反向传播

Backpropagation

梯度下降

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

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

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

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

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

对于图像处理或者语音设别中，一个网络有很多层，每一层会有很多神经元，参数有的会高达100M+
反响传播并不是一个新的优化方法，只不过是可以高效求出左边的vector

链式求导

Case 1

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

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

Case 2

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

$$\begin{tikzcd}
 & \Delta x \arrow[dr] & \\
\Delta s \arrow[ur] \arrow[dr] & & \Delta z \\
 & \Delta y \arrow[ur] &
\end{tikzcd}$$

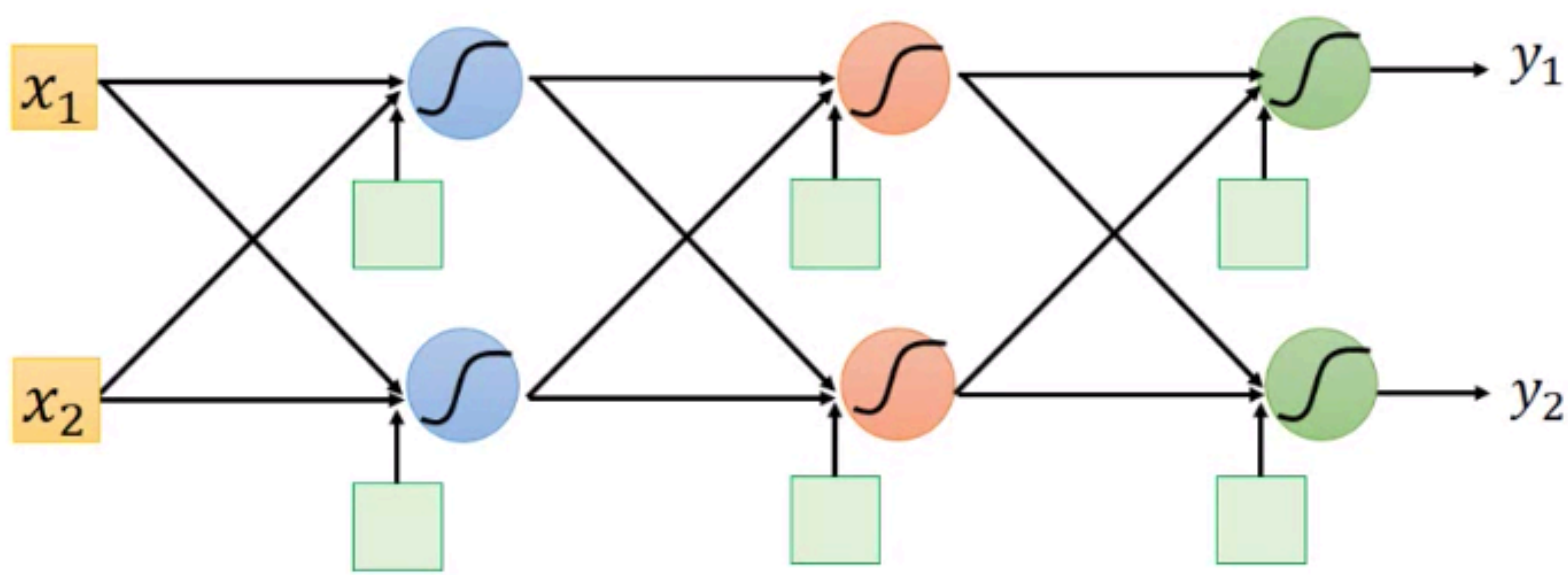
$$\frac{dz}{ds} = \frac{\partial z}{\partial x}\frac{dx}{ds} + \frac{\partial z}{\partial y}\frac{dy}{ds}$$

