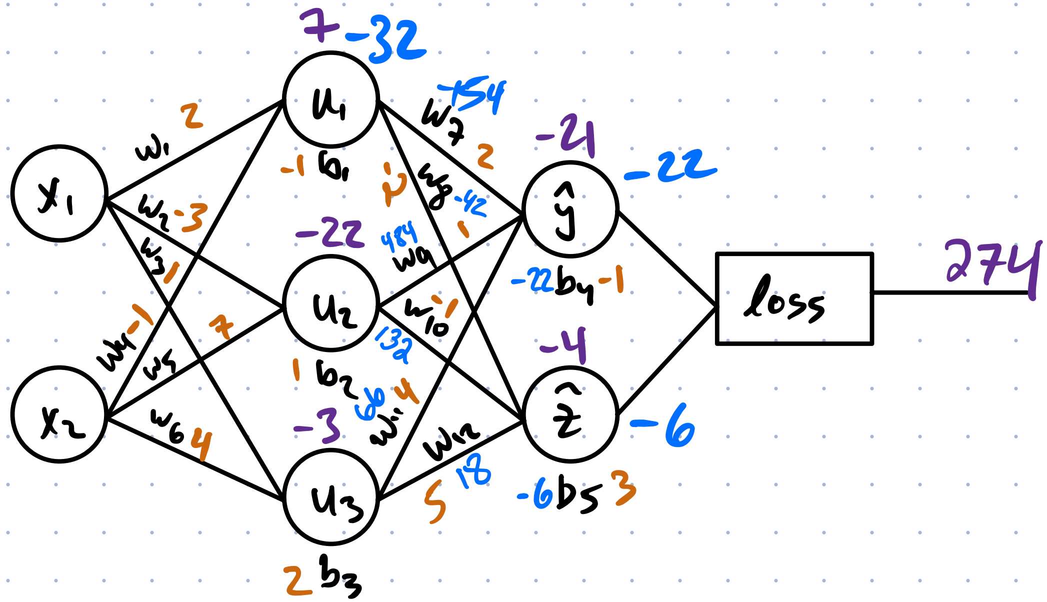
So we have the multivariate chain rule now:

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

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

Sample: (3, -2) → (1, 4)

Work backward from the right.

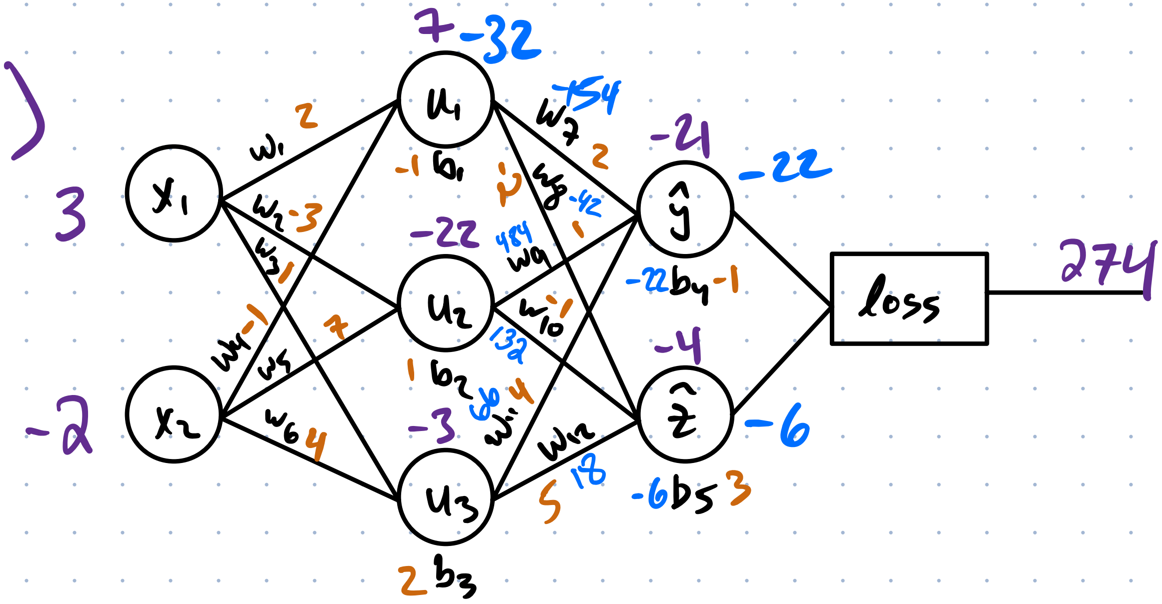


$$\frac{\partial l}{\partial u_1} = w_7 \cdot \frac{\partial l}{\partial \hat{y}} + w_9 \cdot \frac{\partial l}{\partial \hat{z}} = \begin{bmatrix} w_7 \\ w_8 \end{bmatrix} \cdot \begin{bmatrix} \partial l / \partial \hat{y} \\ \partial l / \partial \hat{z} \end{bmatrix}$$

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

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

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



Work backward from the right.

$$\frac{\partial l}{\partial u_1} = w_7 \cdot \frac{\partial l}{\partial \hat{y}} + w_9 \cdot \frac{\partial l}{\partial \hat{z}} = \begin{bmatrix} w_7 \\ w_8 \end{bmatrix} \cdot \begin{bmatrix} \partial l / \partial \hat{y} \\ \partial l / \partial \hat{z} \end{bmatrix}$$

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

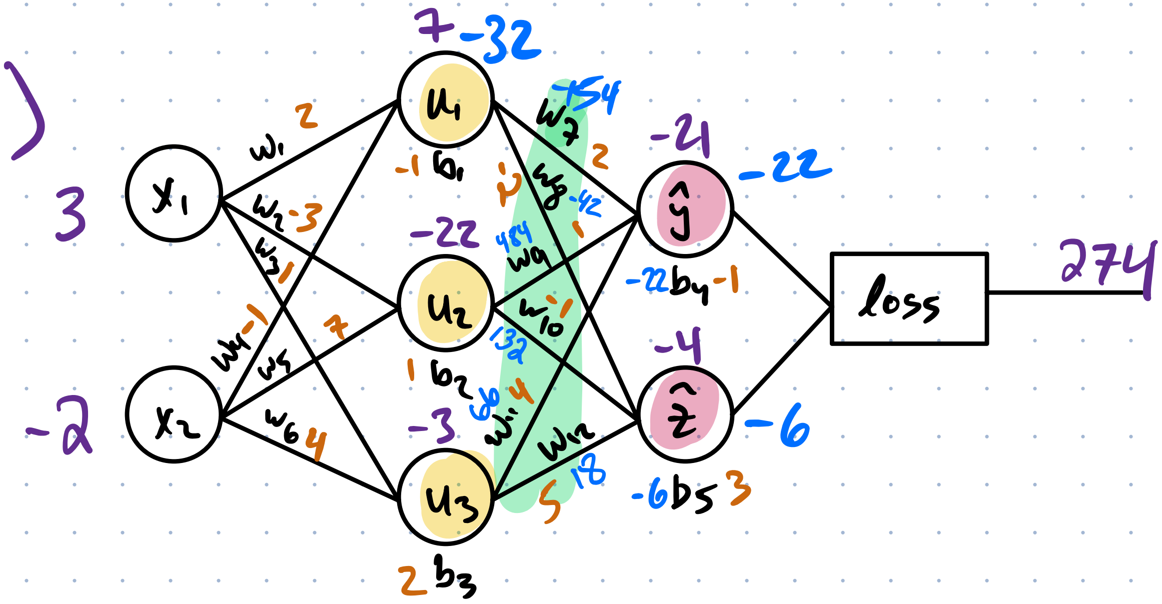
$$\frac{\partial l}{\partial u_3} = \begin{bmatrix} w_{11} \\ w_{12} \end{bmatrix} \begin{bmatrix} \partial l / \partial \hat{y} \\ \partial l / \partial \hat{z} \end{bmatrix}$$

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

d-values = weights · d-values

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

Work backward from the right.

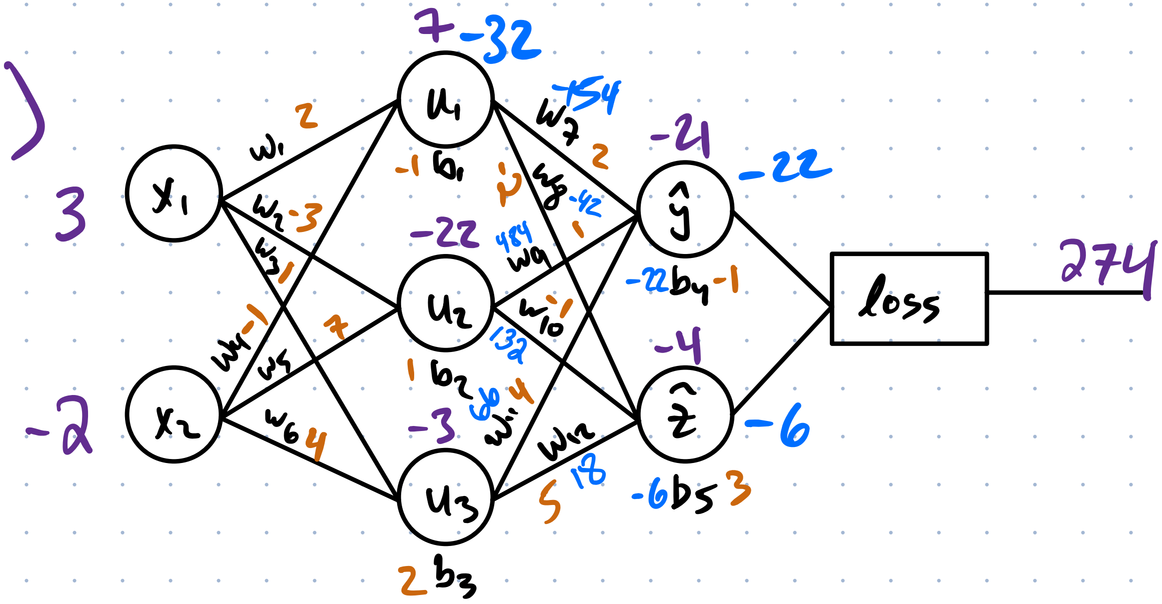


$$\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix} = \begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix} \begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix}$$

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

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

$$\underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$



biases?

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

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

but don't have to do this much work!

