Backpropagation

Gradient Descent

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

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

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

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

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

Millions of parameters ......

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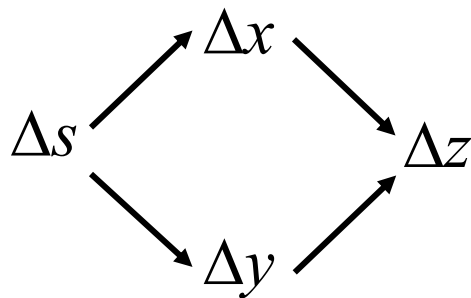
Chain Rule

Case 1 $y = g(x)$ $z = h(y)$

$$\Delta x \rightarrow \Delta y \rightarrow \Delta z \qquad \frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$$
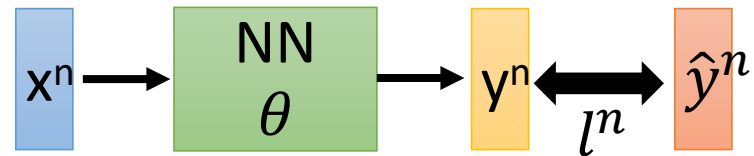
Case 2

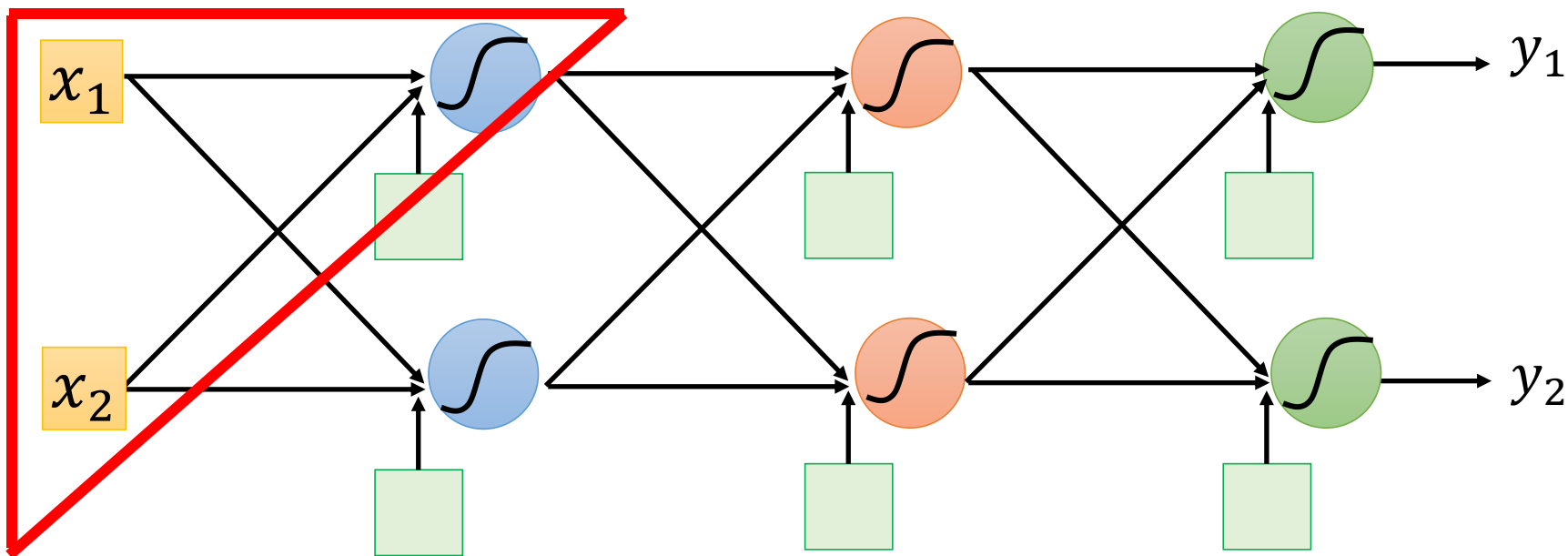$$x = g(s) \qquad y = h(s) \qquad z = k(x, y)$$

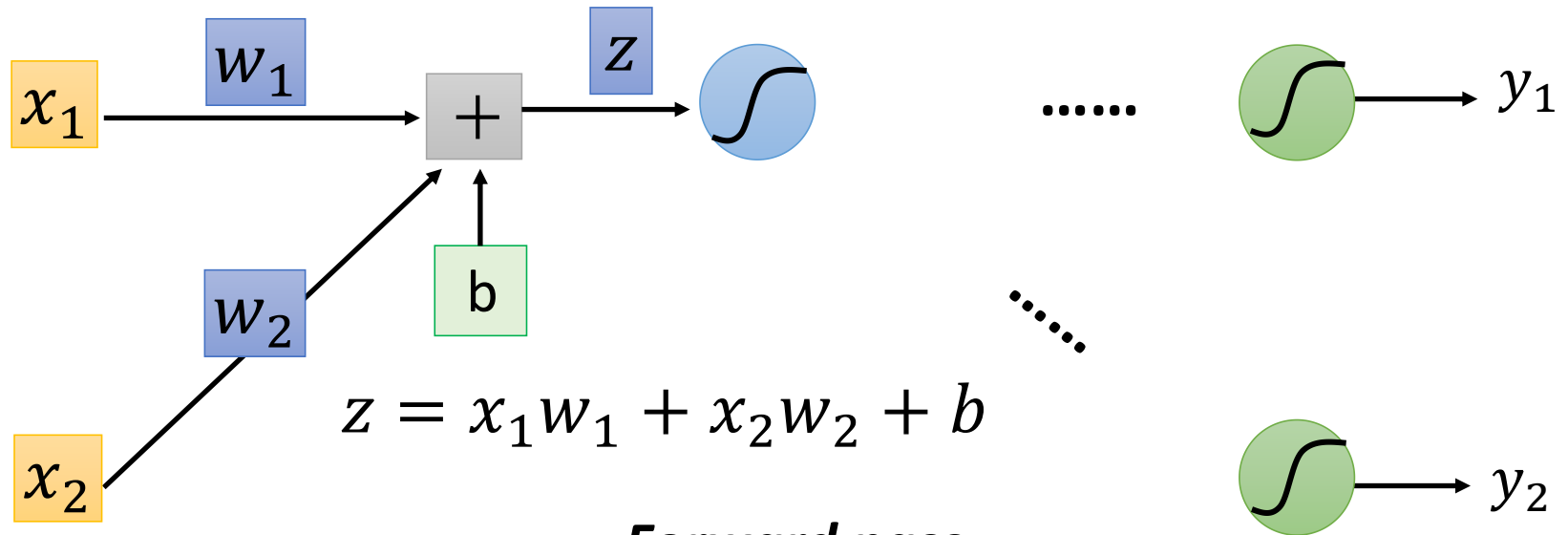$$\frac{dz}{ds} = \frac{\partial z}{\partial x}\frac{dx}{ds} + \frac{\partial z}{\partial y}\frac{dy}{ds}$$

