

Scientific Computing

April 30, 2025

Announcements

→ Homework 6 due Friday, May 2, 11:59pm

→ Final exam is take-home only
assigned Fri, May 2
due Fri, May 9

Today

→ Backpropagation

Office Hours:

Mon + Fri

9:30am - 10:30am

Cudahy 307

Topic 16 - Backpropagation

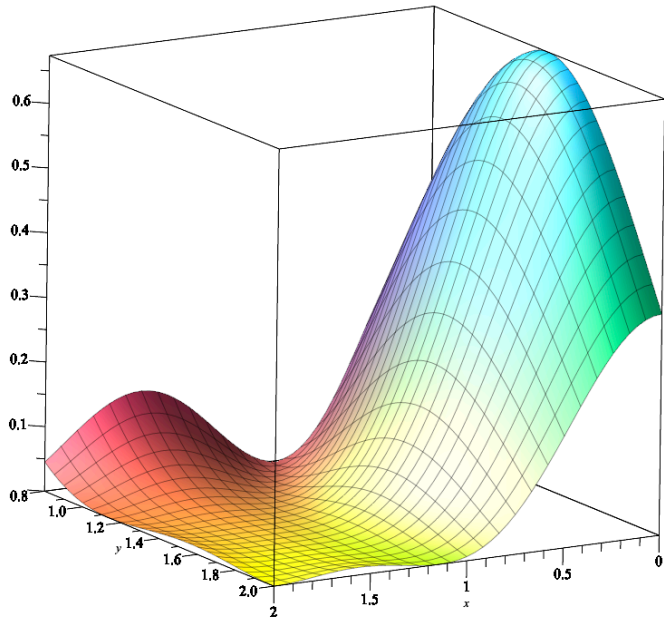
Big picture:

- * Start with a NN with random weights and biases
- * Feed all of the known inputs (training data) through in one batch and get all their outputs
- * Compute the loss between the expected outputs (y) and the actual outputs ($\hat{y}$)
- * Use **CALCULUS** to figure out how to tweak the weights + biases a little to make the loss go down a little.
- * Repeat until your NN is good enough.

CALCULUS

Let's recall the idea of Gradient Descent.

Suppose you have a function $f(x,y)=z$.



The vector $\begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} = \nabla f$ tells you the direction of steepest ascent from any point (x,y) .

The negative version $-\nabla f = \begin{bmatrix} -\frac{\partial f}{\partial x} \\ -\frac{\partial f}{\partial y} \end{bmatrix}$ is the direction of steepest descent.

$$l = \text{NN}(w_1, w_2, \dots, w_m, b_1, b_2, \dots, b_n).$$

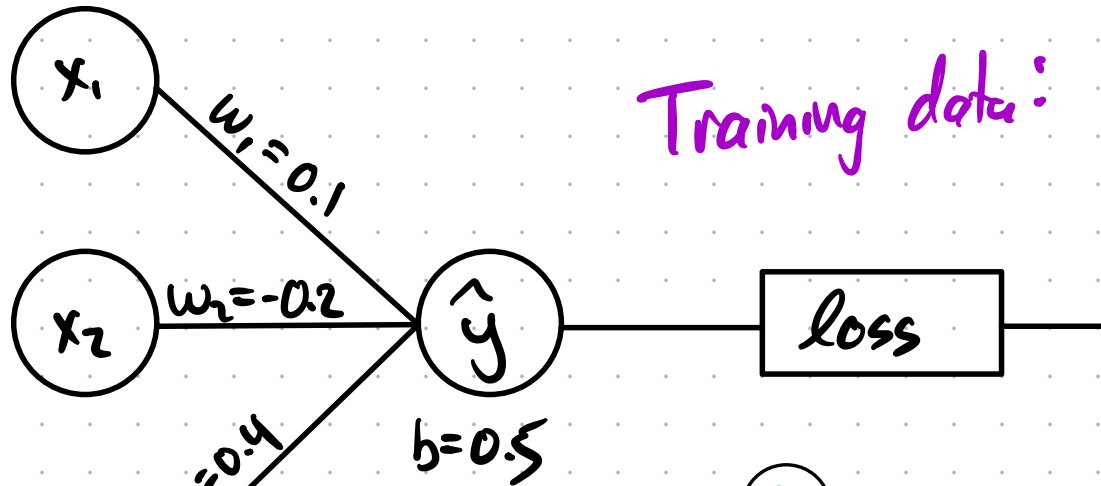
How will we compute

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

CALCULUS

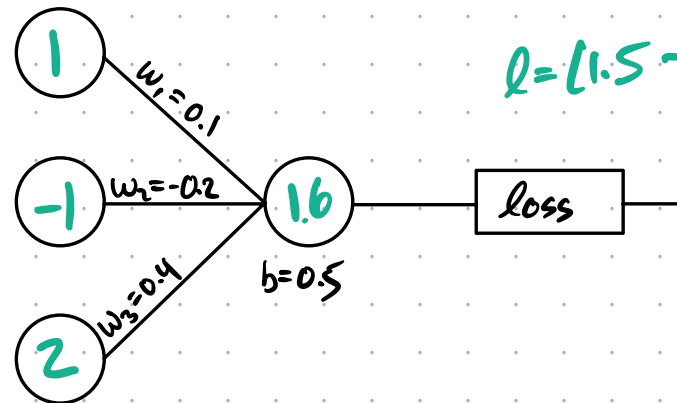
specifically
the CHAIN RULE

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

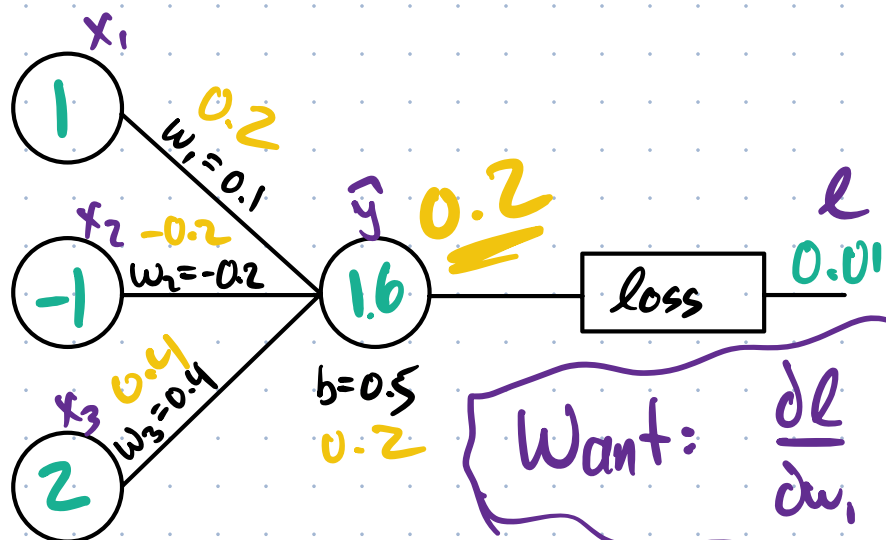


Training data: $(1, -1, 2) \rightarrow 1.5$

$$l = (1.5 - 1.6)^2 = 0.01$$



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



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

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

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

$$\text{So, } \frac{\partial l}{\partial \hat{y}} = 2(\hat{y} - 1.5) = 2 \cdot (0.1) = 0.2$$

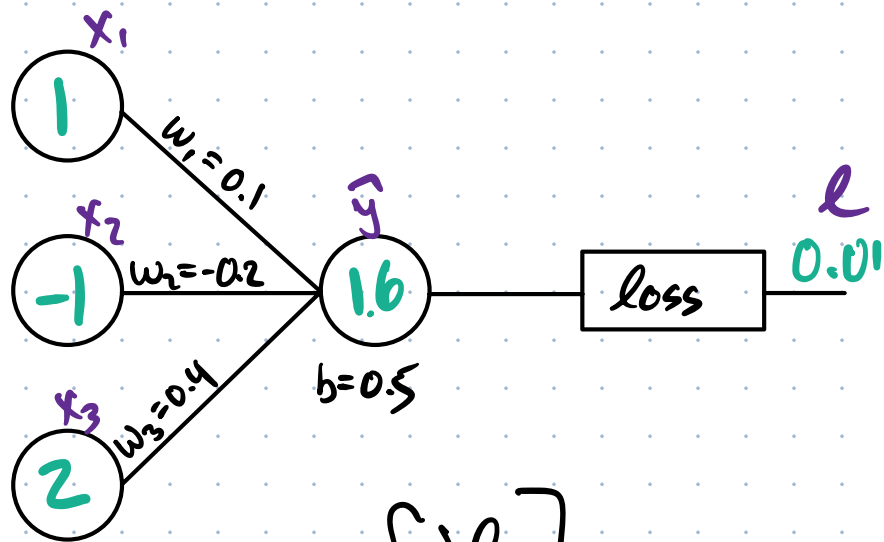
$$\hat{y} = 1 \cdot w_1 - 1 \cdot w_2 + 2 \cdot w_3 + b$$

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

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

chain rule!

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



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

$$-\nabla_{NN} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial w_1} \\ \frac{\partial \mathcal{L}}{\partial w_2} \\ \frac{\partial \mathcal{L}}{\partial w_3} \\ \frac{\partial \mathcal{L}}{\partial b} \end{bmatrix} = \begin{bmatrix} -0.2 \\ 0.2 \\ -0.4 \\ -0.2 \end{bmatrix}$$

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

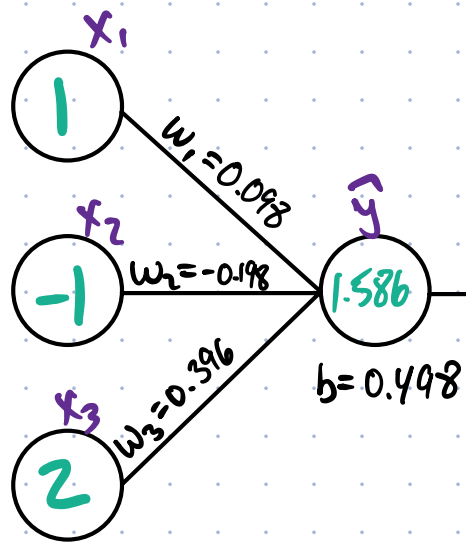
$$w_1: 0.1 \rightarrow 0.098$$

$$w_2: -0.2 \rightarrow -0.198$$

$$w_3: 0.4 \rightarrow 0.396$$

$$b: 0.5 \rightarrow 0.498$$

$$0.1 + \frac{1}{100} \cdot (-0.2) = 0.098$$



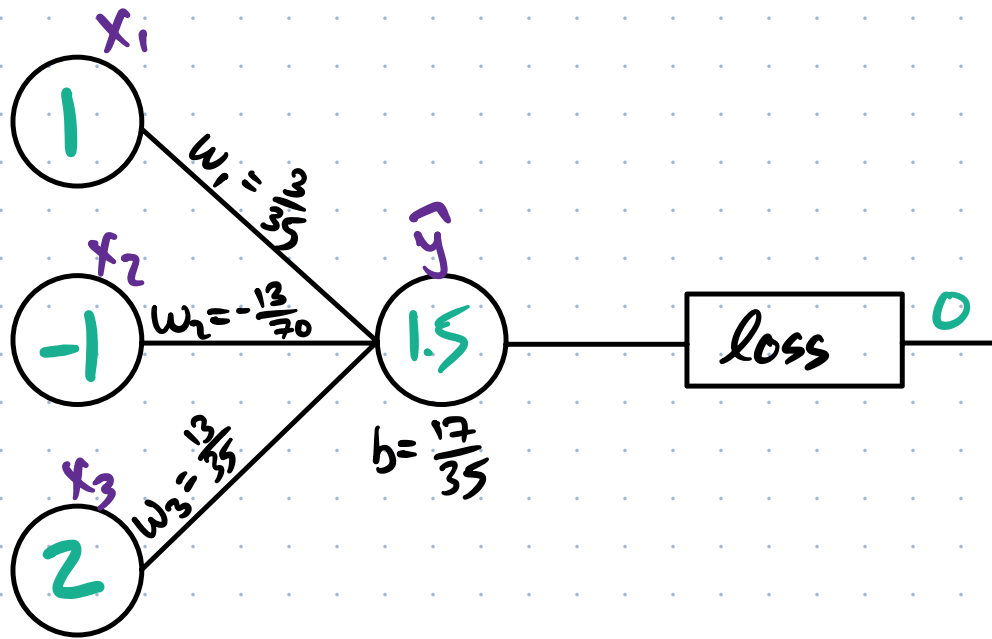
Training data: $(1, -1, 2) \rightarrow 1.5$
(prev. 0.01)

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

* Python demo.

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

After 100 or so repetitions:



Easy because only one piece of training data!

Example 2: Backpropagation on a Batch of data.

Training Data:

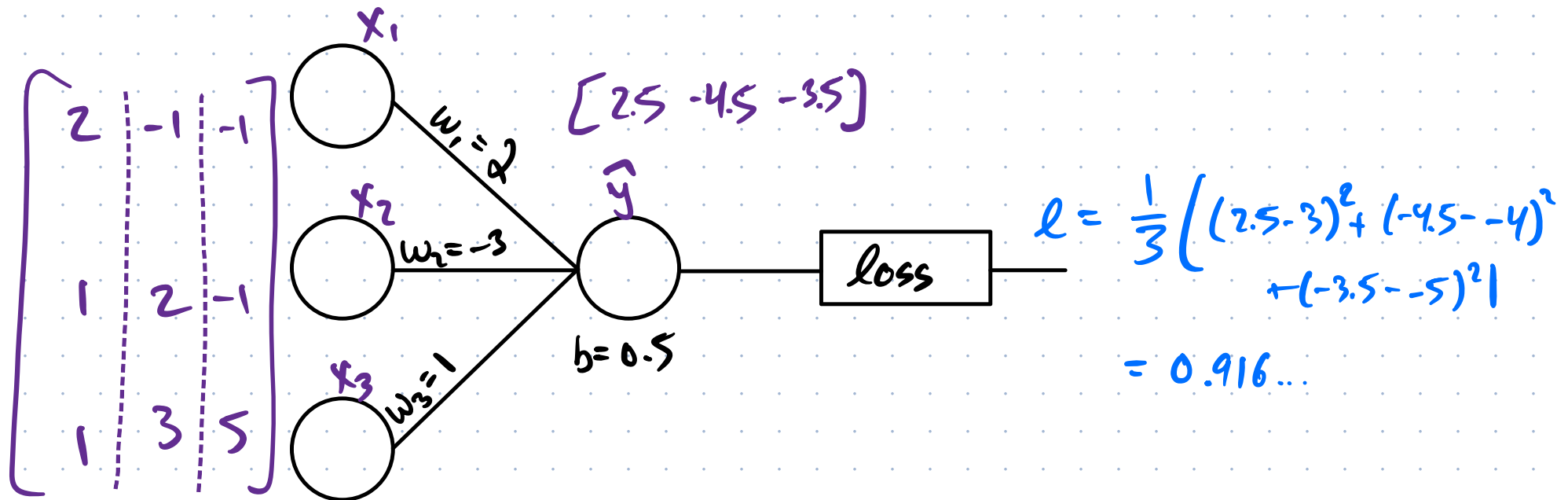
$$(2, 1, 1) \rightarrow 3$$

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

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

Still no AF

no hidden layers



(making the weights whole #s for now)

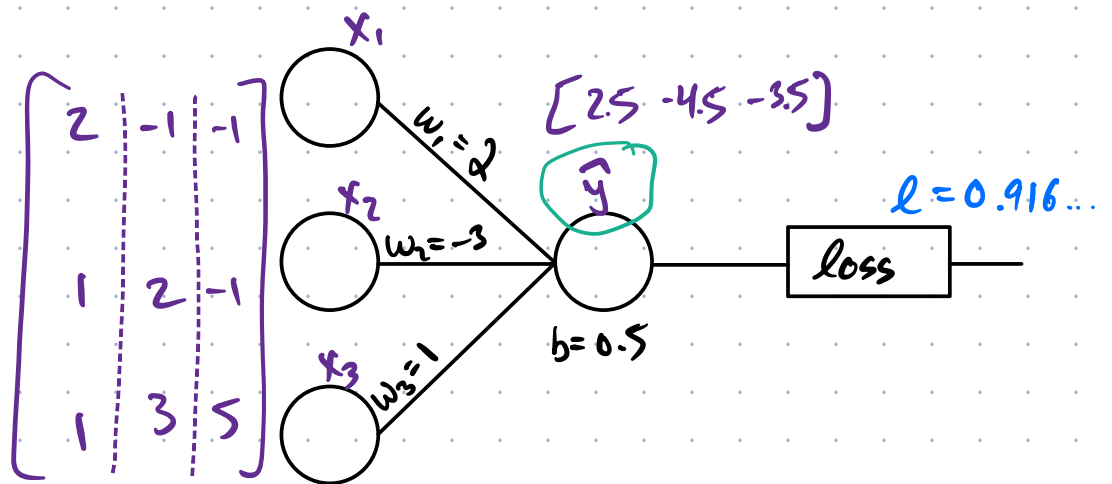
Example 2: Backpropagation on a Batch of data.