反向传播

C 为预测值与真实值之间的距离

最小化L，对w求导

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

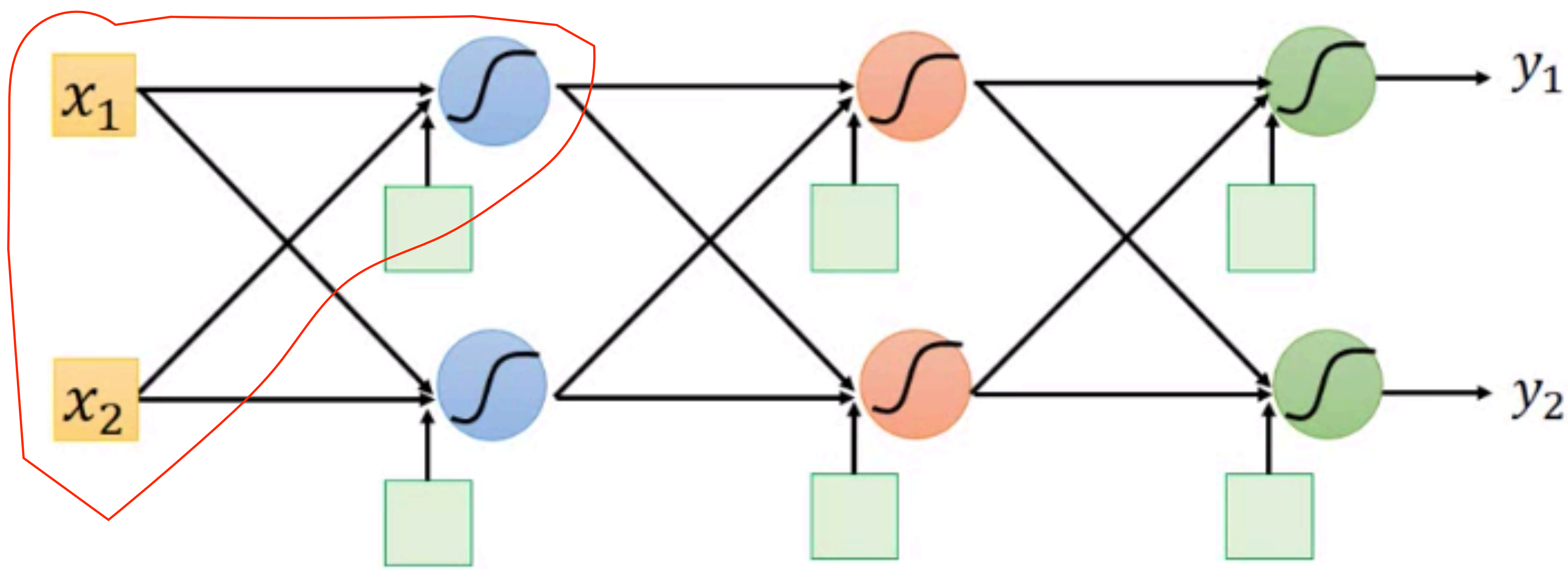


反向传播

C 为预测值与真实值之间的距离

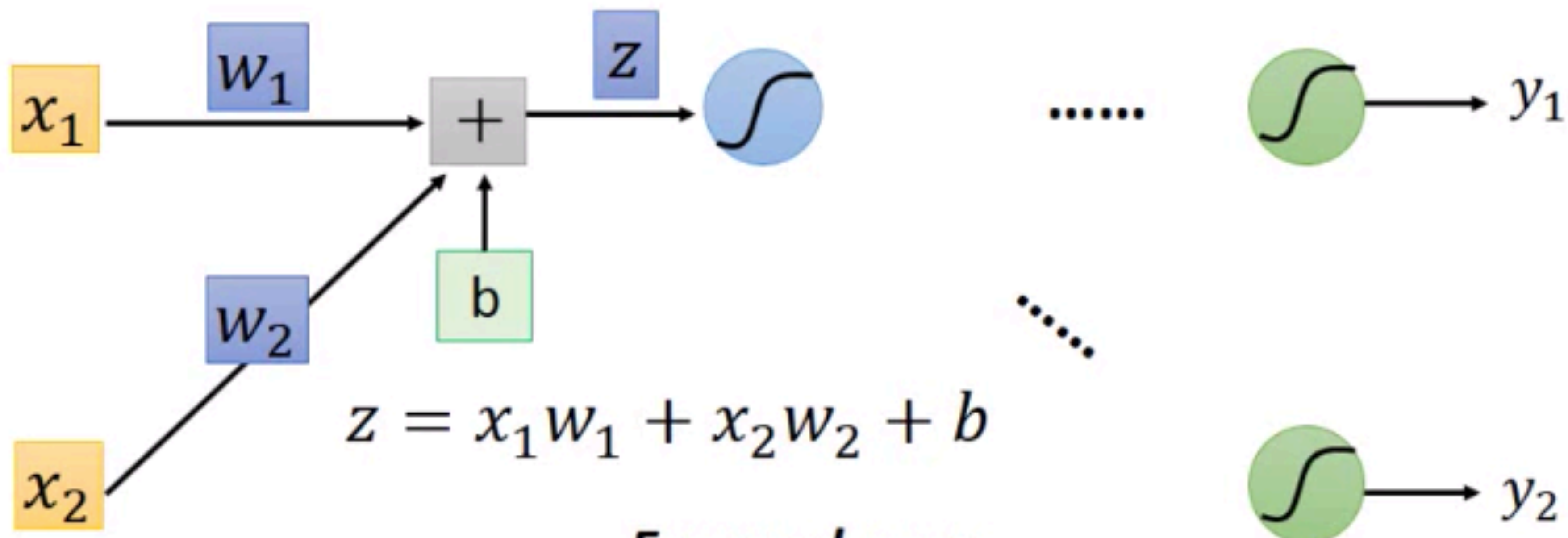
最小化L，对w求导

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



先看这一部分

反向传播



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

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

(Chain rule)

Forward pass:

$\partial z/\partial w$ 对所有参数计算

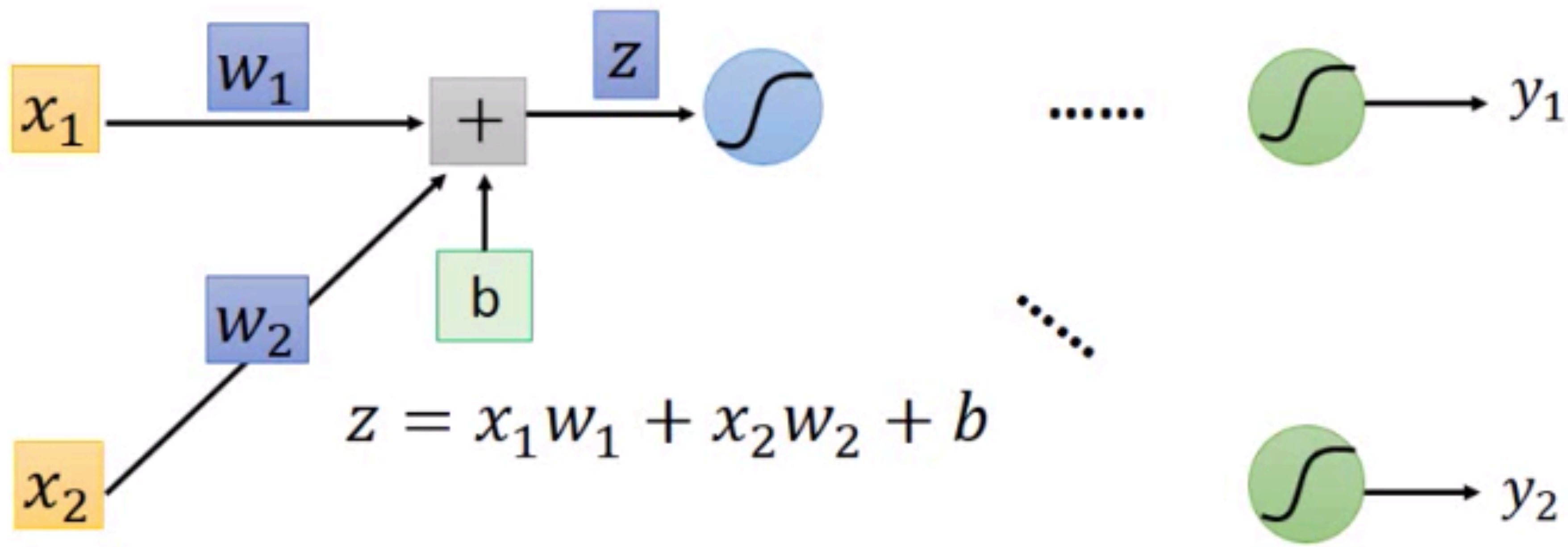
Backward pass:

