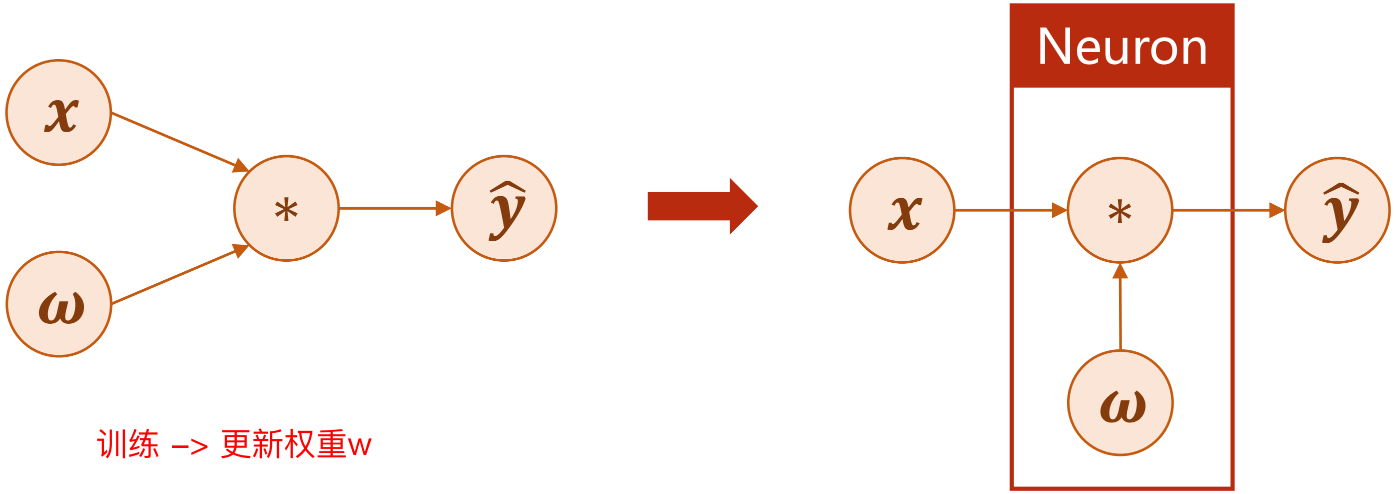
PyTorch Tutorial

04. Back Propagation

Compute gradient in simple network

Linear Model

$$\hat{y} = x * \omega$$



训练 –> 更新权重w

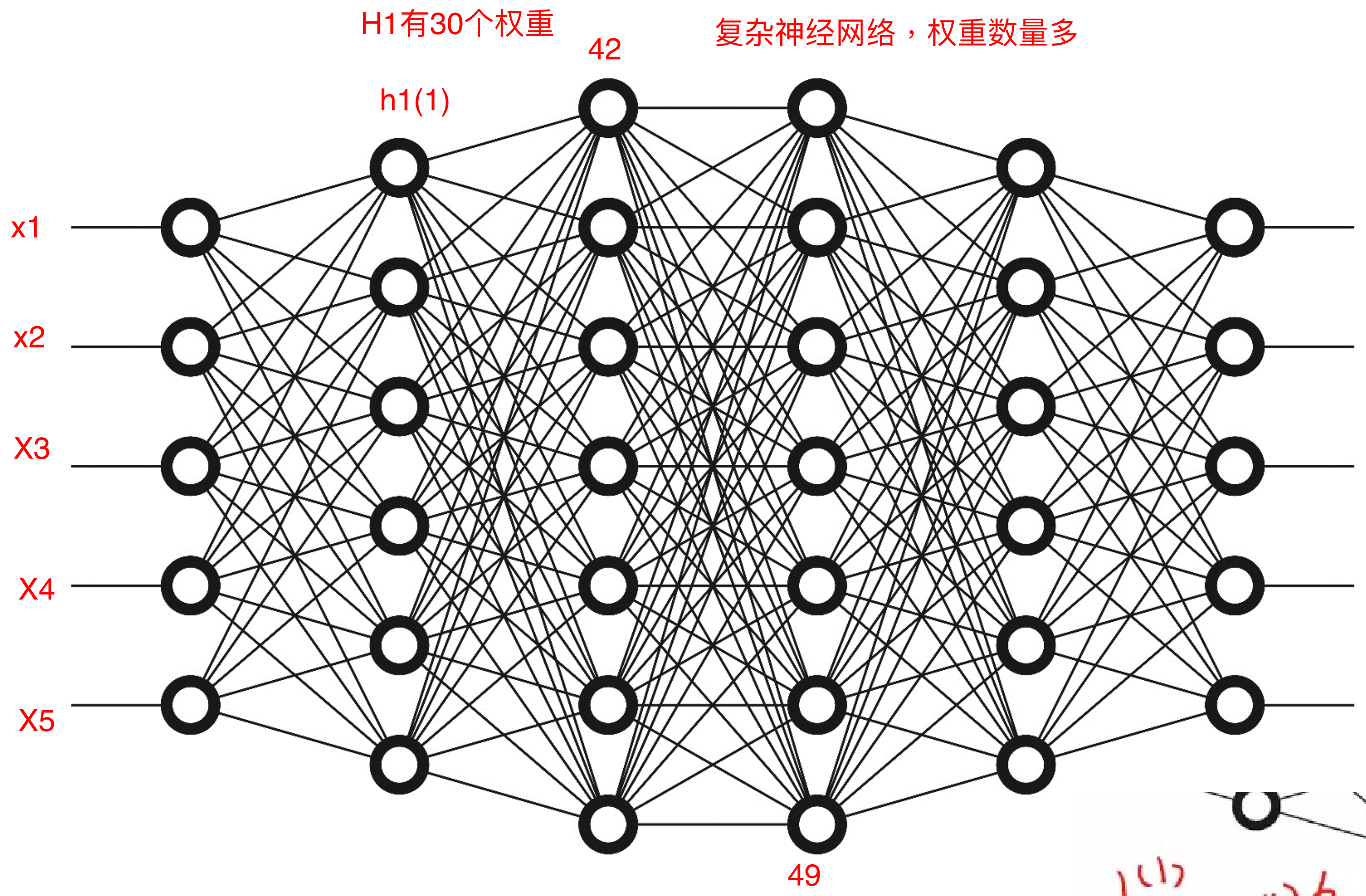
Stochastic Gradient Descent

$$\omega = \omega - \alpha \frac{\partial loss}{\partial \omega}$$

Derivative of Loss Function

$$\frac{\partial loss_n}{\partial \omega} = 2 \cdot x_n \cdot (x_n \cdot \omega - y_n)$$

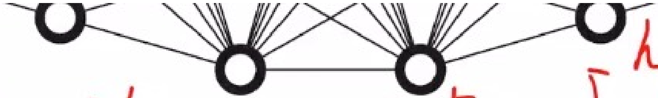
What about the complicated network?