Training Data:

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

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

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



We fed forward, now we backpropagate the derivatives

$$L = \frac{1}{3} ((\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2) = \text{MSE}$$

$$\begin{aligned} \frac{\partial L}{\partial \hat{y}_1} &= \frac{2}{3} (\hat{y}_1 - y_1) \\ &= \frac{2}{3} \cdot (-0.5) \\ &= -\frac{1}{3} \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial \hat{y}_2} &= \frac{2}{3} (\hat{y}_2 - y_2) \\ &= \frac{2}{3} \cdot (-0.5) \\ &= -\frac{1}{3} \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial \hat{y}_3} &= \frac{2}{3} (\hat{y}_3 - y_3) \\ &= \frac{2}{3} \cdot (1.5) \\ &= 1 \end{aligned}$$

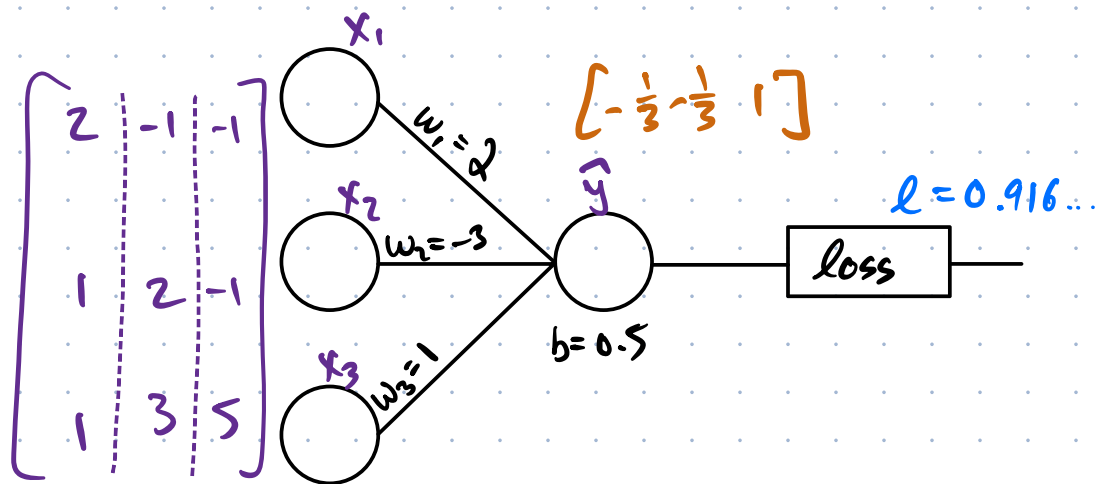
Example 2: Backpropagation on a Batch of data.

Training Data:

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

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

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



We fed forward, now we backpropagate the derivatives

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

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

$$\frac{\partial L}{\partial w_1} = \left(\frac{\partial L}{\partial \hat{y}_1} \right) \cdot \left(\frac{\partial \hat{y}_1}{\partial w_1} \right) = -\frac{1}{3} \cdot 2 = -\frac{2}{3}$$

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

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

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

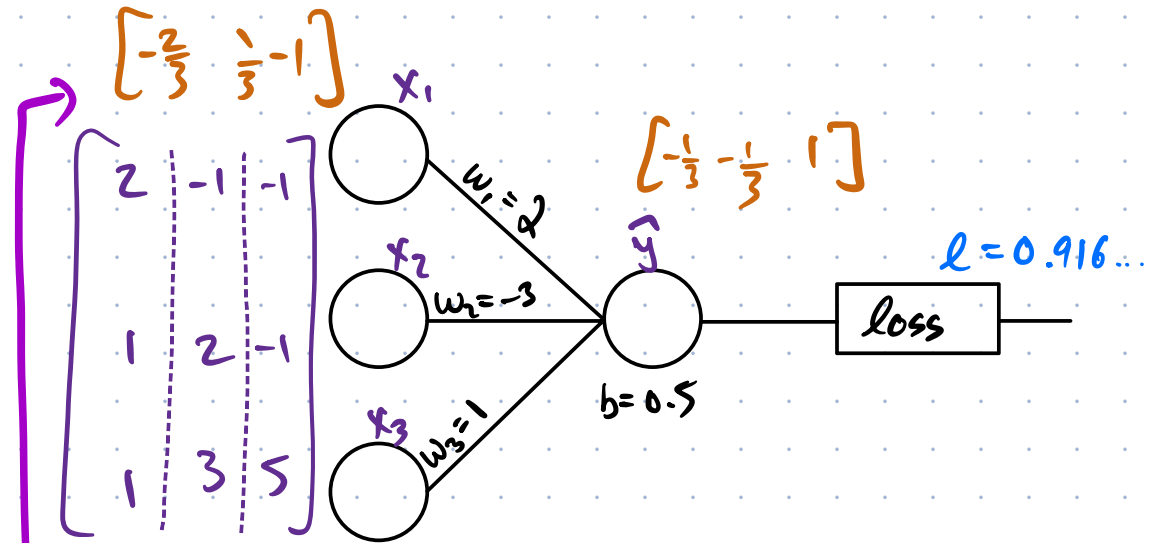
Example 2: Backpropagation on a Batch of data.

Training Data:

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

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

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



We fed forward, now we backpropagate the derivatives

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

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

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

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

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

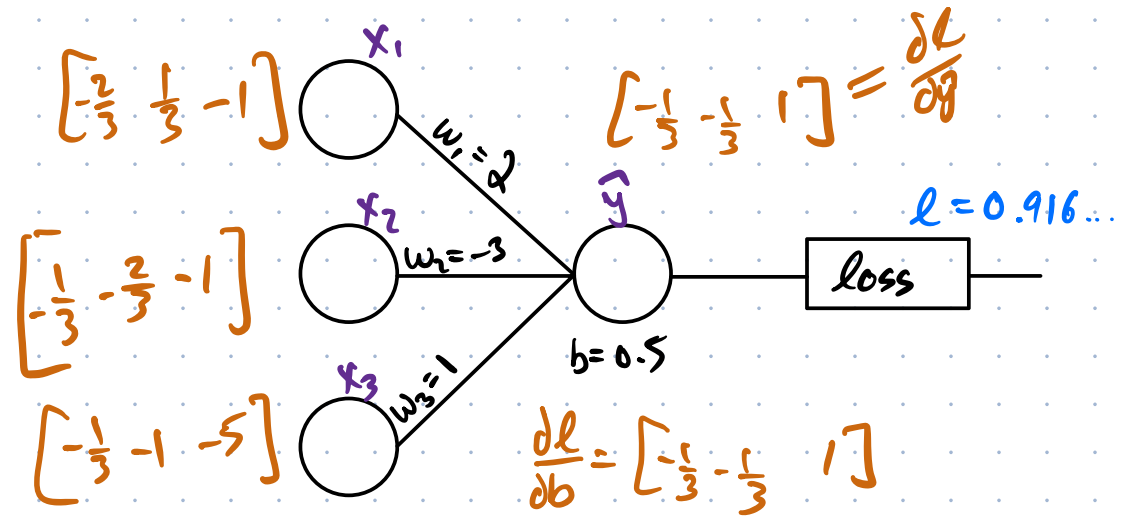
Example 2: Backpropagation on a Batch of data.

Training Data:

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

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

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



We fed forward, now we backpropagate the derivatives

$$\frac{\partial L}{\partial \hat{y}} = [-\frac{1}{3} \ \frac{1}{3} \ 1]$$

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

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

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

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

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

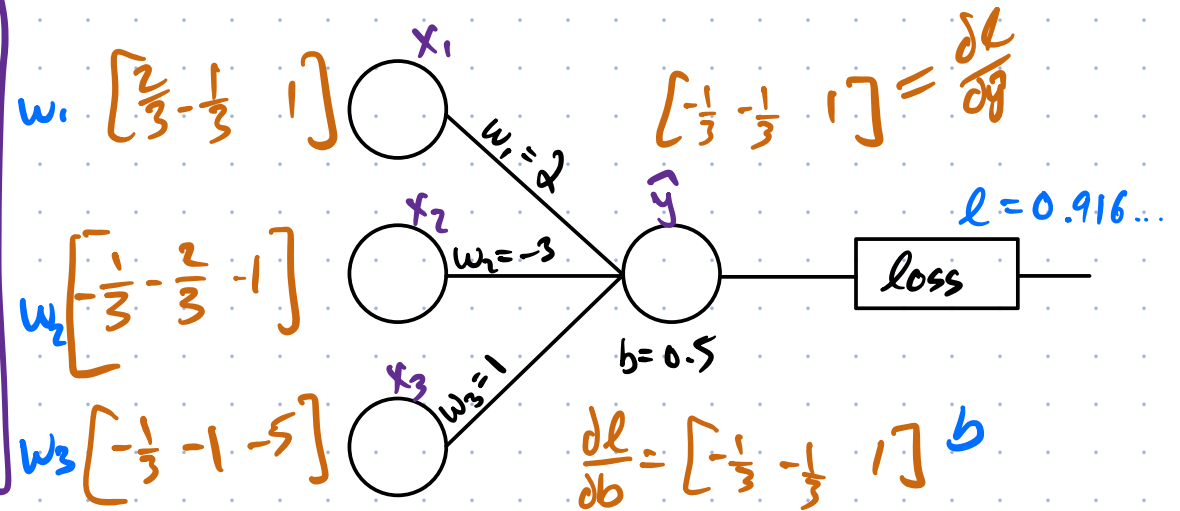
Example 2: Backpropagation on a Batch of data.

Training Data:

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

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

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



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

$$\nabla = \begin{bmatrix} 4/3 \\ -2 \\ -17/3 \\ +1/3 \end{bmatrix} \begin{matrix} w_1 \\ w_2 \\ w_3 \\ b \end{matrix}$$

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.

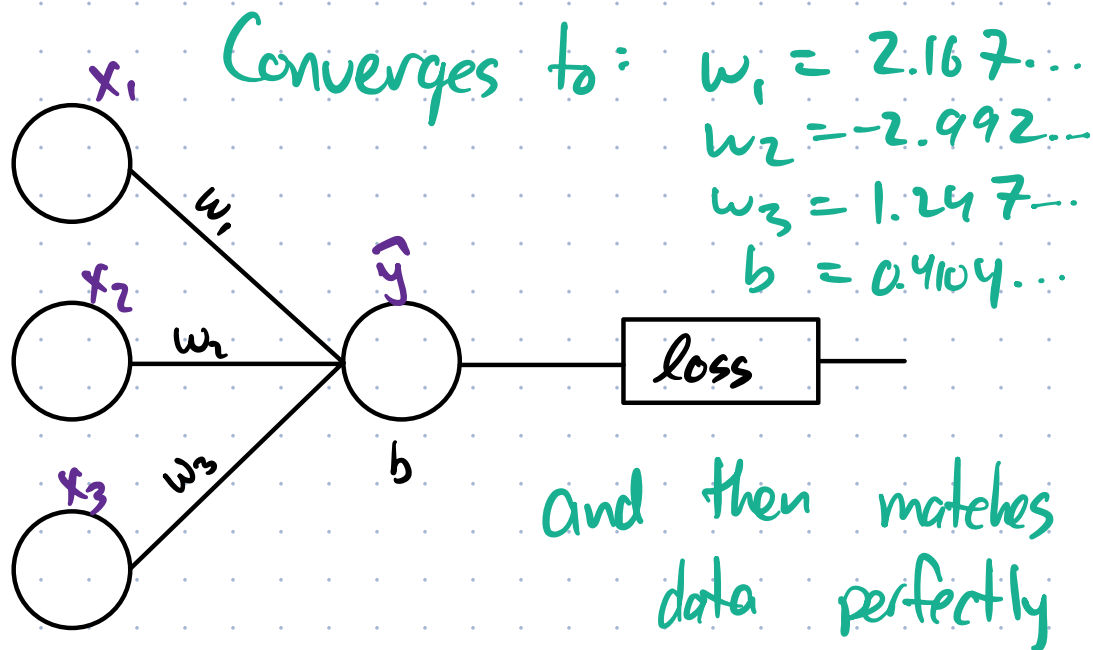
Training Data:

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

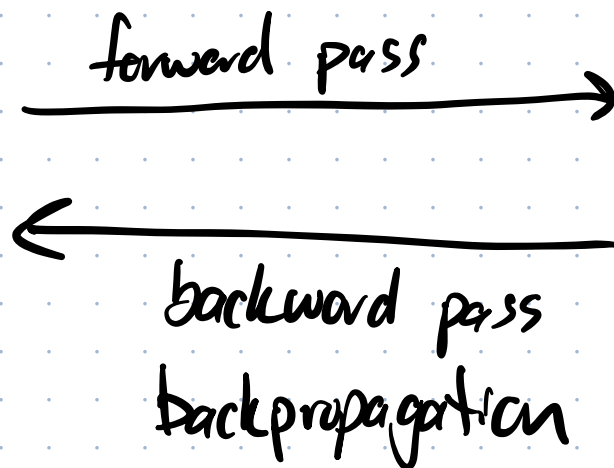
$$(-1, 2, 3) \rightarrow -4$$

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

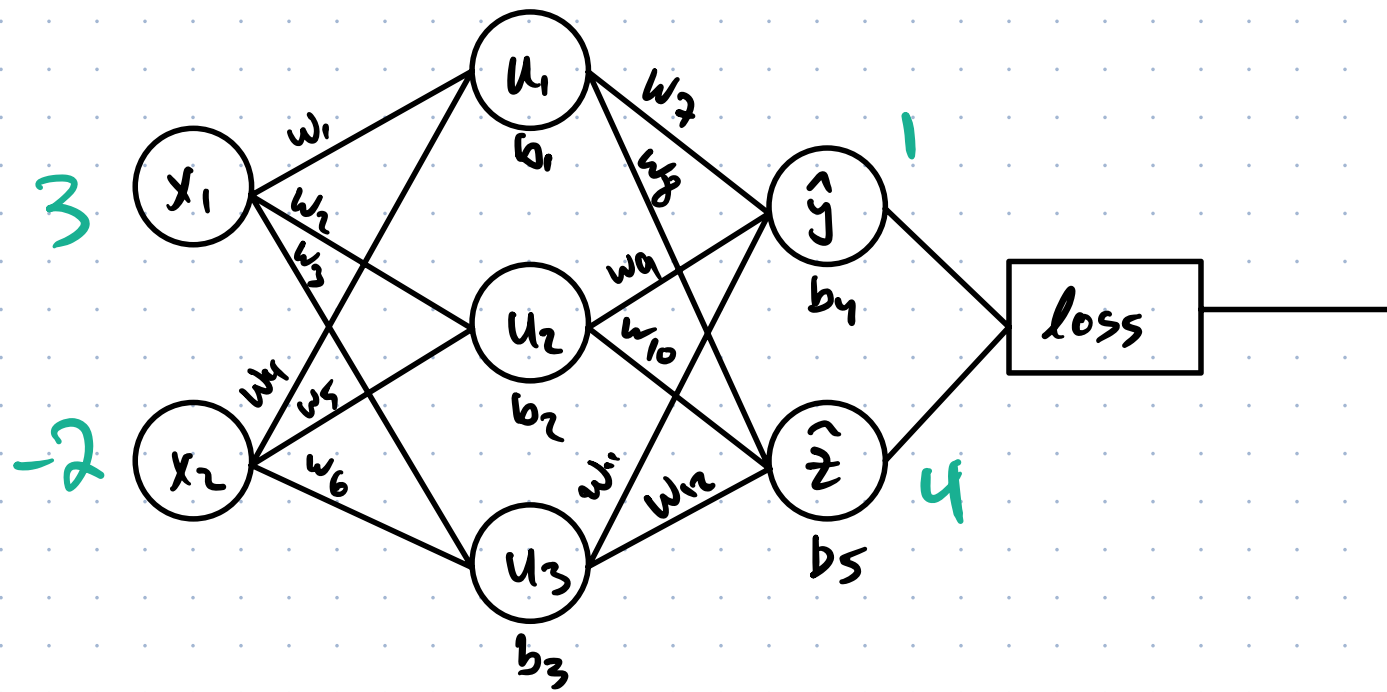
* Demo



- * 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 $-\frac{1}{n} \cdot \nabla$
"optimizers"

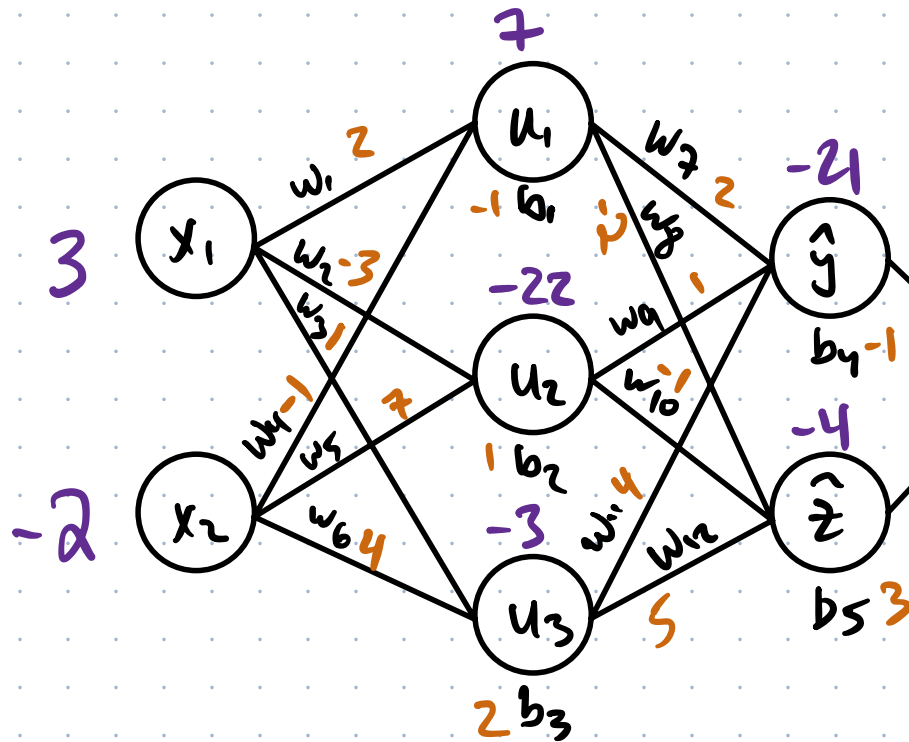


Example 3, with more layers,
still no activation functions
just 1 sample for simplicity → for more, just
add them together
to get the ∇ for
weights and biases,
as before