$\partial C/\partial z$ 对所有激活函数计算

Forward Pass

- 计算$\frac{\partial C}{\partial w}$的过程称为Forward Pass

Forward Pass

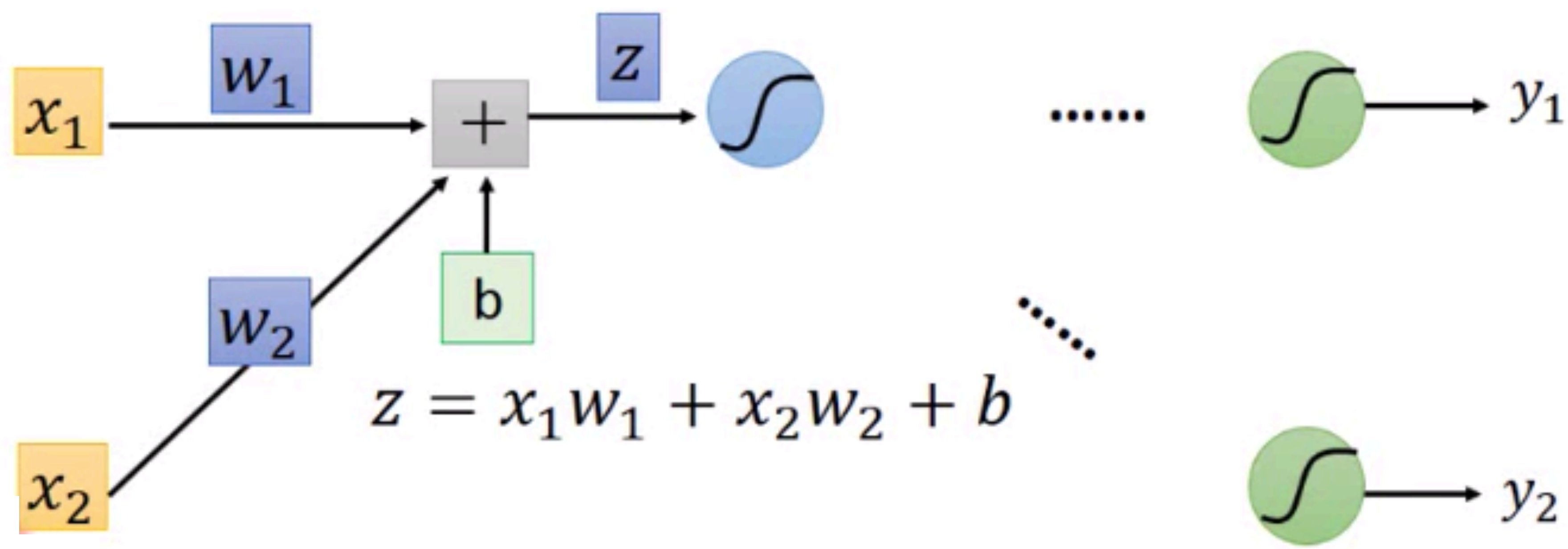


$$z = x_1 w_1 + x_2 w_2 + b$$

$$\frac{\partial C}{\partial w} = ?$$

Forward Pass

对所有参数计算 $\partial z/\partial w$

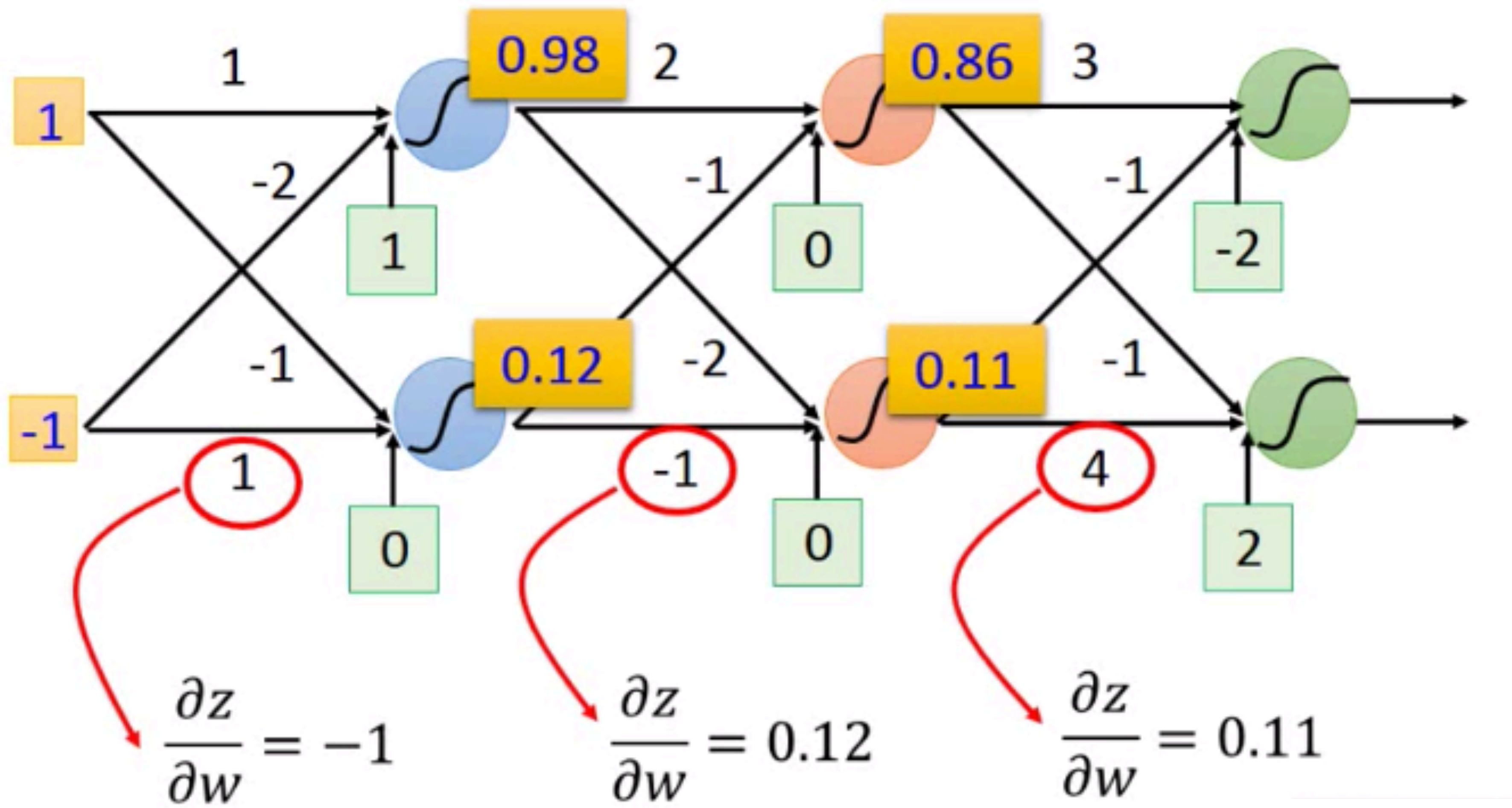


$$z = x_1 w_1 + x_2 w_2 + b$$

$$\left.\begin{aligned} \partial z/\partial w_1 &= ?\ \ x_1 \\ \partial z/\partial w_2 &= ?\ \ x_2 \end{aligned}\right\} \text{就是输入的值}$$