传统求导方式写出来的表达式相当复杂

Gradient

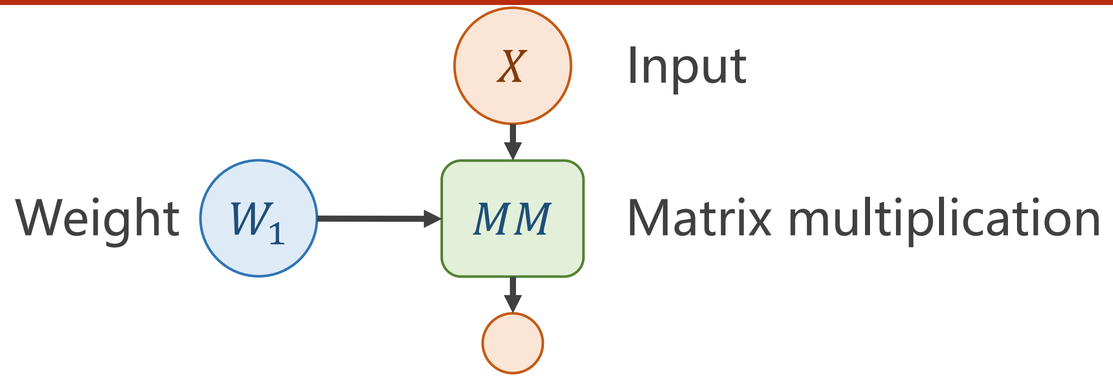
$$\frac{\partial loss}{\partial \omega} = ?$$



$h_1^{(1)} \in \mathbb{R}^6$ $\quad x \in \mathbb{R}^5$

$$\underset{6\times 1}{\overset{h}{[\ ]}} = \underset{6\times 5}{\overset{w}{[\ ]}} \underset{5\times 1}{\overset{x}{[\ ]}}$$

