Backpropagation

Hung-yi Lee
李宏毅

Gradient Descent

Network parameters $\theta = \{w_1, w_2, \cdots, b_1, b_2, \cdots\}$

Starting Parameters $\quad \theta^0 \longrightarrow \theta^1 \longrightarrow \theta^2 \longrightarrow \ldots\ldots$

$$\nabla L(\theta) = \begin{bmatrix} \partial L(\theta)/\partial w_1 \\ \partial L(\theta)/\partial w_2 \\ \vdots \\ \partial L(\theta)/\partial b_1 \\ \partial L(\theta)/\partial b_2 \\ \vdots \end{bmatrix}$$

Compute $\nabla L(\theta^0)$ $\qquad \theta^1 = \theta^0 - \eta \nabla L(\theta^0)$

Compute $\nabla L(\theta^1)$ $\qquad \theta^2 = \theta^1 - \eta \nabla L(\theta^1)$

Millions of parameters ……

To compute the gradients efficiently, we use ***backpropagation***.

Chain Rule

Some preliminaries

Case 1

$$y = g(x) \qquad z = h(y)$$

$$\Delta x \xrightarrow{g} \Delta y \xrightarrow{h} \Delta z \qquad\qquad \frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$$

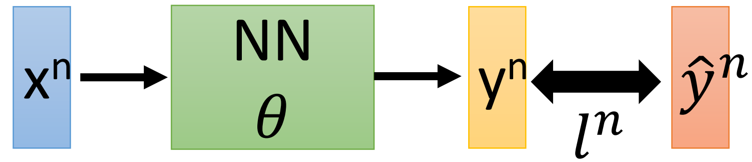
Case 2

$$x = g(s) \qquad y = h(s) \qquad z = k(x, y)$$

$$\begin{tikzcd} & \Delta x \arrow[dr] & \\ \Delta s \arrow[ur, "g"] \arrow[dr, "h"'] & & \Delta z \\ & \Delta y \arrow[ur, "k"] & \end{tikzcd}$$

$$\frac{dz}{ds} = \frac{\partial z}{\partial x}\frac{dx}{ds} + \frac{\partial z}{\partial y}\frac{dy}{ds}$$

Backpropagation

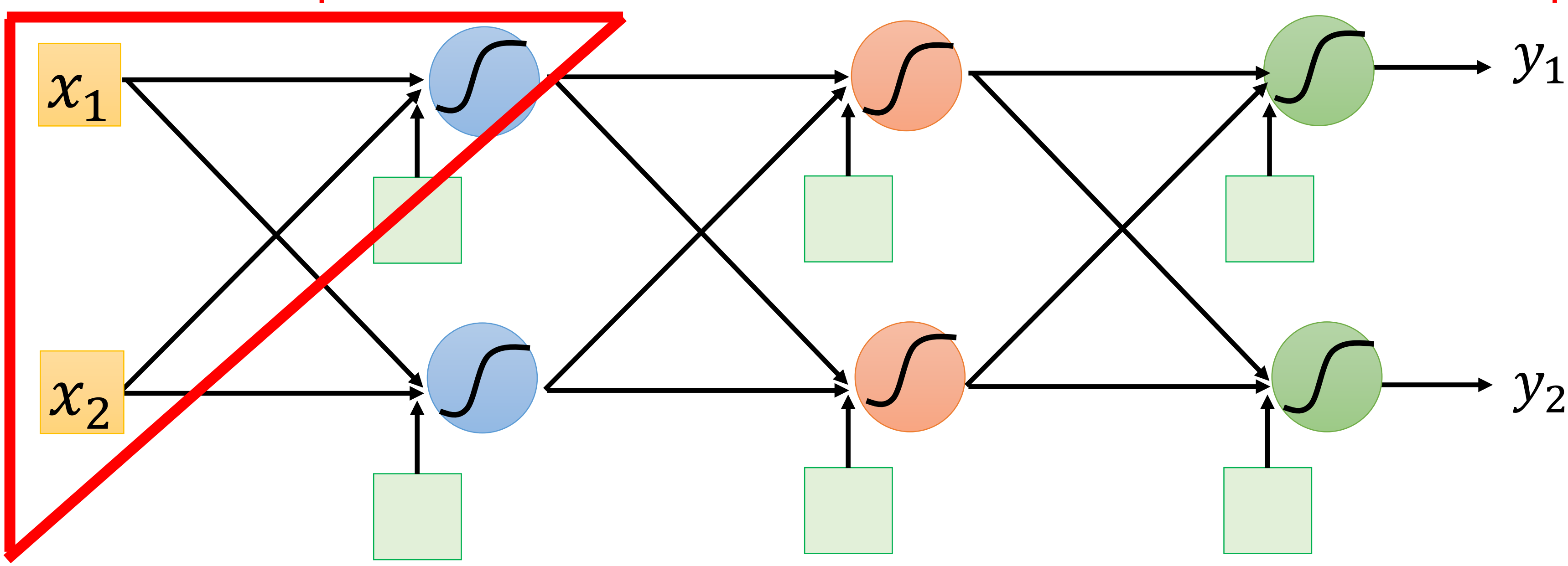


Ignore the superscript n here.
(One data at a time.)