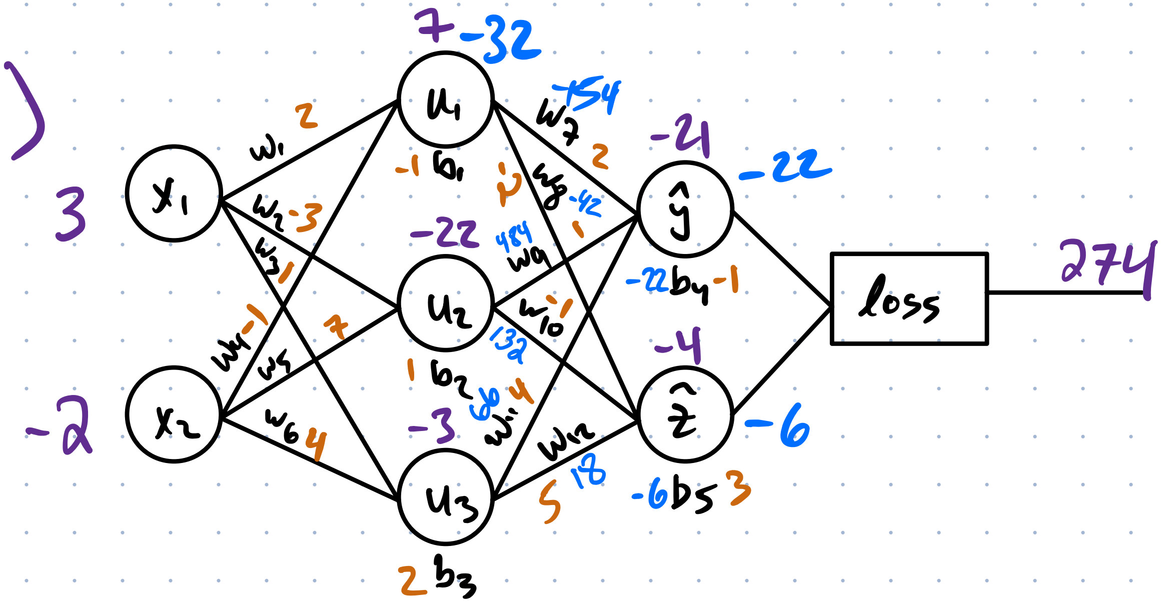
because b_1 directly only affects u_1

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

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

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

d-values   weights   d-values



biases?

$$\frac{\partial l}{\partial b_1} = \frac{\partial l}{\partial u_1} \cdot \frac{\partial u_1}{\partial b_1}$$

$$= \frac{\partial l}{\partial u_1}$$

$u_1 = \cdots + b$

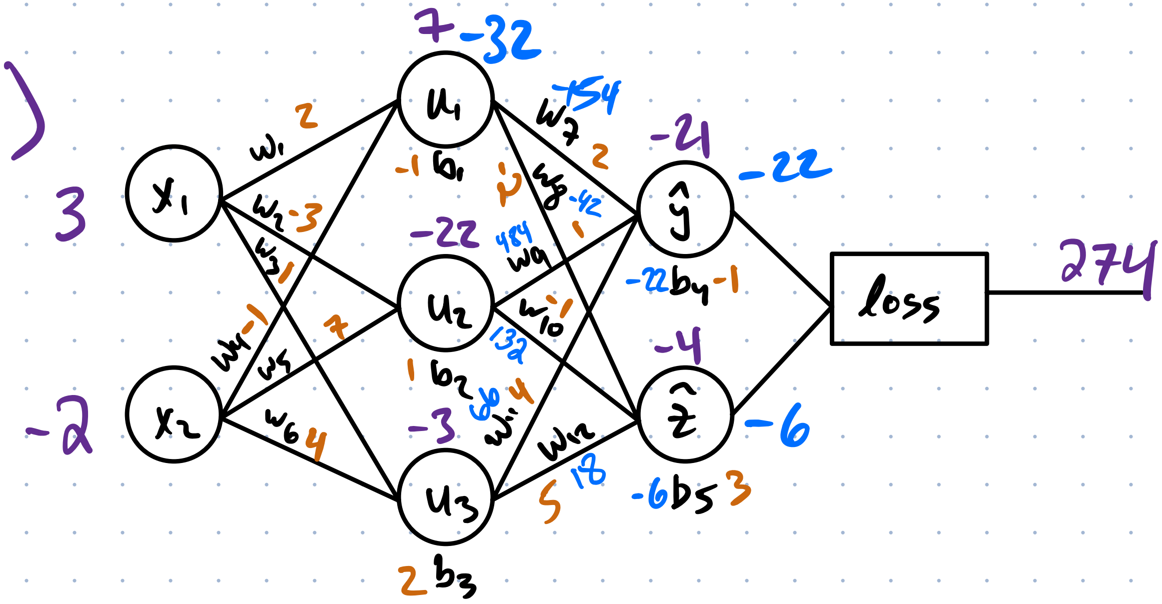
so, $\frac{du_1}{db} = 1$

So
$$\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix} = \begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}$$

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

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

d-values   weights   d-values



biases?

$$\frac{\partial l}{\partial b_1} = \frac{\partial l}{\partial u_1} \cdot \frac{\partial u_1}{\partial b} = \frac{\partial l}{\partial u_1}$$

$u_1 = \dots + b$

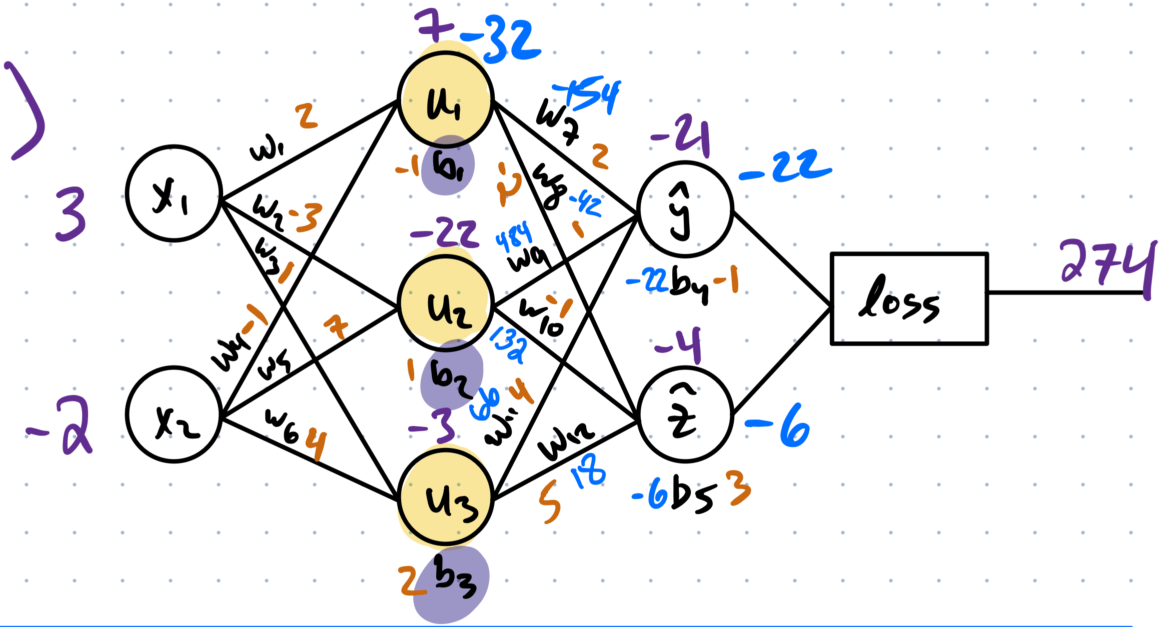
so, $\frac{du_1}{db} = 1$

So $$\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix} = \begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}$$

Sample: (3, -2) → (1, 4)

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

d-values   weights   d-values

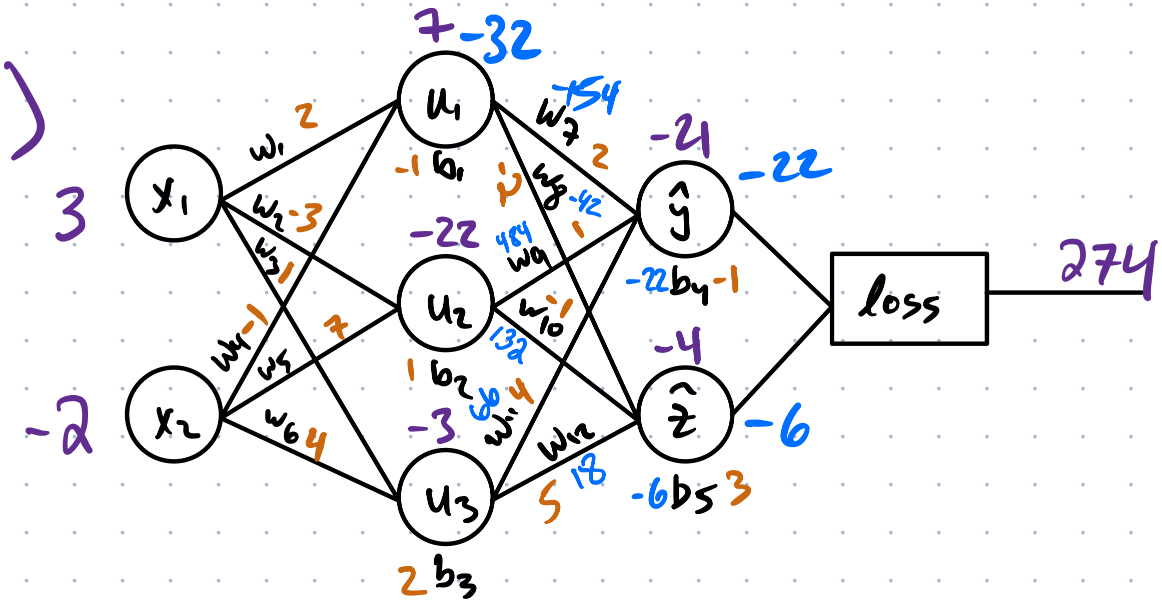


$$\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix} = \begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}$$

d-biases   d-values

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

$$\underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$

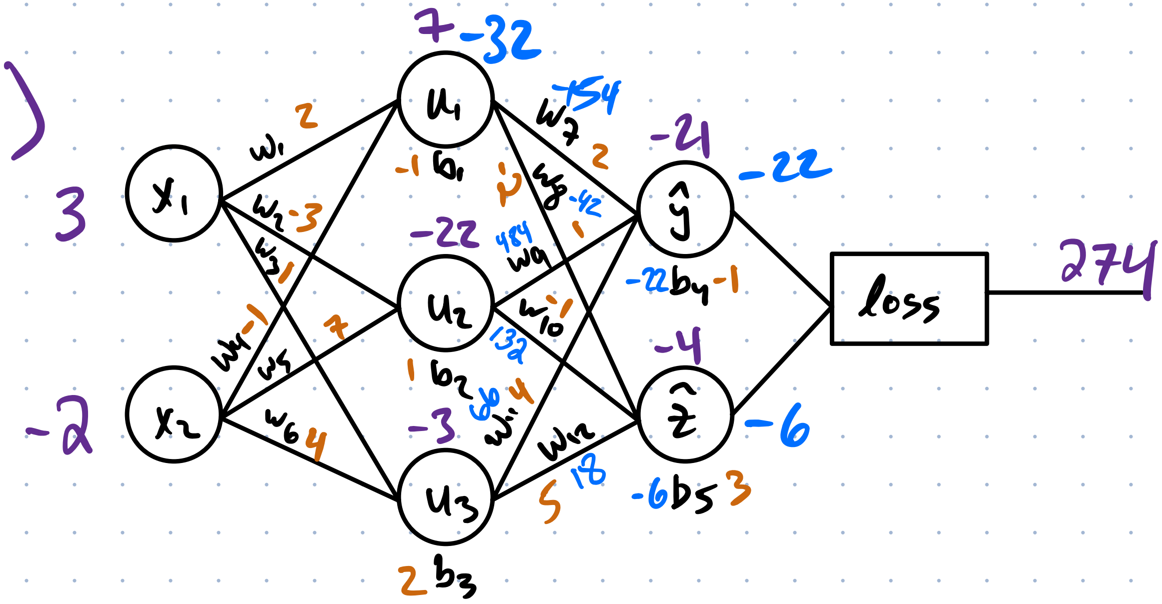


$$\underbrace{\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix}}_{\text{d-biases}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}}$$

Weights?

$$\frac{\partial l}{\partial w_1} = \frac{\partial l}{\partial u_1} \cdot \frac{\partial u_1}{\partial w_1} = x_1$$

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



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

d-values weights d-values

$$\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix} = \begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}$$

d-biases d-values

Weights?