Forward Pass

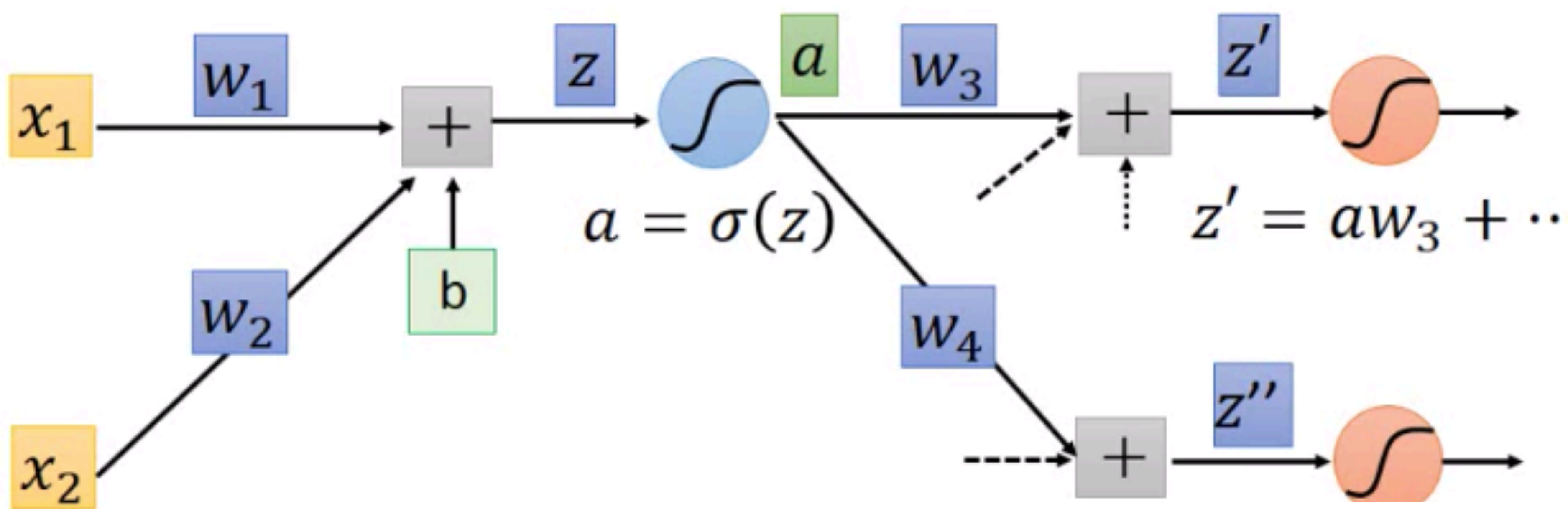
对所有参数计算 $\partial z/\partial w$



Backward Pass

- 计算$\frac{\partial C}{\partial z}$的过程称为Backward Pass

Backward Pass

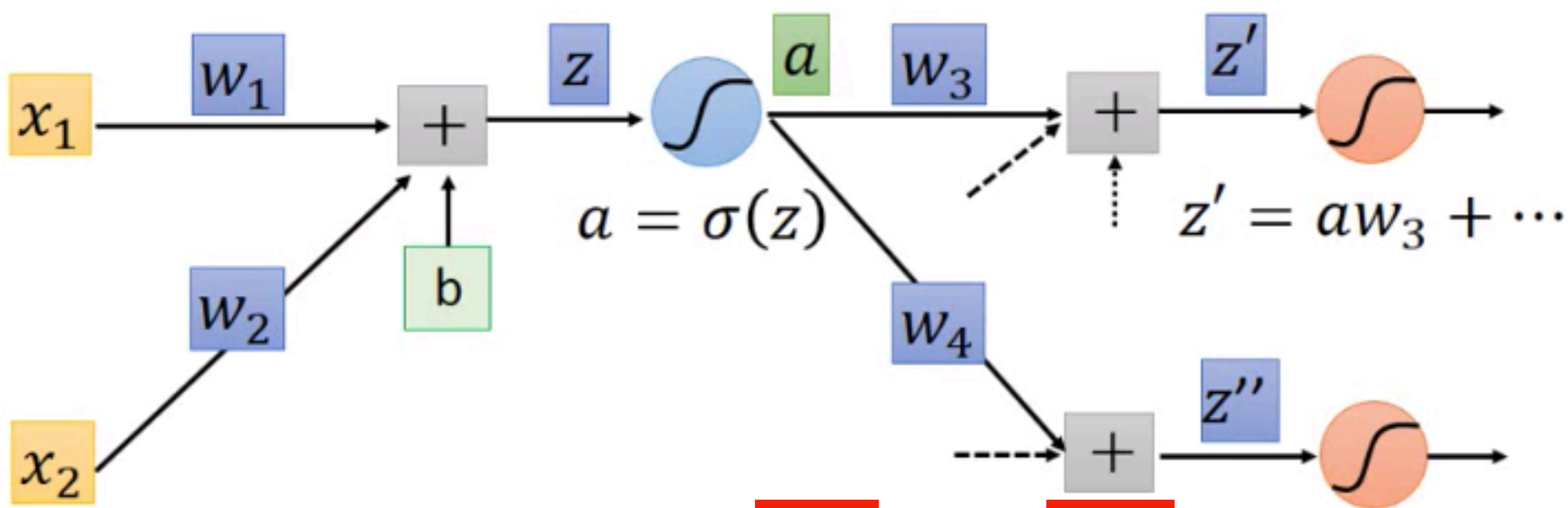


$\frac{\partial C}{\partial z}$ 比较复杂，我们先用chain rull 对这一项做一下拆解

Backward Pass

对所有激活函数计算 $\partial C/\partial z$

a: 为激活函数的输出
实际上a可能对应多个输出

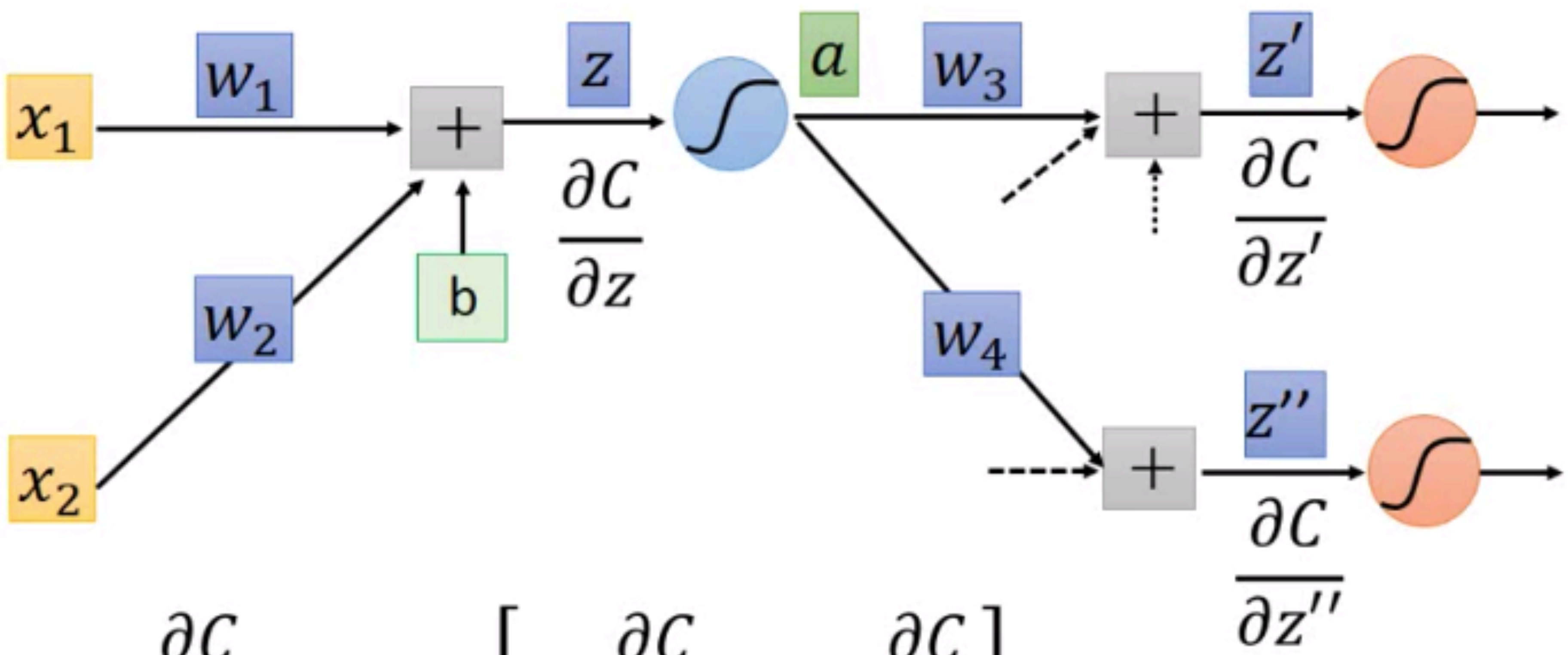