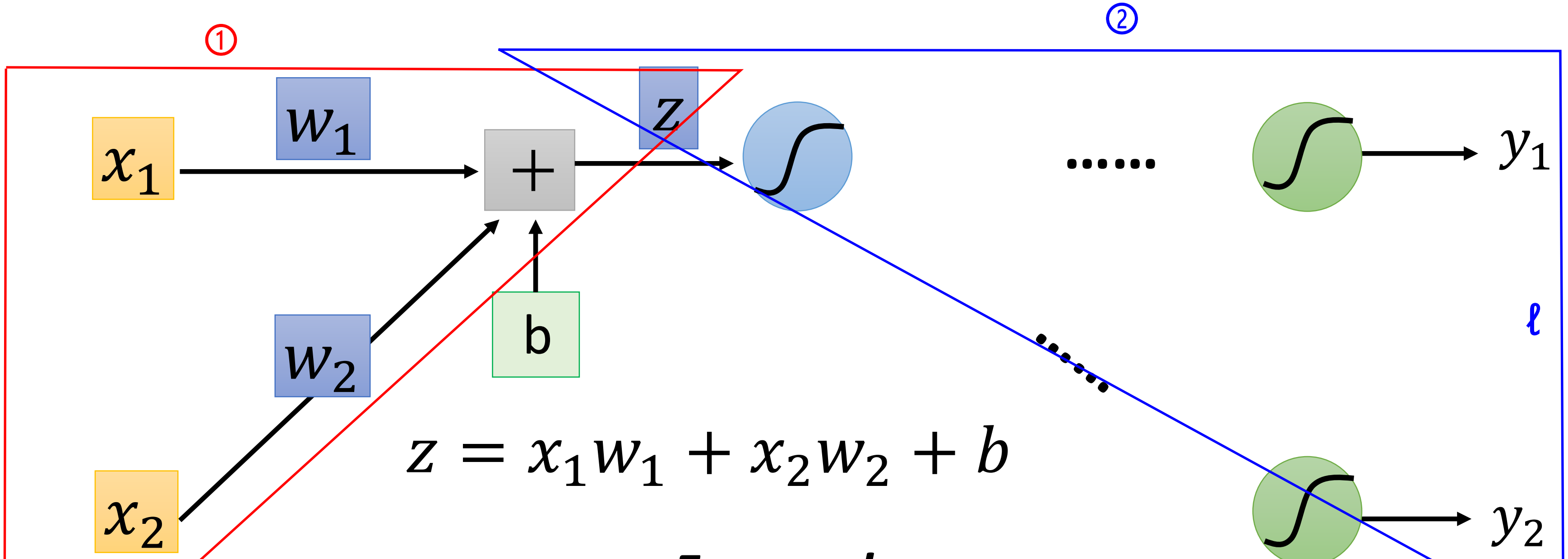
$$L(\theta)=\sum_{n=1}^{N} l^n(\theta) \quad \Rightarrow \quad \frac{\partial L(\theta)}{\partial w}=\sum_{n=1}^{N}\boxed{\frac{\partial l^n(\theta)}{\partial w}}$$

We can just understand this part.

Look at this part first.



Backpropagation



$$z = x_1 w_1 + x_2 w_2 + b$$

$$\frac{\partial l}{\partial w} = ? \quad \frac{\partial z}{\partial w} \frac{\partial l}{\partial z}$$

⓪ ① ②

(Chain rule)