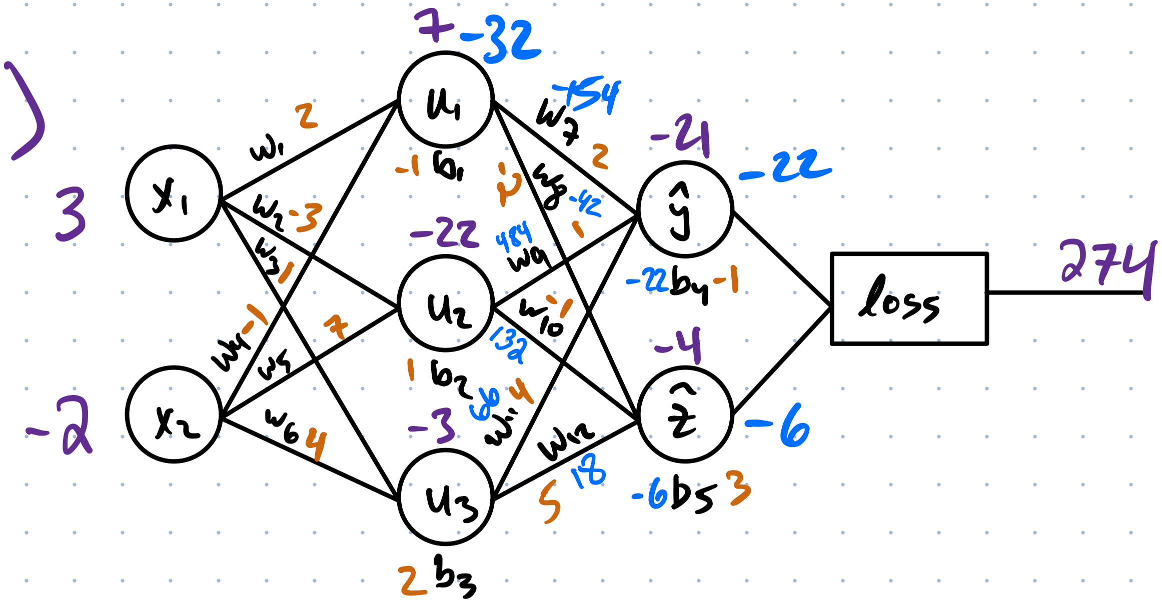
$$\frac{\partial l}{\partial w_1} = \frac{\partial l}{\partial u_1} \cdot \frac{\partial u_1}{\partial w_1} = x_1$$

$$\frac{\partial l}{\partial w_2} = \frac{\partial l}{\partial u_2} \cdot \frac{\partial u_2}{\partial w_2} = x_1$$

$$\frac{\partial l}{\partial w_3} = \frac{\partial l}{\partial u_3} \cdot \frac{\partial u_3}{\partial w_3} = x_1$$

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

$$\underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$



$$\underbrace{\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix}}_{\text{d-biases}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}}$$

Weights?

$$\frac{\partial l}{\partial w_1} = \frac{\partial l}{\partial u_1} \cdot \frac{\partial u_1}{\partial w_1} = x_1$$

$$\frac{\partial l}{\partial w_2} = \frac{\partial l}{\partial u_2} \cdot \frac{\partial u_2}{\partial w_2} = x_1$$

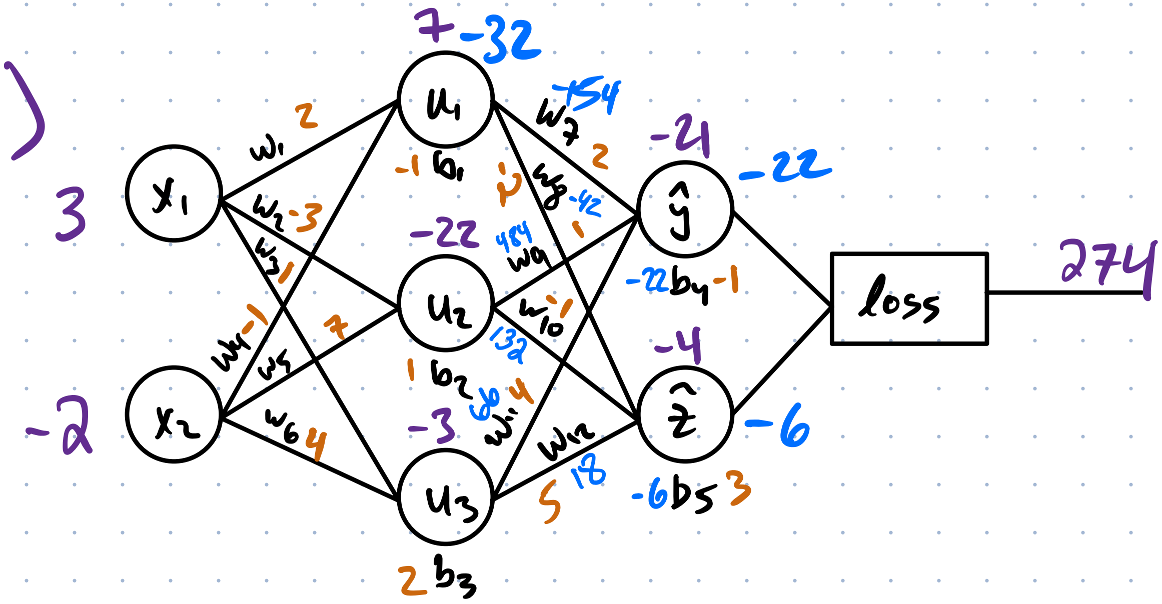
$$\frac{\partial l}{\partial w_3} = \frac{\partial l}{\partial u_3} \cdot \frac{\partial u_3}{\partial w_3} = x_1$$

$$\frac{\partial l}{\partial w_4} = \frac{\partial l}{\partial u_1} \cdot x_2$$

$$\frac{\partial l}{\partial w_5} = \frac{\partial l}{\partial u_2} \cdot x_2$$

$$\frac{\partial l}{\partial w_6} = \frac{\partial l}{\partial u_3} \cdot x_2$$

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



$$\underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix}}_{\text{d-biases}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}}$$

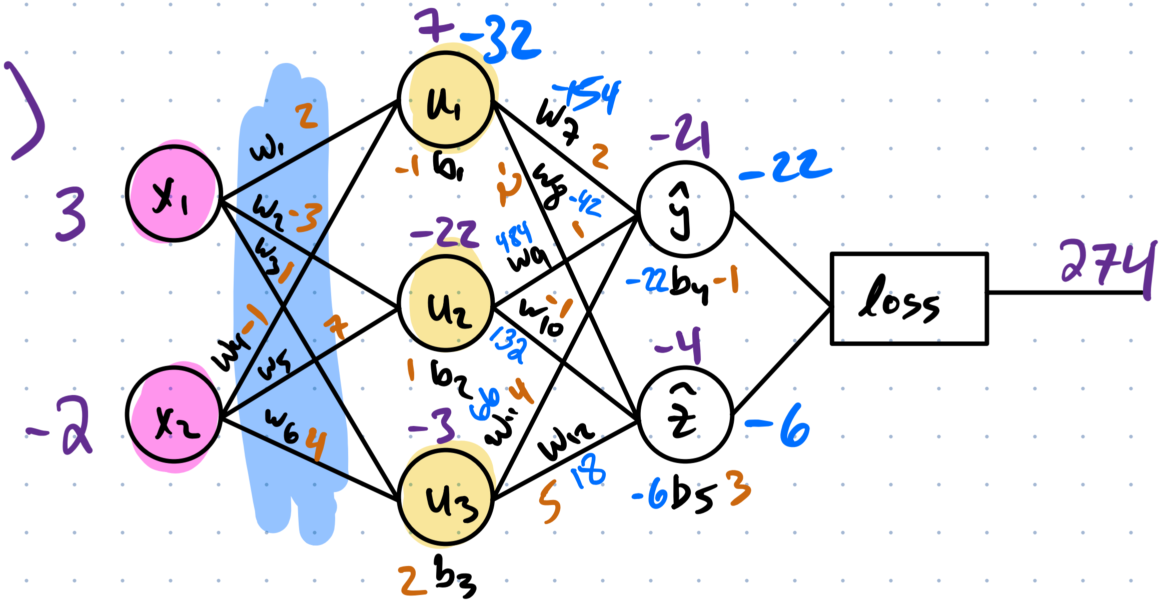
$$\frac{\partial l}{\partial w_1} = \frac{\partial l}{\partial u_1} \cdot \frac{\partial u_1}{\partial w_1} = x_1 \qquad \frac{\partial l}{\partial w_4} = \frac{\partial l}{\partial u_1} \cdot x_2$$

$$\frac{\partial l}{\partial w_2} = \frac{\partial l}{\partial u_2} \cdot \frac{\partial u_2}{\partial w_2} = x_1 \qquad \frac{\partial l}{\partial w_5} = \frac{\partial l}{\partial u_2} \cdot x_2$$

$$\frac{\partial l}{\partial w_3} = \frac{\partial l}{\partial u_3} \cdot \frac{\partial u_3}{\partial w_3} = x_1 \qquad \frac{\partial l}{\partial w_6} = \frac{\partial l}{\partial u_3} \cdot x_2$$

$$\begin{bmatrix} \partial l/\partial w_1 & \partial l/\partial w_4 \\ \partial l/\partial w_2 & \partial l/\partial w_5 \\ \partial l/\partial w_3 & \partial l/\partial w_6 \end{bmatrix} = \begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix}$$

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



$$\underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix}}_{\text{d-biases}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial w_1 & \partial l/\partial w_4 \\ \partial l/\partial w_2 & \partial l/\partial w_5 \\ \partial l/\partial w_3 & \partial l/\partial w_6 \end{bmatrix}}_{\text{d-weights}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} \underbrace{\begin{bmatrix} x_1 & x_2 \end{bmatrix}}_{\text{values}}$$

$$\underbrace{\begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial L/\partial \hat{y} \\ \partial L/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$

$$\underbrace{\begin{bmatrix} \partial L/\partial w_1 & \partial L/\partial w_4 \\ \partial L/\partial w_2 & \partial L/\partial w_5 \\ \partial L/\partial w_3 & \partial L/\partial w_6 \end{bmatrix}}_{\text{d-weights}} = \underbrace{\begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix}}_{\text{d-values}} \underbrace{\begin{bmatrix} x_1 & x_2 \end{bmatrix}}_{\text{values}}$$