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

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



$$\frac{1}{2}((-1-1)^2 + (-4-4)^2) = \frac{1}{2}((-2)^2 + (-8)^2) = \frac{1}{2}(4 + 64) = \frac{1}{2}(68) = 34$$

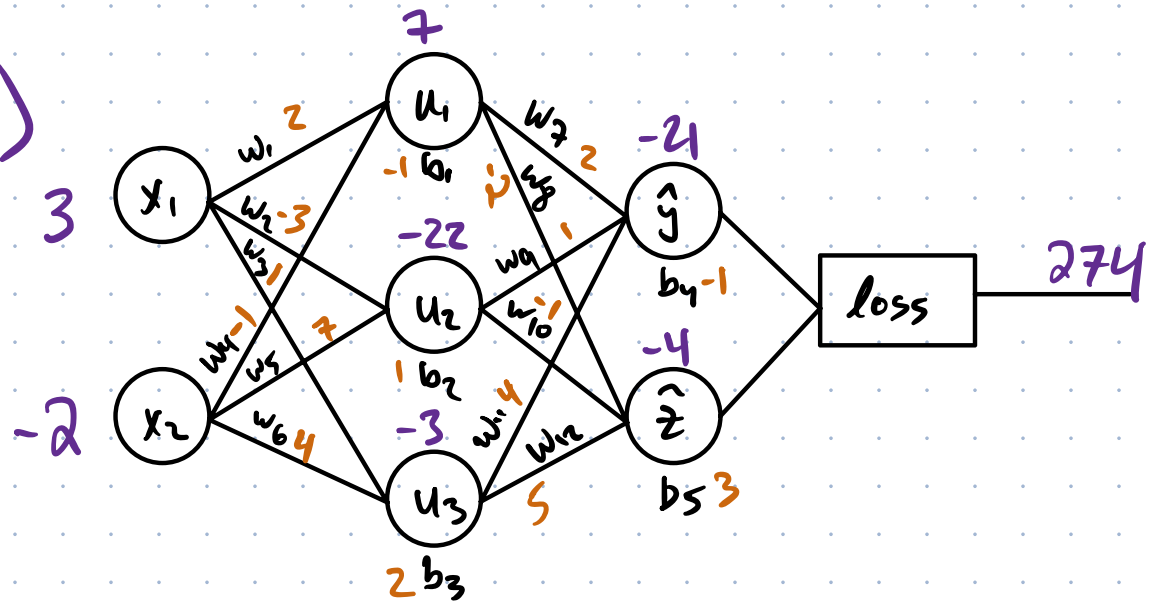
Backprop: Find:

$$\frac{\partial \mathcal{L}}{\partial w_1}, \frac{\partial \mathcal{L}}{\partial w_2}, \dots, \frac{\partial \mathcal{L}}{\partial w_{12}}, \frac{\partial \mathcal{L}}{\partial b_1}, \frac{\partial \mathcal{L}}{\partial b_2}, \dots, \frac{\partial \mathcal{L}}{\partial b_5}$$

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

Work backward from
the right.

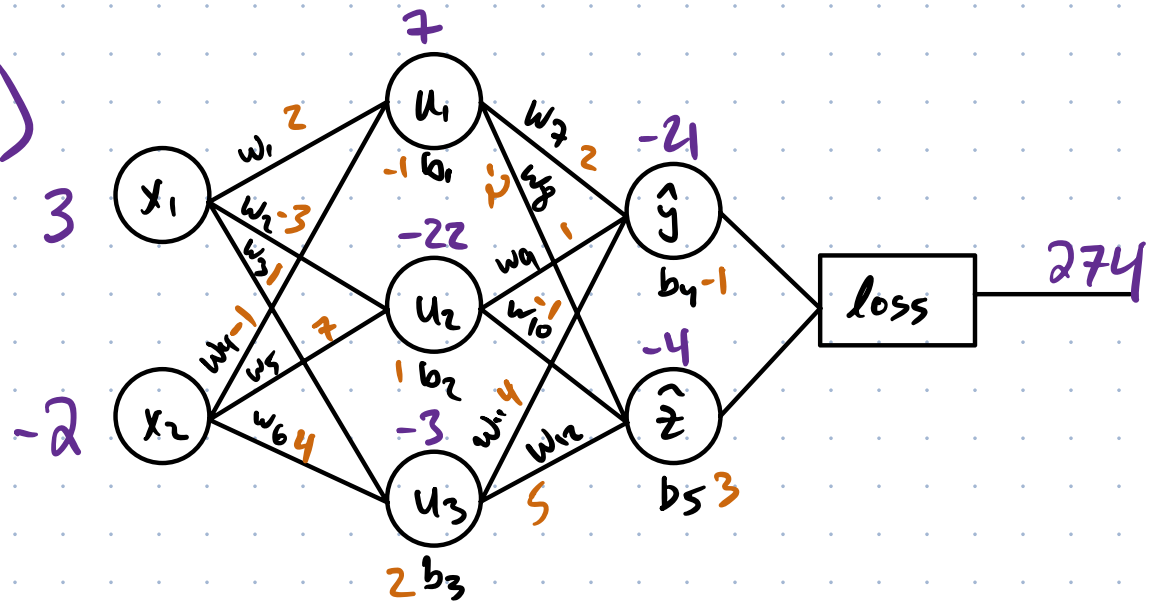
$$\frac{\partial \mathcal{L}}{\partial \hat{y}}?$$



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

Work backward from the right.

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



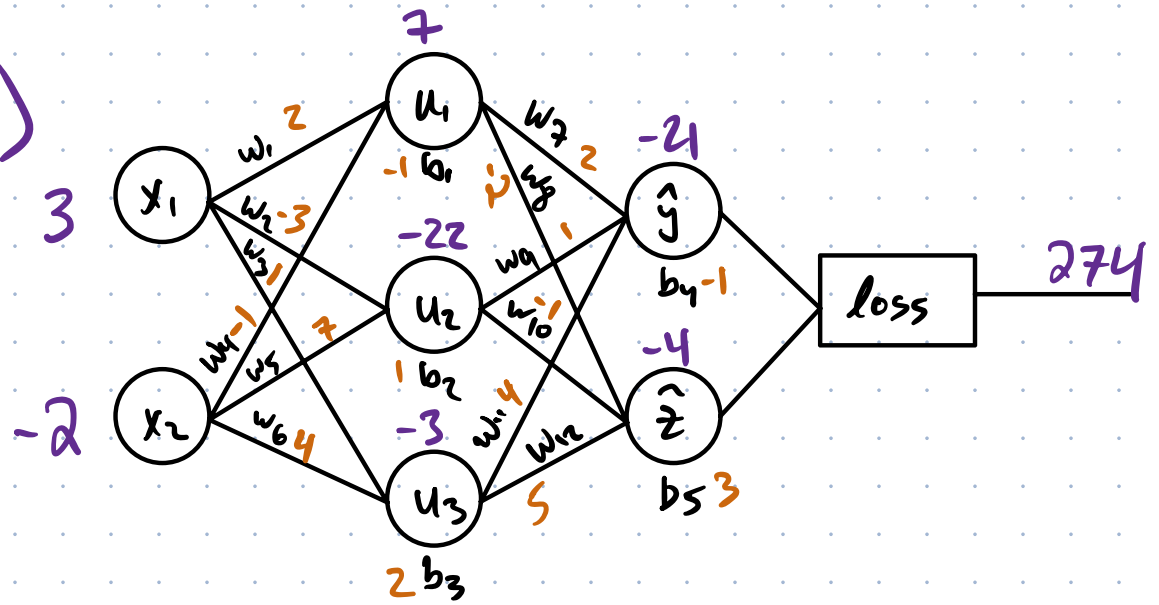
$$\ell = \frac{1}{2} ((\hat{y} - 1)^2 + (\hat{z} - 4)^2)$$

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

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

Work backward from the right.

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



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

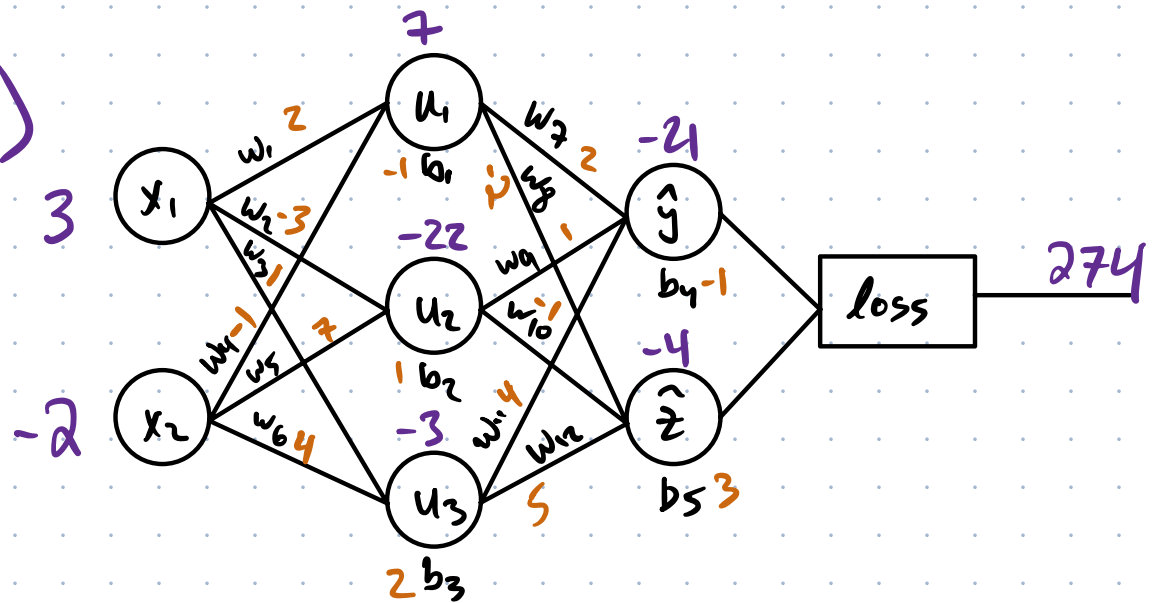
$$\frac{\partial \ell}{\partial \hat{z}} = \hat{z} - 4 = -2 - 4 = -6$$

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

Work backward from the right.

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

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



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

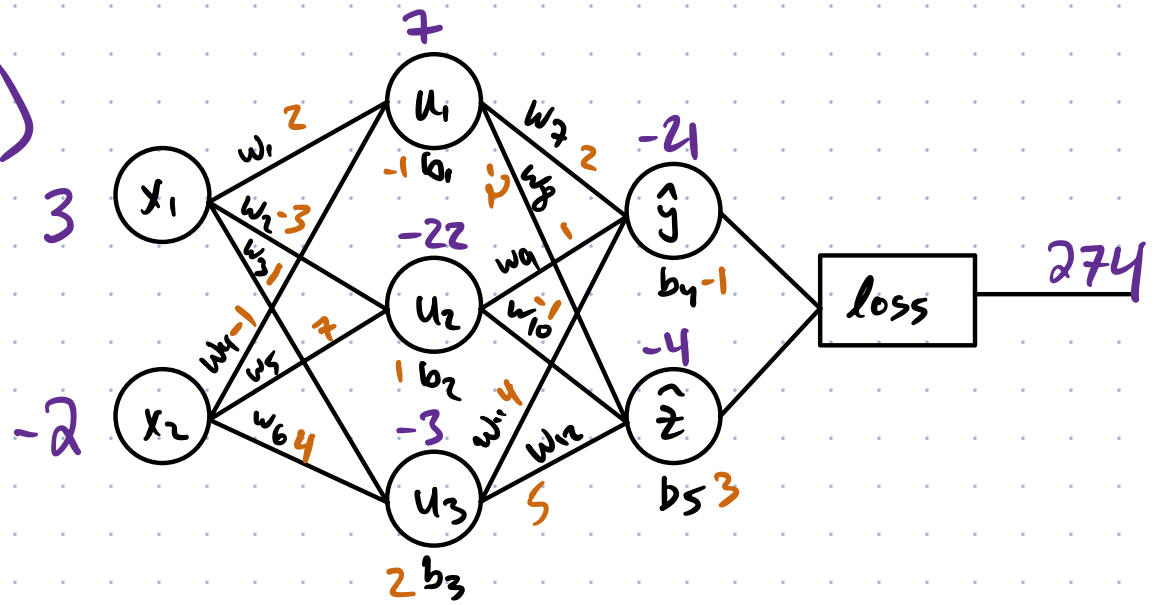
$$\Rightarrow \frac{\partial \hat{y}}{\partial w_7} = u_1 = 7$$

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

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

Work backward from the right.

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

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

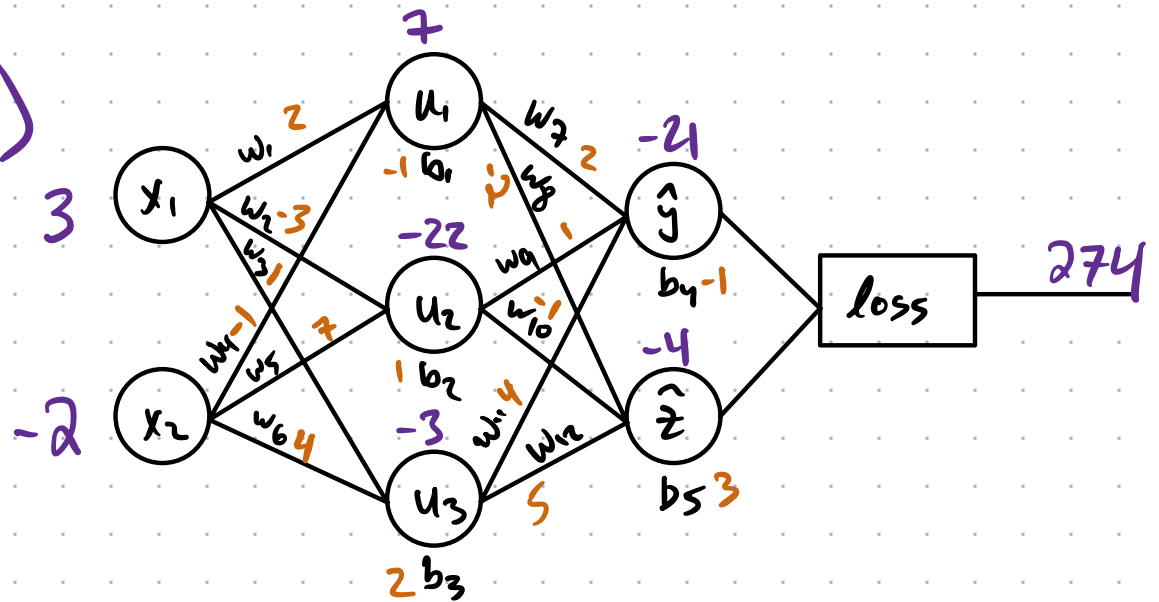
$$\frac{\partial \ell}{\partial w_9} = (-22)(-22) = 484$$

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

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

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

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



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

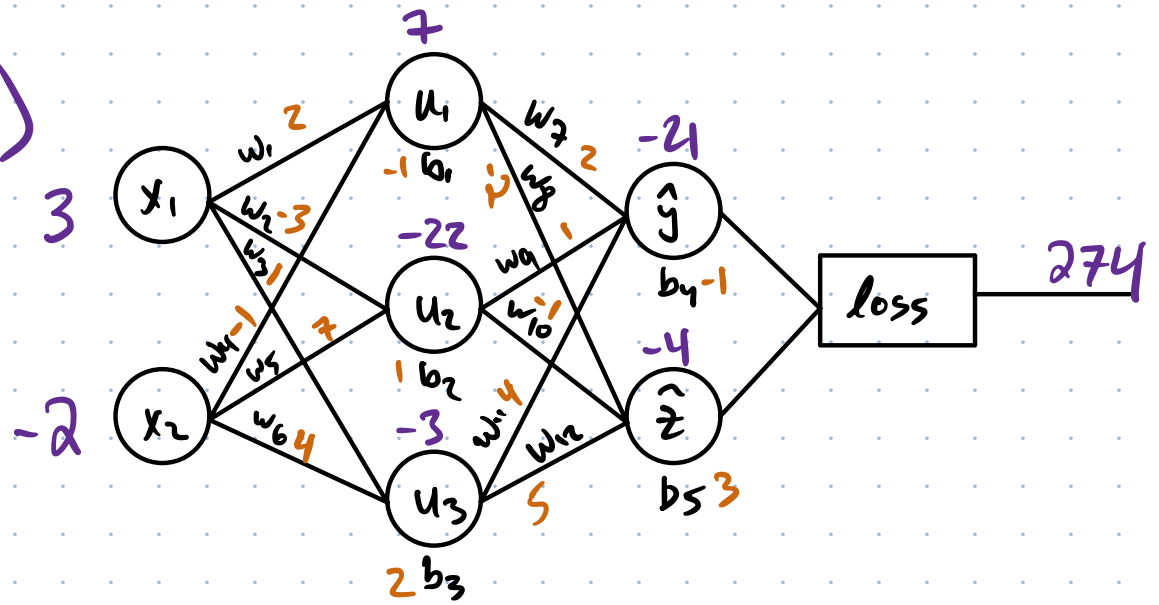
Work backward from the right.