Forward pass: easy

Compute $\partial z / \partial w$ for all parameters

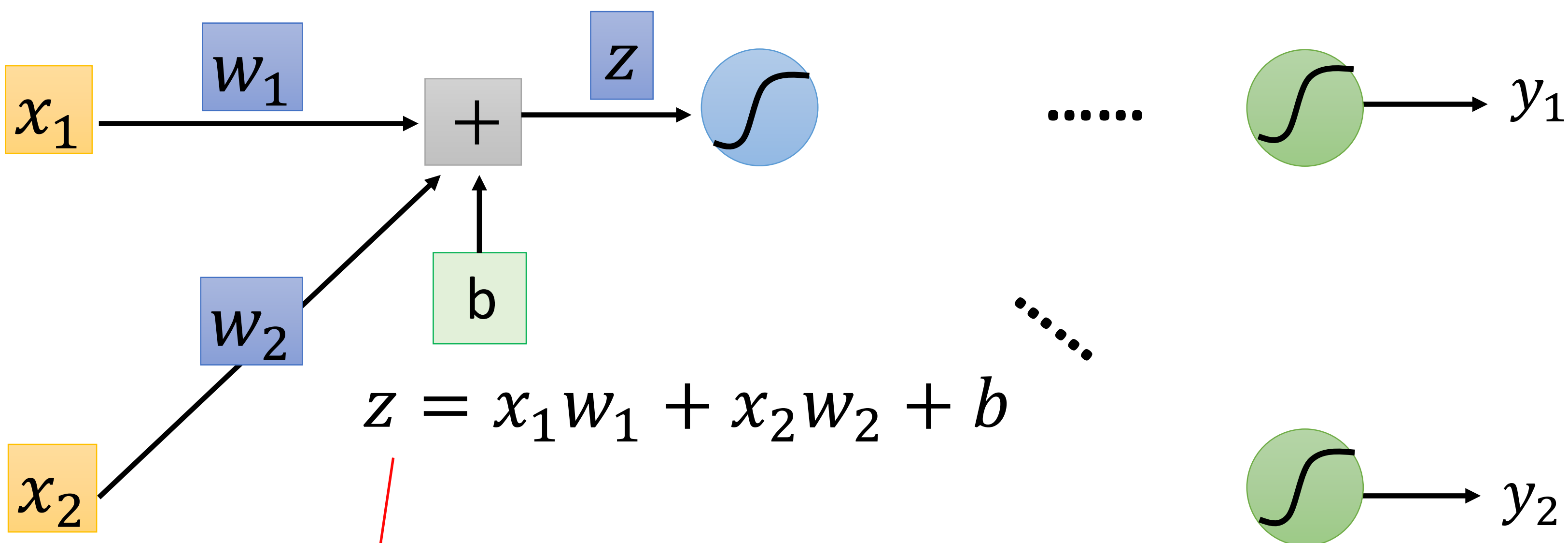
Backward pass: complex

Compute $\partial l / \partial z$ for all activation function inputs z

Backpropagation – Forward pass

①

Compute $\boxed{\partial z/\partial w}$ for all parameters



$$z = x_1 w_1 + x_2 w_2 + b$$

$\partial z/\partial w_1 =?\ \ x_1$

$\partial z/\partial w_2 =?\ \ x_2$

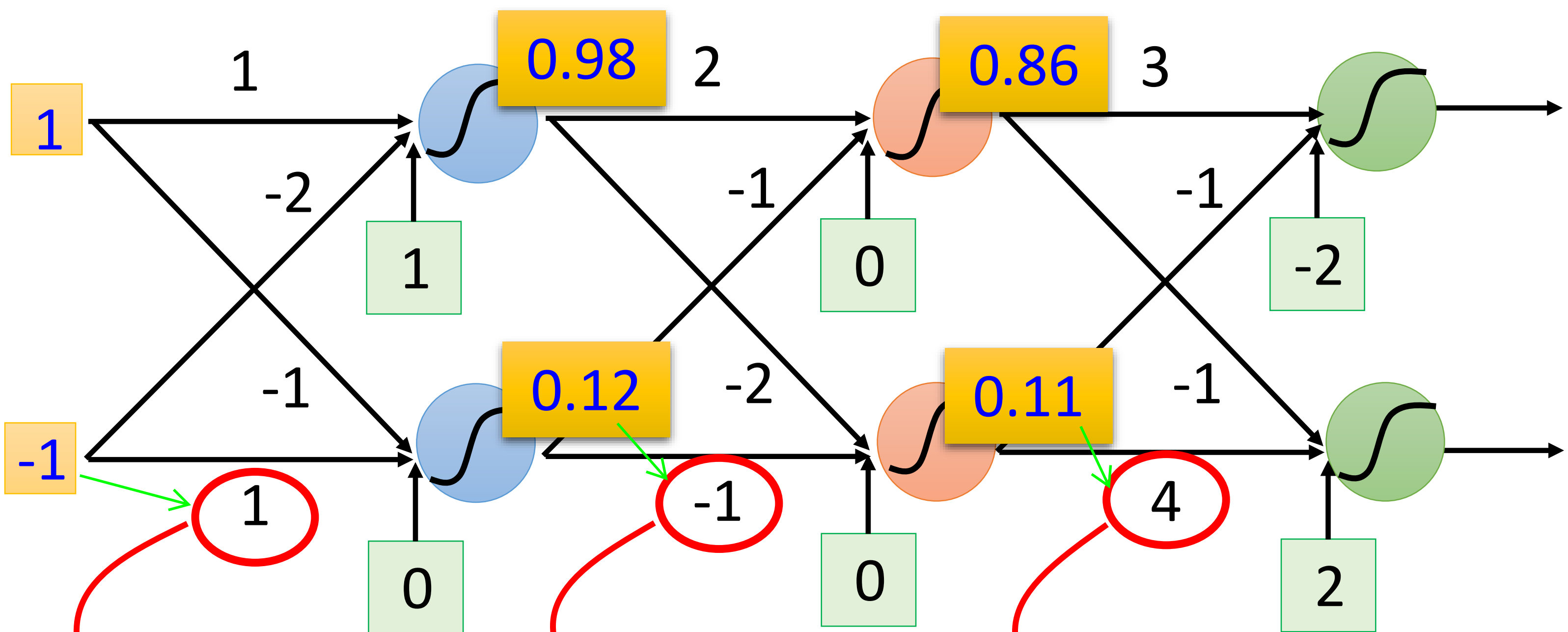
The value of the input connected by the weight

Backpropagation – Forward pass

①

Compute $\boxed{\partial z/\partial w}$ for all parameters

It's just the input.



$$\frac{\partial z}{\partial w} = -1$$

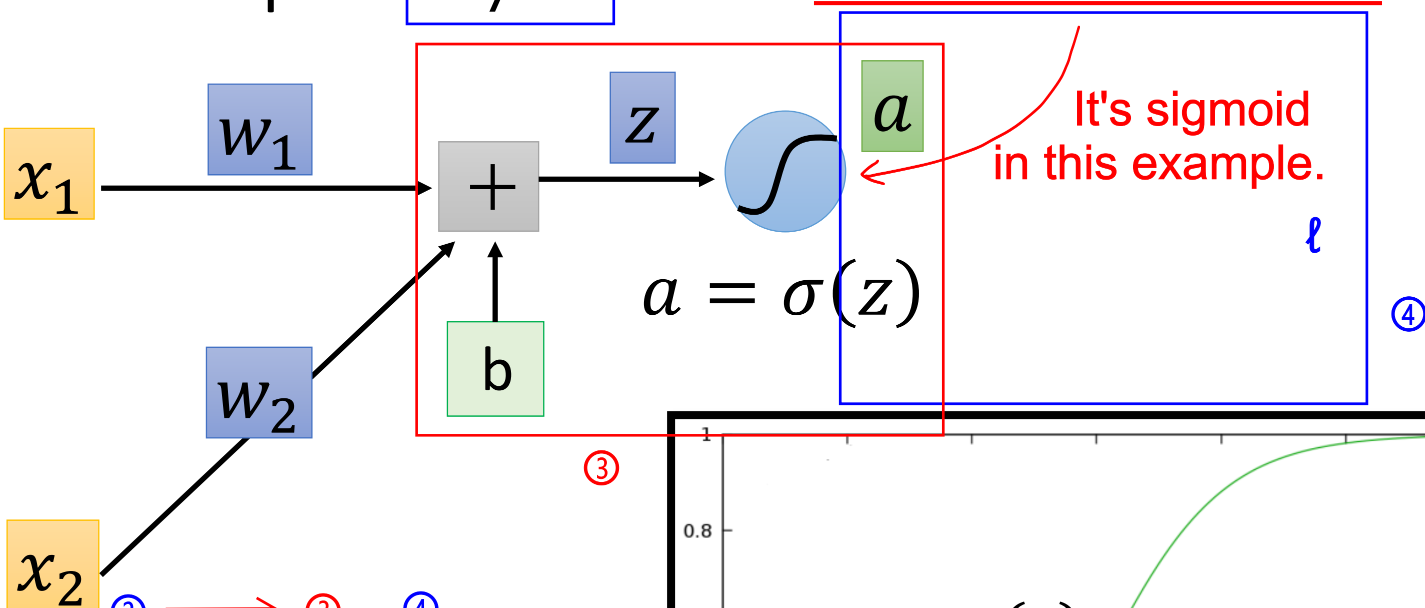
$$\frac{\partial z}{\partial w} = 0.12$$

$$\frac{\partial z}{\partial w} = 0.11$$

Backpropagation – Backward pass

②

Compute $\boxed{\partial l / \partial z}$ for all activation function inputs z ⇒ ③

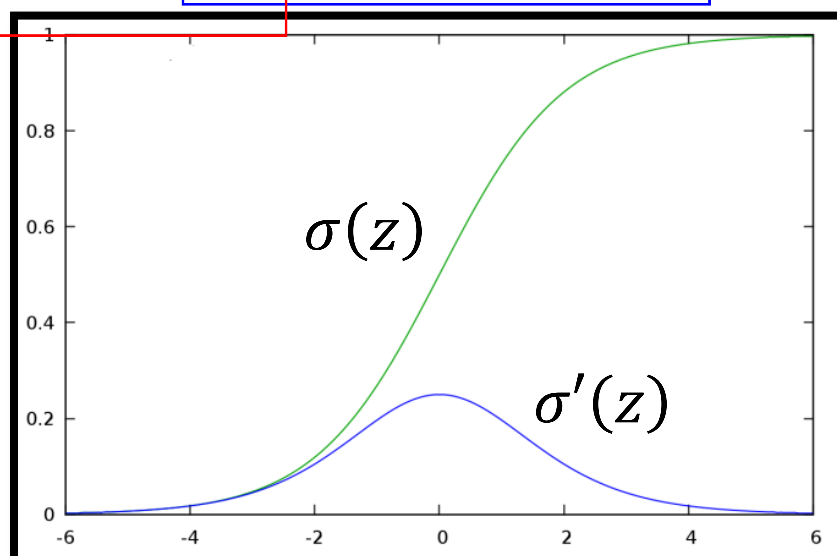


② → ③ ④

$$\boxed{\frac{\partial l}{\partial z}} = \boxed{\frac{\partial a}{\partial z}}\boxed{\frac{\partial l}{\partial a}}$$ complex