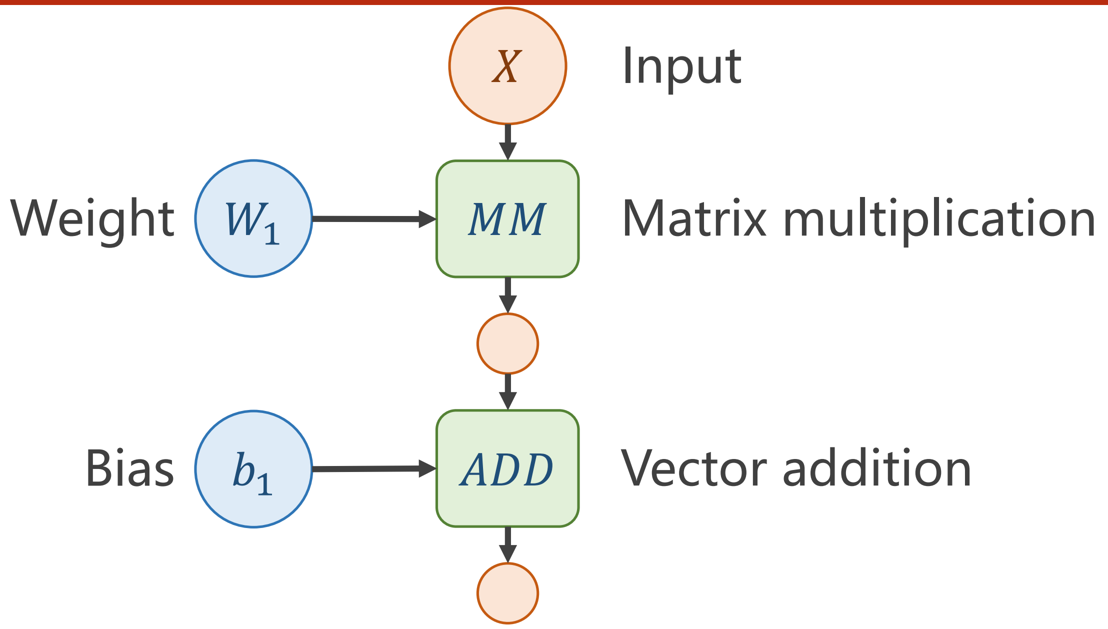
Computational Graph



A two layer neural network

$$\hat{y} = W_2(\underline{W_1 \cdot X} + b_1) + b_2$$

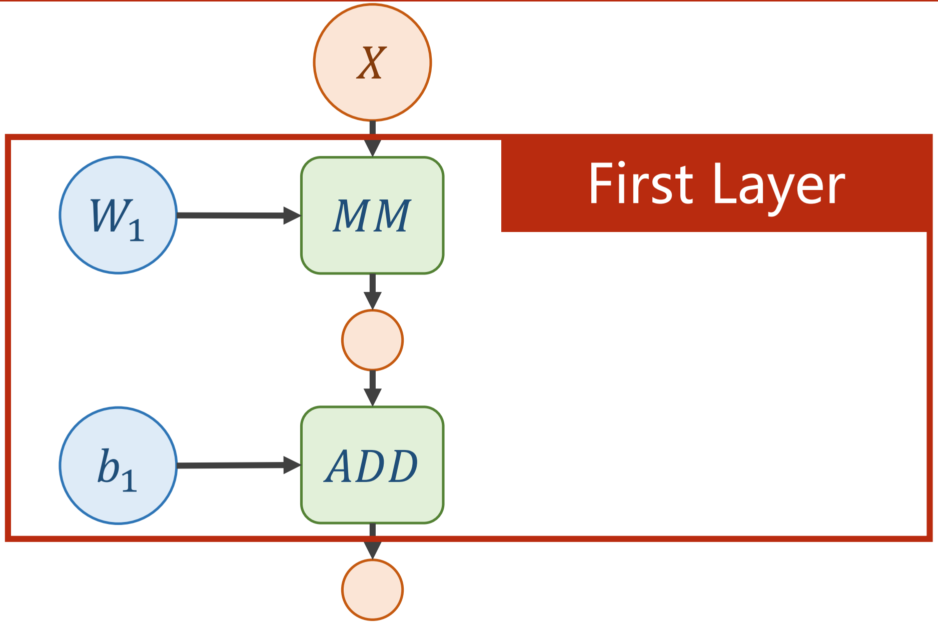
Computational Graph



A two layer neural network

$$\hat{y} = W_2(\underline{W_1 \cdot X + b_1}) + b_2$$

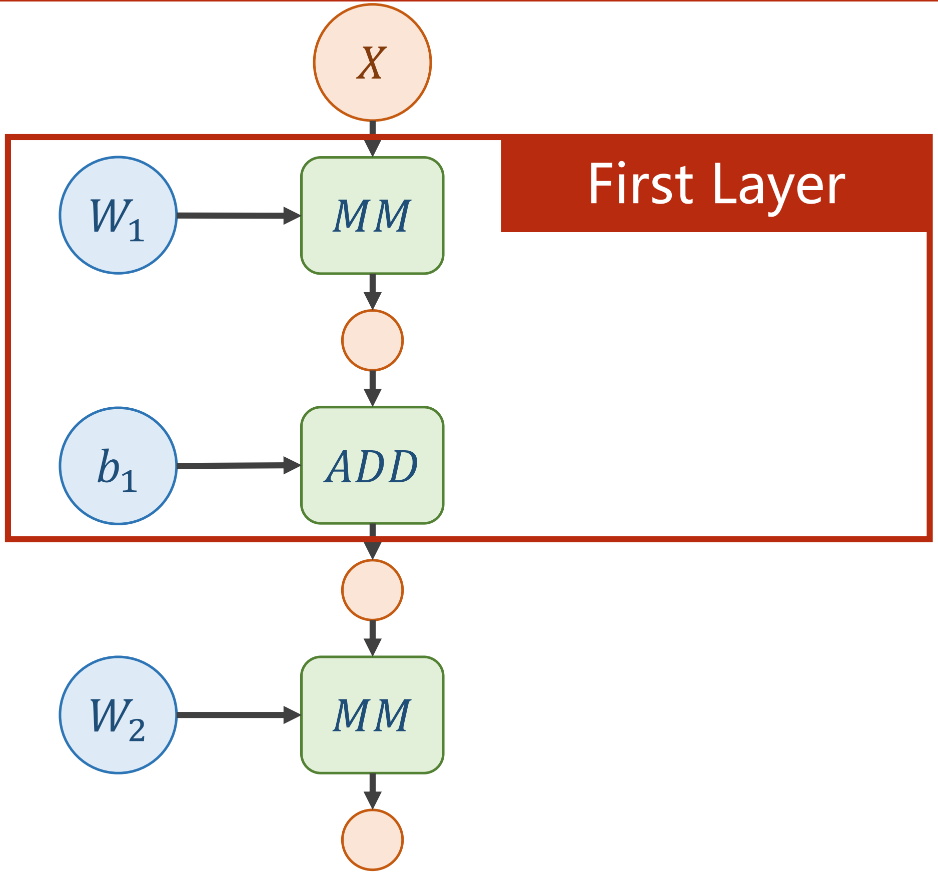
Computational Graph



A two layer neural network

$$\hat{y} = W_2(\underline{W_1 \cdot X + b_1}) + b_2$$

Computational Graph



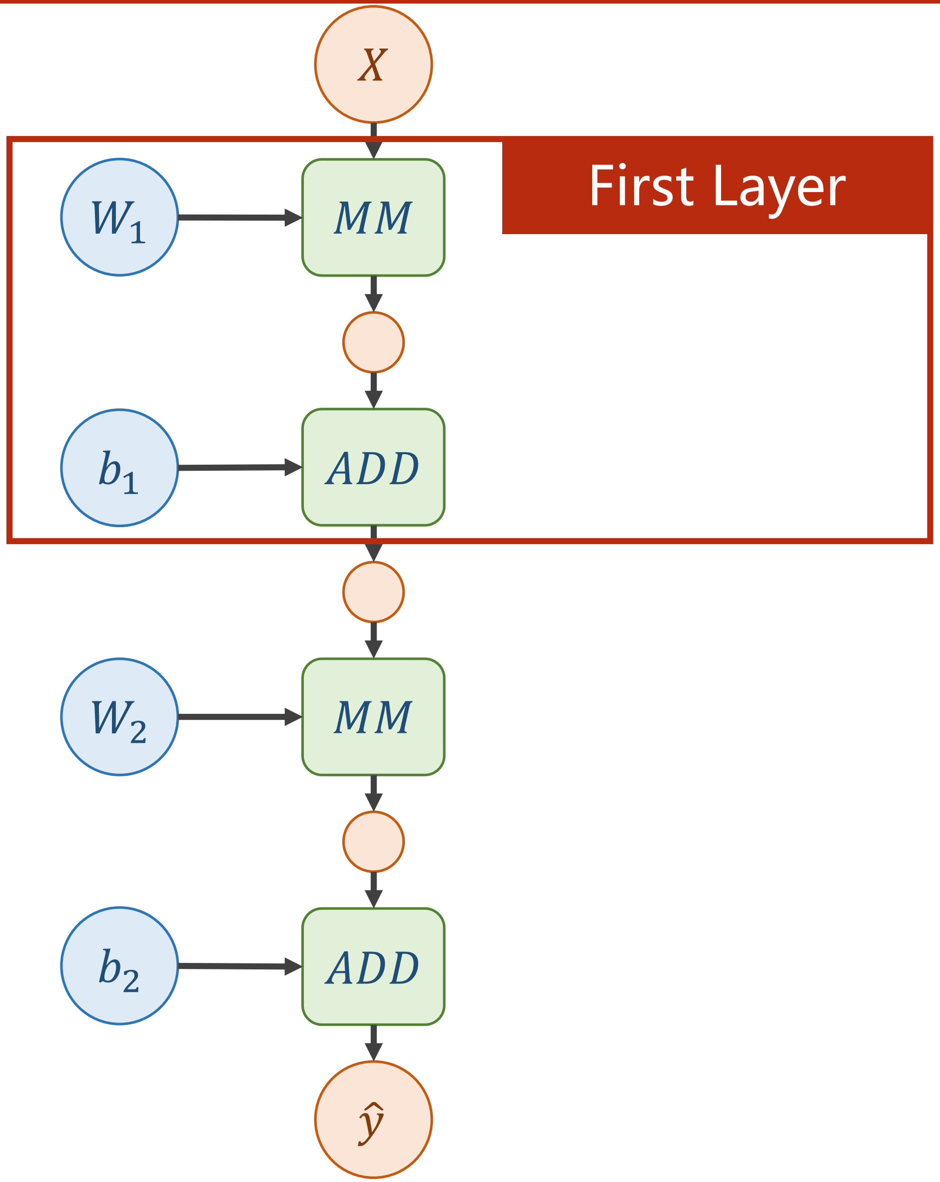
A two layer neural network

$$\hat{y} = \underline{W_2(W_1 \cdot X + b_1)} + b_2$$

Computational Graph

A two layer neural network

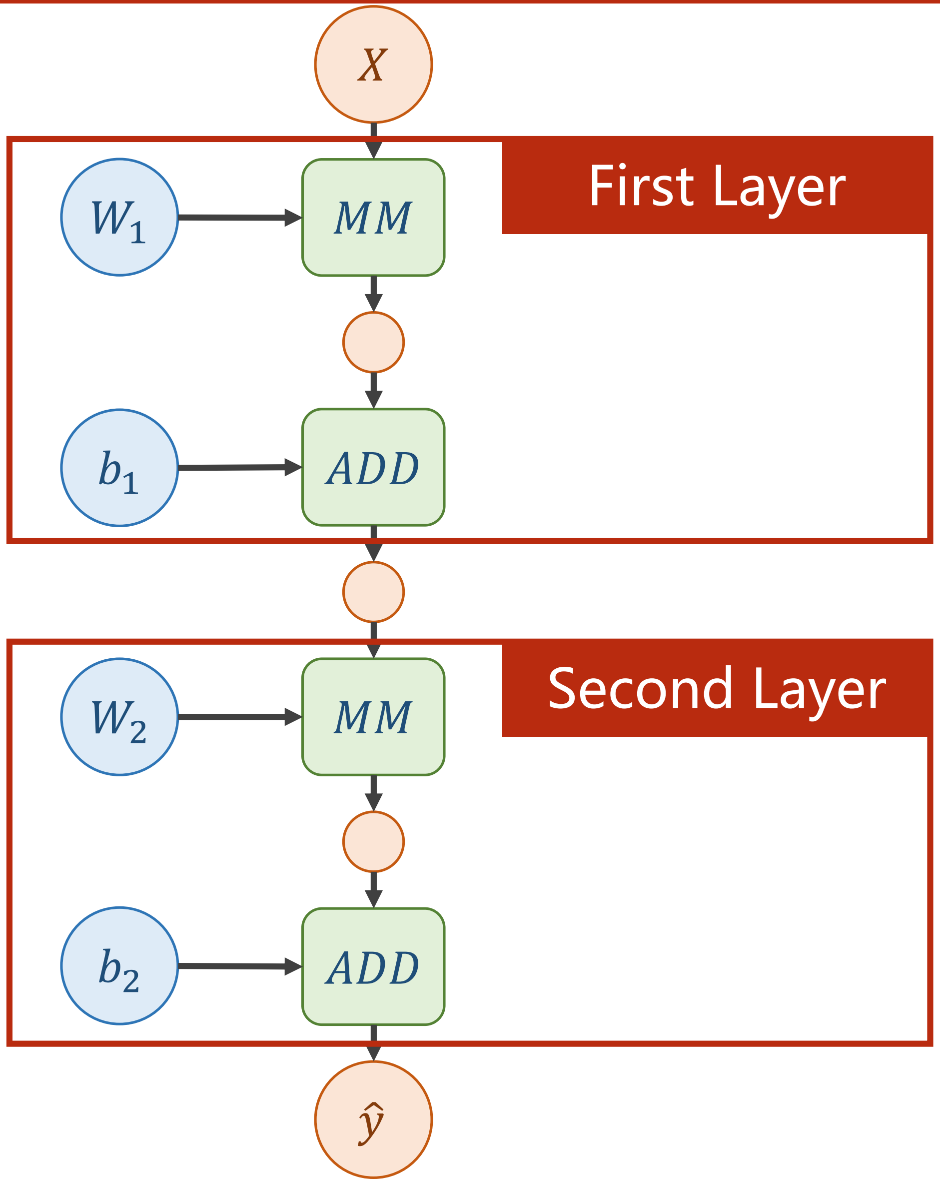
$$\hat{y} = \underline{W_2(W_1 \cdot X + b_1) + b_2}$$