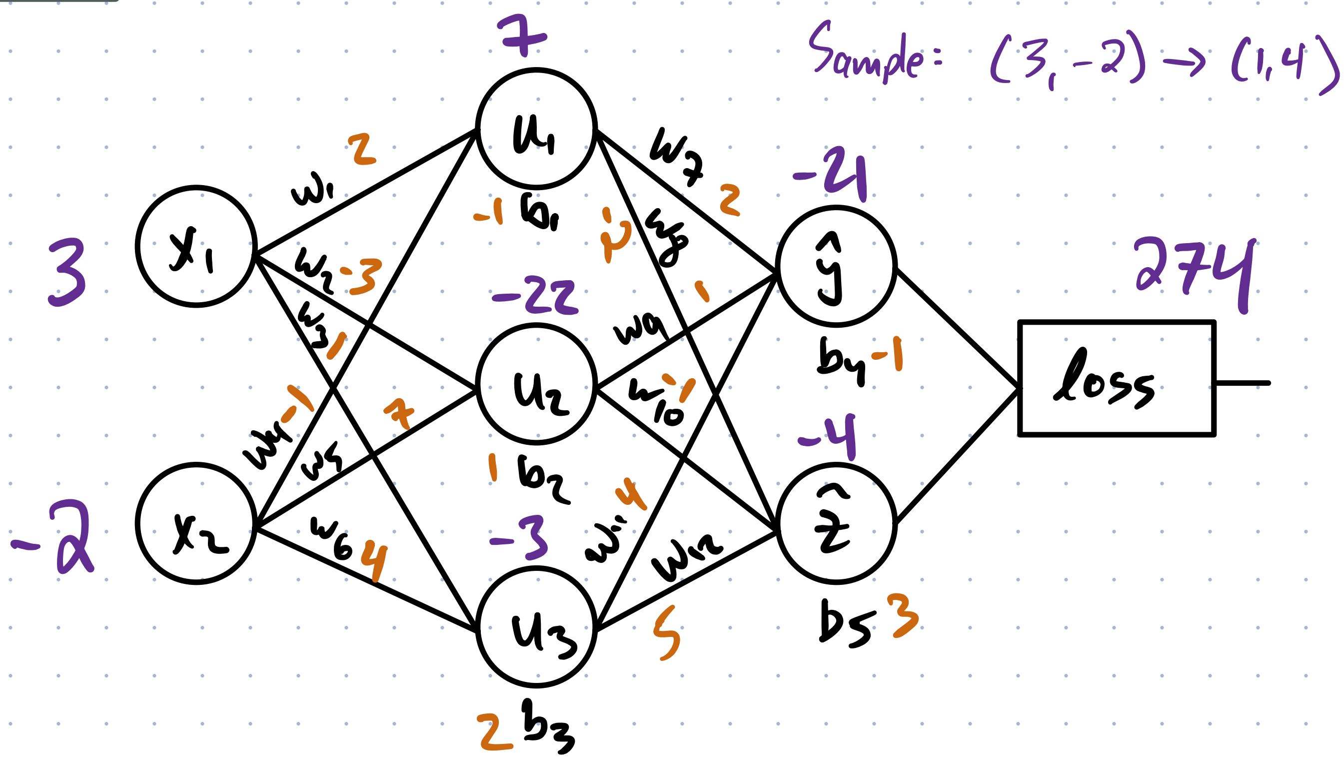
$$\underbrace{\begin{bmatrix} \partial L/\partial b_1 \\ \partial L/\partial b_2 \\ \partial L/\partial b_3 \end{bmatrix}}_{\text{d-biases}} = \underbrace{\begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix}}_{\text{d-values}}$$

These three rules tell how to compute the derivative of the weights, biases, and neuron values based on the weights, neuron values, and derivative of the neuron values from the layer ahead.

$$\underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial w_1 & \partial l/\partial w_4 \\ \partial l/\partial w_2 & \partial l/\partial w_5 \\ \partial l/\partial w_3 & \partial l/\partial w_6 \end{bmatrix}}_{\text{d-weights}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} \underbrace{\begin{bmatrix} x_1 & x_2 \end{bmatrix}}_{\text{values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix}}_{\text{d-biases}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}}$$

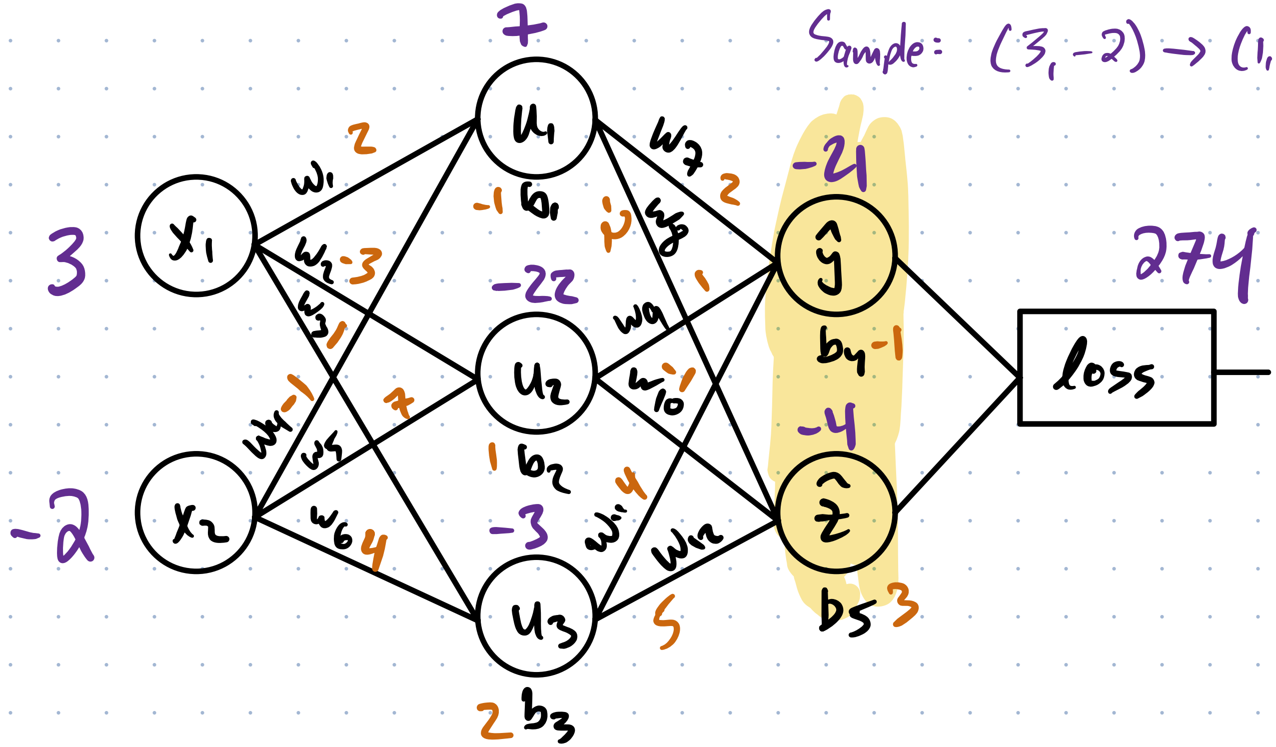


$$\underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial w_1 & \partial l/\partial w_4 \\ \partial l/\partial w_2 & \partial l/\partial w_5 \\ \partial l/\partial w_3 & \partial l/\partial w_6 \end{bmatrix}}_{\text{d-weights}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} \underbrace{\begin{bmatrix} x_1 & x_2 \end{bmatrix}}_{\text{values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix}}_{\text{d-biases}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}}$$

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

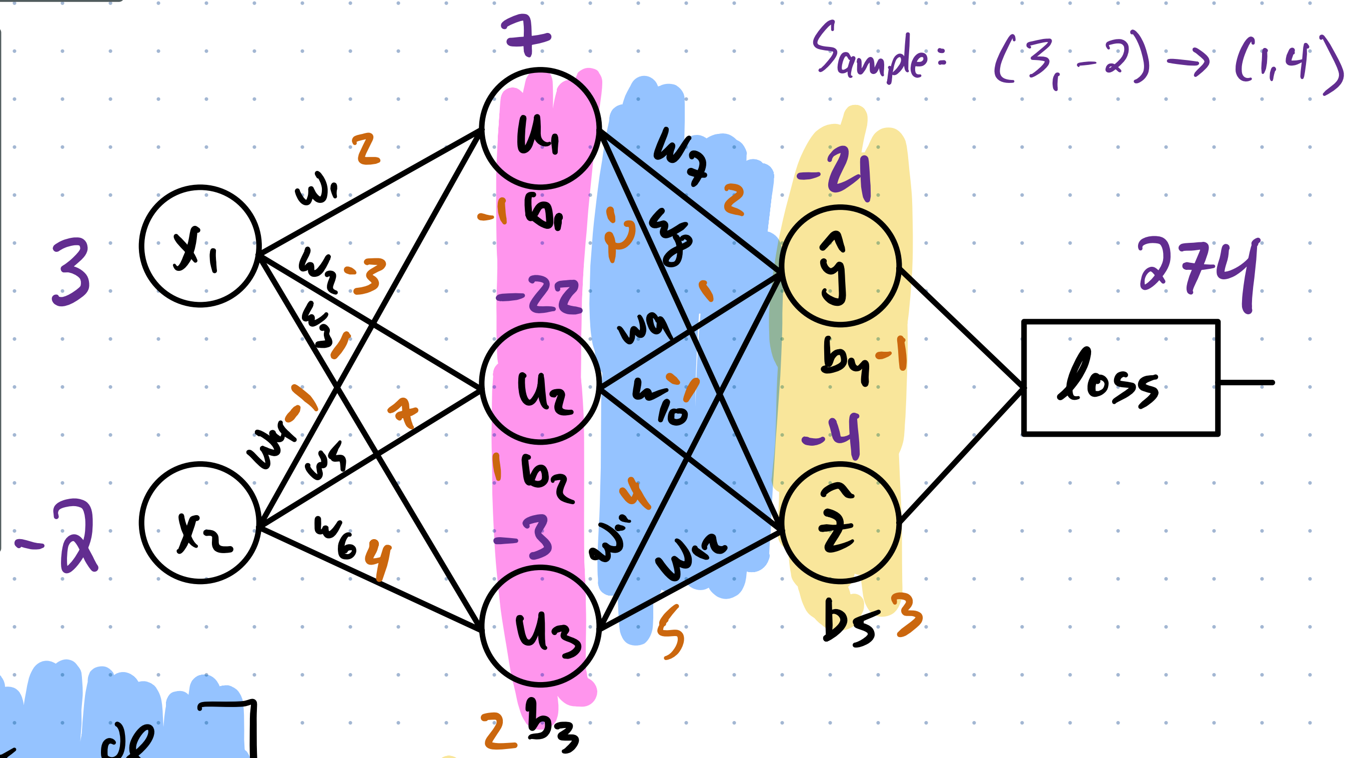


$$\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial L/\partial \hat{y} \\ \partial L/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$

$$\underbrace{\begin{bmatrix} \partial L/\partial w_1 & \partial L/\partial w_4 \\ \partial L/\partial w_2 & \partial L/\partial w_5 \\ \partial L/\partial w_3 & \partial L/\partial w_6 \end{bmatrix}}_{\text{d-weights}} = \underbrace{\begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix}}_{\text{d-values}} \underbrace{\begin{bmatrix} x_1 & x_2 \end{bmatrix}}_{\text{values}}$$

$$\underbrace{\begin{bmatrix} \partial L/\partial b_1 \\ \partial L/\partial b_2 \\ \partial L/\partial b_3 \end{bmatrix}}_{\text{d-biases}} = \underbrace{\begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix}}_{\text{d-values}}$$



$$\begin{bmatrix} \partial L/\partial \hat{y} \\ \partial L/\partial \hat{z} \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix}$$