⇘ easy

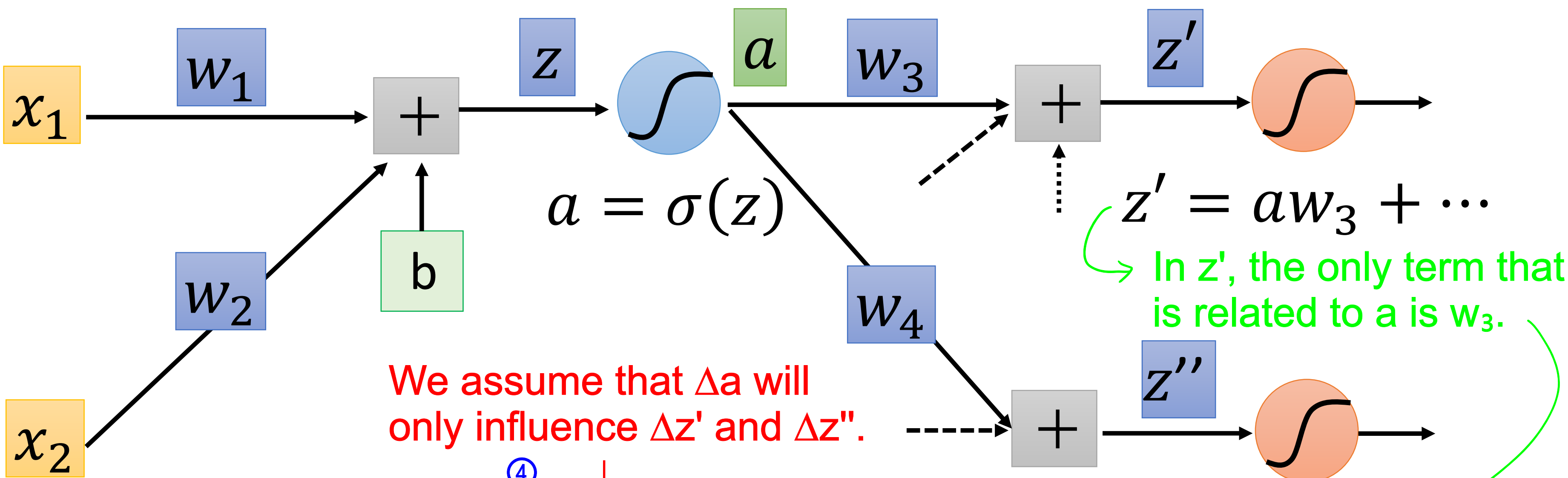
⇒ $\sigma'(z)$



Backpropagation – Backward pass

②

Compute $\boxed{\partial l / \partial z}$ from the output layer ⇒ ④



④

$$\frac{\partial l}{\partial z} = \frac{\partial a}{\partial z} \frac{\partial l}{\partial a} \qquad \boxed{\frac{\partial l}{\partial a}} = \frac{\partial z'}{\partial a} \frac{\partial l}{\partial z'} + \frac{\partial z''}{\partial a} \frac{\partial l}{\partial z''}$$ (Chain rule)

w_3 ? w_4 ?