Computational Graph

A two layer neural network

$$\hat{y} = \underline{W_2(W_1 \cdot X + b_1) + b_2}$$

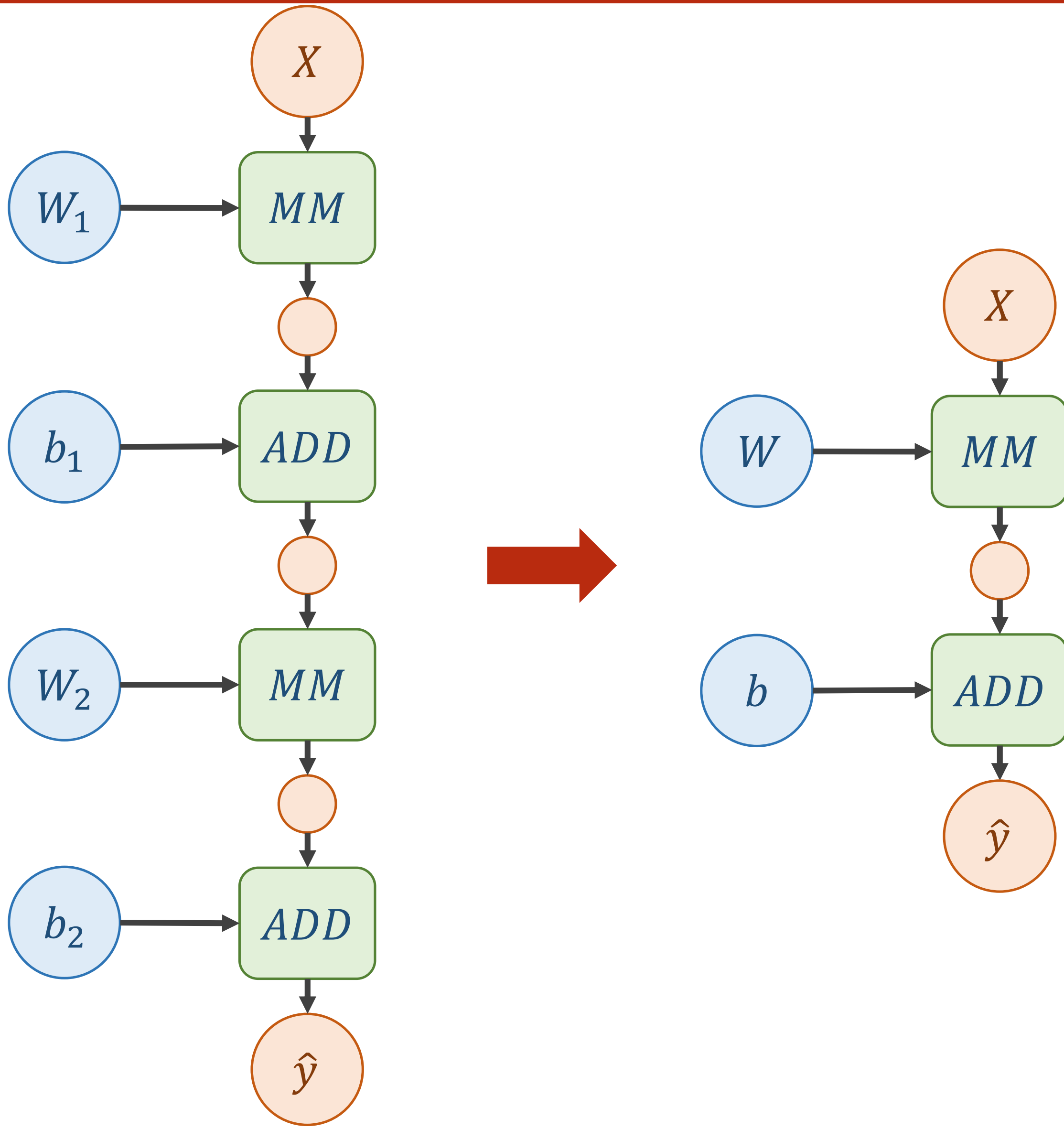


What problem about this two layer neural network?

A two layer neural network

$$\hat{y} = W_2(W_1 \cdot X + b_1) + b_2$$
$$= W_2 \cdot W_1 \cdot X + (W_2 b_1 + b_2)$$
$$= W \cdot X + b$$

化简之后发现不论几层最终的形式都是 y = wx+b



What problem about this two layer neural network?