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

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

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

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



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

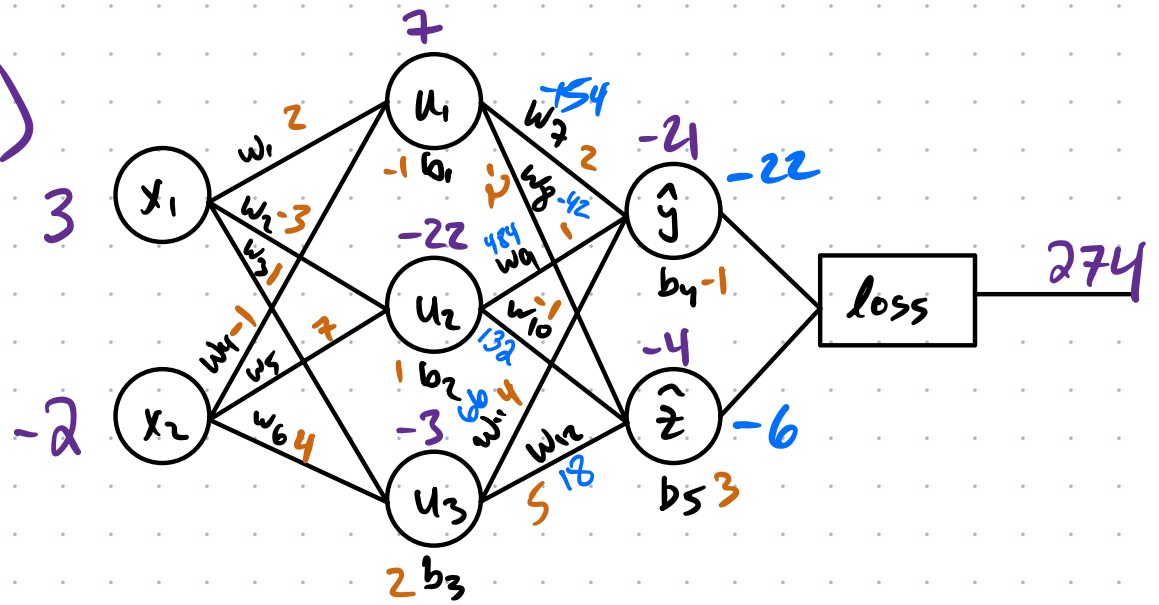
Work backward from the right.

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

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

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

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



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

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

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

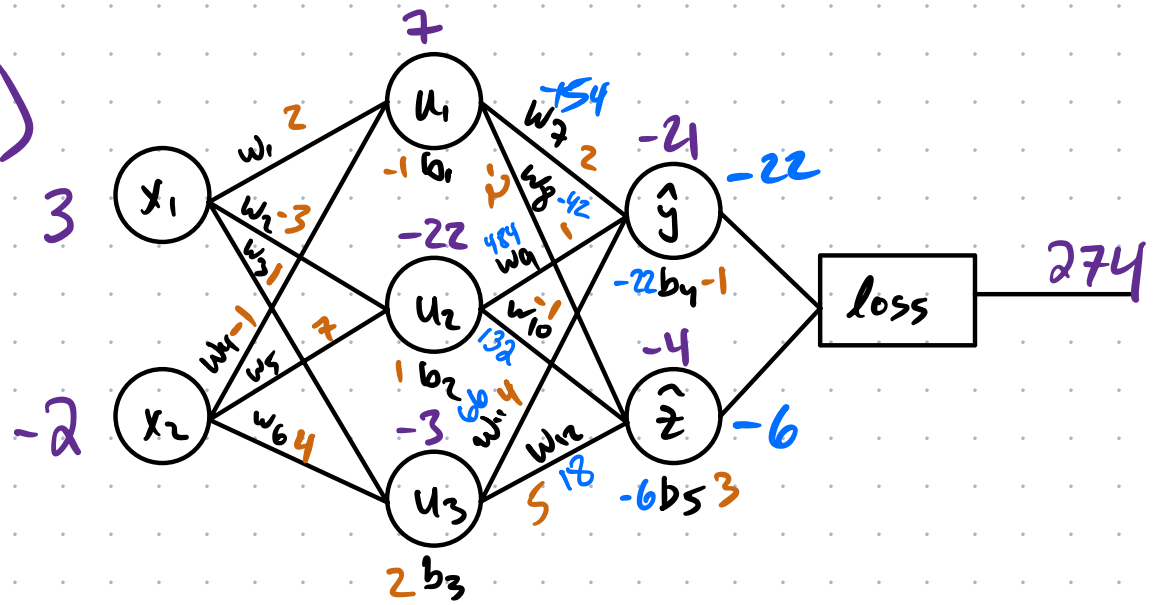
Work backward from the right.

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

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

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

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

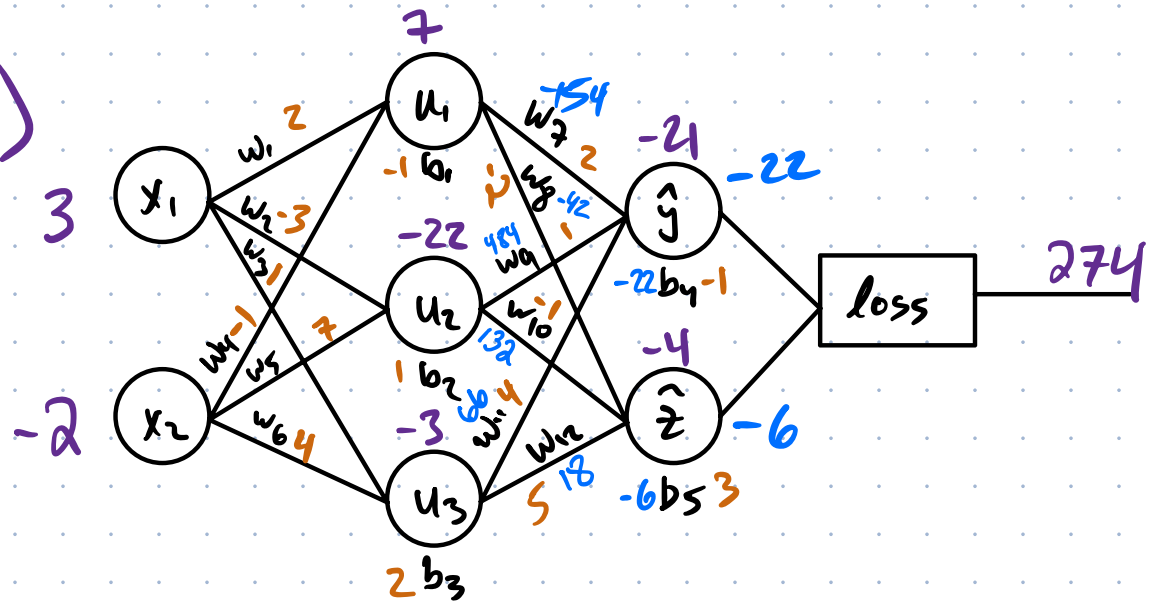


$$\frac{\partial \ell}{\partial b_4} = -22 \quad \frac{\partial \ell}{\partial b_5} = -6$$

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

Work backward from the right.

Next:



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

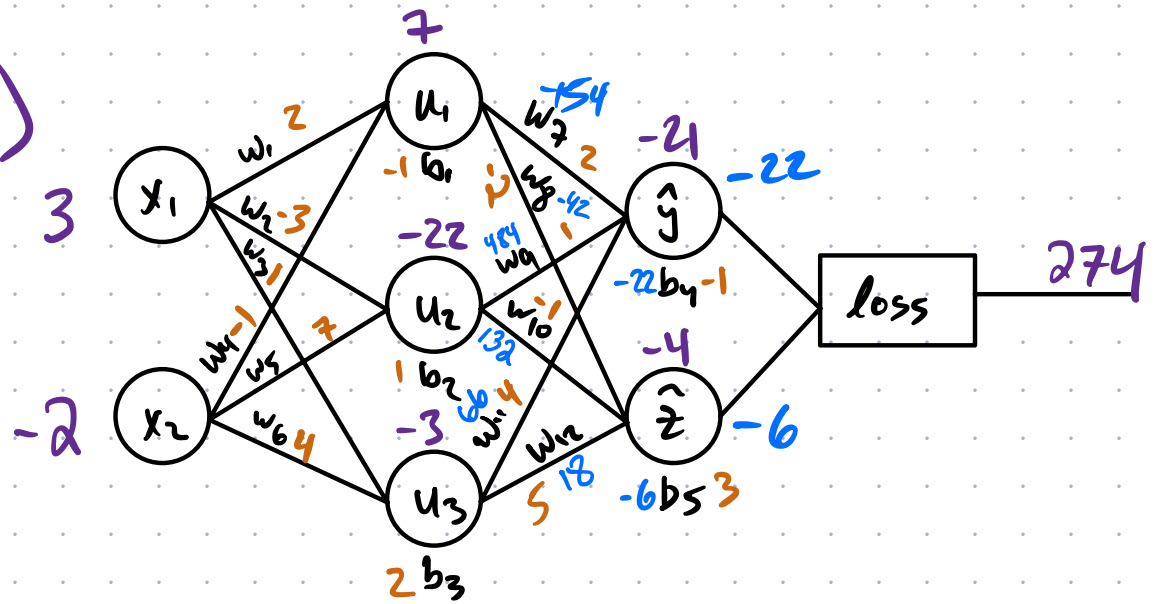
$$u_1 = w_1 x_1 + w_4 x_2 + b_1$$

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

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

Work backward from
the right.

Next:



$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial u_1} \cdot \frac{\partial u_1}{\partial w_1}$$

not computed yet

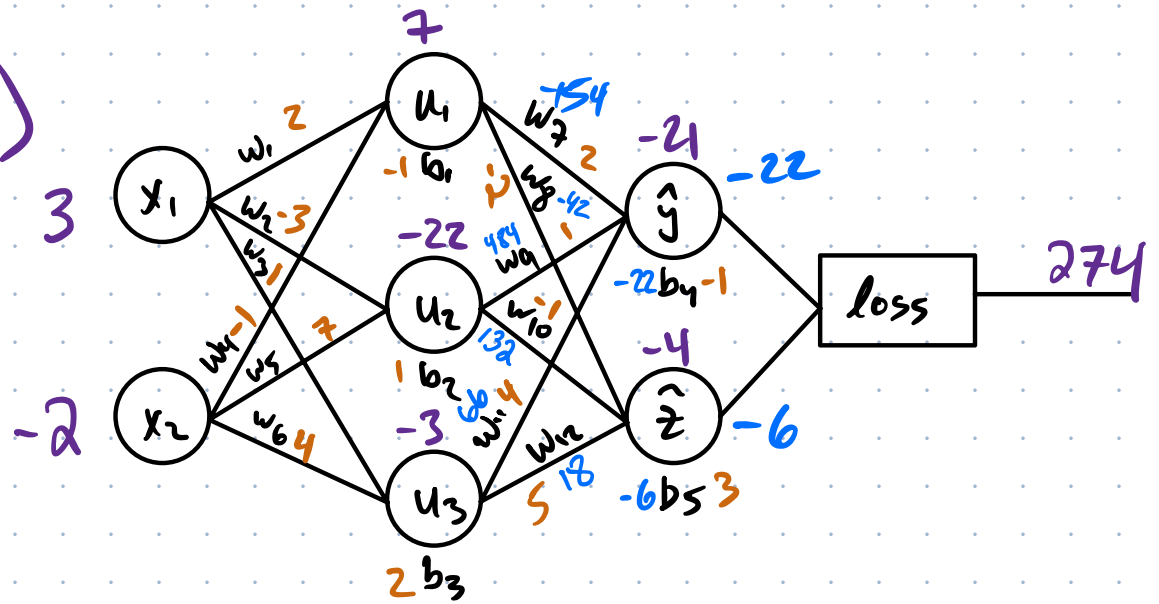
$$u_1 = w_1 x_1 + w_2 x_2 + b_1$$

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

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

Work backward from the right.

Next:



$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial u_1} \cdot \frac{\partial u_1}{\partial w_1}$$

not computed yet

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

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

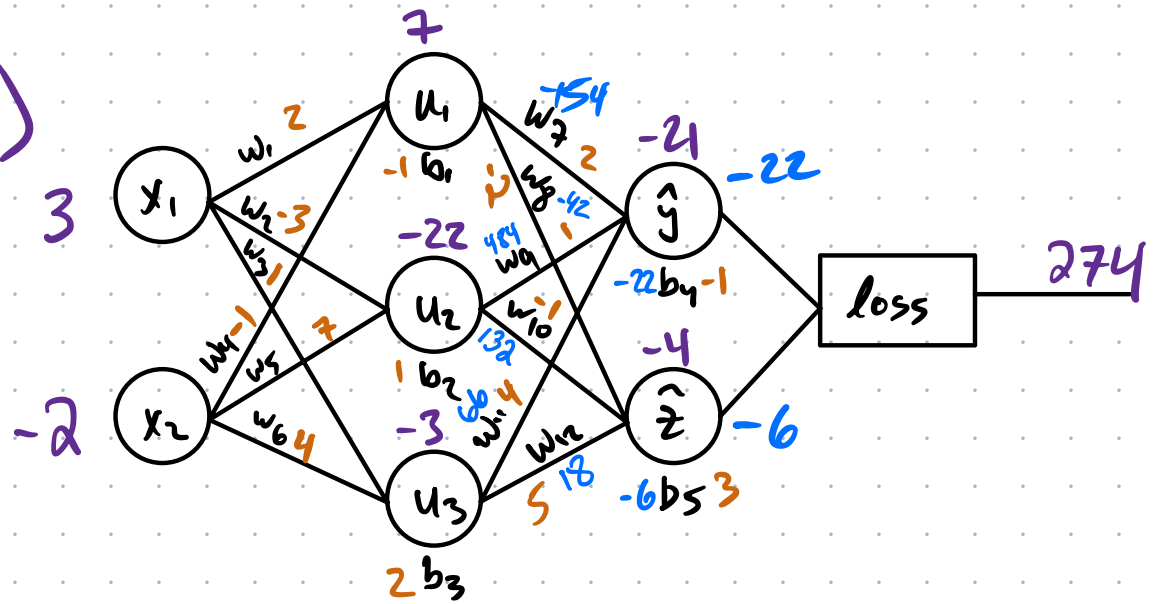
Now we see that with hidden layers, we also need derivs of the neuron values to keep passing backward

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

Work backward from the right.

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

?

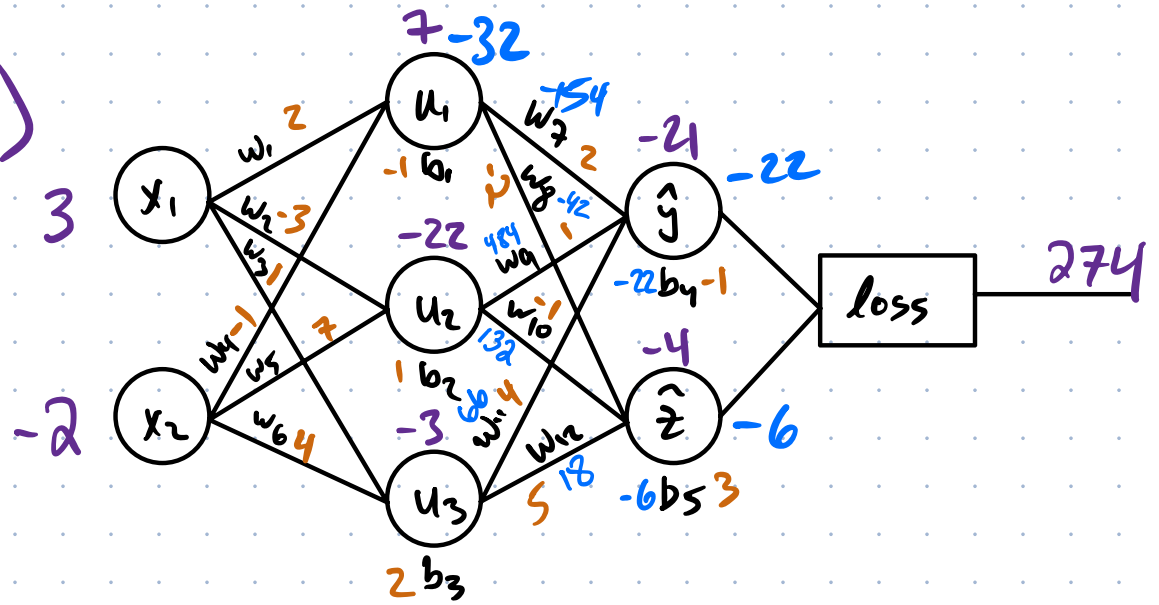


Sample: $(3, -2) \rightarrow (1, 4)$

Work backward from the right.