Backpropagation

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

Backpropagation

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

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

(Chain rule)

Forward pass:
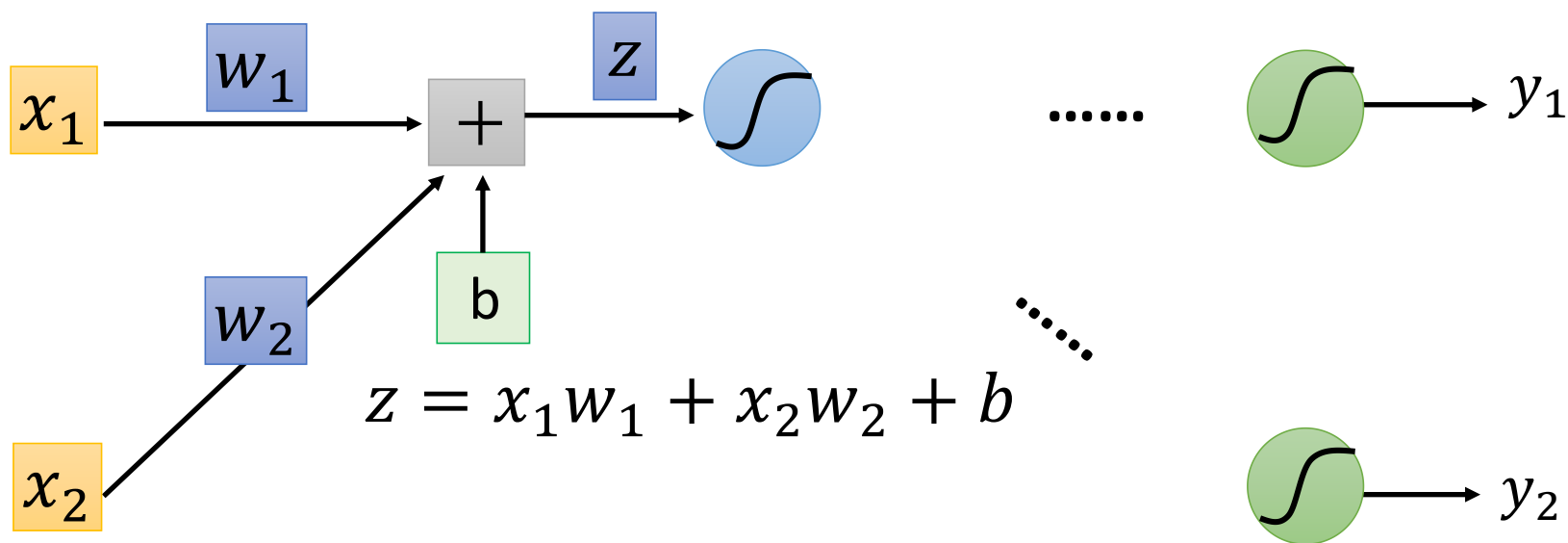
Compute $\partial z / \partial w$ for all parameters

Backward pass:

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

Backpropagation – Forward pass

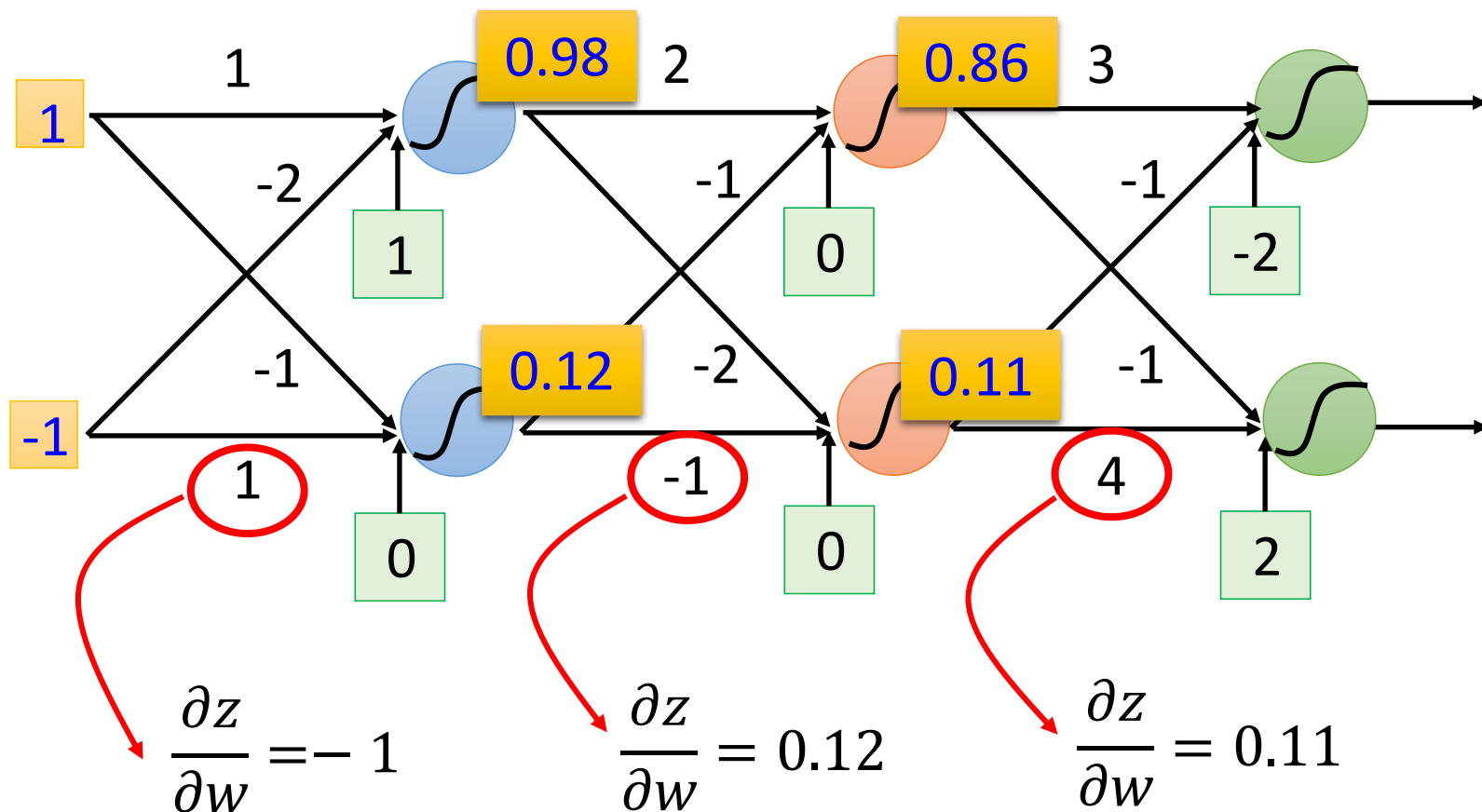Compute $\partial z / \partial w$ for all parameters