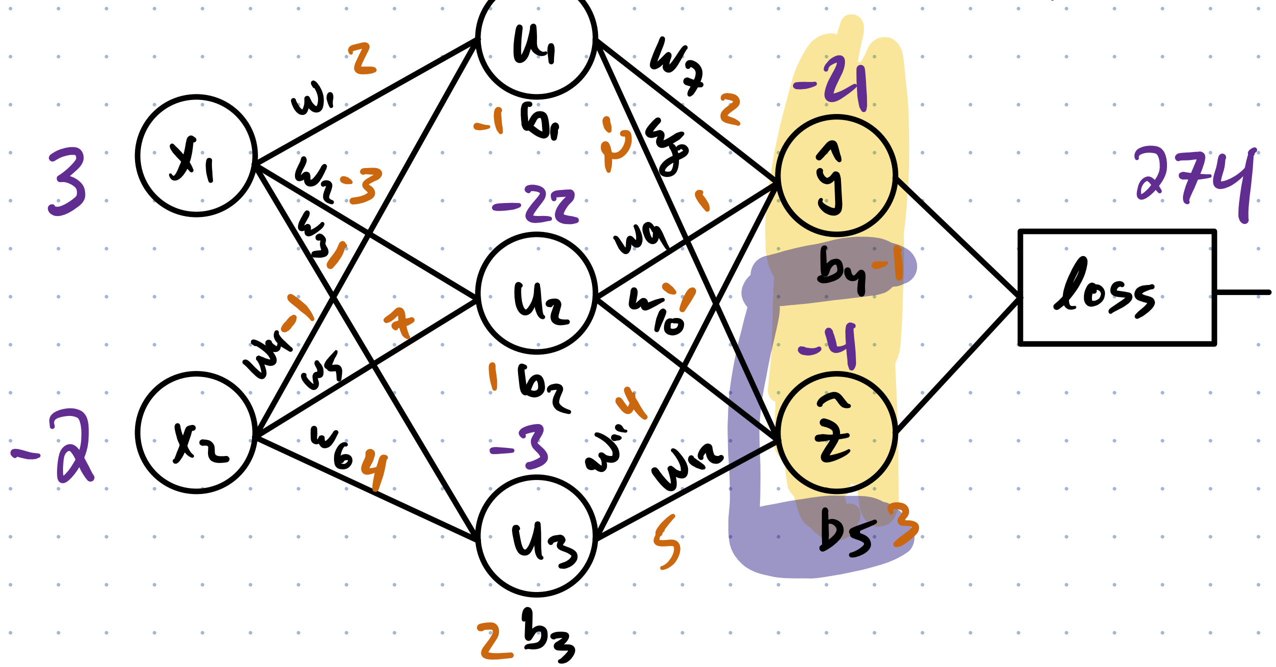
$$\begin{bmatrix} \frac{\partial L}{\partial w_7} & \frac{\partial L}{\partial w_9} & \frac{\partial L}{\partial w_{11}} \\ \partial L/w_8 & \partial L/\partial w_{10} & \partial L/\partial w_{12} \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix} \begin{bmatrix} 7 & -22 & -3 \end{bmatrix} = \begin{bmatrix} -154 & 484 & 66 \\ -42 & 132 & 18 \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial w_1 & \partial l/\partial w_4 \\ \partial l/\partial w_2 & \partial l/\partial w_5 \\ \partial l/\partial w_3 & \partial l/\partial w_6 \end{bmatrix}}_{\text{d-weights}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} \underbrace{\begin{bmatrix} x_1 & x_2 \end{bmatrix}}_{\text{values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix}}_{\text{d-biases}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}}$$

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



$$\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix}$$

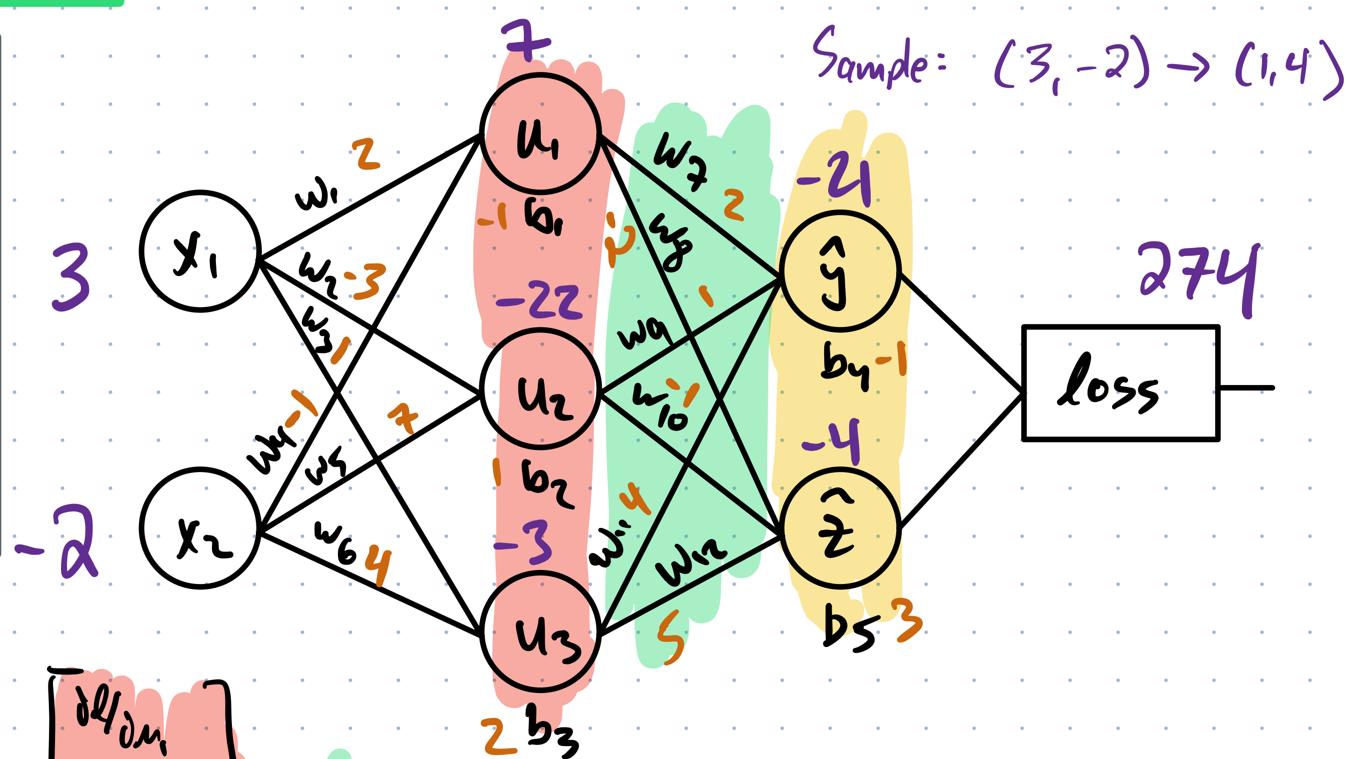
$$\begin{bmatrix} \frac{\partial l}{\partial w_7} & \frac{\partial l}{\partial w_9} & \frac{\partial l}{\partial w_{11}} \\ \partial l/\partial w_8 & \partial l/\partial w_{10} & \partial l/\partial w_{12} \end{bmatrix} = \begin{bmatrix} -154 & 484 & 66 \\ -42 & 132 & 18 \end{bmatrix}$$

$$\begin{bmatrix} \partial l/\partial b_4 \\ \partial l/\partial b_5 \end{bmatrix} = \begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial w_1 & \partial l/\partial w_4 \\ \partial l/\partial w_2 & \partial l/\partial w_5 \\ \partial l/\partial w_3 & \partial l/\partial w_6 \end{bmatrix}}_{\text{d-weights}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} \underbrace{\begin{bmatrix} x_1 & x_2 \end{bmatrix}}_{\text{values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix}}_{\text{d-biases}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}}$$



$$\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix} \qquad \begin{bmatrix} \partial l/\partial b_4 \\ \partial l/\partial b_5 \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix}$$

$$\begin{bmatrix} \partial l/\partial w_7 & \partial l/\partial w_9 & \partial l/\partial w_{11} \\ \partial l/\partial w_8 & \partial l/\partial w_{10} & \partial l/\partial w_{12} \end{bmatrix} = \begin{bmatrix} -154 & 484 & 66 \\ -42 & 132 & 18 \end{bmatrix}$$

$$\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix} = \begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix} \begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix} = \begin{bmatrix} 2 & -2 \\ 1 & -1 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} -22 \\ -6 \end{bmatrix} = \begin{bmatrix} -32 \\ -16 \\ -118 \end{bmatrix}$$

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

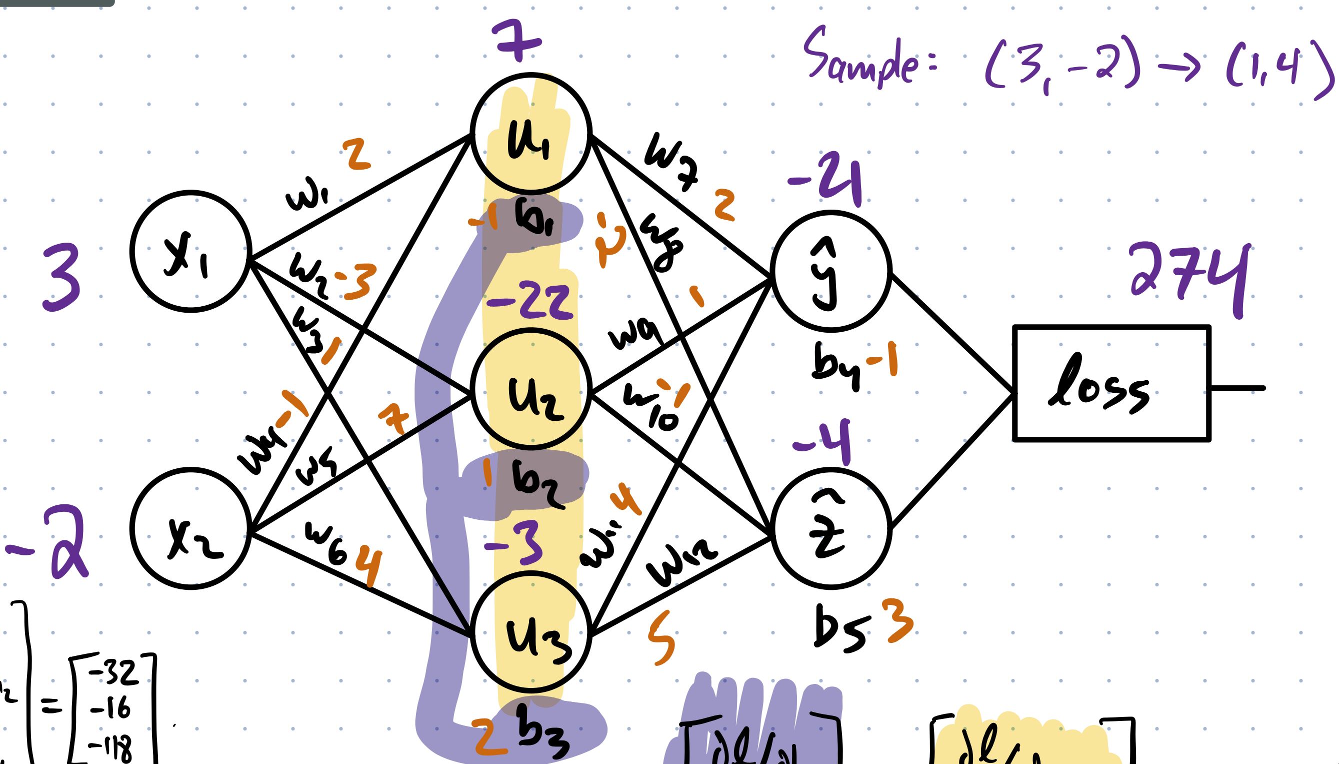
d-values   weights   d-values

$$\begin{bmatrix} \partial L/\partial w_1 & \partial L/\partial w_4 \\ \partial L/\partial w_2 & \partial L/\partial w_5 \\ \partial L/\partial w_3 & \partial L/\partial w_6 \end{bmatrix} = \begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix}$$