Forward pass

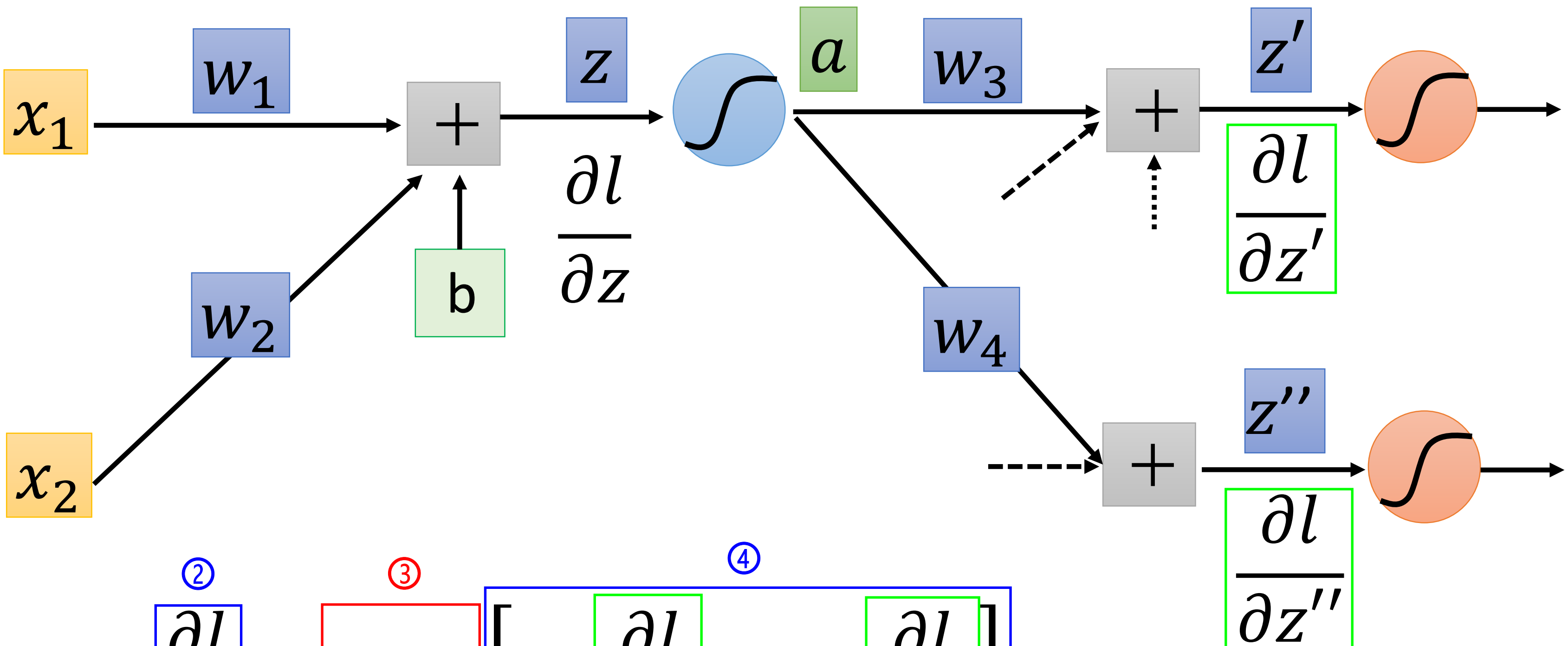
Backward pass

Assumed it's known

Backpropagation – Backward pass

②

Compute $\boxed{\partial l / \partial z}$ from the output layer ⇒ ④

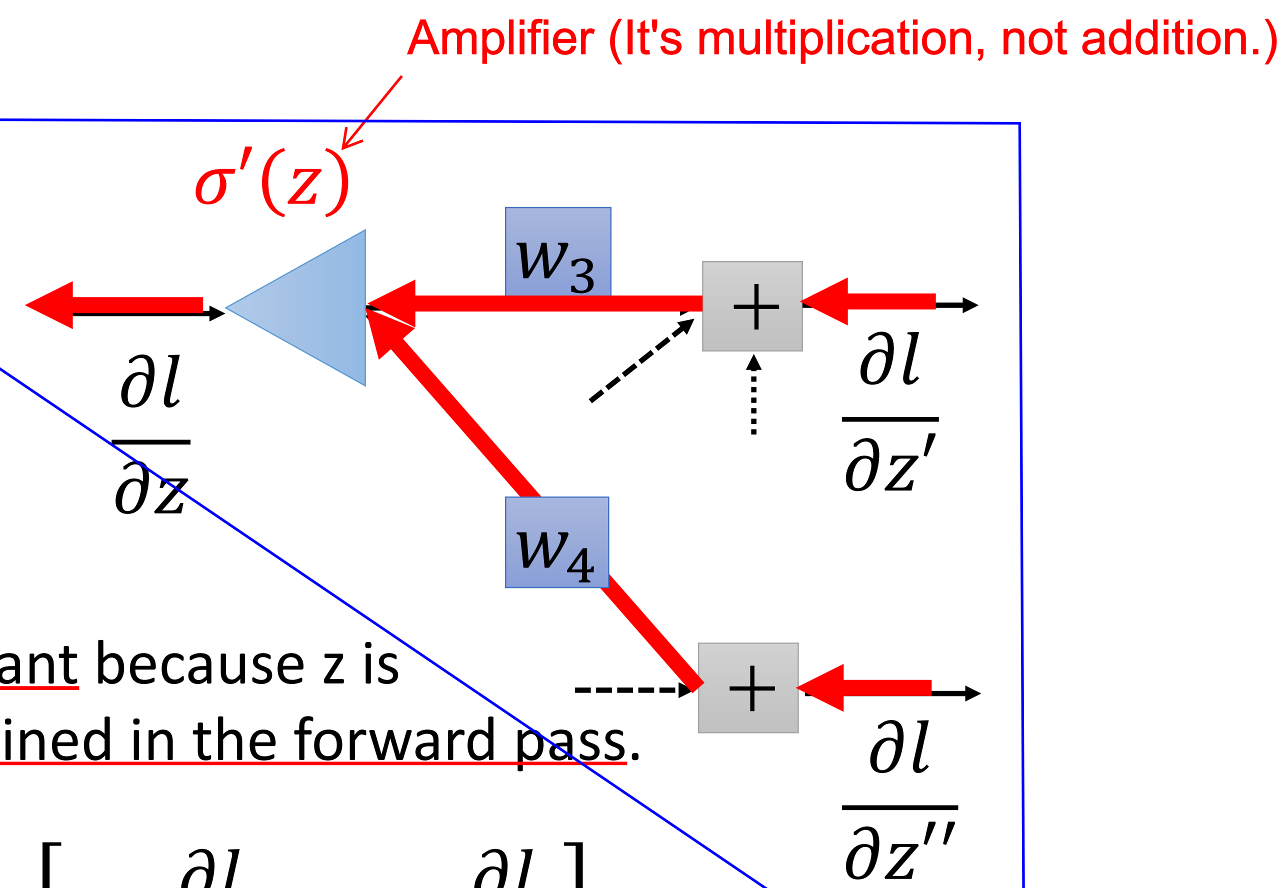


② ③ ④

$$\underbrace{\boxed{\frac{\partial l}{\partial z}}}_{②} = \underbrace{\boxed{\sigma'(z)}}_{③} \underbrace{\boxed{\left[w_3 \boxed{\frac{\partial l}{\partial z'}} + w_4 \boxed{\frac{\partial l}{\partial z''}} \right]}}_{④}$$

If we know the latter, we can calculate the front.