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

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

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

The value of the input connected by the weight

Backpropagation – Forward pass
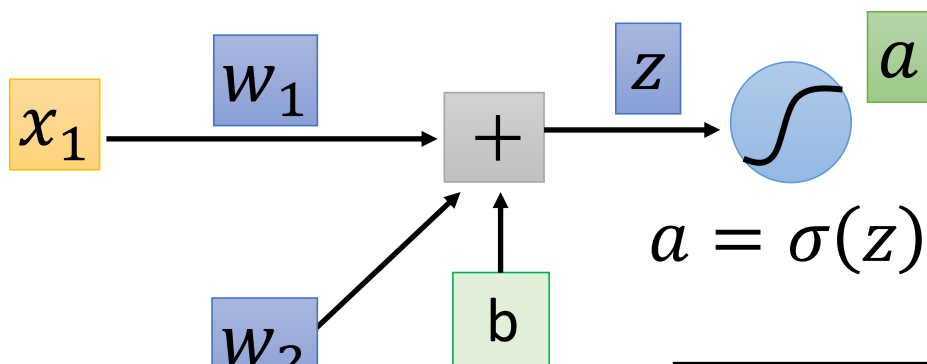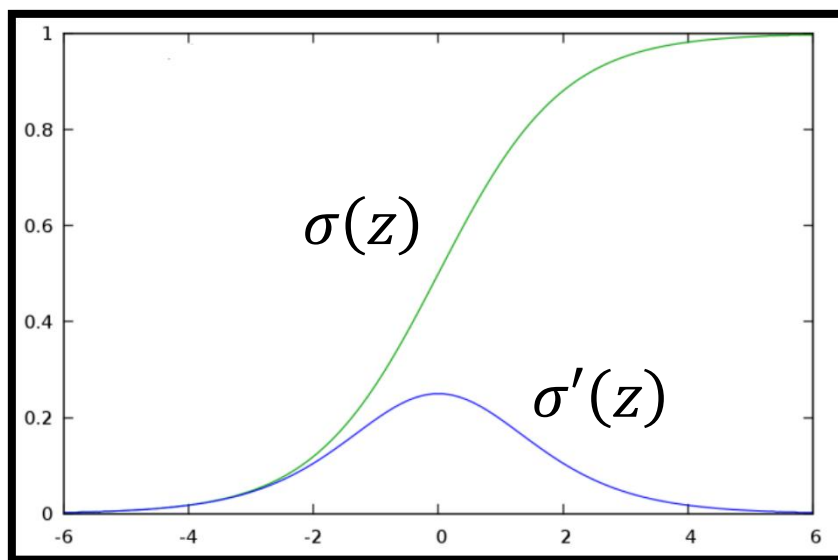
Compute $\partial z / \partial w$ for all parameters

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

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

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

Backpropagation – Backward pass

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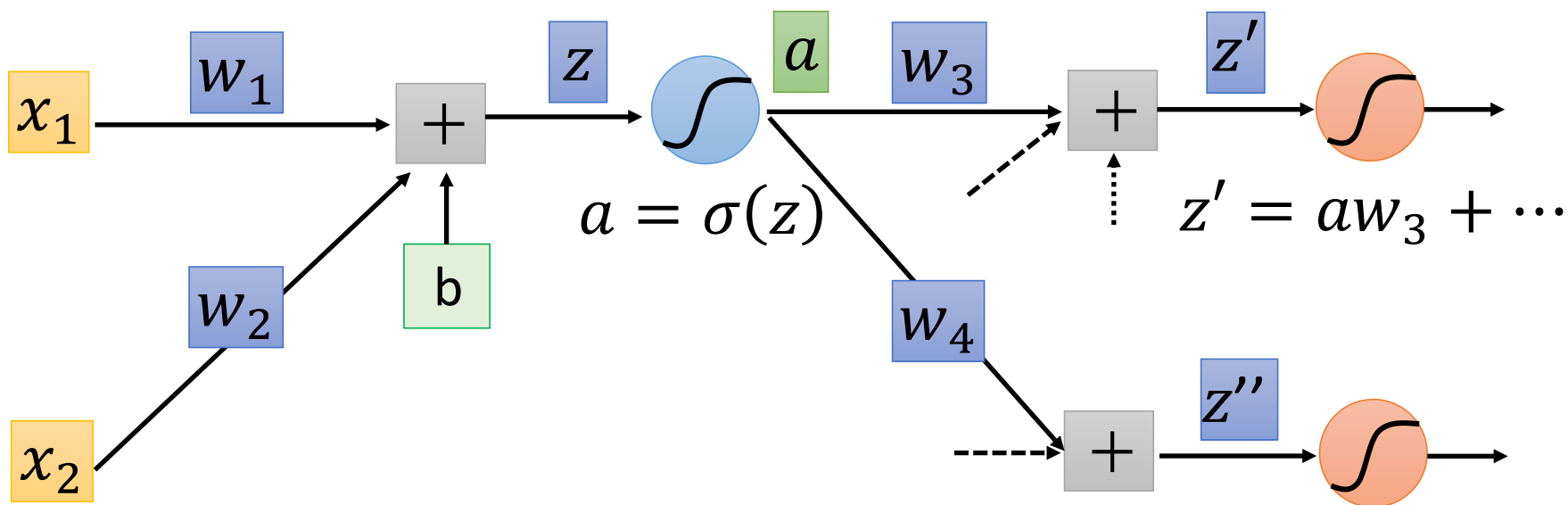