d-weights   d-values   values

$$\begin{bmatrix} \partial L/\partial b_1 \\ \partial L/\partial b_2 \\ \partial L/\partial b_3 \end{bmatrix} = \begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix}$$

d-biases   d-values



$$\begin{bmatrix} \partial L/\partial \hat{y} \\ \partial L/\partial \hat{z} \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix} \qquad \begin{bmatrix} \partial L/\partial b_4 \\ \partial L/\partial b_5 \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix} \qquad \begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix} = \begin{bmatrix} -32 \\ -16 \\ -118 \end{bmatrix}$$

$$\begin{bmatrix} \frac{\partial L}{\partial w_7} & \frac{\partial L}{\partial w_9} & \frac{\partial L}{\partial w_{11}} \\ \partial L/\partial w_8 & \partial L/\partial w_{10} & \partial L/\partial w_{12} \end{bmatrix} = \begin{bmatrix} -154 & 484 & 66 \\ -42 & 132 & 18 \end{bmatrix}$$

$$\begin{bmatrix} \partial L/\partial b_1 \\ \partial L/\partial b_2 \\ \partial L/\partial b_3 \end{bmatrix} = \begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix} = \begin{bmatrix} -32 \\ -16 \\ -118 \end{bmatrix}$$

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

d-values weights d-values

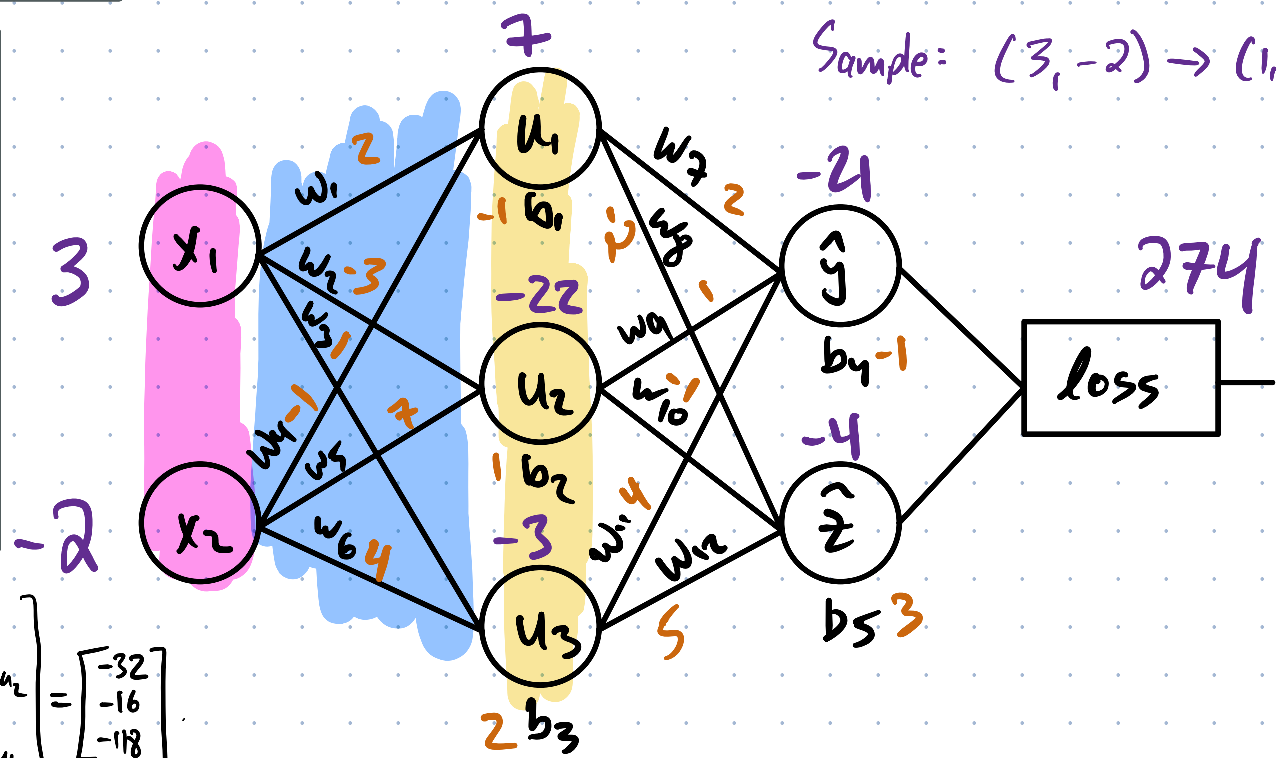
$$\begin{bmatrix} \partial L/\partial w_1 & \partial L/\partial w_4 \\ \partial L/\partial w_2 & \partial L/\partial w_5 \\ \partial L/\partial w_3 & \partial L/\partial w_6 \end{bmatrix} = \begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix}$$

d-weights d-values values

$$\begin{bmatrix} \partial L/\partial b_1 \\ \partial L/\partial b_2 \\ \partial L/\partial b_3 \end{bmatrix} = \begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix}$$

d-biases d-values

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



$$\begin{bmatrix} \partial L/\partial \hat{y} \\ \partial L/\partial \hat{z} \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix} \qquad \begin{bmatrix} \partial L/\partial b_4 \\ \partial L/\partial b_5 \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix} \qquad \begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix} = \begin{bmatrix} -32 \\ -16 \\ -118 \end{bmatrix}$$

$$\begin{bmatrix} \frac{\partial L}{\partial w_7} & \frac{\partial L}{\partial w_9} & \frac{\partial L}{\partial w_{11}} \\ \partial L/\partial w_8 & \partial L/\partial w_{10} & \partial L/\partial w_{12} \end{bmatrix} = \begin{bmatrix} -154 & 484 & 66 \\ -42 & 132 & 18 \end{bmatrix} \qquad \begin{bmatrix} \partial L/\partial b_1 \\ \partial L/\partial b_2 \\ \partial L/\partial b_3 \end{bmatrix} = \begin{bmatrix} -32 \\ -16 \\ -118 \end{bmatrix}$$

$$\begin{bmatrix} \partial L/\partial w_1 & \partial L/\partial w_4 \\ \partial L/\partial w_2 & \partial L/\partial w_5 \\ \partial L/\partial w_3 & \partial L/\partial w_6 \end{bmatrix} = \begin{bmatrix} -32 \\ -16 \\ -118 \end{bmatrix} \begin{bmatrix} 3 & -2 \end{bmatrix} = \begin{bmatrix} -96 & 64 \\ -48 & 32 \\ -354 & 236 \end{bmatrix}$$

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

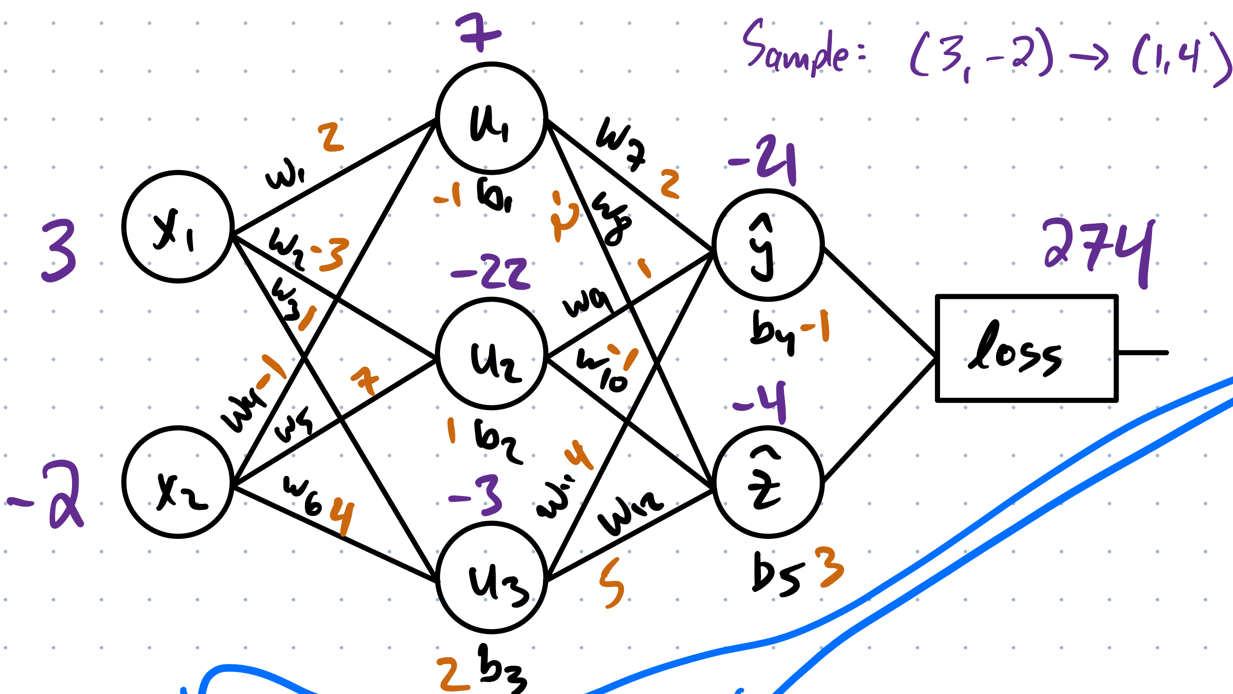
d-values weights d-values

$$\begin{bmatrix} \partial L/\partial w_1 & \partial L/\partial w_4 \\ \partial L/\partial w_2 & \partial L/\partial w_5 \\ \partial L/\partial w_3 & \partial L/\partial w_6 \end{bmatrix} = \begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix}$$

d-weights d-values values

$$\begin{bmatrix} \partial L/\partial b_1 \\ \partial L/\partial b_2 \\ \partial L/\partial b_3 \end{bmatrix} = \begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix}$$

d-biases d-values



And now backprop is done.

* We don't need $\frac{\partial L}{\partial x_1}$ or $\frac{\partial L}{\partial x_2}$.

* We don't need or anymore, they were just middle steps.