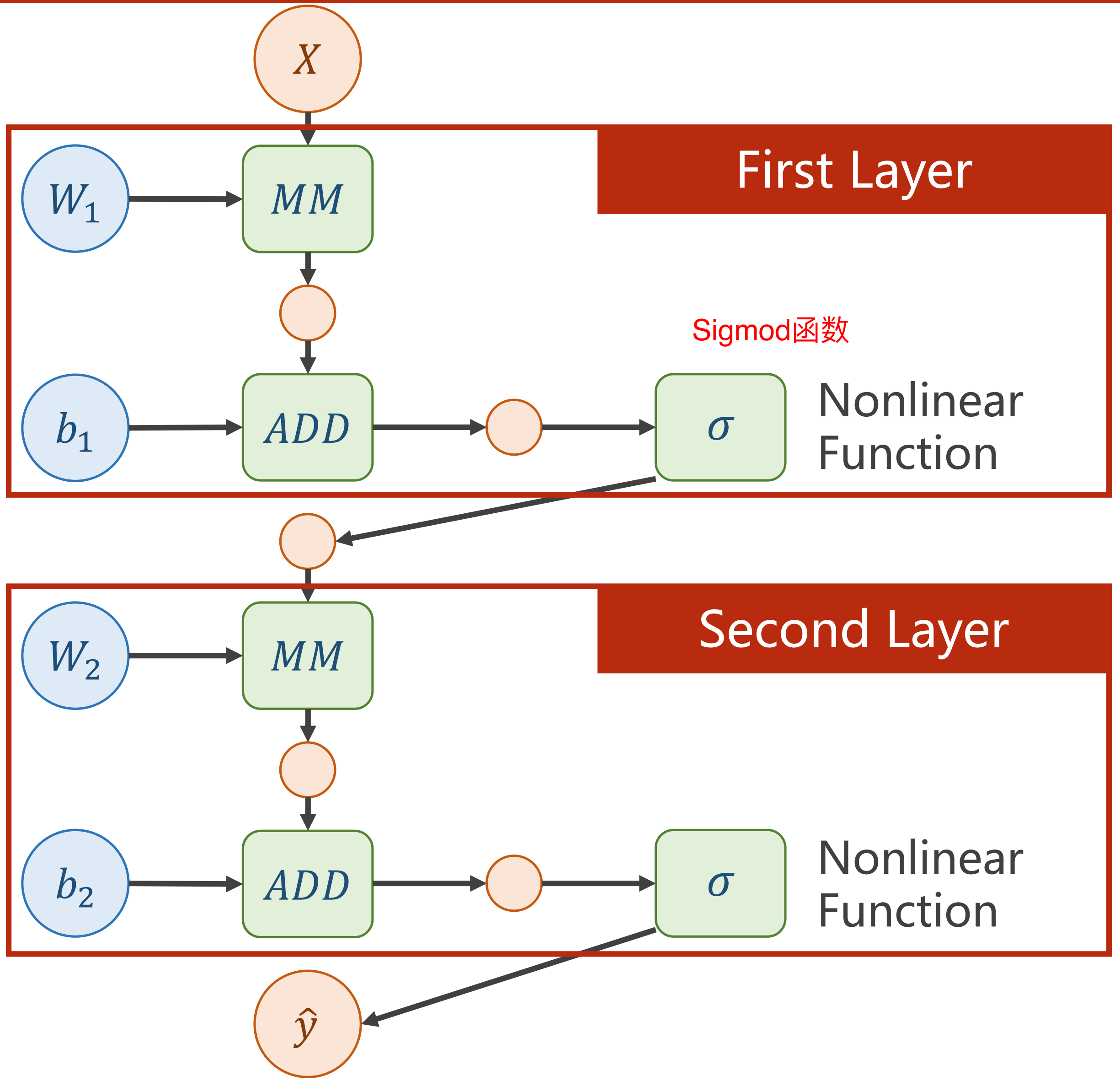
A two layer neural network

$$\hat{y} = W_2(W_1 \cdot X + b_1) + b_2$$
$$= W_2 \cdot W_1 \cdot X + (W_2 b_1 + b_2)$$
$$= W \cdot X + b$$

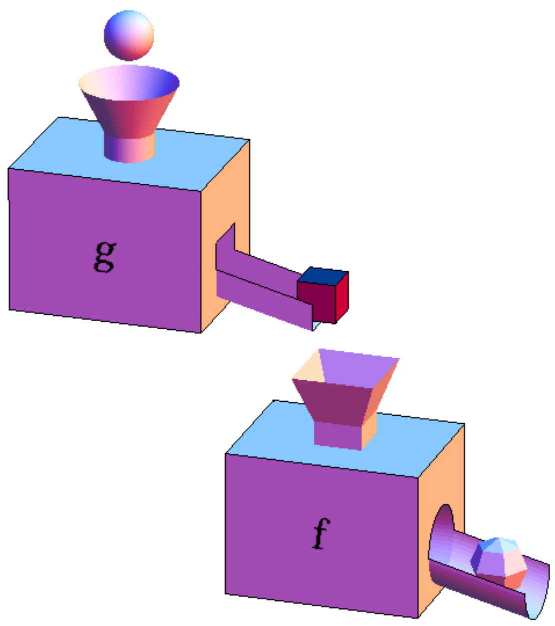
在每层之间增加非线性函数变换（激活函数）。

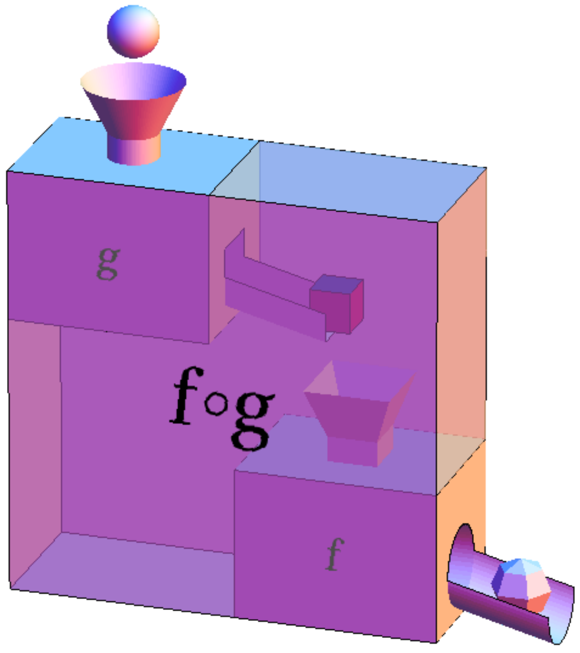
A nonlinear function is required by each layer.

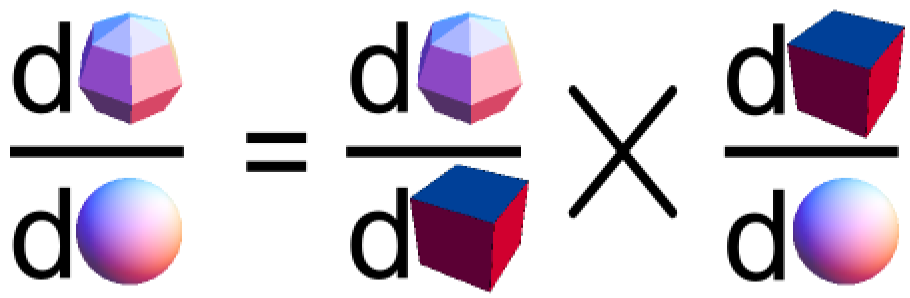
We shall talk about this later.



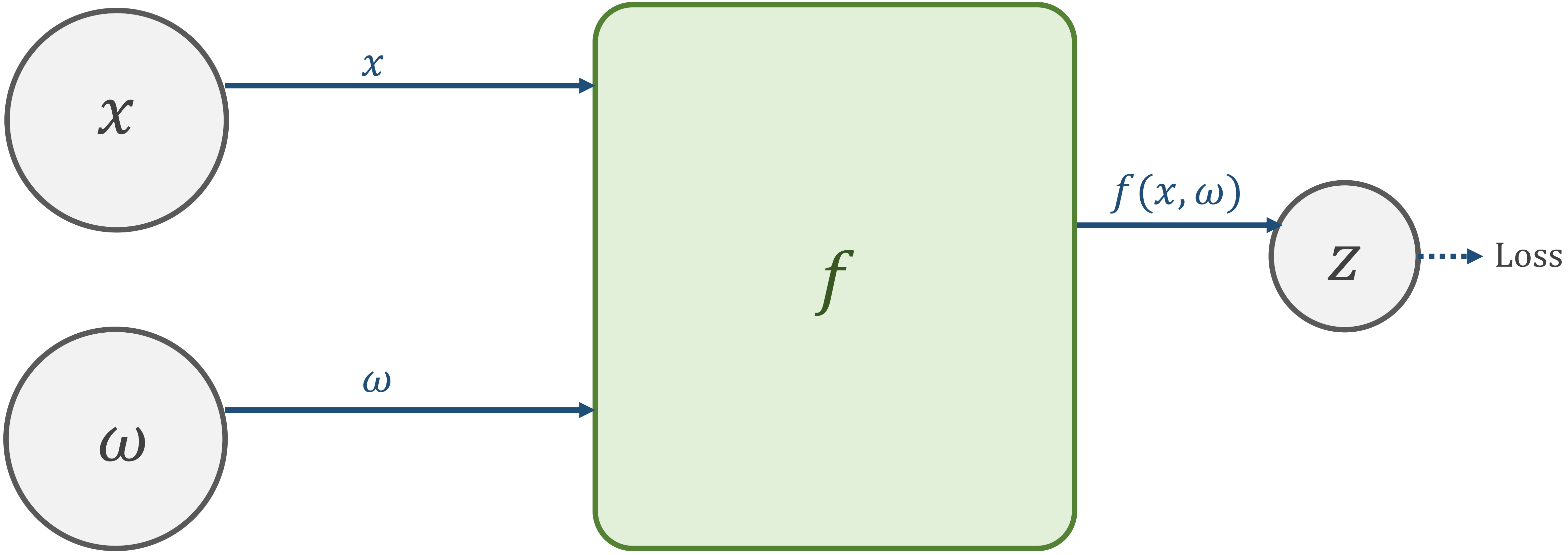
The composition of functions and Chain Rule