$$a = \sigma(z)$$

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

$$\Rightarrow \sigma'(z)$$

Backpropagation – Backward pass

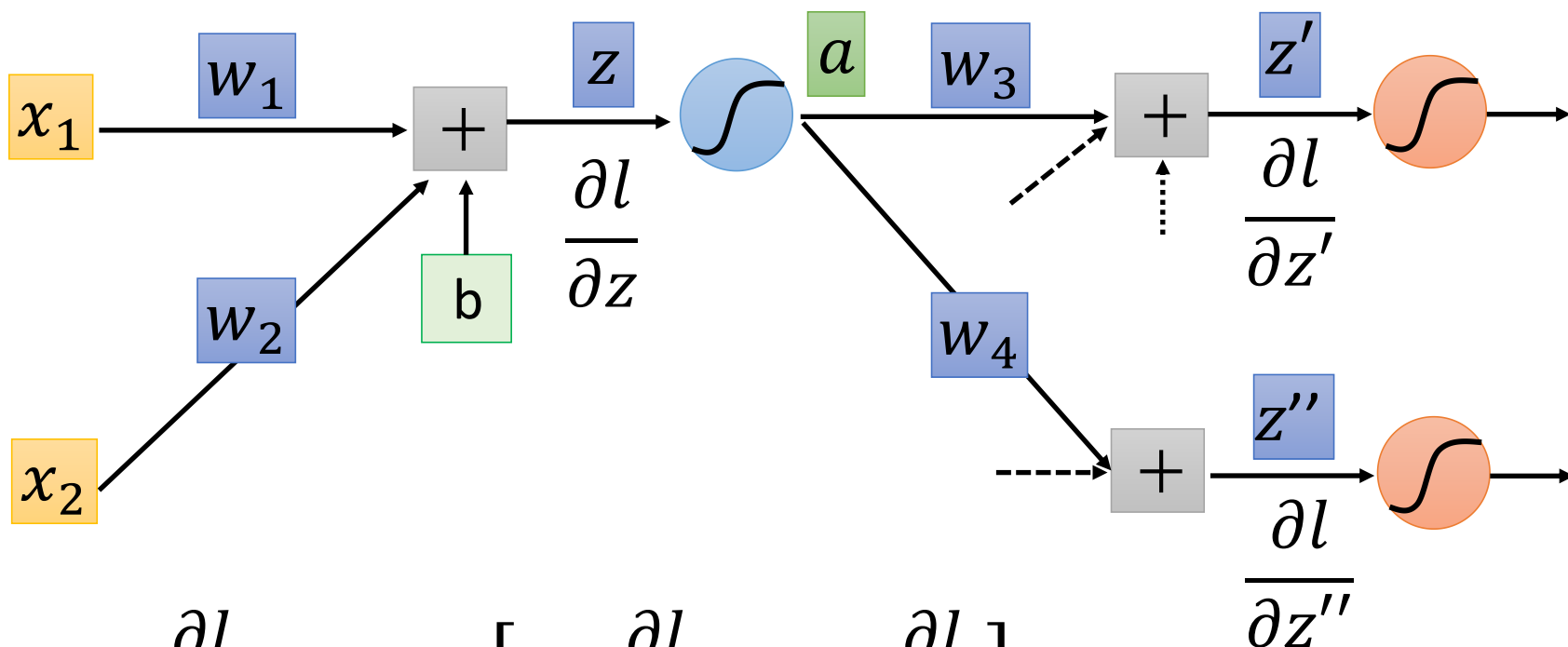Compute $\partial l / \partial z$ for all activation function inputs z

$$\frac{\partial l}{\partial z} = \frac{\partial a}{\partial z}\frac{\partial l}{\partial a} \qquad \frac{\partial l}{\partial a} = \frac{\partial z'}{\partial a}\frac{\partial l}{\partial z'} + \frac{\partial z''}{\partial a}\frac{\partial l}{\partial z''} \text{ (Chain rule)}$$
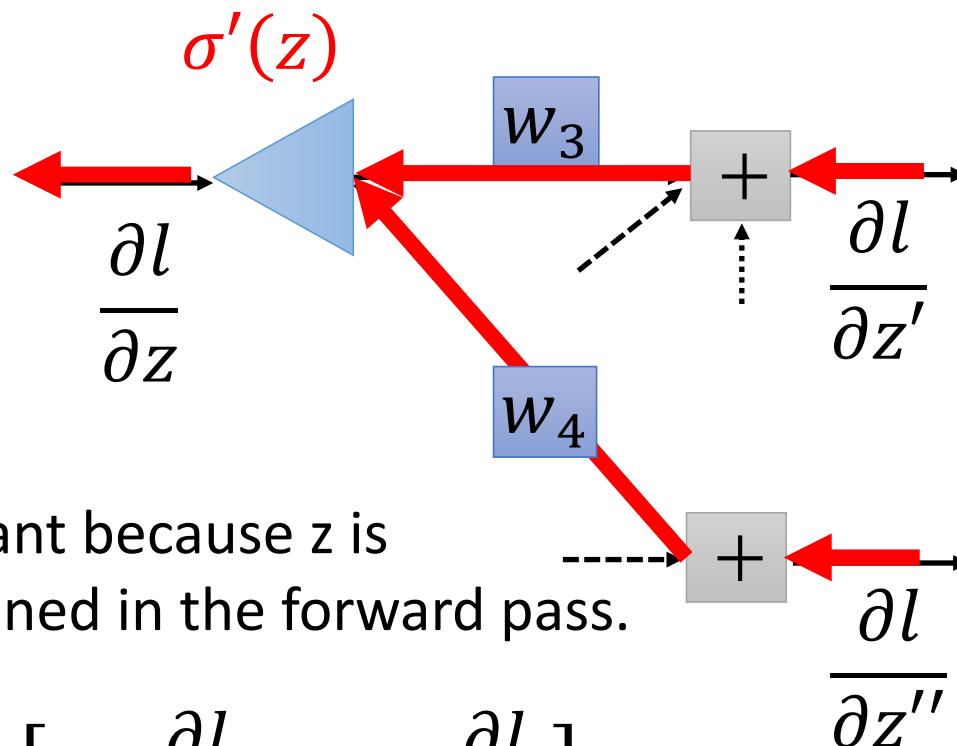
Assumed it's known