$$\frac{\partial C}{\partial z} = \frac{\partial a}{\partial z}\frac{\partial C}{\underline{\partial a}} \qquad \frac{\partial C}{\partial a} = \frac{\partial z'}{\underline{\partial a}}\boxed{\frac{\partial C}{\partial z'}} + \frac{\partial z''}{\underline{\partial a}}\boxed{\frac{\partial C}{\partial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

w_3 w_4 假设我们已经知道如何求这个了

Backward Pass

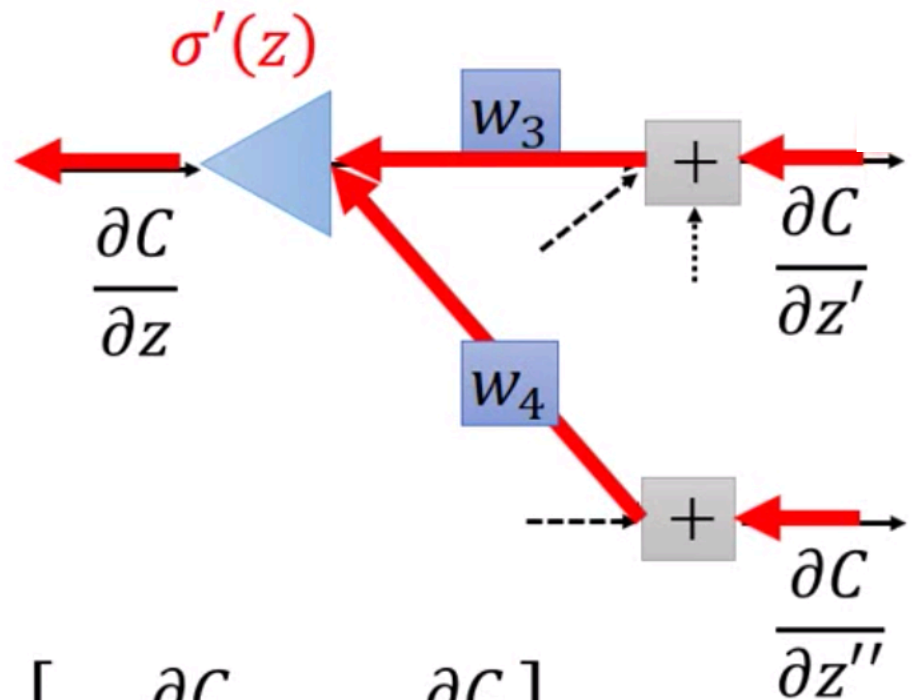
对所有激活函数计算 $\partial C/\partial z$



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

Backward Pass

下面的式子与途中的反向传播是等价的



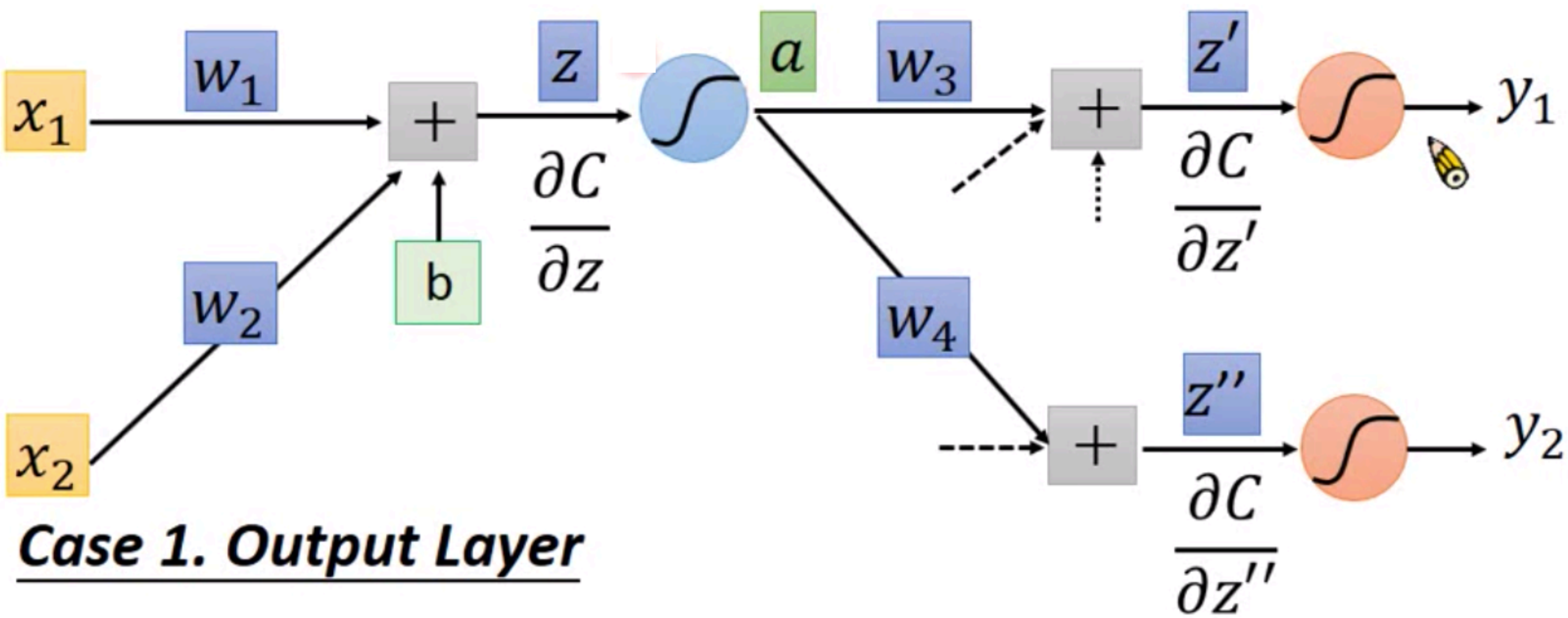
$$\frac{\partial C}{\partial z} = \sigma'(z)\left[w_3\frac{\partial C}{\partial z'} + w_4\frac{\partial C}{\partial z''}\right]$$