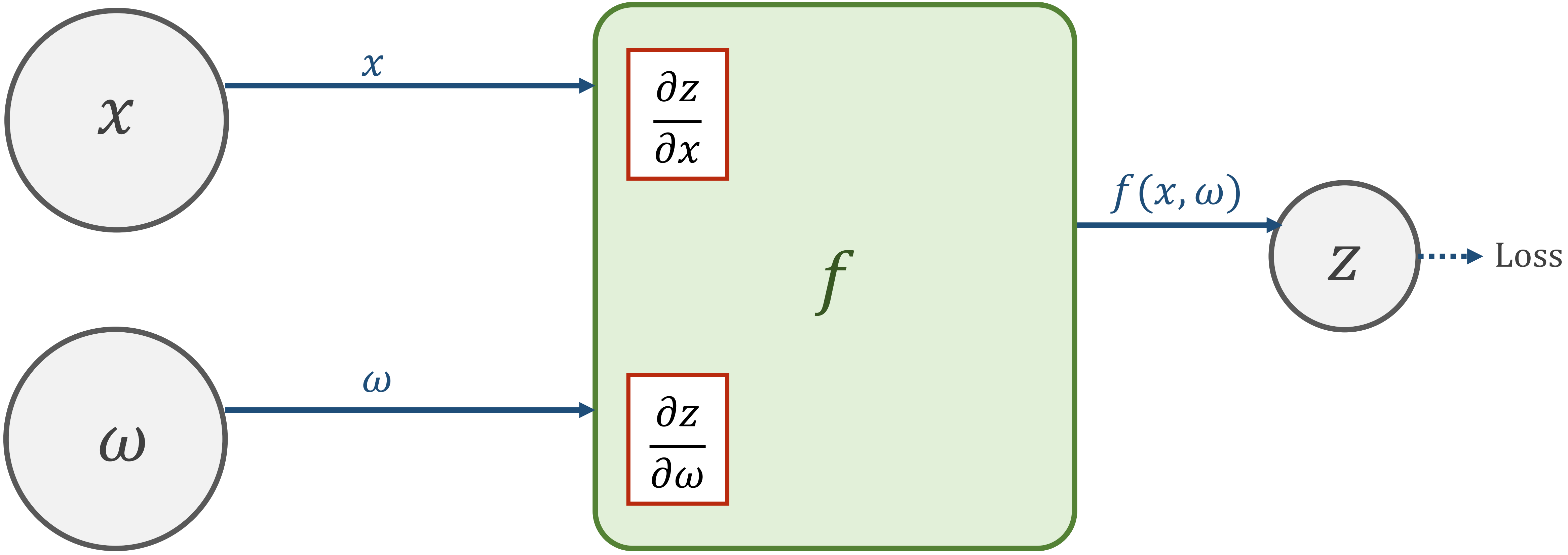




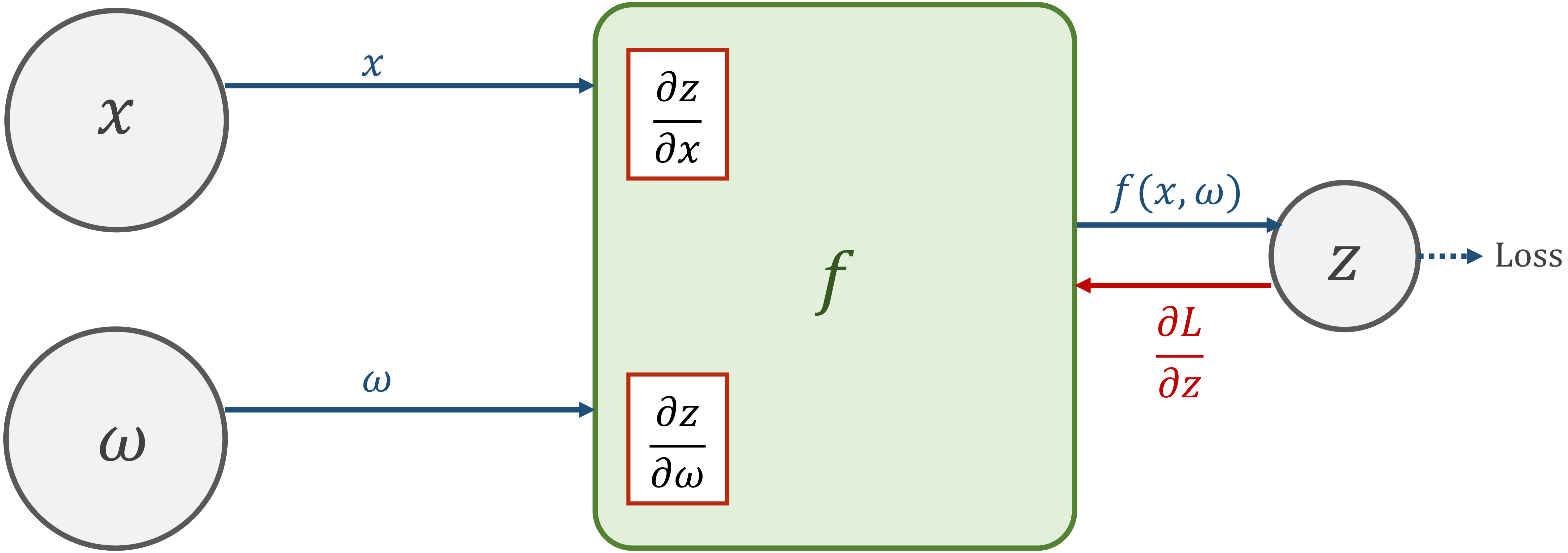
Chain Rule – 1. Create Computational Graph (Forward)