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

?



u_1 affects the loss along two paths:

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

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

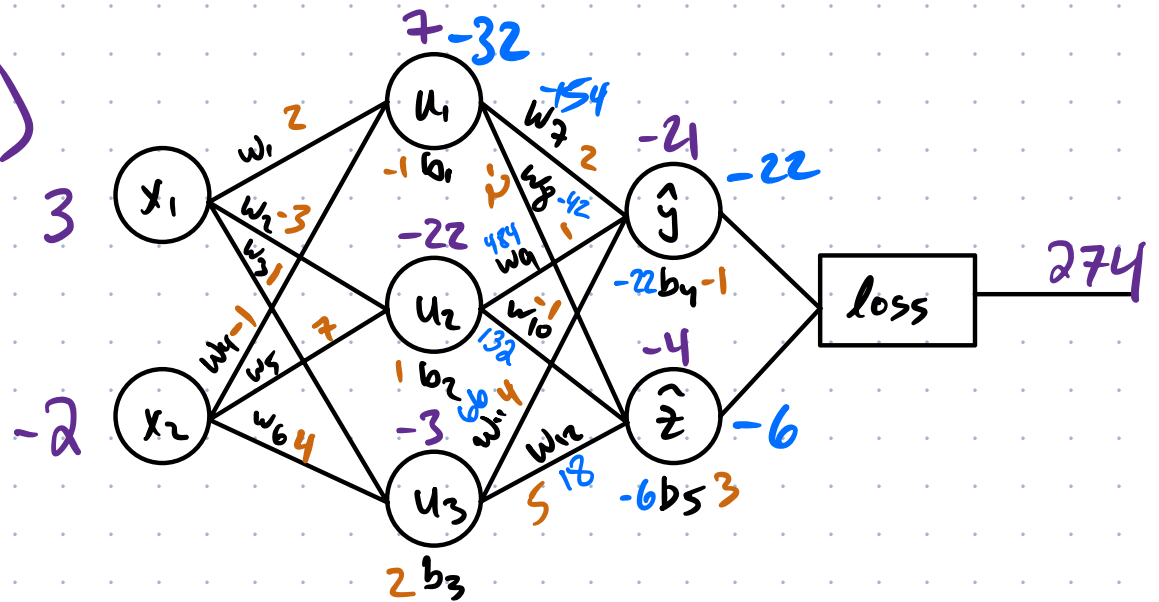
So we have the multivariate chain rule now:

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

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

Work backward from the right.

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



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

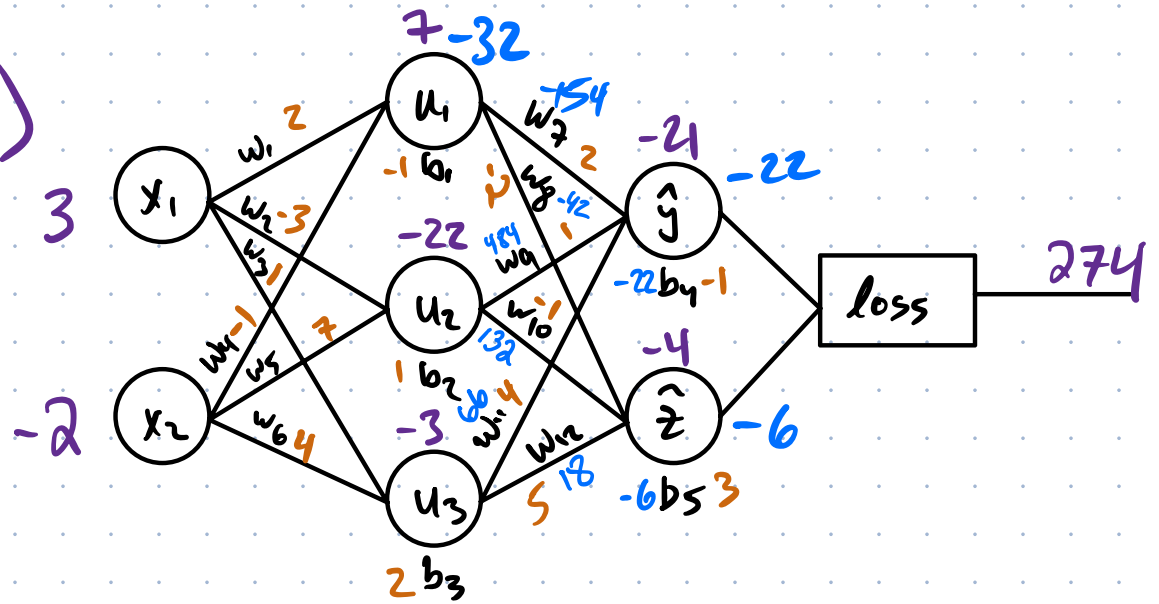
So we have the multivariate chain rule now:

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

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

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

Work backward from the right.



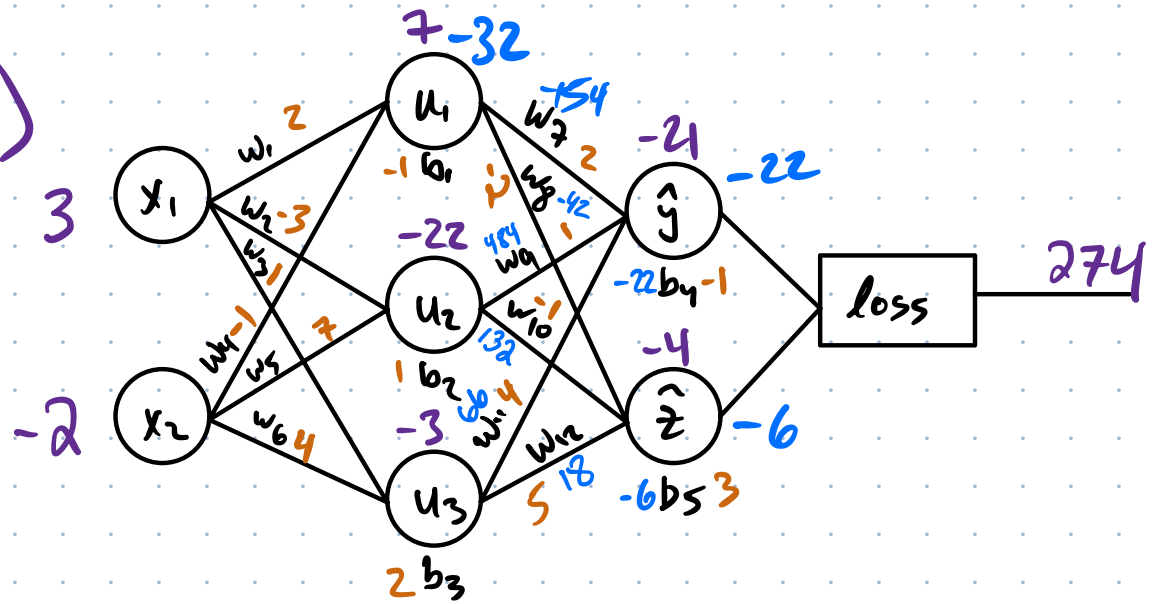
$$\frac{\partial l}{\partial u_1} = w_7 \cdot \frac{\partial l}{\partial \hat{y}} + w_9 \cdot \frac{\partial l}{\partial \hat{z}} = \begin{bmatrix} w_7 \\ w_9 \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial l}{\partial \hat{y}} \\ \frac{\partial l}{\partial \hat{z}} \end{bmatrix}$$

$$\frac{\partial l}{\partial u_2} = \begin{bmatrix} w_8 \\ w_{10} \end{bmatrix} \begin{bmatrix} \frac{\partial l}{\partial \hat{y}} \\ \frac{\partial l}{\partial \hat{z}} \end{bmatrix}$$

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

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

Work backward from the right.



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

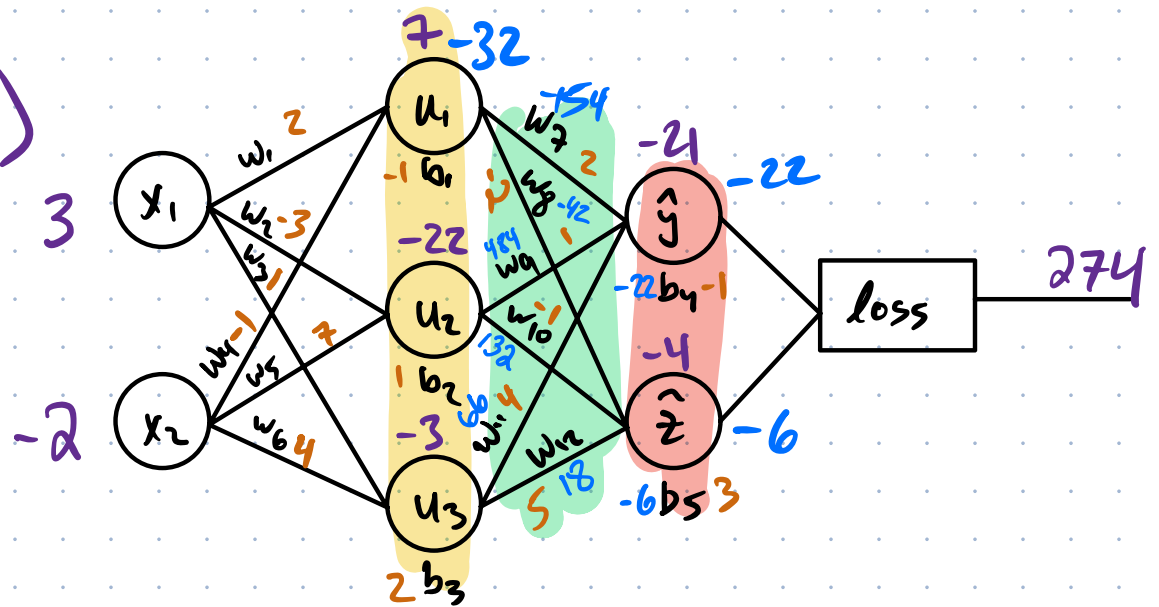
$$\frac{\partial l}{\partial u_2} = \begin{bmatrix} w_8 \\ w_{10} \end{bmatrix} \begin{bmatrix} \partial l / \partial \hat{y} \\ \partial l / \partial \hat{z} \end{bmatrix}$$

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

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

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

Work backward from the right.

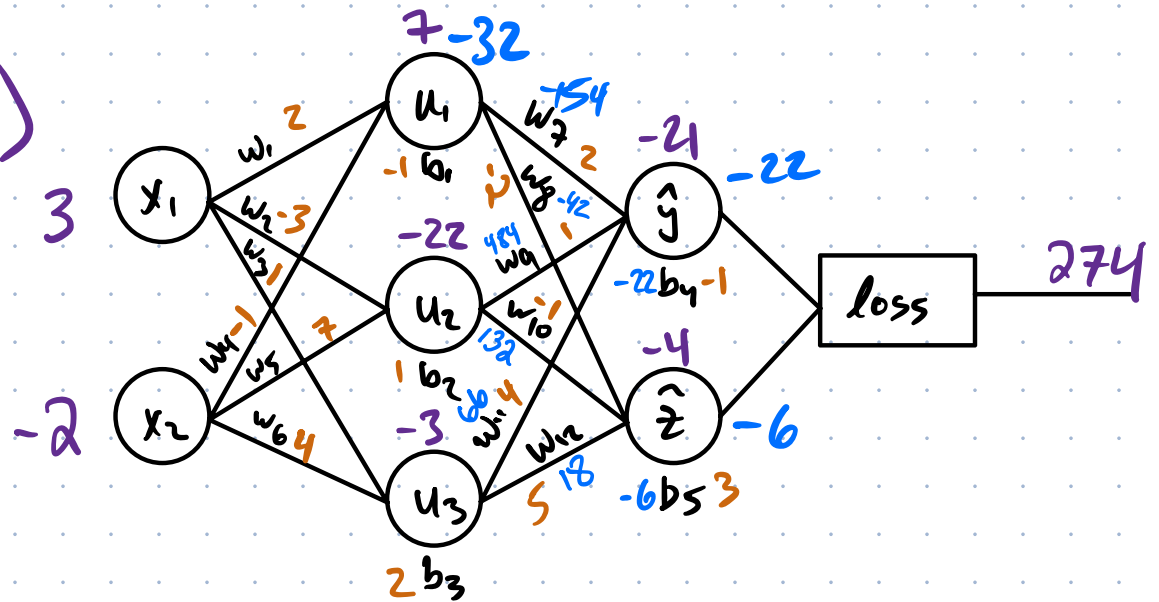


$$\begin{bmatrix} \partial l / \partial u_1 \\ \partial l / \partial u_2 \\ \partial l / \partial u_3 \end{bmatrix} = \begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix} \begin{bmatrix} \partial l / \partial \hat{y} \\ \partial l / \partial \hat{z} \end{bmatrix}$$

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

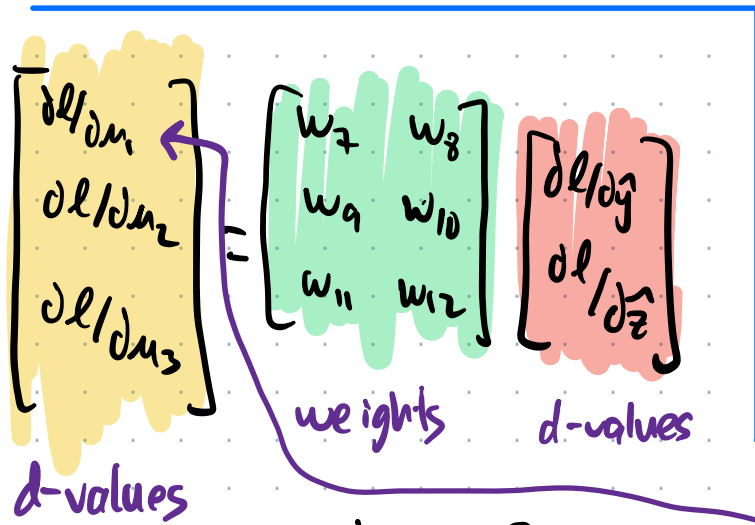
Work backward from the right.



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

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



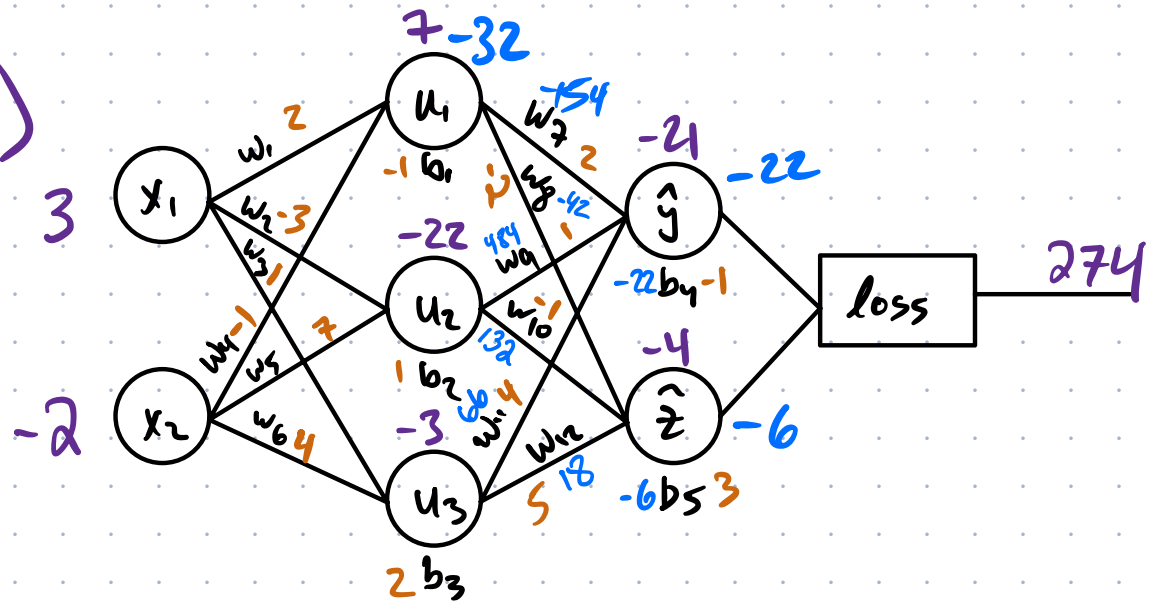
biases?

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

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

So

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



$u_1 = \dots + b$
 so, $\frac{du_1}{db} = 1$

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

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

d-values weights d-values

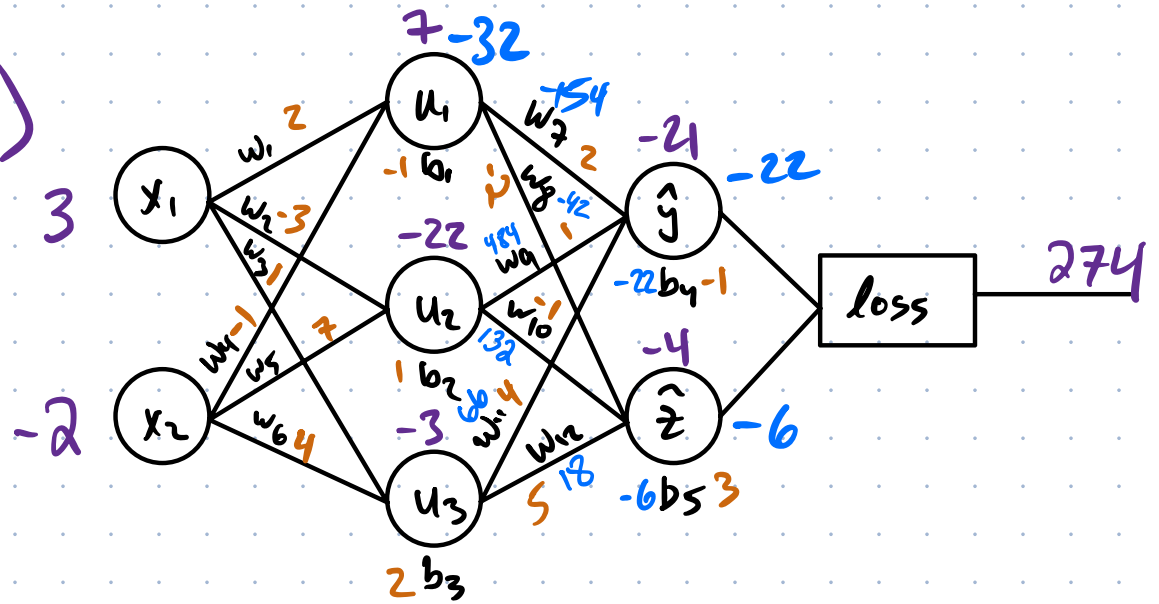
biases?