* Let's rearrange the rest.

$$\begin{bmatrix} \partial L/\partial \hat{y} \\ \partial L/\partial \hat{z} \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix} \qquad \begin{bmatrix} \partial L/\partial b_4 \\ \partial L/\partial b_5 \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix} \qquad \begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix} = \begin{bmatrix} -32 \\ -16 \\ -118 \end{bmatrix}$$

$$\begin{bmatrix} \frac{\partial L}{\partial w_7} & \frac{\partial L}{\partial w_9} & \frac{\partial L}{\partial w_{11}} \\ \partial L/\partial w_8 & \partial L/\partial w_{10} & \partial L/\partial w_{12} \end{bmatrix} = \begin{bmatrix} -154 & 484 & 66 \\ -42 & 132 & 18 \end{bmatrix} \qquad \begin{bmatrix} \partial L/\partial b_1 \\ \partial L/\partial b_2 \\ \partial L/\partial b_3 \end{bmatrix} = \begin{bmatrix} -32 \\ -16 \\ -118 \end{bmatrix}$$

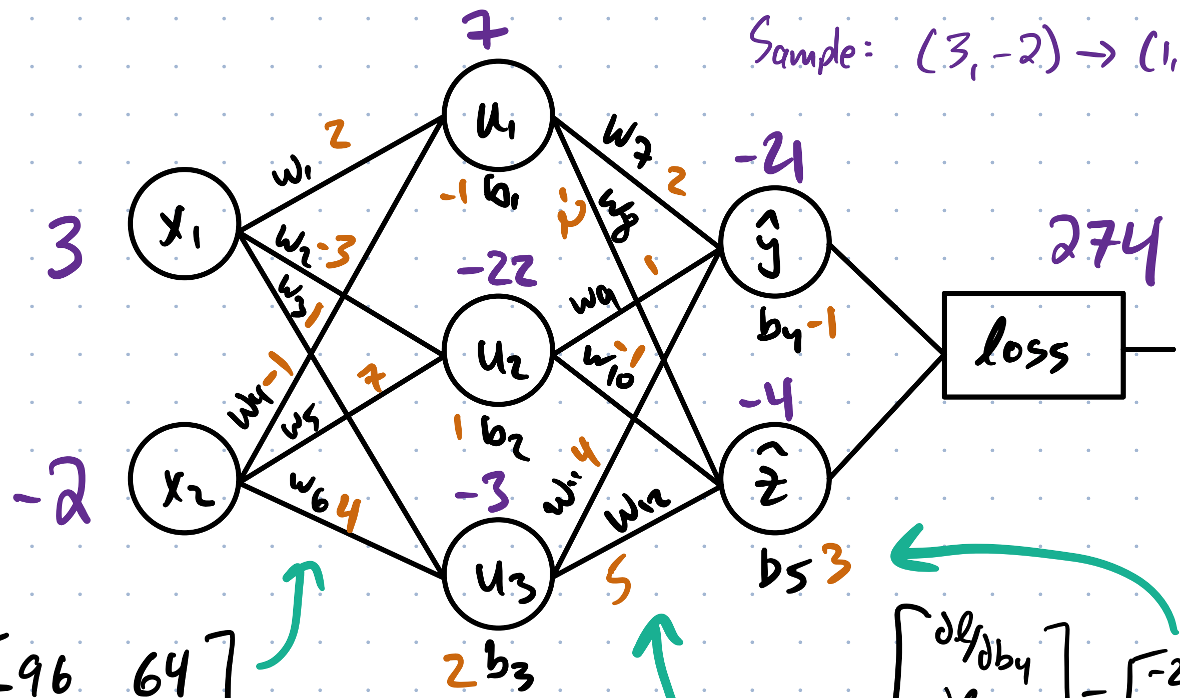
$$\begin{bmatrix} \partial L/\partial w_1 & \partial L/\partial w_4 \\ \partial L/\partial w_2 & \partial L/\partial w_5 \\ \partial L/\partial w_3 & \partial L/\partial w_6 \end{bmatrix} = \begin{bmatrix} -96 & 64 \\ -48 & 32 \\ -354 & 236 \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial w_1 & \partial l/\partial w_4 \\ \partial l/\partial w_2 & \partial l/\partial w_5 \\ \partial l/\partial w_3 & \partial l/\partial w_6 \end{bmatrix}}_{\text{d-weights}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} \underbrace{\begin{bmatrix} x_1 & x_2 \end{bmatrix}}_{\text{values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix}}_{\text{d-biases}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}}$$

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



$$\begin{bmatrix} \partial l/\partial w_1 & \partial l/\partial w_4 \\ \partial l/\partial w_2 & \partial l/\partial w_5 \\ \partial l/\partial w_3 & \partial l/\partial w_6 \end{bmatrix} = \begin{bmatrix} -96 & 64 \\ -48 & 32 \\ -354 & 236 \end{bmatrix}$$

$$\begin{bmatrix} \partial l/\partial b_4 \\ \partial l/\partial b_5 \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix}$$

$$\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix} = \begin{bmatrix} -32 \\ -16 \\ -118 \end{bmatrix}$$

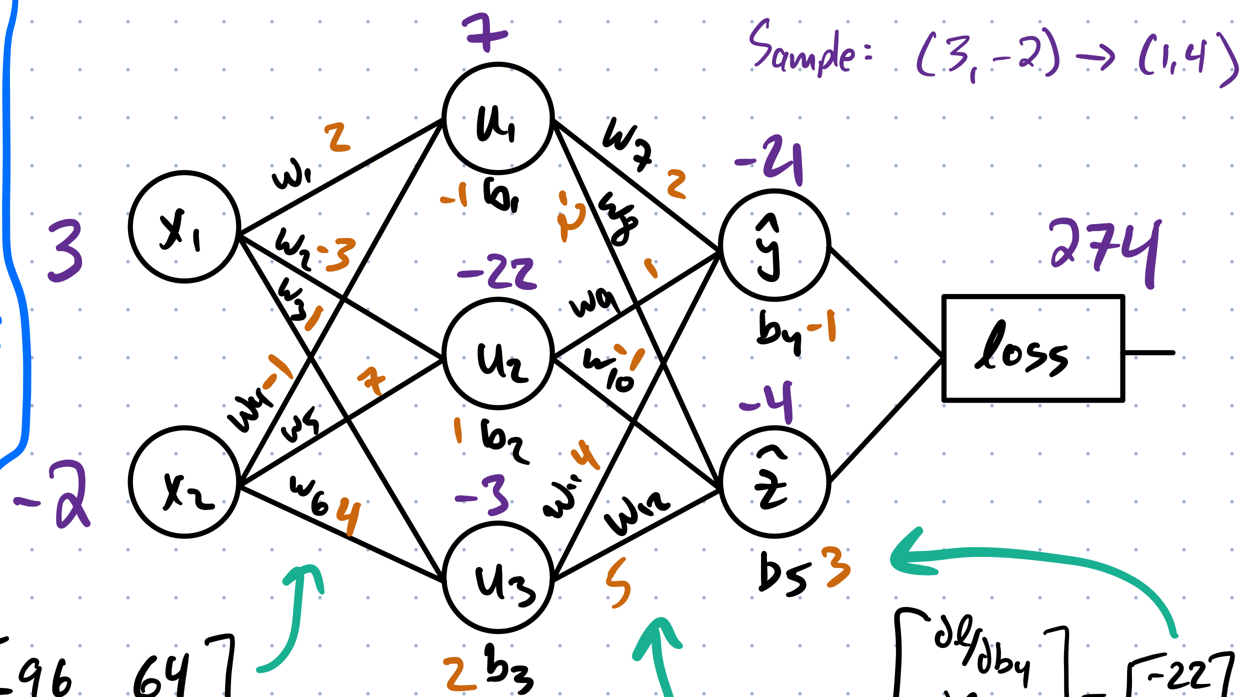
$$\begin{bmatrix} \frac{\partial l}{\partial w_7} & \frac{\partial l}{\partial w_9} & \frac{\partial l}{\partial w_{11}} \\ \partial l/\partial w_8 & \partial l/\partial w_{10} & \partial l/\partial w_{12} \end{bmatrix} = \begin{bmatrix} -154 & 484 & 66 \\ -42 & 132 & 18 \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial l/\partial \hat{y} \\ \partial l/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial w_1 & \partial l/\partial w_4 \\ \partial l/\partial w_2 & \partial l/\partial w_5 \\ \partial l/\partial w_3 & \partial l/\partial w_6 \end{bmatrix}}_{\text{d-weights}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}} \underbrace{\begin{bmatrix} x_1 & x_2 \end{bmatrix}}_{\text{values}}$$

$$\underbrace{\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix}}_{\text{d-biases}} = \underbrace{\begin{bmatrix} \partial l/\partial u_1 \\ \partial l/\partial u_2 \\ \partial l/\partial u_3 \end{bmatrix}}_{\text{d-values}}$$

We have the gradient now! Adjust all biases by $-\frac{1}{100}\cdot\nabla$, and repeat a lot.



$$\begin{bmatrix} \partial l/\partial w_1 & \partial l/\partial w_4 \\ \partial l/\partial w_2 & \partial l/\partial w_5 \\ \partial l/\partial w_3 & \partial l/\partial w_6 \end{bmatrix} = \begin{bmatrix} -96 & 64 \\ -48 & 32 \\ -354 & 236 \end{bmatrix}$$

$$\begin{bmatrix} \partial l/\partial b_4 \\ \partial l/\partial b_5 \end{bmatrix} = \begin{bmatrix} -22 \\ -6 \end{bmatrix}$$

$$\begin{bmatrix} \partial l/\partial b_1 \\ \partial l/\partial b_2 \\ \partial l/\partial b_3 \end{bmatrix} = \begin{bmatrix} -32 \\ -16 \\ -118 \end{bmatrix}$$

$$\begin{bmatrix} \frac{\partial l}{\partial w_7} & \frac{\partial l}{\partial w_9} & \frac{\partial l}{\partial w_{11}} \\ \partial l/\partial w_8 & \partial l/\partial w_{10} & \partial l/\partial w_{12} \end{bmatrix} = \begin{bmatrix} -154 & 484 & 66 \\ -42 & 132 & 18 \end{bmatrix}$$

Two more things in this topic:

* How does the previous example change with batching?
* Adding in activation functions.