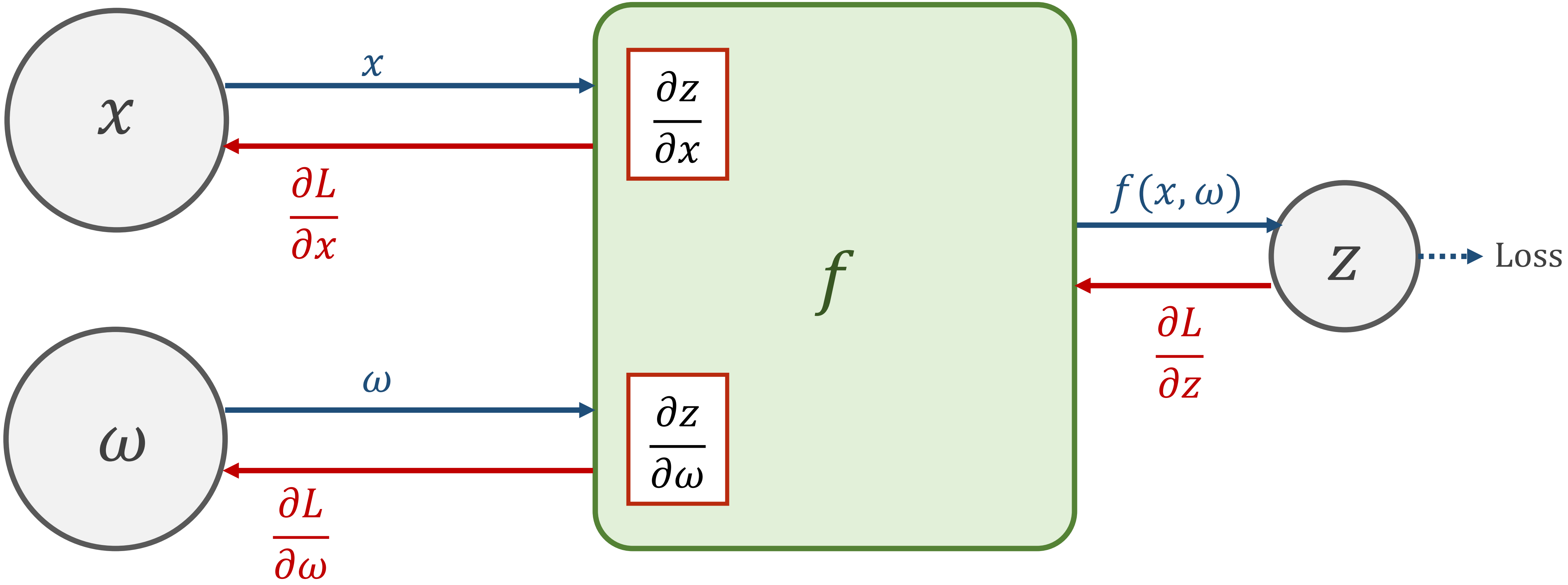
Chain Rule – 2. Local Gradient



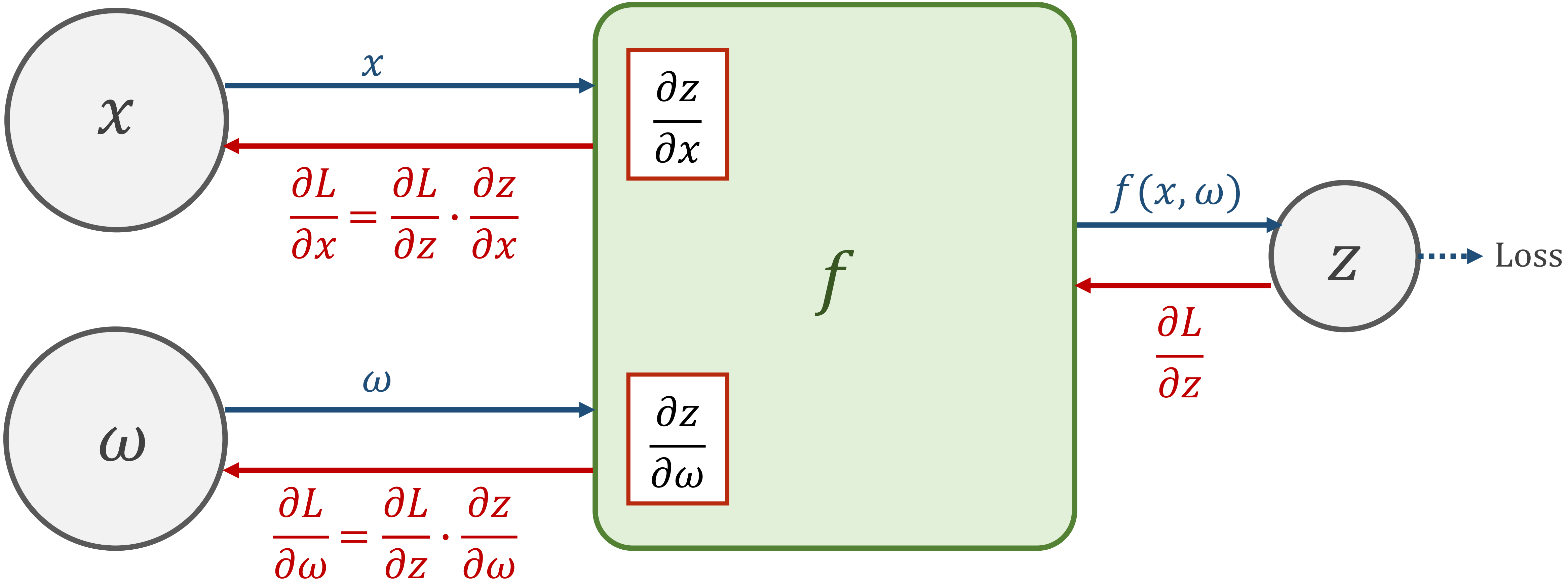
Chain Rule – 3. Given gradient from successive node



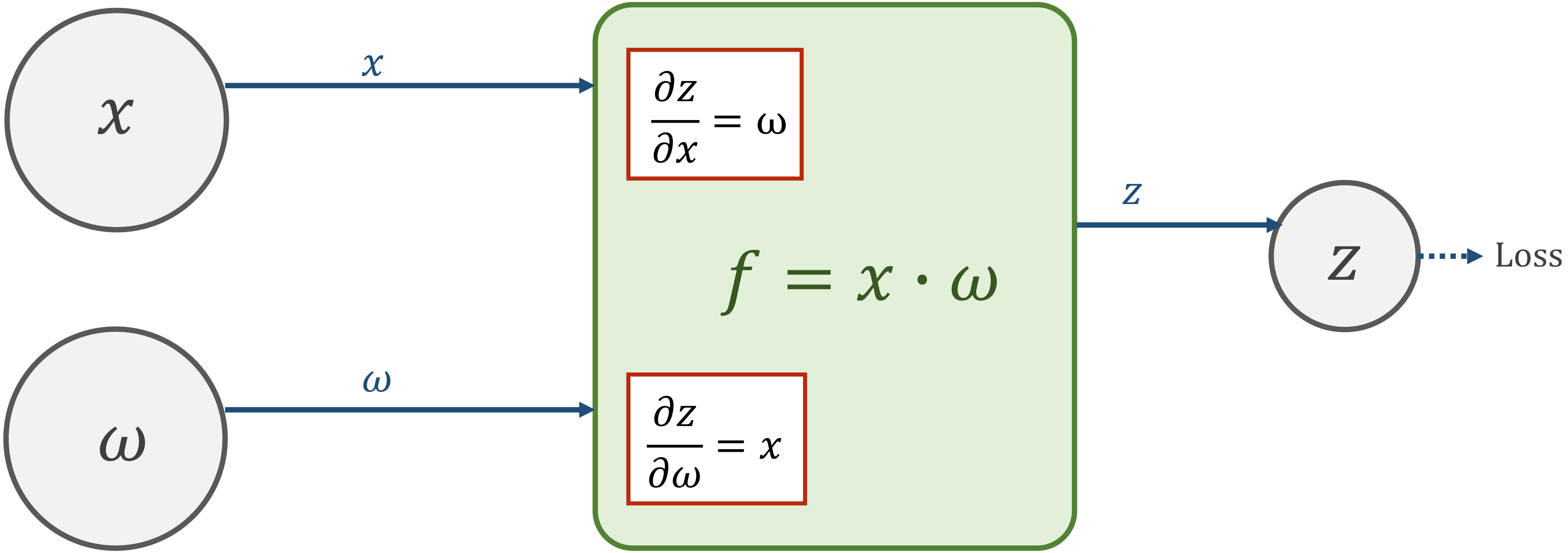
Chain Rule – 4. Use chain rule to compute the gradient (Backward)