Backpropagation − Backward pass

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

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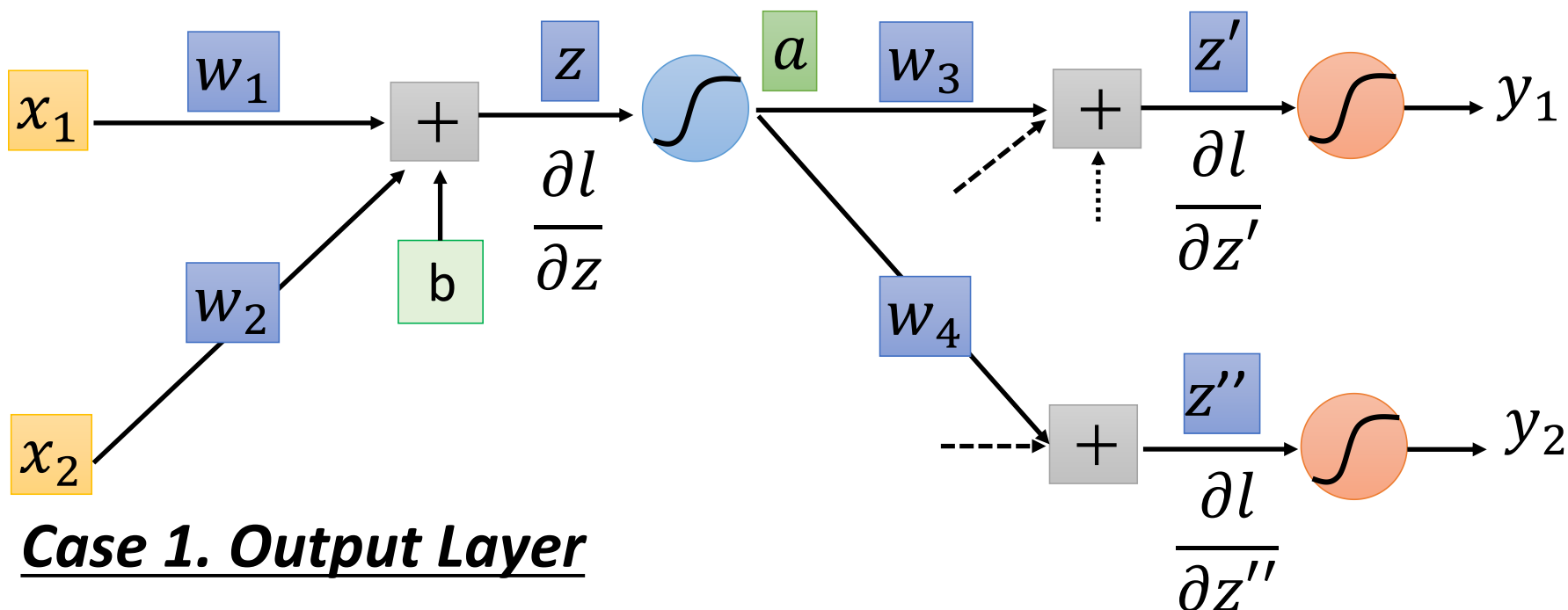
Backpropagation – Backward pass

$\sigma'(z)$

w_3

$+$

$\dfrac{\partial l}{\partial z}$

$\dfrac{\partial l}{\partial z'}$

w_4

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

$+$

$\dfrac{\partial l}{\partial z''}$

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

Backpropagation – Backward pass

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

Case 1. Output Layer

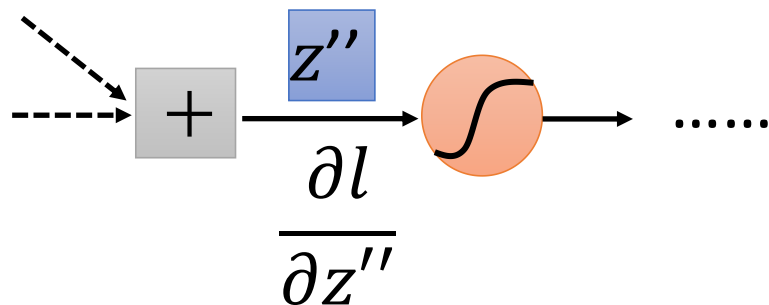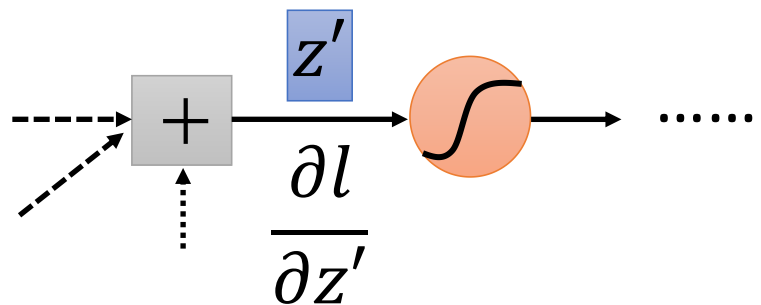$$\frac{\partial l}{\partial z'} = \frac{\partial y_1}{\partial z'} \frac{\partial l}{\partial y_1} \qquad \frac{\partial l}{\partial z''} = \frac{\partial y_2}{\partial z''} \frac{\partial l}{\partial y_2}$$

Done!