Different topics:

* What do you do with the gradient?
  - Moving $-\frac{1}{100}\nabla$ is very simplistic.
  - Incorporate momentum
  - These are called "optimizers"

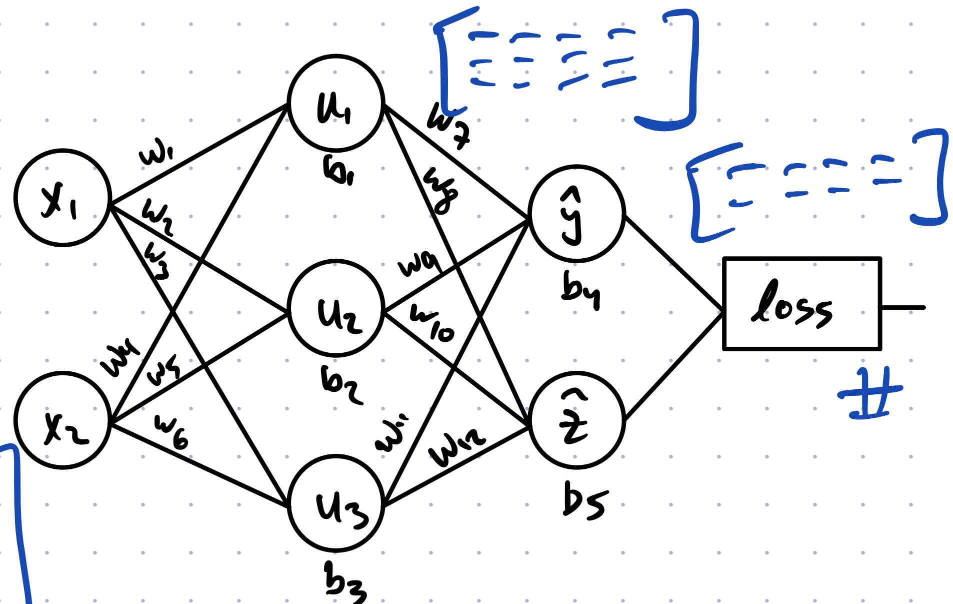
  Adam optimizer
* How to structure a network for your particular task.

$$\underbrace{\begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix}}_{\text{d-values}} = \underbrace{\begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix}}_{\text{weights}} \underbrace{\begin{bmatrix} \partial L/\partial \hat{y} \\ \partial L/\partial \hat{z} \end{bmatrix}}_{\text{d-values}}$$

$$\underbrace{\begin{bmatrix} \partial L/\partial w_1 & \partial L/\partial w_4 \\ \partial L/\partial w_2 & \partial L/\partial w_5 \\ \partial L/\partial w_3 & \partial L/\partial w_6 \end{bmatrix}}_{\text{d-weights}} = \underbrace{\begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix}}_{\text{d-values}} \underbrace{\begin{bmatrix} x_1 & x_2 \end{bmatrix}}_{\text{values}}$$

$$\underbrace{\begin{bmatrix} \partial L/\partial b_1 \\ \partial L/\partial b_2 \\ \partial L/\partial b_3 \end{bmatrix}}_{\text{d-biases}} = \underbrace{\begin{bmatrix} \partial L/\partial u_1 \\ \partial L/\partial u_2 \\ \partial L/\partial u_3 \end{bmatrix}}_{\text{d-values}}$$

What changes when batching inputs?



The values of the layers are matrices instead of vectors. So d-values are also matrices.

d-biases are then column sums of d-values (averaging)

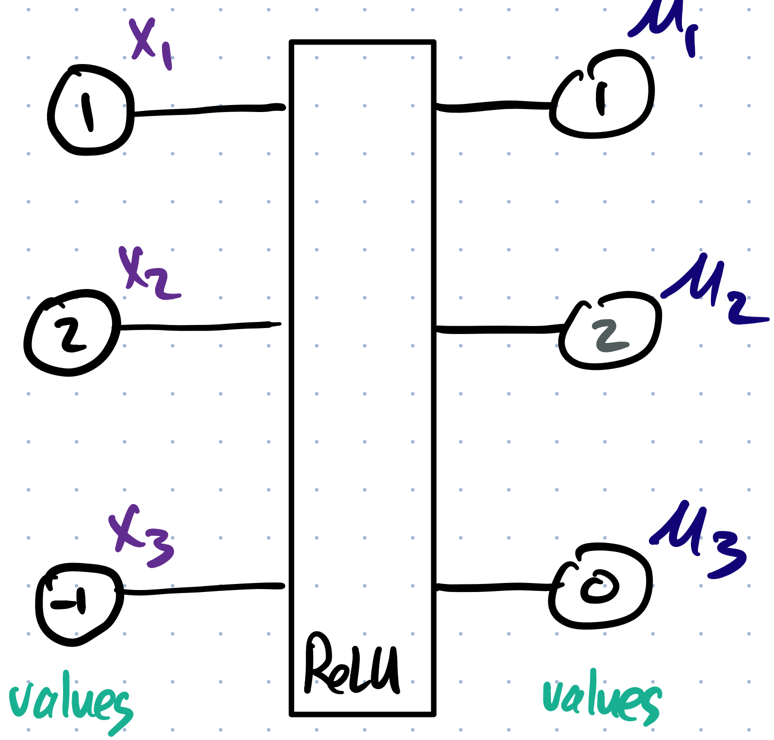
d-weights is a matrix times a matrix (more averaging)

Activation Functions

* Easy derivatives that go between the layers.

$$\text{ReLU}(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

$$\frac{\partial \text{ReLU}}{\partial x} = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$ (ignore the discontinuity)



Activation Functions

* Easy derivatives that go between the layers.

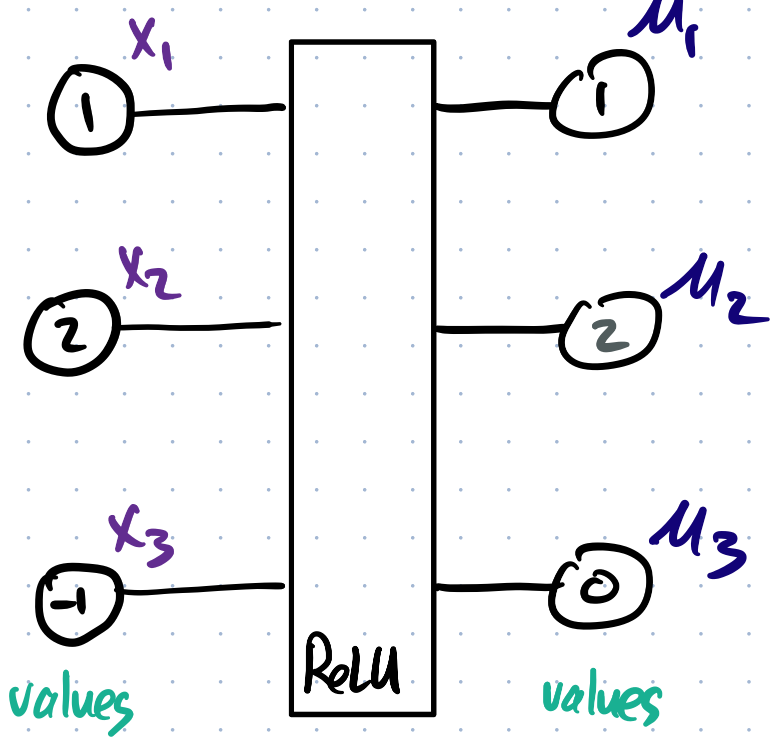
$$\mathrm{ReLU}(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

$$\frac{\partial \mathrm{ReLU}}{\partial x} = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$ (ignore the discontinuity)

$$\frac{\partial u_1}{\partial x_1} = 1$$

$$\frac{\partial u_2}{\partial x_2} = 1$$

$$\frac{\partial u_3}{\partial x_3} = 0$$



Activation Functions

* Easy derivatives that go between the layers.

Sigmoid, $\sigma(x) = \dfrac{1}{1+e^{-x}}$

$$\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \sigma(x)\cdot(1-\sigma(x))$$

$$\frac{1}{1+e^{-x}}\cdot\left(1-\frac{1}{1+e^{-x}}\right)$$

$$\frac{1}{1+e^{-x}}\cdot\left(\frac{1+e^{-x}}{1+e^{-x}}-\frac{1}{1+e^{-x}}\right) = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$\underbrace{\sigma'(x) = \sigma(x)\cdot(1-\sigma(x))}$$

you can compute the derivative of $\sigma(x)$ using only its value!

Activation Functions

* Tanh

$$g(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

$$g'(z) = \cdots = 1 - g(z)^2 .$$

Another trick!

Activation Functions

Classification

* Categorical Cross-Entropy Loss
* Softmax

There's a great trick here where we do the softmax activation and then CCE loss in as one big function instead of 2, and their joint derivative is much nicer than the individual ones.

→ Details in the code and in the Neural Networks From Scratch book.

Layer Class:

```python
def forward_pass(self, batch_inputs):
    """
    Compute the value for the neurons in this layer given the input
    from the previous layer.
    """
    self.inputs = batch_inputs
    return np.dot(self.weights, batch_inputs) + self.biases
```

```python
def backward_pass(self, dvalues):
    """
    Compute the gradient of the loss with respect to the weights and biases
    feeding into this layer and the neuron values of previous layer (aka, the
    inputs into this layer). Input is the gradient with respect to neuron values
    in this layer. Think of this like we are pushing out the gradients from us
    into the previous layer.
    """
    # Compute the gradient of the loss with respect to the weights and biases
    self.dweights = np.dot(dvalues, self.inputs.T)
    self.dbiases = np.sum(dvalues, axis=1, keepdims=True)

    # Compute the gradient of the loss with respect to the inputs
    self.dinputs = np.dot(self.weights.T, dvalues)
    return self.dinputs
```

transpose

ReLU Activation

```python
def forward_pass(self, batch_values):
    """
    Compute the ReLU activation for the input values.
    """
    self.inputs = batch_values
    return np.maximum(0, batch_values)

def backward_pass(self, dvalues):
    """
    Compute the gradient of the loss with respect to the input values.
    """
    # Gradient of ReLU is 1 for positive values, 0 for negative values
    self.dvalues = dvalues * (self.inputs > 0)
    return self.dvalues
```

Sigmoid Activation

```python
def forward_pass(self, batch_values):
    """
    Compute the sigmoid activation for the input values.
    Sigmoid function: f(x) = 1 / (1 + e^(-x))
    """
    self.inputs = batch_values
    # in this case we save the outputs because we can reuse them for
    # the backward pass
    self.outputs = 1 / (1 + np.exp(-batch_values))
    return self.outputs

def backward_pass(self, dvalues):
    """
    Compute the gradient of the loss with respect to the input values.
    The derivative of the sigmoid function is f'(x) = f(x) * (1 - f(x))
    """
    self.dvalues = dvalues * (self.outputs * (1 - self.outputs))
    return self.dvalues
```

$\sigma(x) \cdot (1 - \sigma(x))$

Neural Network Training Loop

input data

```python
def forward_pass(self, batch_data):
    """
    Perform a forward pass through the network by passing the data through
    each layer.
    """
    input_data = batch_data
    for layer in self.layers:
        input_data = layer.forward_pass(input_data)
    return input_data

def backward_pass(self, dvalues):
    """
    Perform a backward pass through the network given gradient at output.
    """
    grad = dvalues
    for layer in reversed(self.layers):
        grad = layer.backward_pass(grad)
```

Plenty more details we're skipping for now.

Done!

* We know how to make a NN match the training data a little bit better.

* Pass it all through in a forward pass $\longrightarrow$
  then compute the gradients going backward $\longleftarrow$

* Update weights + biases based on those gradients ("optimizers")
  then repeat.

Well, one more thing: **mini batching**

* Usually you don't send <u>all</u> the training data through in one big batch.

* Shuffle the order of the data. Then do forward and backward passes of 64 (or whatever) input/output pairs at a time until all have been used.

* Shuffle again and repeat.

- Faster, less memory usage, and often trains better!

(3) Bike sharing (there is a "bike-share-explanation.md" markdown file

Data from a bike rental service in Washington, D.C.

Date + Time
Season
Holiday
Workday
Weather
Temperature
"Feels like" temp.
Humidity
Wind Speed

rentals per day (registered / casual)