是一个常数，因为在forward pass阶段中已经计算出来了

Backward Pass

对所有激活函数计算 $\partial C/\partial z$

case1: $z^{'}, z^{''}$后面接的是输出层



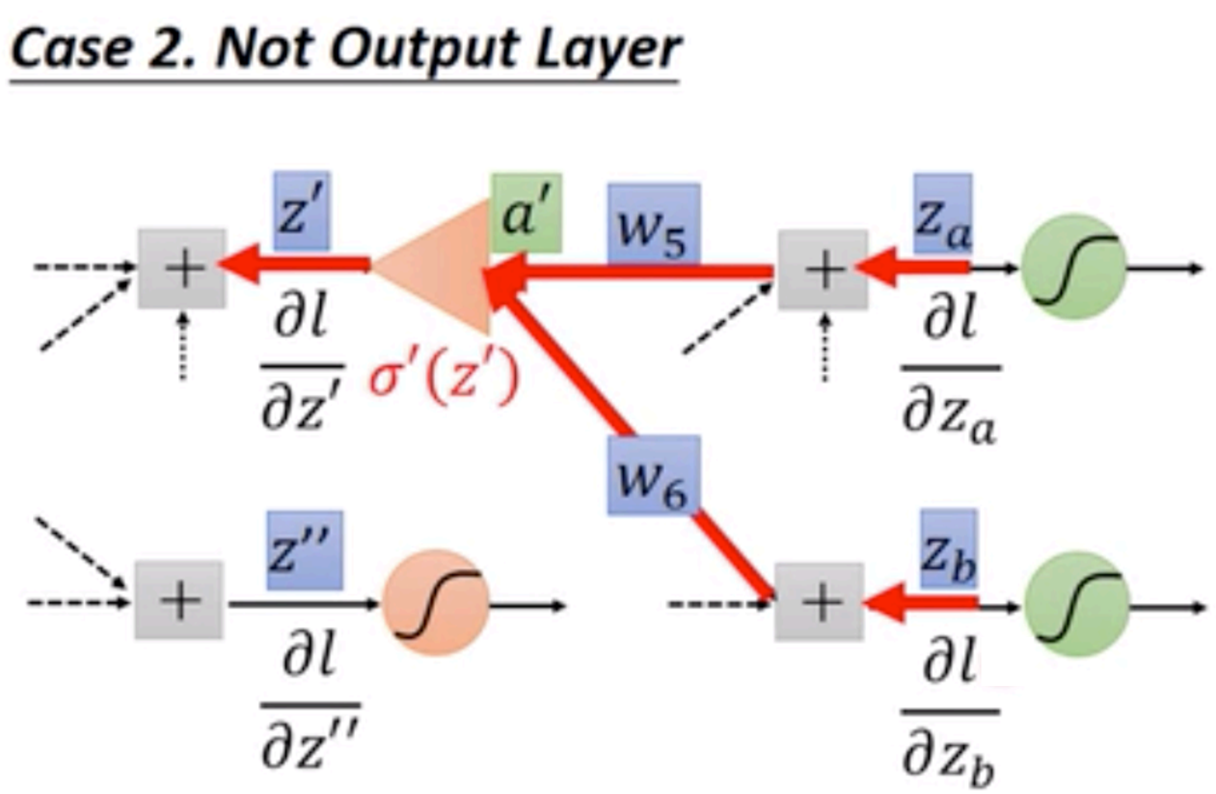
Case 1. Output Layer

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

Backward Pass

对所有激活函数计算 $\partial C/\partial z$

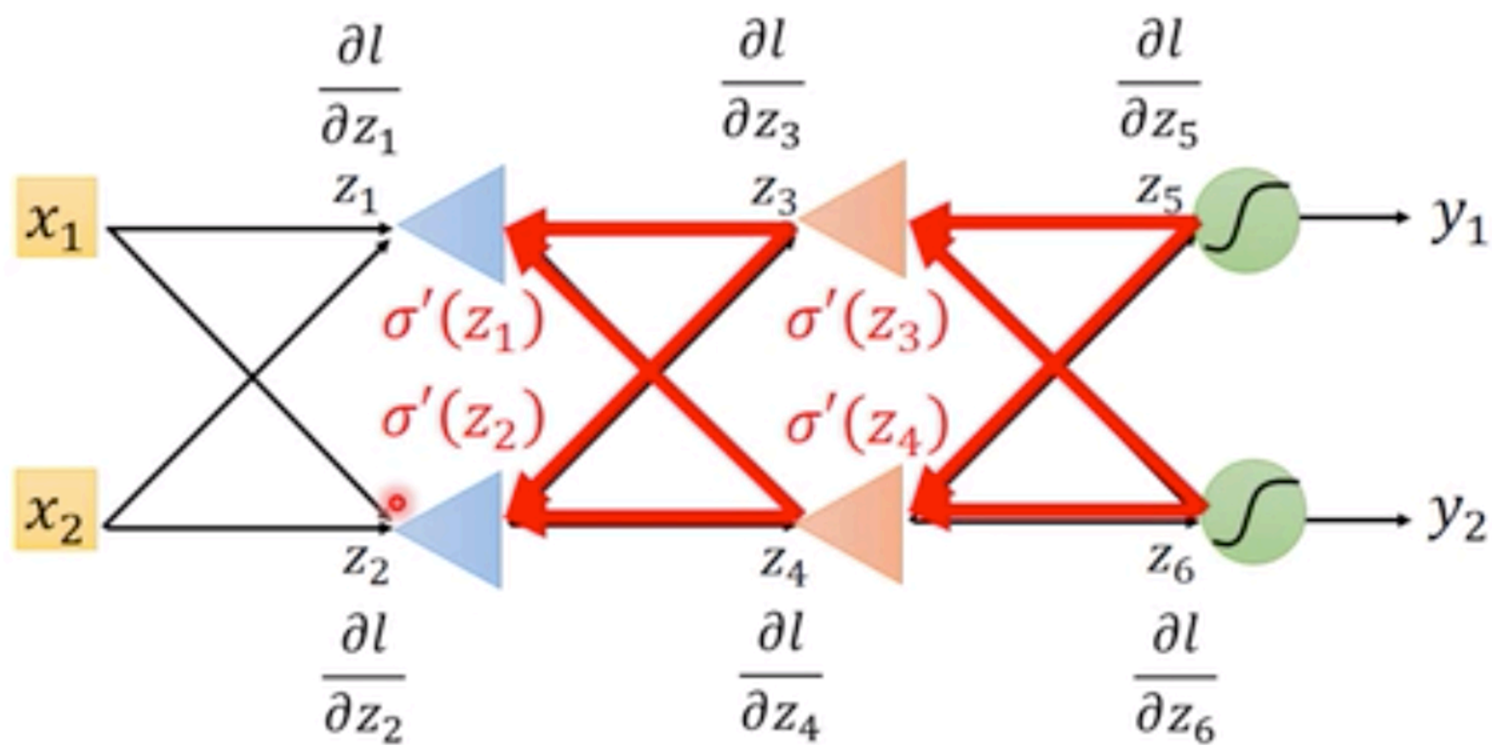