Backpropagation − Backward pass

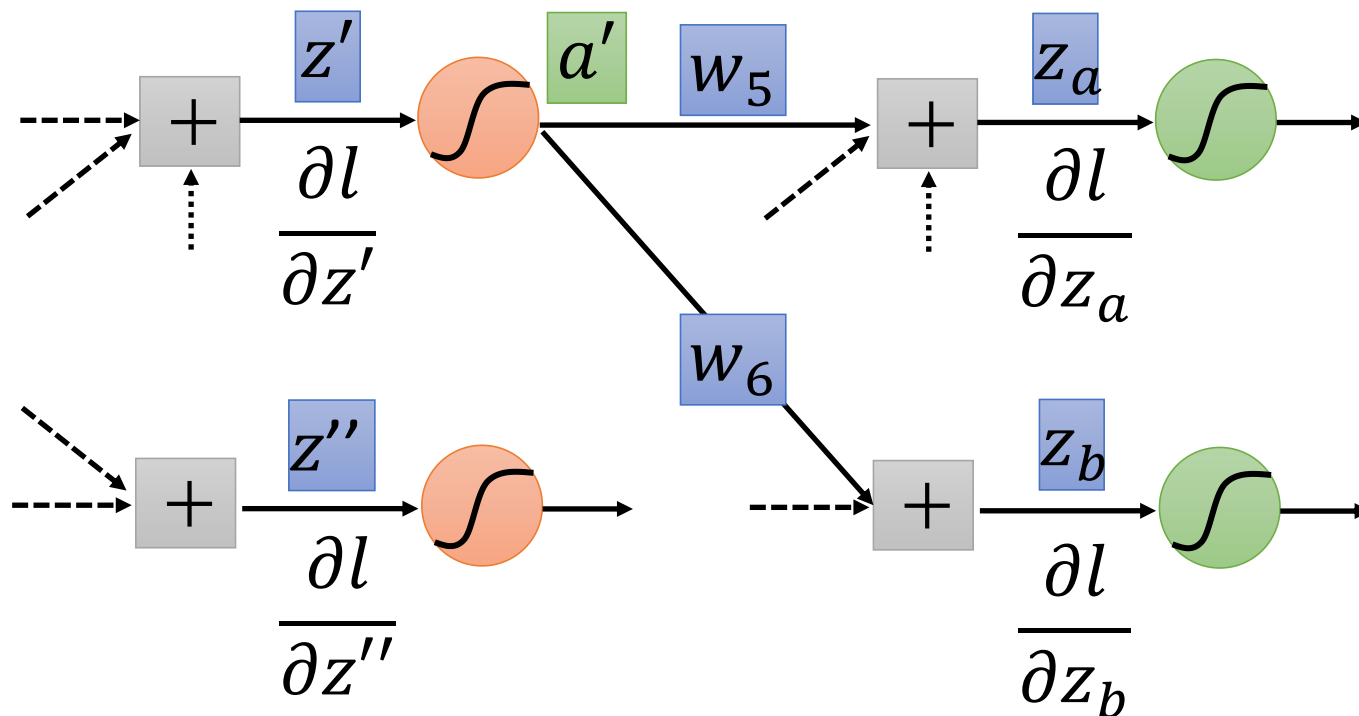Compute $\partial l / \partial z$ for all activation function inputs z

Case 2. Not Output Layer

Backpropagation – Backward pass

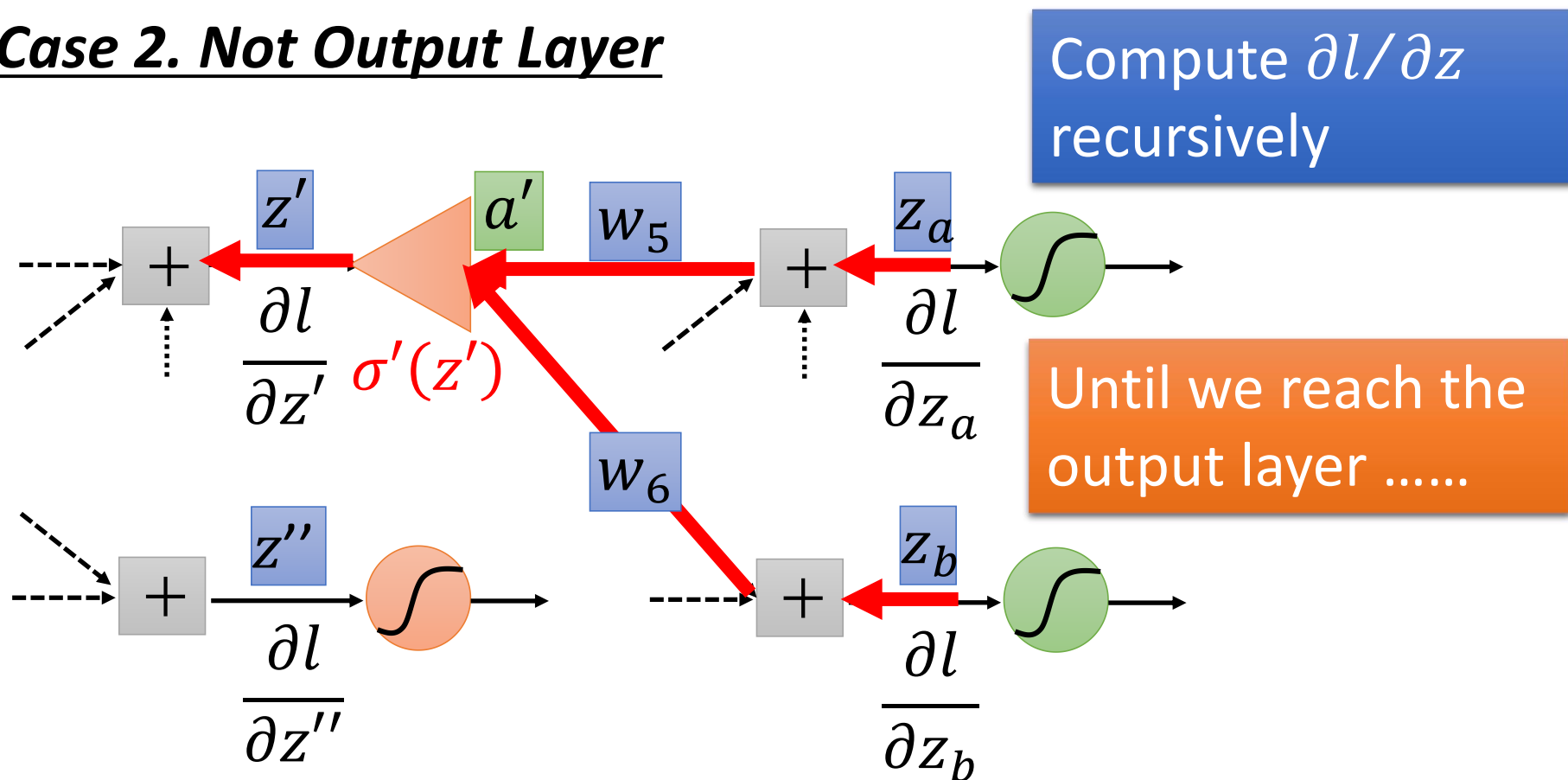Compute $\partial l / \partial z$ for all activation function inputs z