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

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

So

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



$u_1 = \dots + b$
 so, $\frac{\partial u_1}{\partial b} = 1$

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

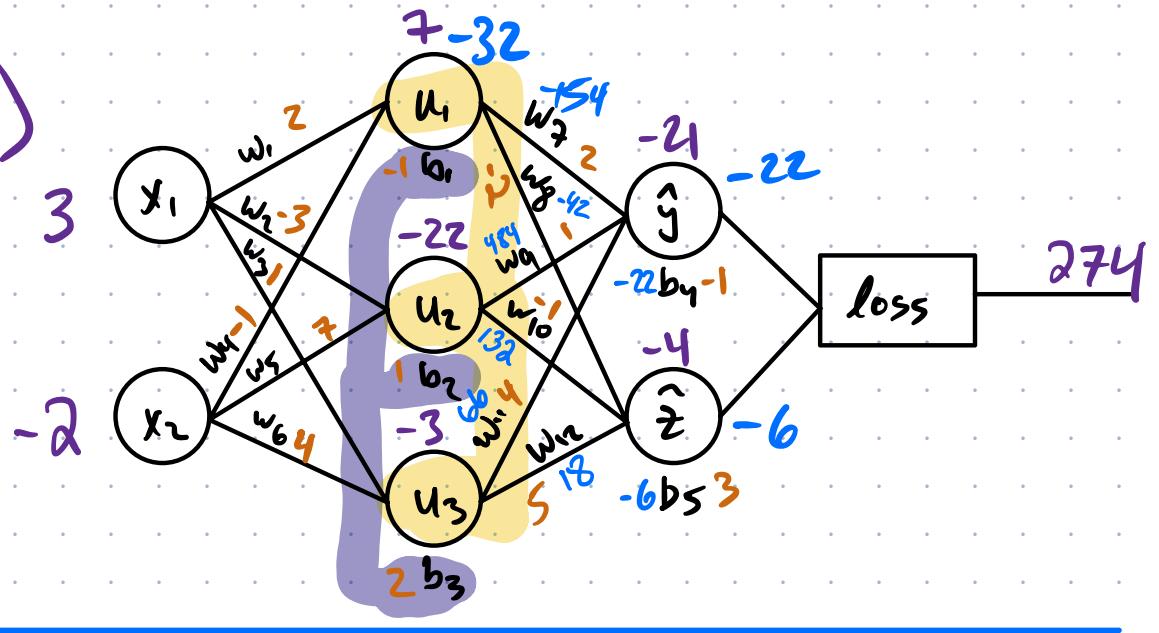
$$\begin{bmatrix} \partial l / \partial u_1 \\ \partial l / \partial u_2 \\ \partial l / \partial u_3 \end{bmatrix} = \begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix} \begin{bmatrix} \partial l / \partial \hat{y} \\ \partial l / \partial \hat{z} \end{bmatrix}$$

weights d-values

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

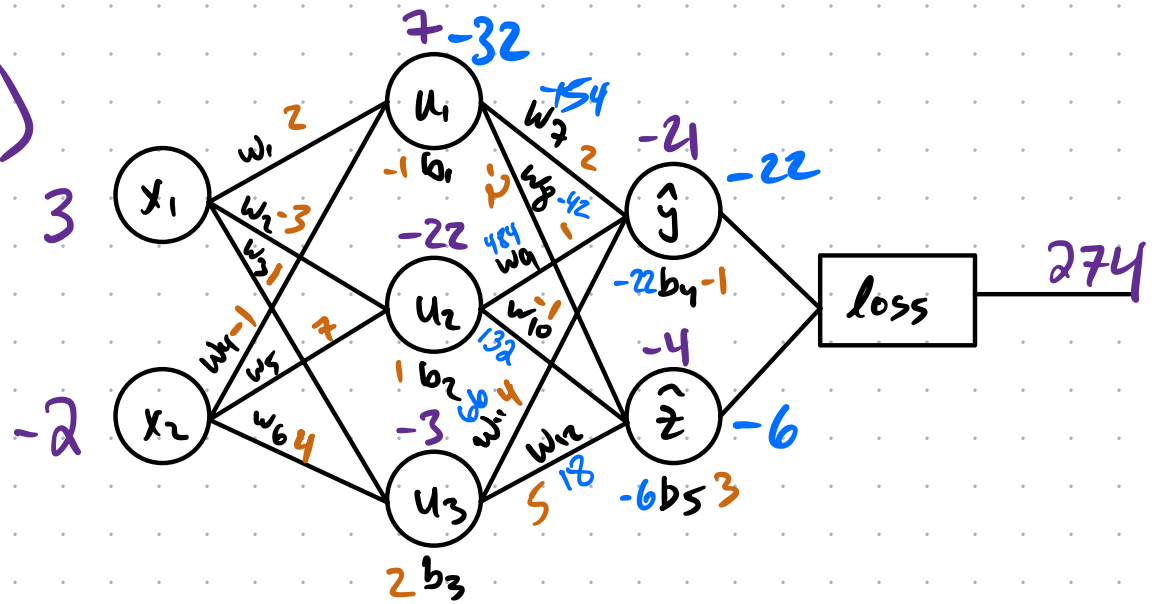
$$\begin{bmatrix} \partial l / \partial u_1 \\ \partial l / \partial u_2 \\ \partial l / \partial u_3 \end{bmatrix} = \begin{bmatrix} w_7 & w_8 \\ w_9 & w_{10} \\ w_{11} & w_{12} \end{bmatrix} \begin{bmatrix} \partial l / \partial \hat{y} \\ \partial l / \partial \hat{z} \end{bmatrix}$$

weights d-values

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

$$\frac{\partial l}{\partial w_1} = \frac{\partial l}{\partial u_1} \cdot \left(\frac{\partial u_1}{\partial w_1} \right) = x_1$$

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

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

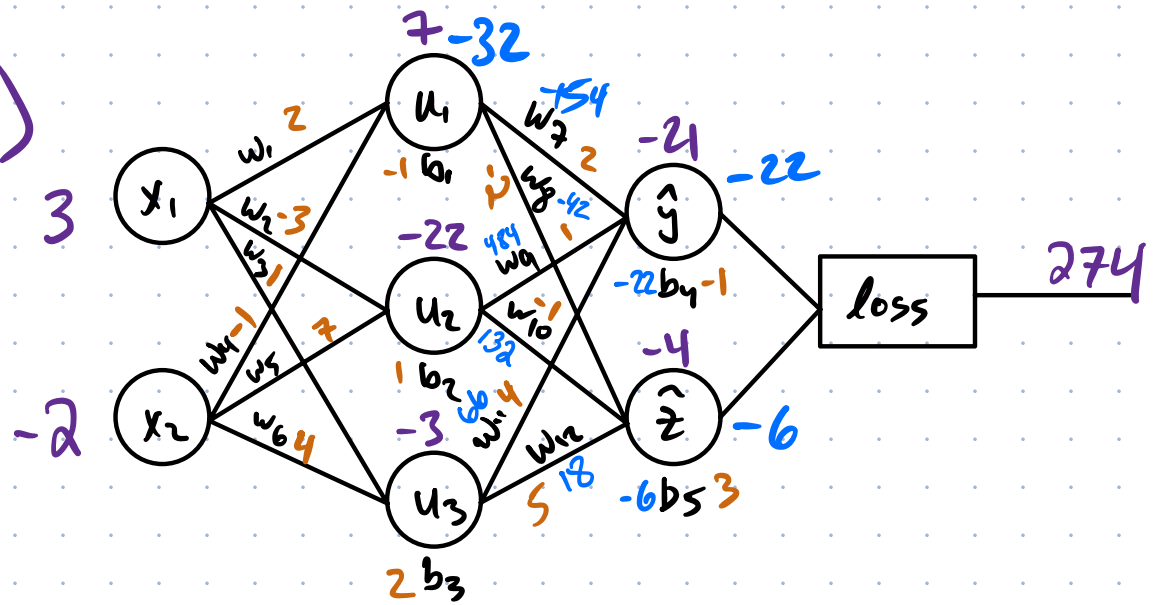
weights d-values

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?



$$\frac{\partial l}{\partial w_1} = \frac{\partial l}{\partial u_1} \cdot \left(\frac{\partial u_1}{\partial w_1} \right) = x_1$$

$$\frac{\partial l}{\partial w_2} = \frac{\partial l}{\partial u_2} \cdot \left(\frac{\partial u_2}{\partial w_2} \right) = x_1$$

$$\frac{\partial l}{\partial w_3} = \frac{\partial l}{\partial u_3} \cdot \left(\frac{\partial u_3}{\partial w_3} \right) = x_1$$

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

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

weights d-values

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

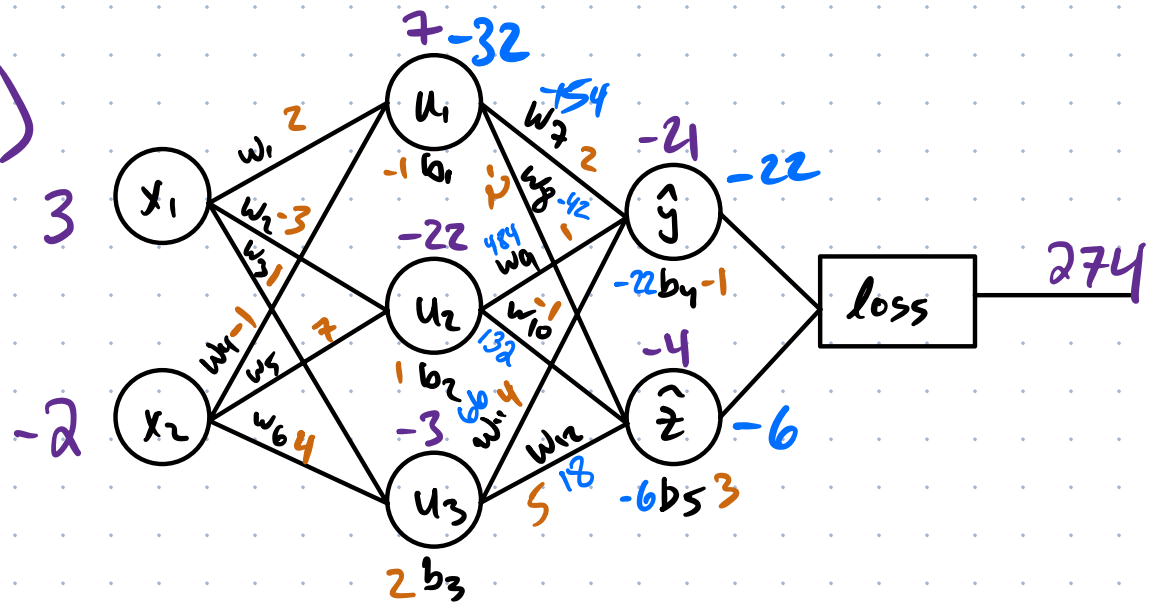
Weights?

$$\begin{aligned} \frac{\partial l}{\partial w_1} &= \frac{\partial l}{\partial u_1} \cdot \left(\frac{\partial u_1}{\partial w_1} \right) = x_1 \\ \frac{\partial l}{\partial w_2} &= \frac{\partial l}{\partial u_1} \cdot \left(\frac{\partial u_1}{\partial w_2} \right) = x_1 \\ \frac{\partial l}{\partial w_3} &= \frac{\partial l}{\partial u_1} \cdot \left(\frac{\partial u_1}{\partial w_3} \right) = x_1 \end{aligned}$$

$$\frac{\partial l}{\partial w_4} = \frac{\partial l}{\partial u_2} \cdot x_2$$

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

$$\frac{\partial l}{\partial w_6} = \frac{\partial l}{\partial u_2} \cdot x_2$$



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

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

weights d-values

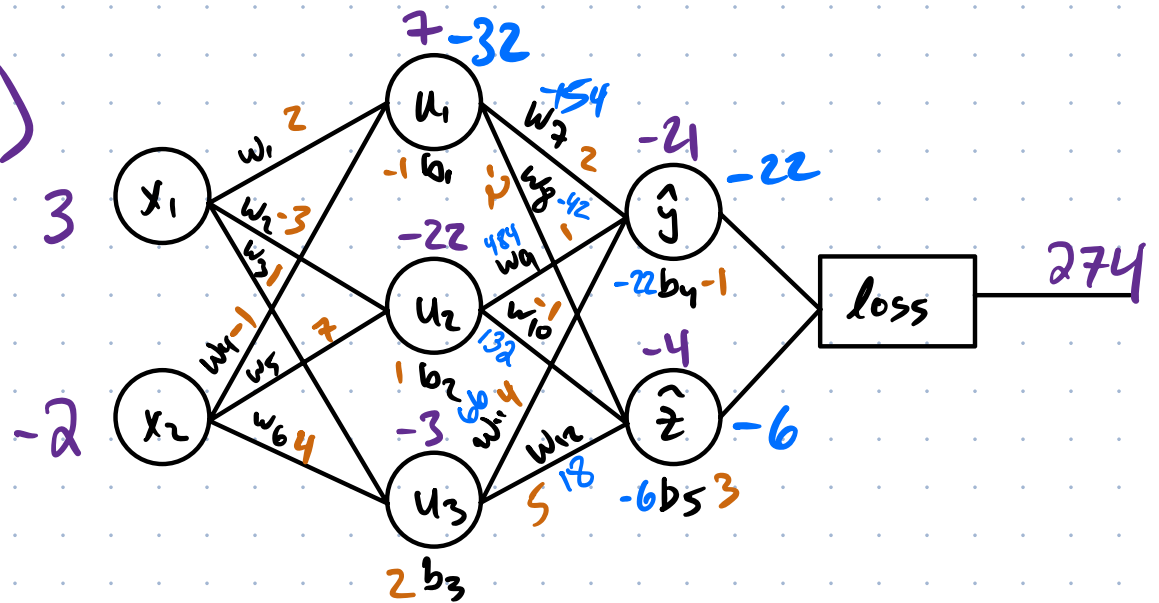
d-values

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

d-biases d-values

$$\begin{aligned} \frac{\partial l}{\partial w_1} &= \frac{\partial l}{\partial u_1} \cdot \left(\frac{\partial u_1}{\partial w_1} \right) = x_1 & \frac{\partial l}{\partial w_4} &= \frac{\partial l}{\partial u_1} \cdot x_2 \\ \frac{\partial l}{\partial w_2} &= \frac{\partial l}{\partial u_1} \cdot \left(\frac{\partial u_1}{\partial w_2} \right) = x_1 & \frac{\partial l}{\partial w_5} &= \frac{\partial l}{\partial u_1} \cdot x_2 \\ \frac{\partial l}{\partial w_3} &= \frac{\partial l}{\partial u_1} \cdot \left(\frac{\partial u_1}{\partial w_3} \right) = x_1 & \frac{\partial l}{\partial w_6} &= \frac{\partial l}{\partial u_1} \cdot x_2 \end{aligned}$$

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

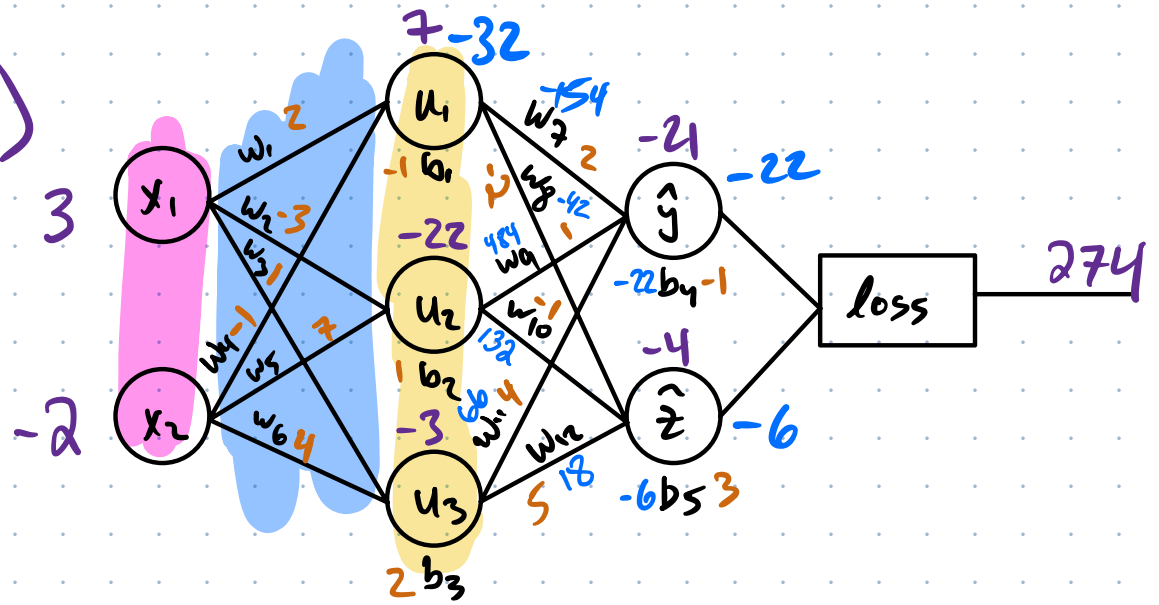


Sample: $(3, -2) \rightarrow (1, 4)$

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

weights d-values

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

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

d-weights d-values values

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

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

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.