case2: $z^{'}, z^{''}$后面接的不是输出层



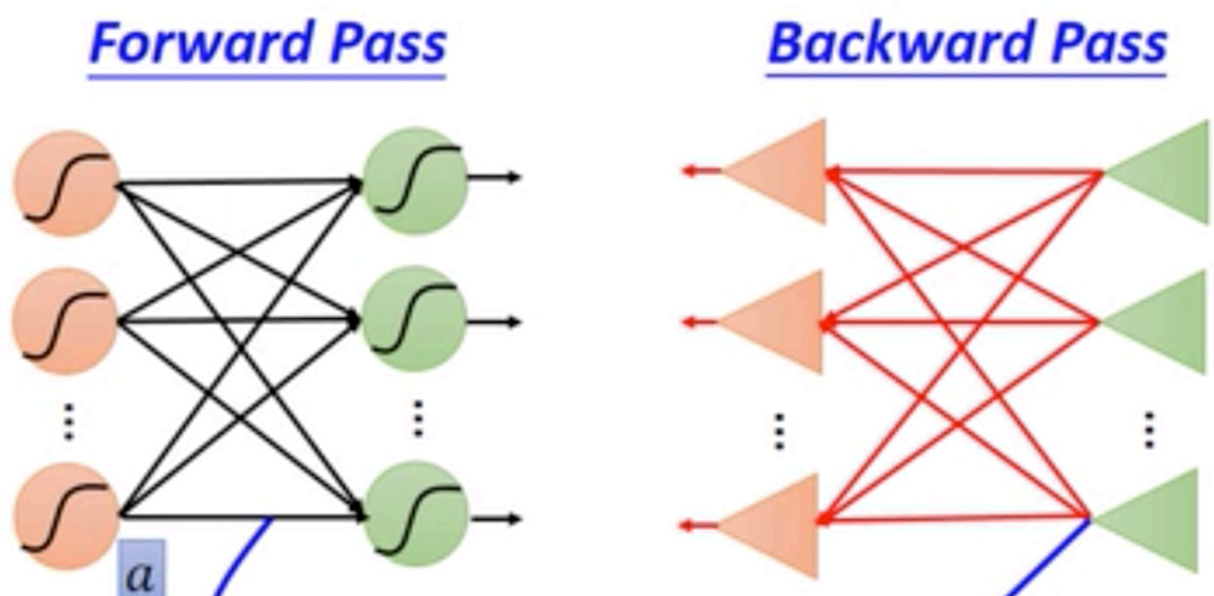
如果绿色的激活函数后面是输出层，那么我们就可以按照case1计算。
如果后面不是输出层，就按照case2的方法继续运算，知道遇到输出层

Backward Pass

如果从反向传播的角度来看，计算量不是不是很大



Summary



$$\frac{\partial z}{\partial w} = a \quad \text{X} \quad \frac{\partial l}{\partial z} \quad = \frac{\partial l}{\partial w}$$