Backpropagation – Backward pass

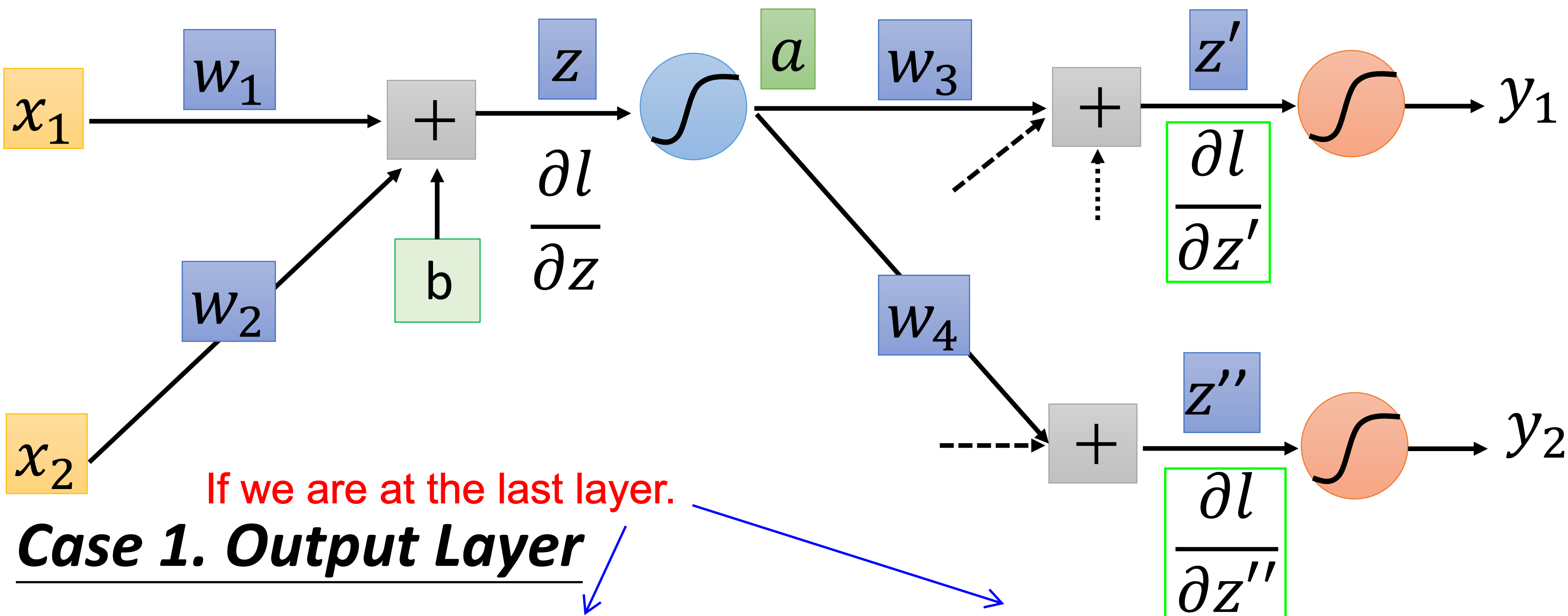


$\sigma'(z)$ is a constant because z is already determined in the forward pass.

$$\frac{\partial l}{\partial z} = \sigma'(z)\left[w_3 \frac{\partial l}{\partial z'} + w_4 \frac{\partial l}{\partial z''}\right]$$

Backpropagation – Backward pass

Compute $\partial l / \partial z$ from the output layer ⇒ ④



If we are at the last layer.

Case 1. Output Layer

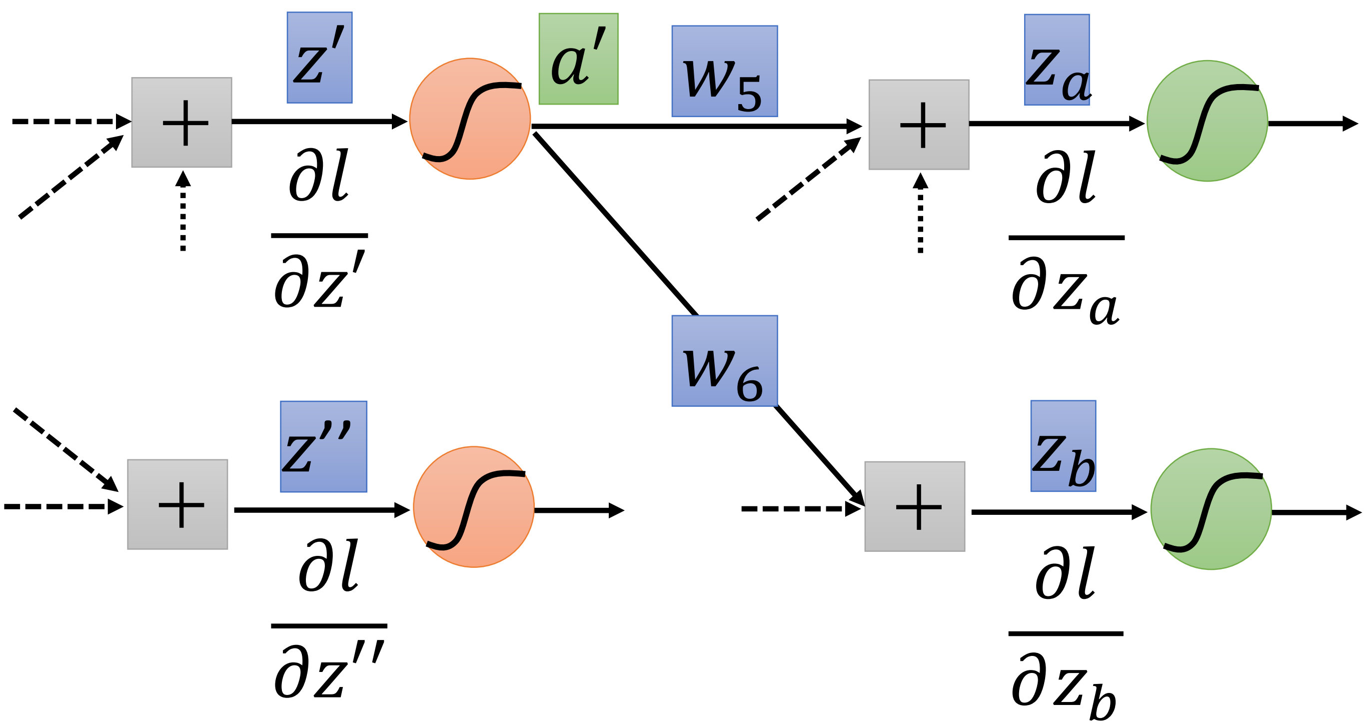
$$\boxed{\frac{\partial l}{\partial z'}} = \underbrace{\boxed{\frac{\partial y_1}{\partial z'}}}_{\text{easy}} \underbrace{\boxed{\frac{\partial l}{\partial y_1}}}_{\text{known}} \qquad \boxed{\frac{\partial l}{\partial z''}} = \underbrace{\boxed{\frac{\partial y_2}{\partial z''}}}_{\text{easy}} \underbrace{\boxed{\frac{\partial l}{\partial y_2}}}_{\text{known}}$$

Done!

Backpropagation – Backward pass

Compute $\partial l / \partial z$ from the output layer ⇒ ④

Case 2. Not Output Layer If we are NOT at the last layer.