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

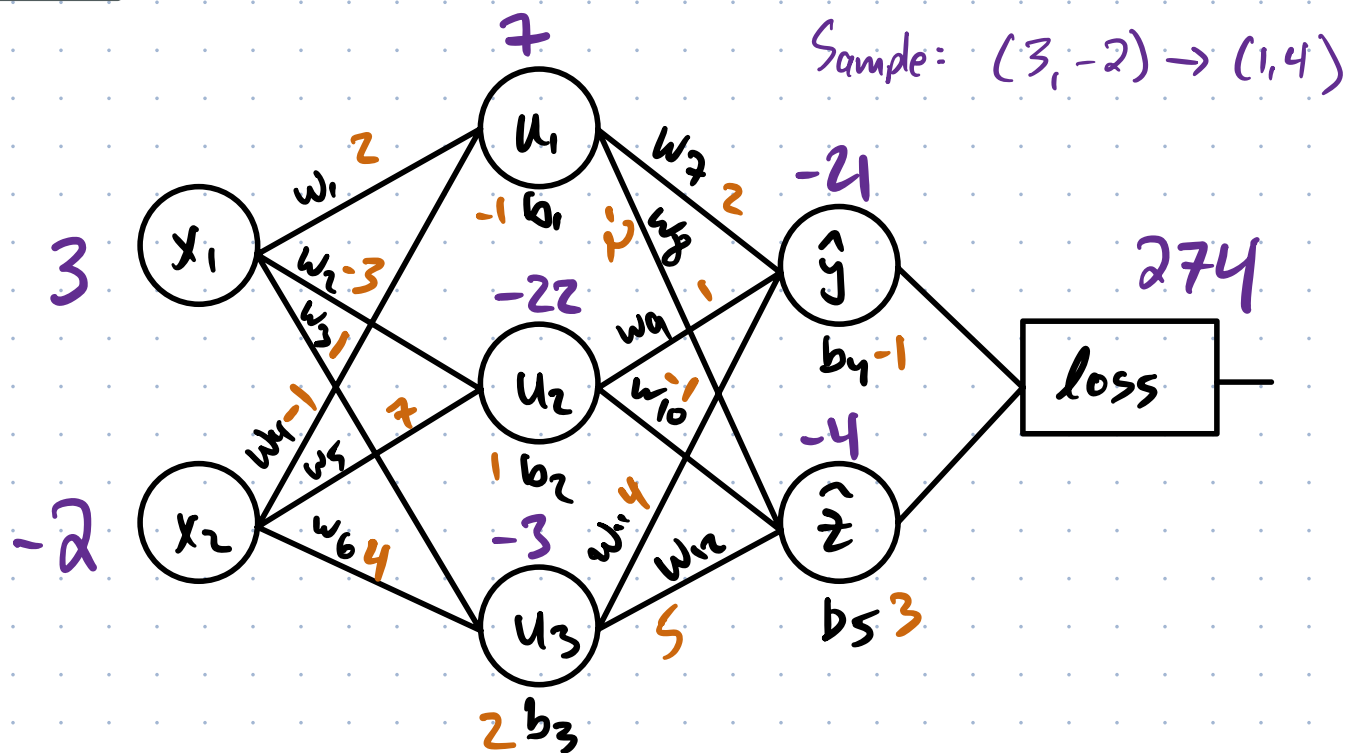
d-values weights d-values

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

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



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

d-values weights d-values

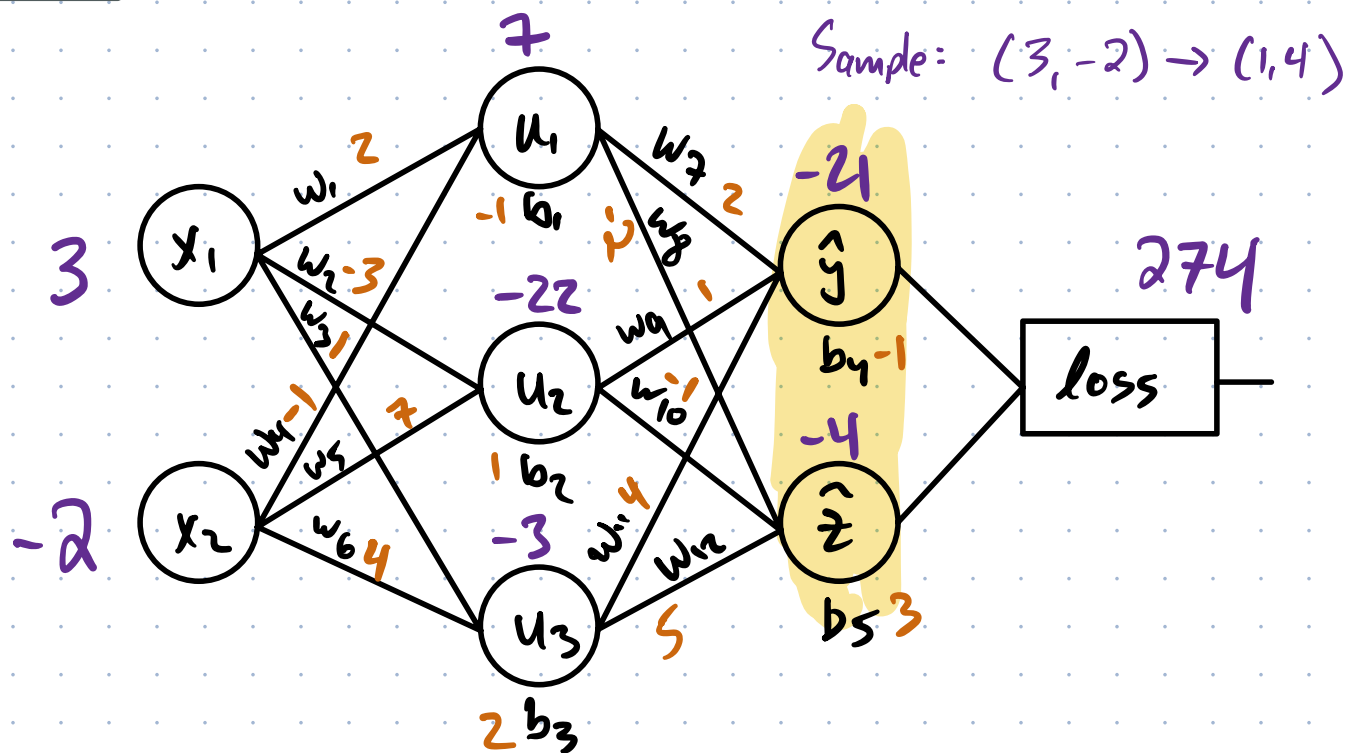
$$\begin{bmatrix} \partial l / \partial w_1 & \partial l / \partial w_4 \\ \partial l / \partial w_2 & \partial l / \partial w_5 \\ \partial l / \partial w_3 & \partial l / \partial w_6 \end{bmatrix} = \begin{bmatrix} \partial l / \partial u_1 \\ \partial l / \partial u_2 \\ \partial l / \partial u_3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