Case 2. Not Output Layer

Backpropagation – Backward pass

Compute $\partial l / \partial z$ for all activation function inputs z
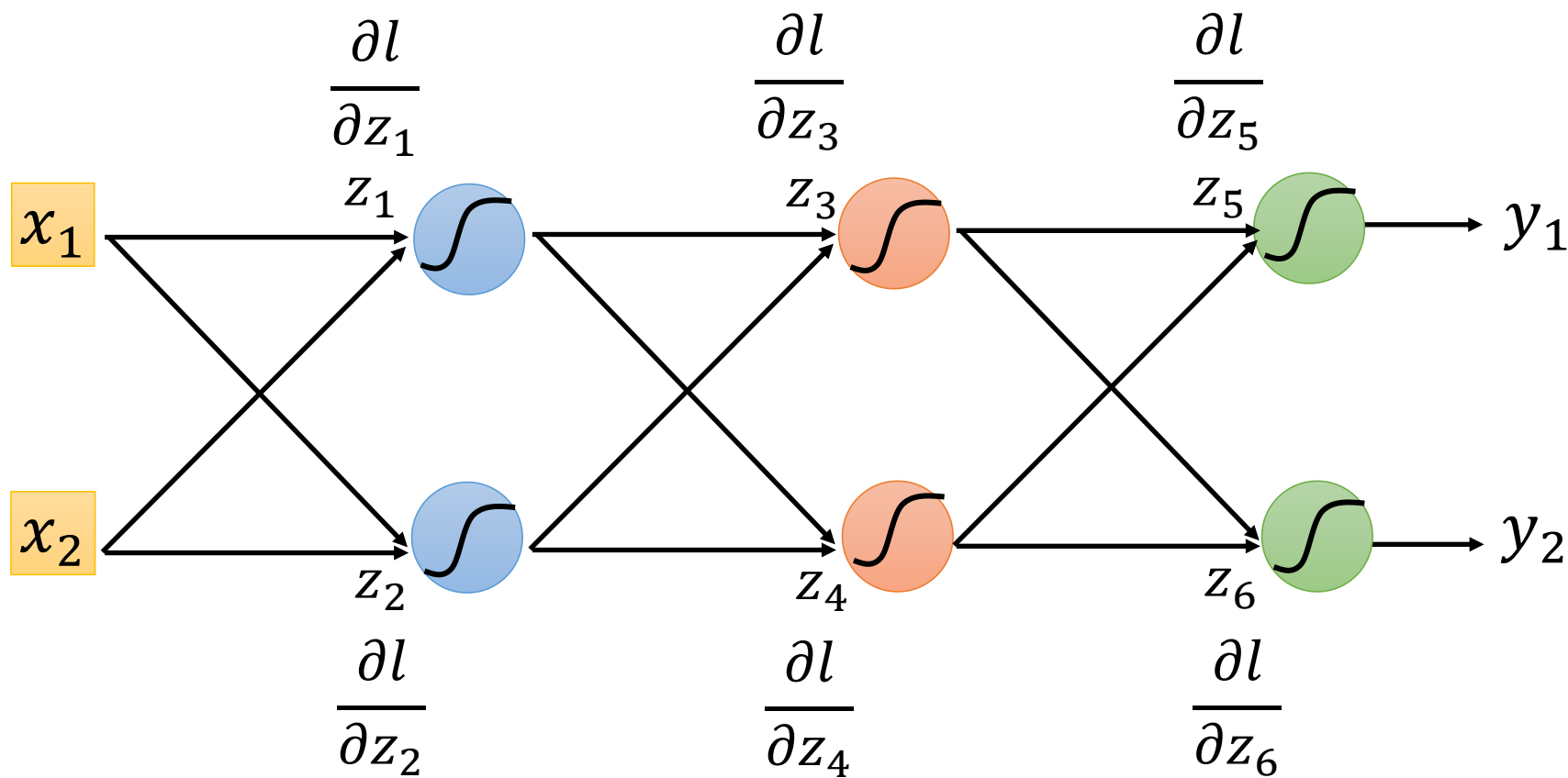
Case 2. Not Output Layer