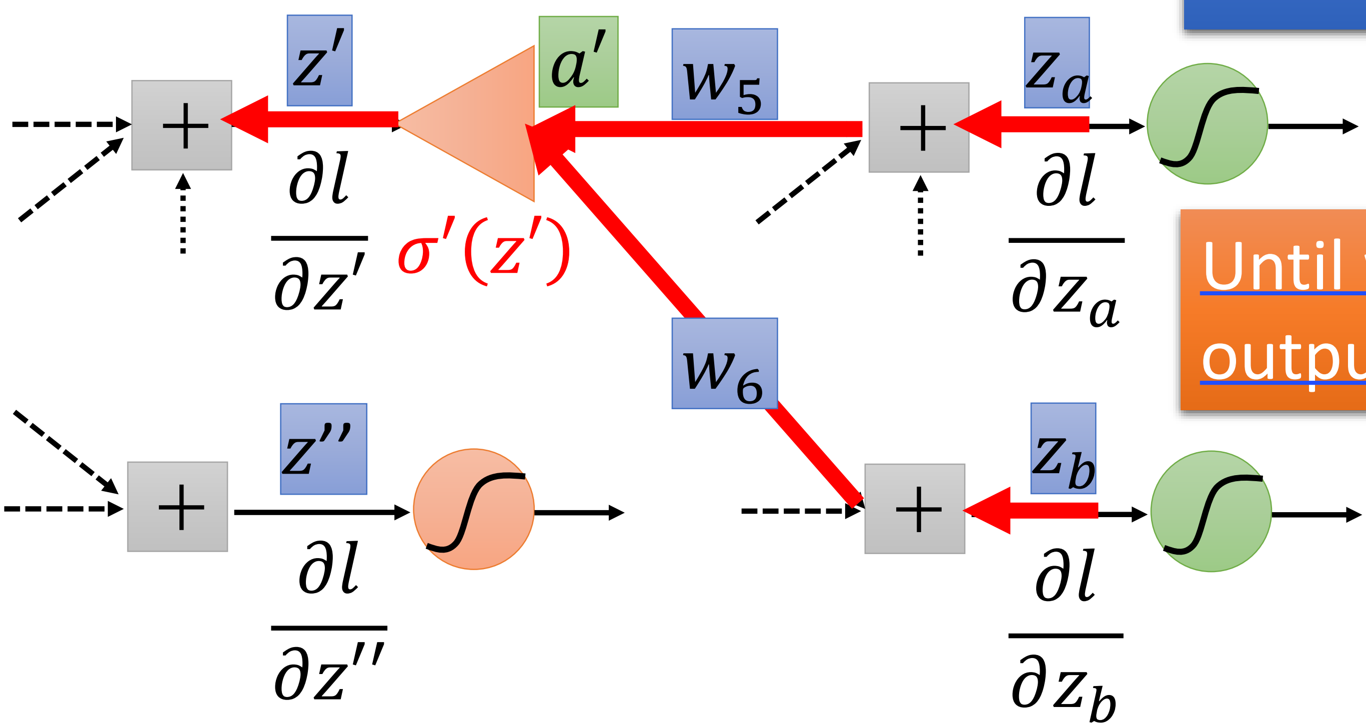


Backpropagation – Backward pass

Compute $\partial l / \partial z$ from the output layer ⇒ ④

Case 2. Not Output Layer

Compute $\partial l / \partial z$ recursively



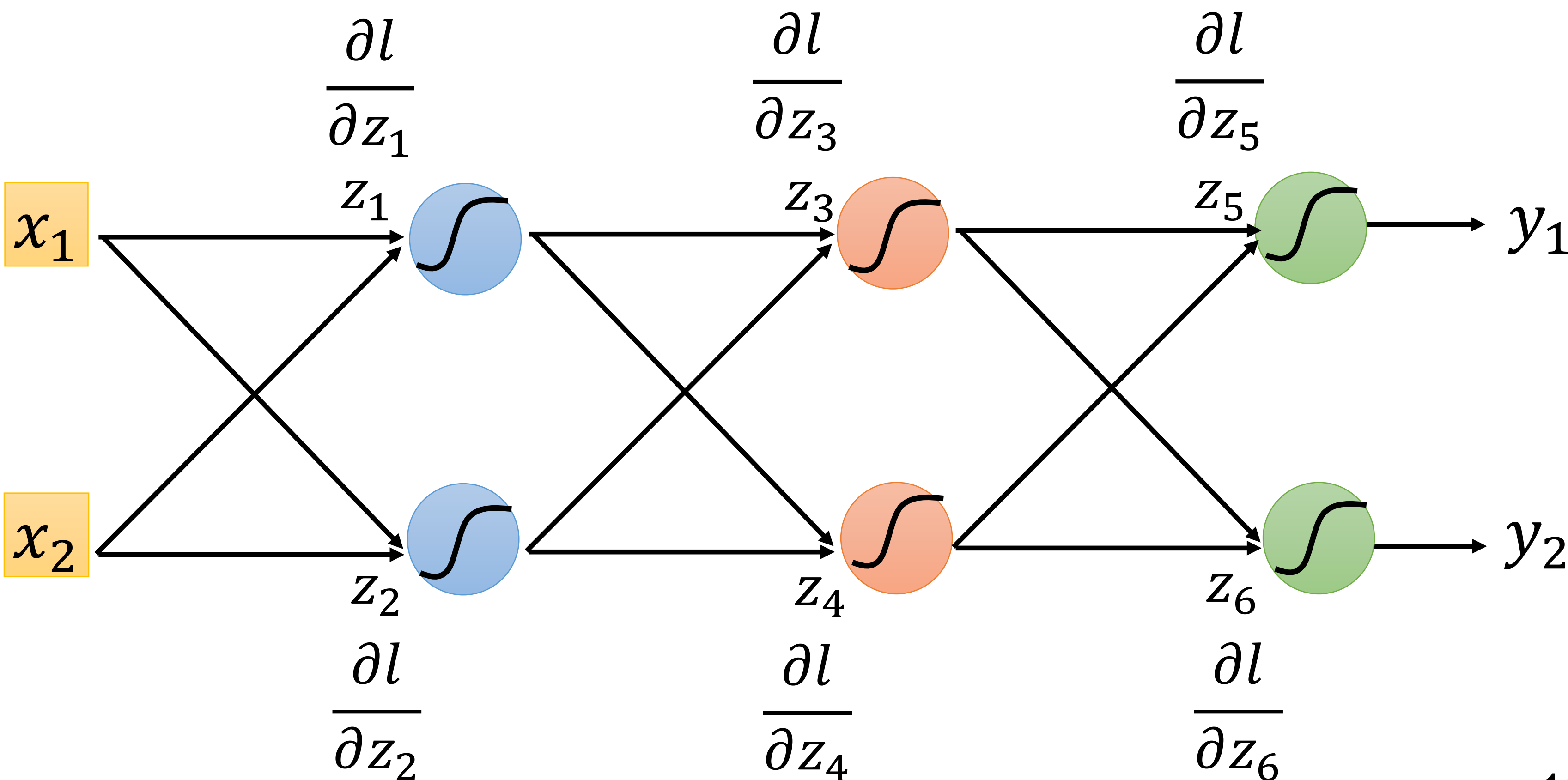
Until we reach the output layer ……

Backpropagation – Backward Pass

Forward pass

Compute $\partial l / \partial z$ for all activation function inputs z ⇒ ③

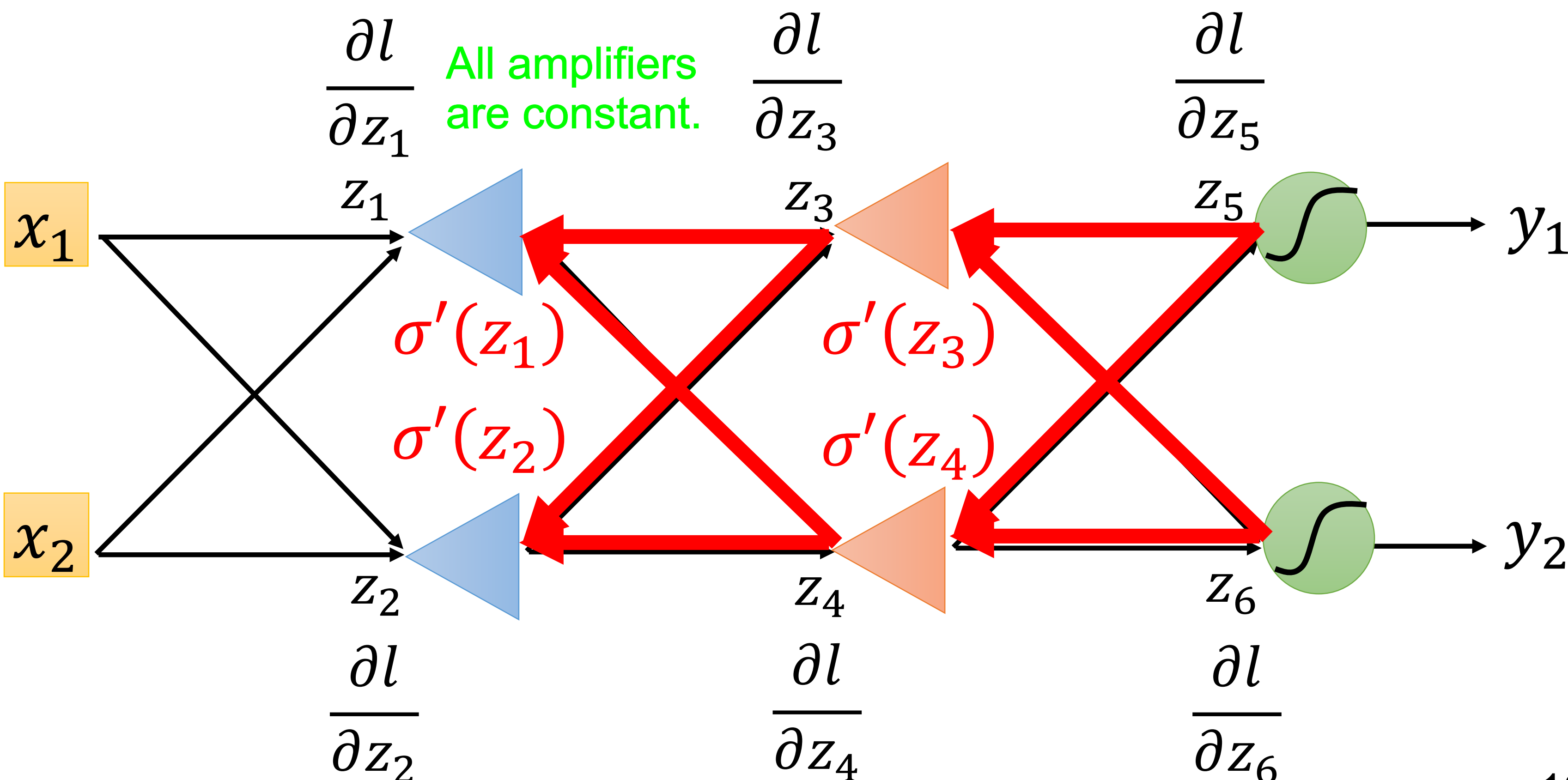
Compute $\partial l / \partial z$ from the output layer ⇒ ④ Backward pass