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

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



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

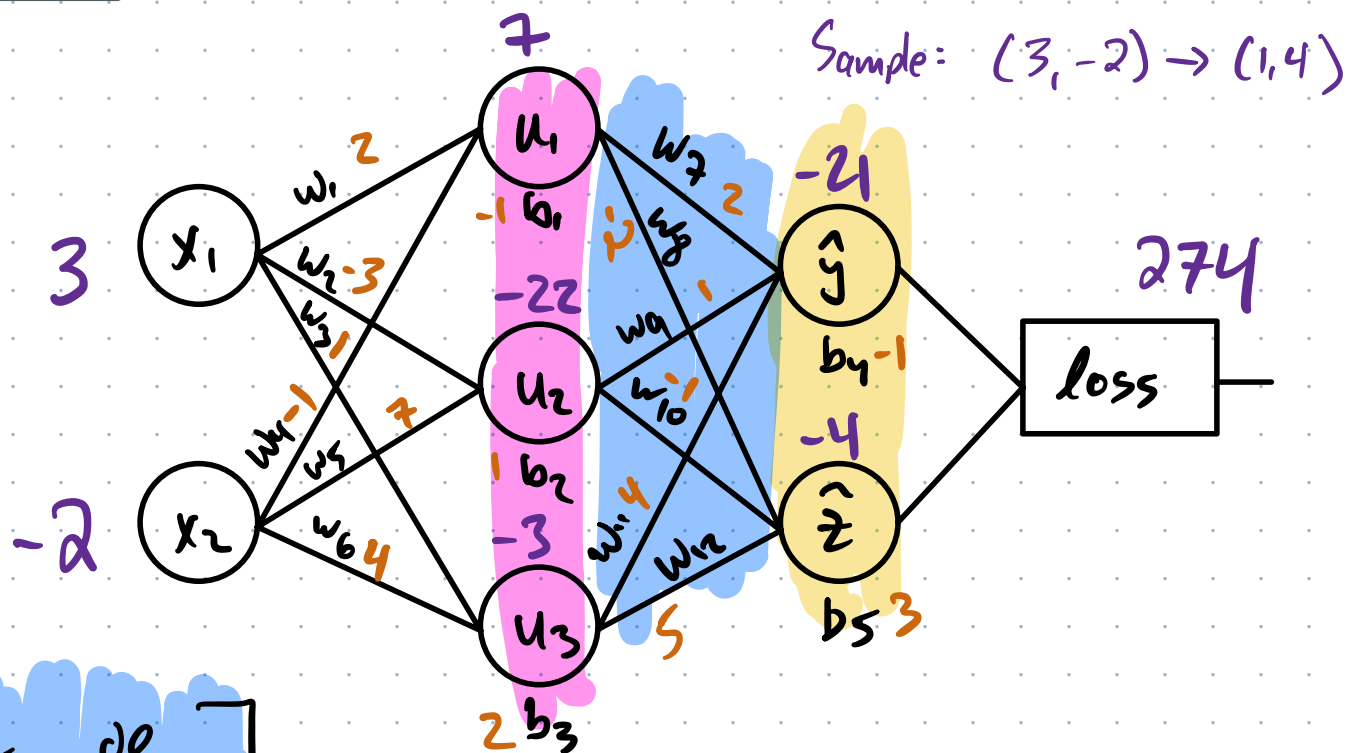
d-values weights d-values

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

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



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

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

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

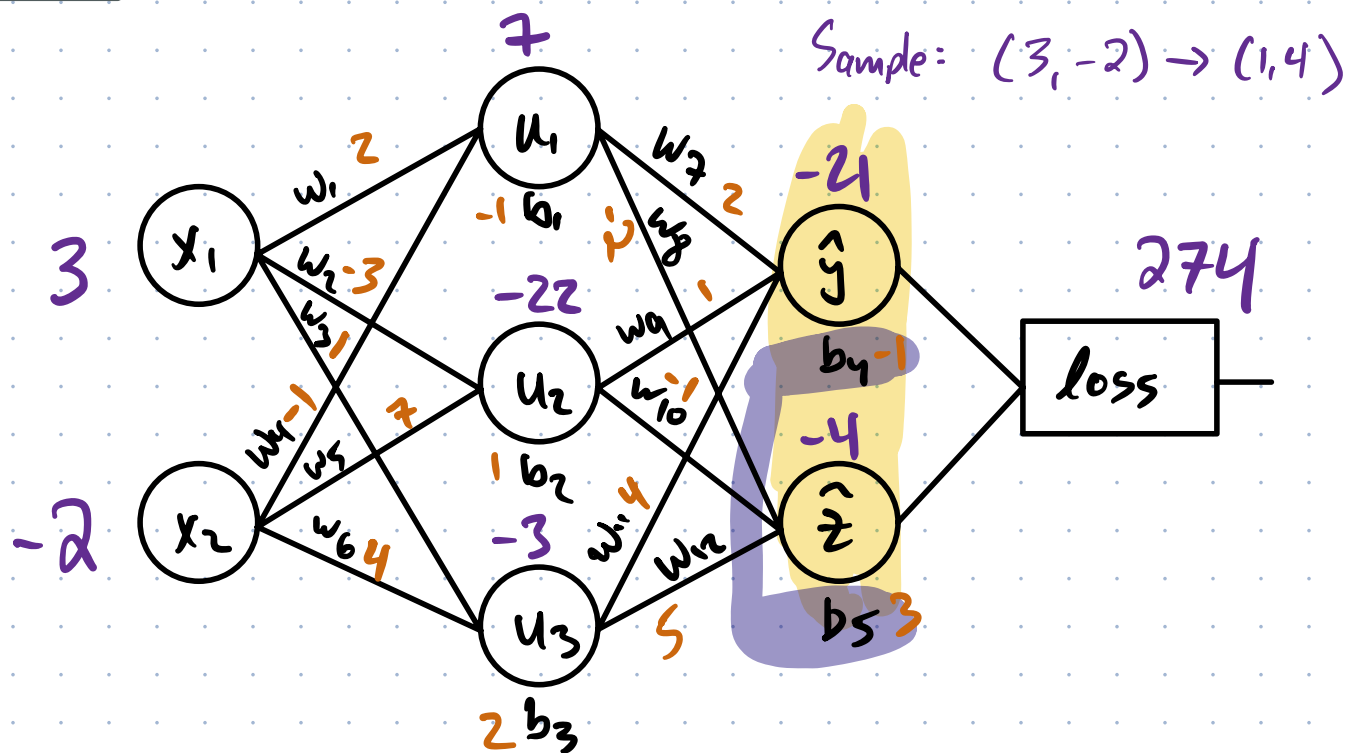
d-values weights d-values

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

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



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

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

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

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

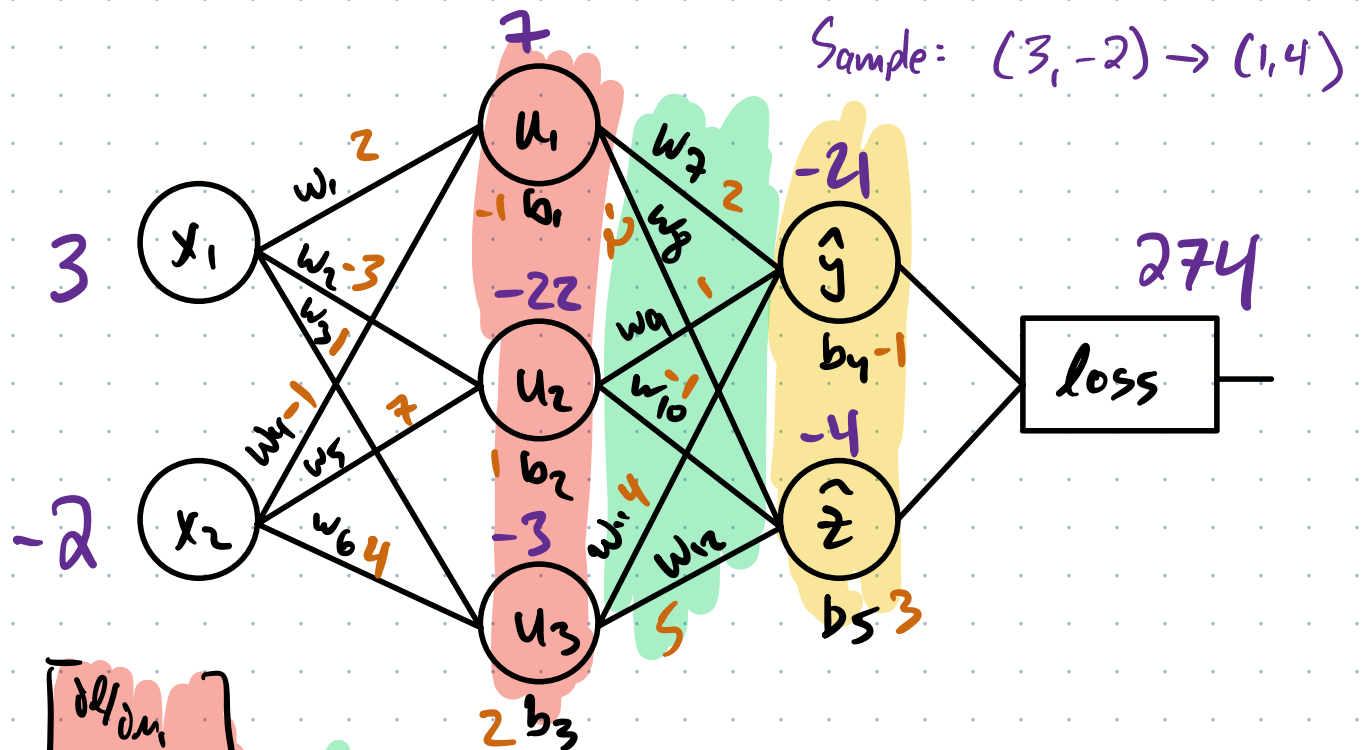
d-values weights d-values

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

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



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

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

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

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

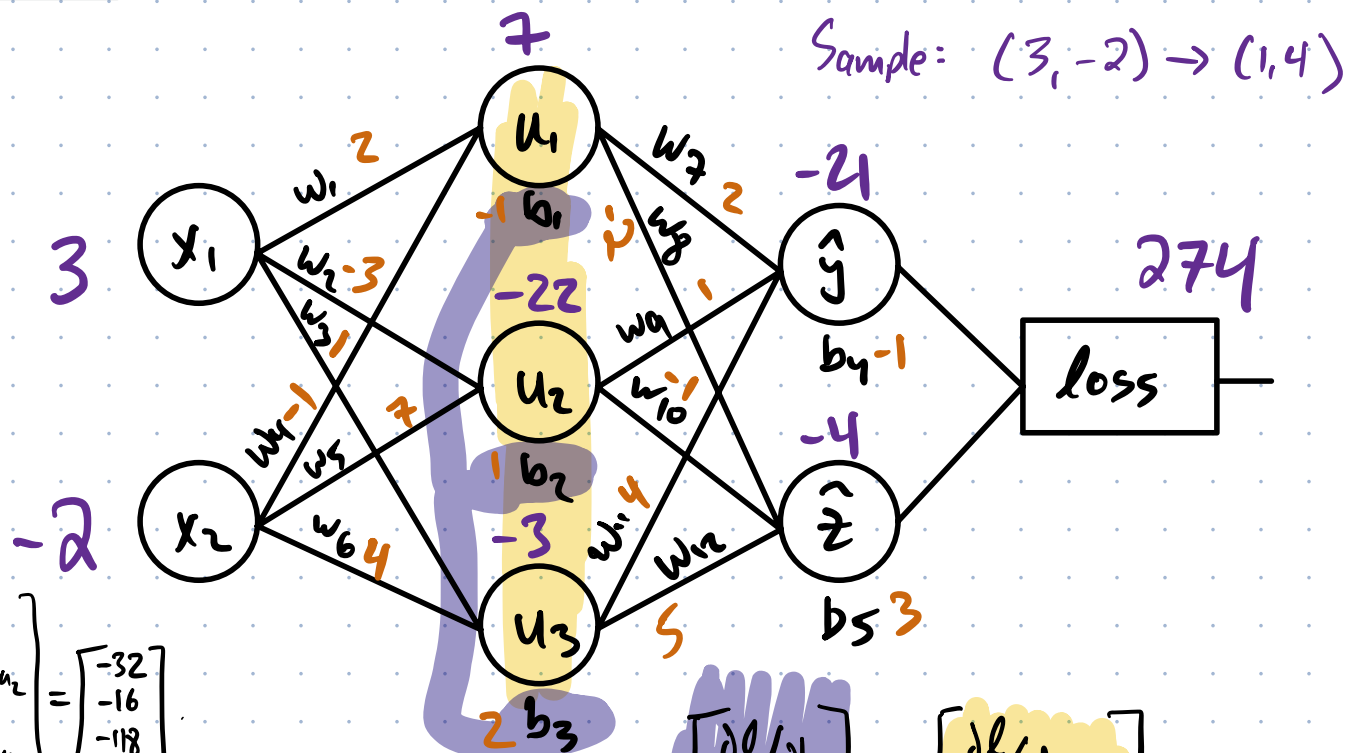
d-values weights d-values

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

d-weights d-values values

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

d-biases d-values



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

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

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

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

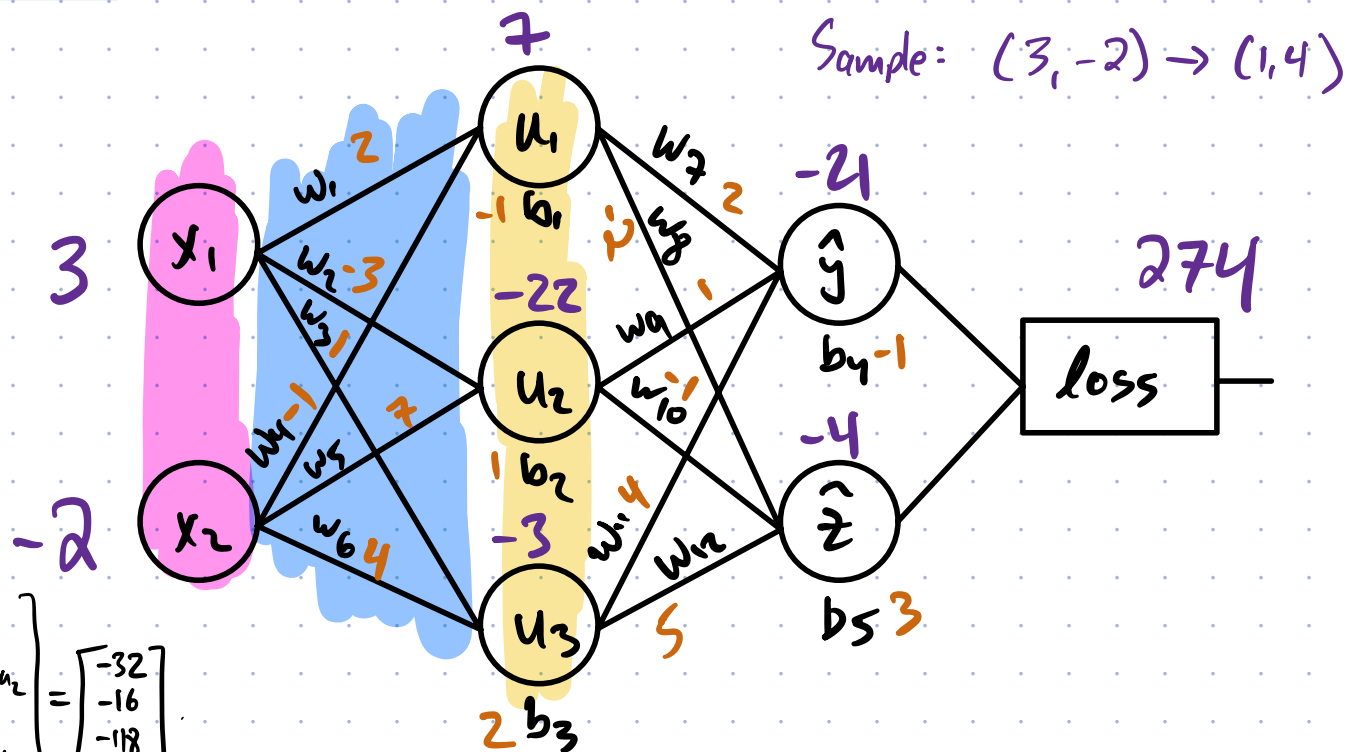
d-values weights d-values

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

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



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

$$\begin{bmatrix} \frac{\partial L}{\partial w_7} & \frac{\partial L}{\partial w_8} & \frac{\partial L}{\partial w_{11}} \\ \frac{\partial L}{\partial w_9} & \frac{\partial L}{\partial w_{10}} & \frac{\partial L}{\partial w_{12}} \end{bmatrix} = \begin{bmatrix} -154 & 484 & 66 \\ -42 & 132 & 18 \end{bmatrix} \quad \begin{bmatrix} \partial L / \partial b_1 \\ \partial L / \partial b_2 \\ \partial L / \partial b_3 \end{bmatrix} = \begin{bmatrix} -32 \\ -16 \\ -118 \end{bmatrix} \quad \begin{bmatrix} \partial L / \partial w_1 & \partial L / \partial w_4 \\ \partial L / \partial w_2 & \partial L / \partial w_5 \\ \partial L / \partial w_3 & \partial L / \partial w_6 \end{bmatrix} = \begin{bmatrix} -32 \\ -16 \\ -118 \end{bmatrix} \quad \begin{bmatrix} 3 & -2 \end{bmatrix} = \begin{bmatrix} -96 & 64 \\ -48 & 32 \\ -354 & 236 \end{bmatrix}$$

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

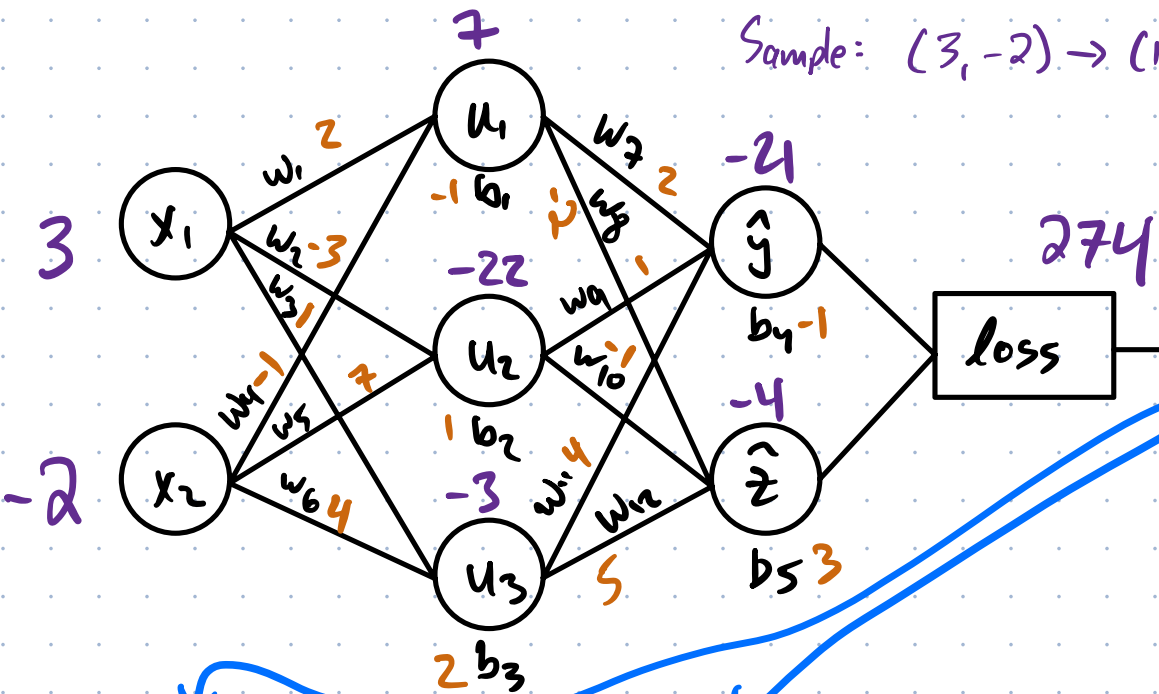
d-values weights d-values

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

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.

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

$$\begin{bmatrix} \frac{\partial L}{\partial w_7} & \frac{\partial L}{\partial w_8} & \frac{\partial L}{\partial w_{11}} \\ \frac{\partial L}{\partial w_9} & \frac{\partial L}{\partial w_{10}} & \frac{\partial L}{\partial w_{12}} \end{bmatrix} = \begin{bmatrix} -154 & 484 & 66 \\ -42 & 132 & 18 \end{bmatrix} \quad \begin{bmatrix} \partial L / \partial b_1 \\ \partial L / \partial b_2 \\ \partial L / \partial b_3 \end{bmatrix} = \begin{bmatrix} -32 \\ -16 \\ -118 \end{bmatrix} \quad \begin{bmatrix} \partial L / \partial w_1 & \partial L / \partial w_4 \\ \partial L / \partial w_2 & \partial L / \partial w_5 \\ \partial L / \partial w_3 & \partial L / \partial w_6 \end{bmatrix} = \begin{bmatrix} -96 & 64 \\ -48 & 32 \\ -354 & 236 \end{bmatrix}$$

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

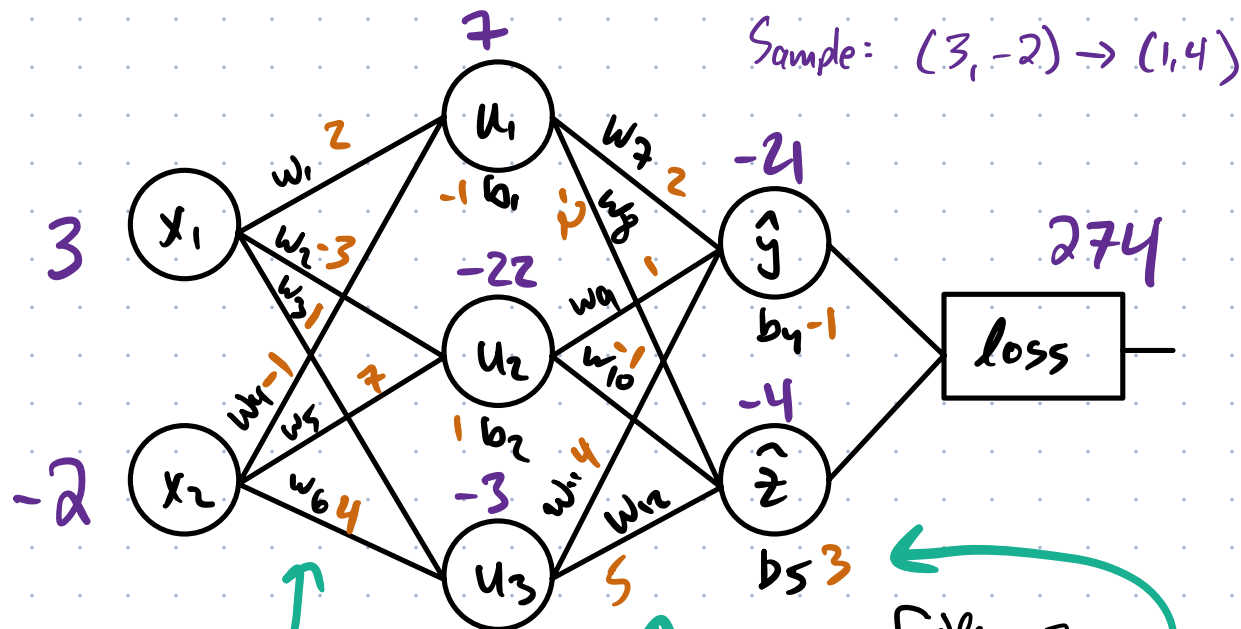
d-values weights d-values

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

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



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

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

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

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

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

d-values weights d-values

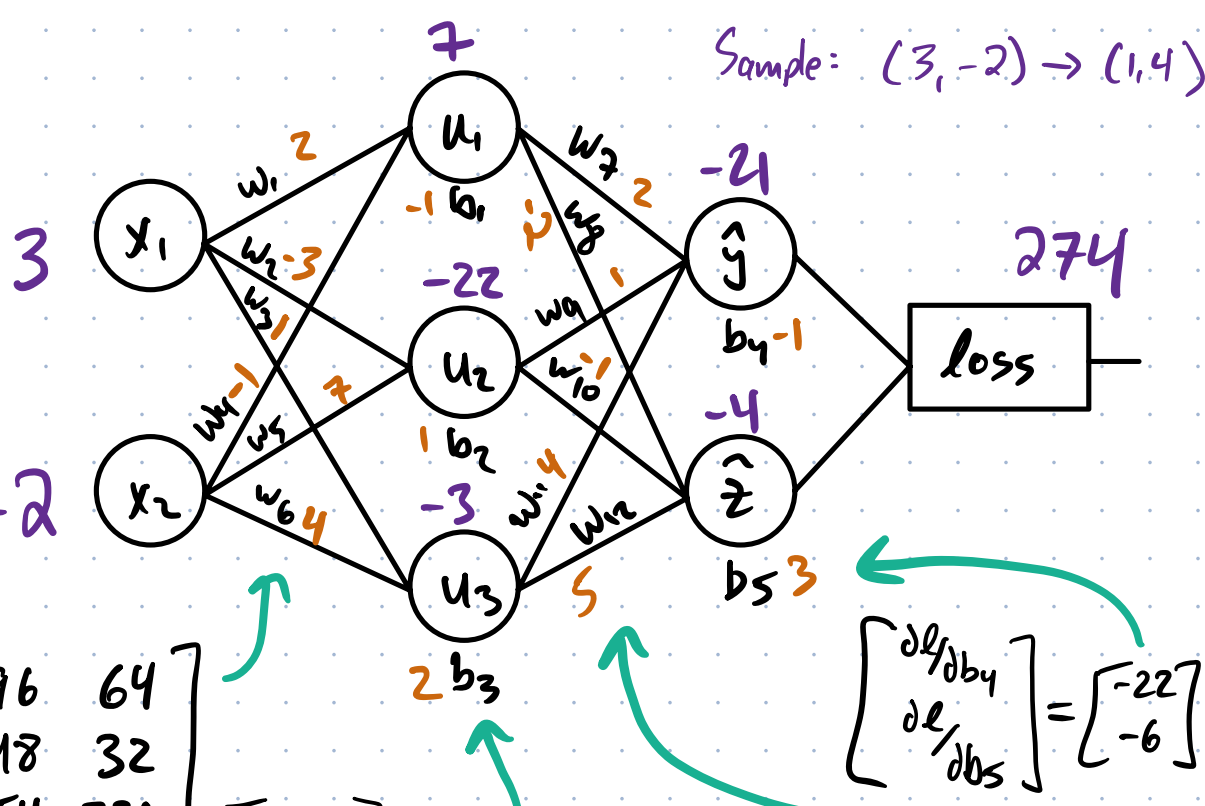
$$\begin{bmatrix} \partial L / \partial w_1 & \partial L / \partial w_4 \\ \partial L / \partial w_2 & \partial L / \partial w_5 \\ \partial L / \partial w_3 & \partial L / \partial w_6 \end{bmatrix} = \begin{bmatrix} \partial L / \partial u_1 \\ \partial L / \partial u_2 \\ \partial L / \partial u_3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

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

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



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

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

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

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".
- * How to structure a network for your particular task.