Compute $\partial l / \partial z$ recursively

Until we reach the output layer ……

Backpropagation – Backward Pass

Compute $\partial l / \partial z$ for all activation function inputs z
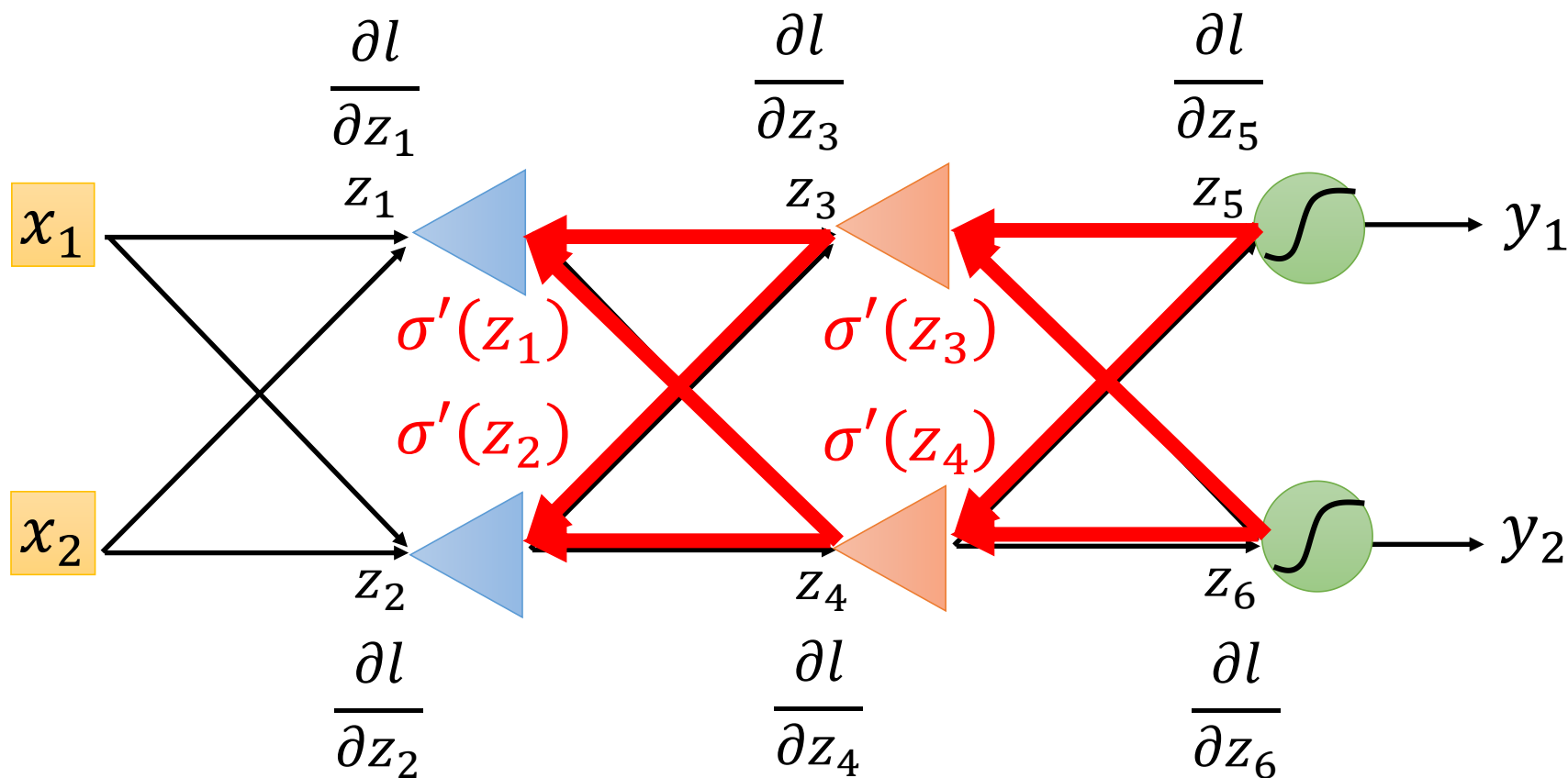
Compute $\partial l / \partial z$ from the output layer

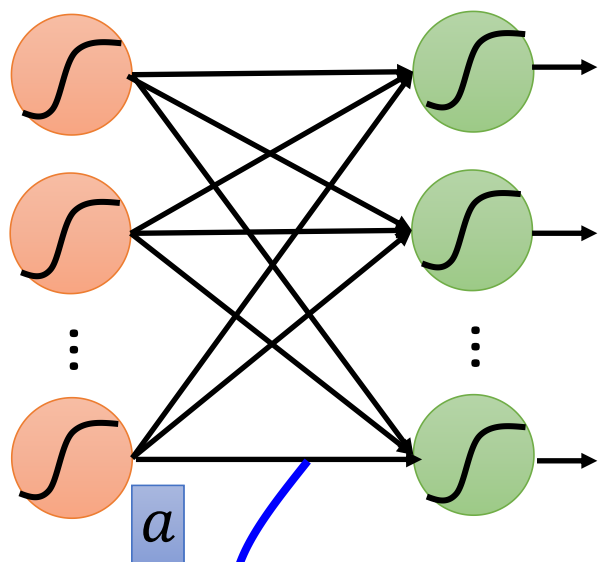# Backpropagation – Backward Pass

Compute $\partial l / \partial z$ for all activation function inputs z
Compute $\partial l / \partial z$ from the output layer
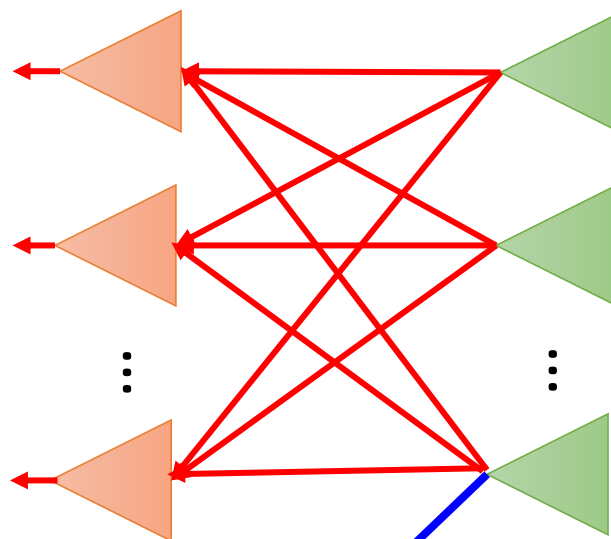
Backpropagation – Summary

Forward Pass

Backward Pass

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

X

$$\frac{\partial l}{\partial z}$$

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

for all w