Backpropagation – Backward Pass

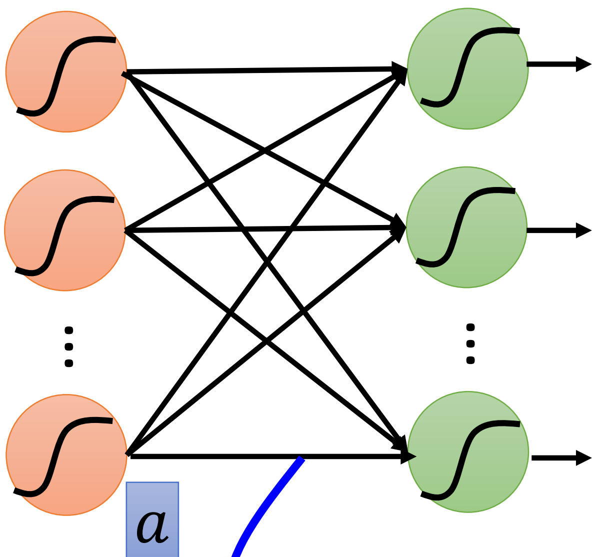
Compute $\partial l/\partial z$ for all activation function inputs z ⇒ ③

Compute $\partial l/\partial z$ from the output layer ⇒ ④

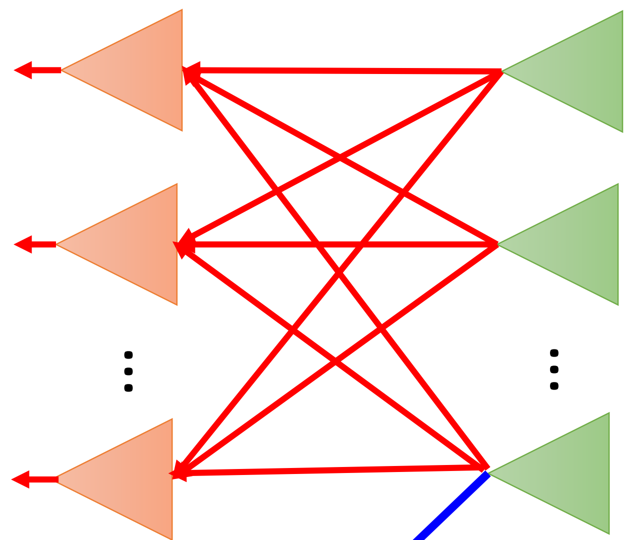


Backpropagation – Summary

Forward Pass



Backward Pass



$$\underset{\text{easy}}{\overset{①}{\boxed{\frac{\partial z}{\partial w}}}} = a \quad \times \quad \underset{\text{complex}}{\overset{②}{\boxed{\frac{\partial l}{\partial z}}}} \quad = \overset{⓪}{\boxed{\frac{\partial l}{\partial w}}}$$

for all w