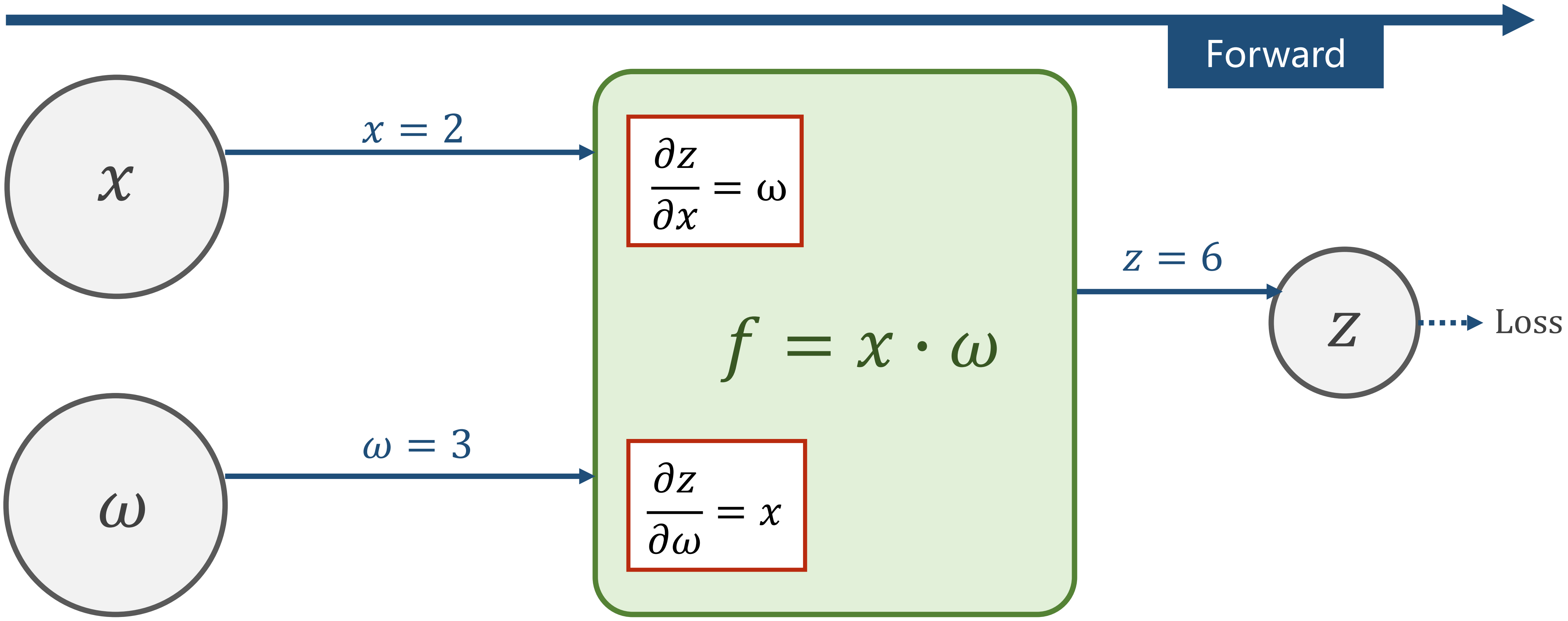
Chain Rule – 4. Use chain rule to compute the gradient (Backward)



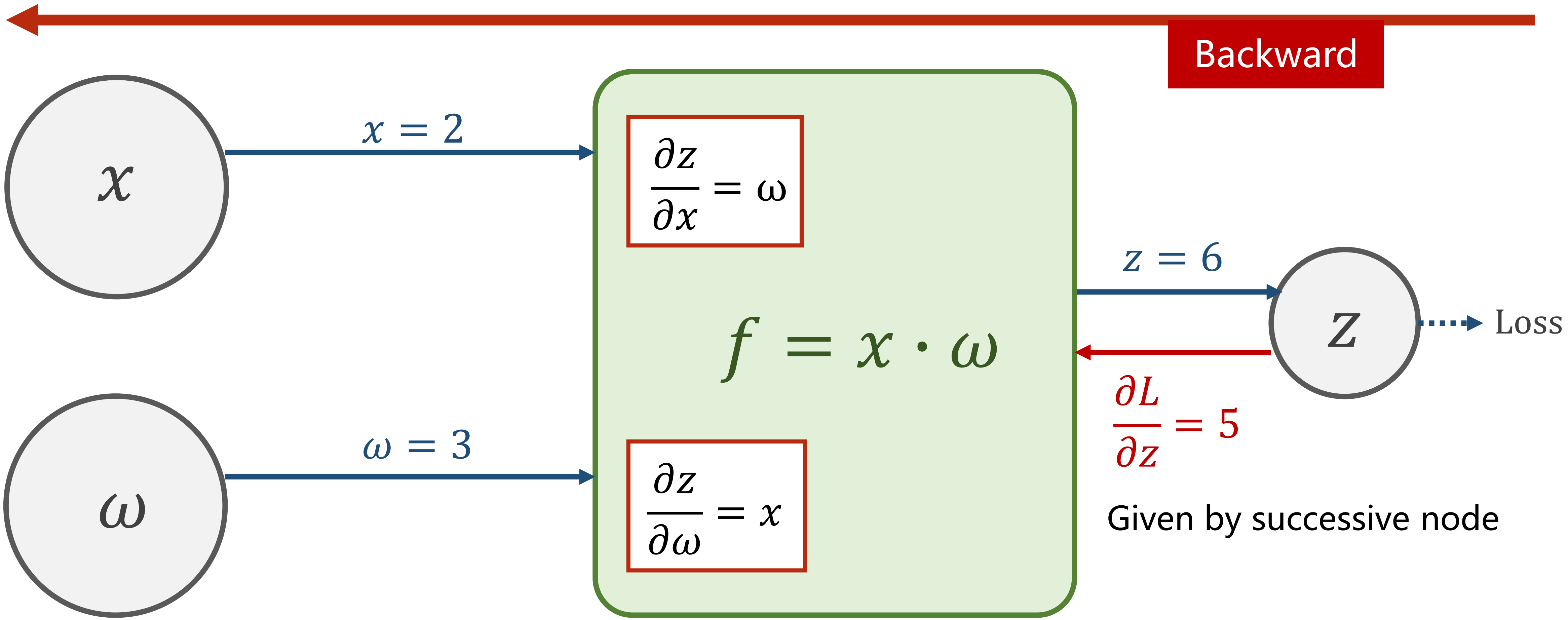
Example: $f = x \cdot \omega$



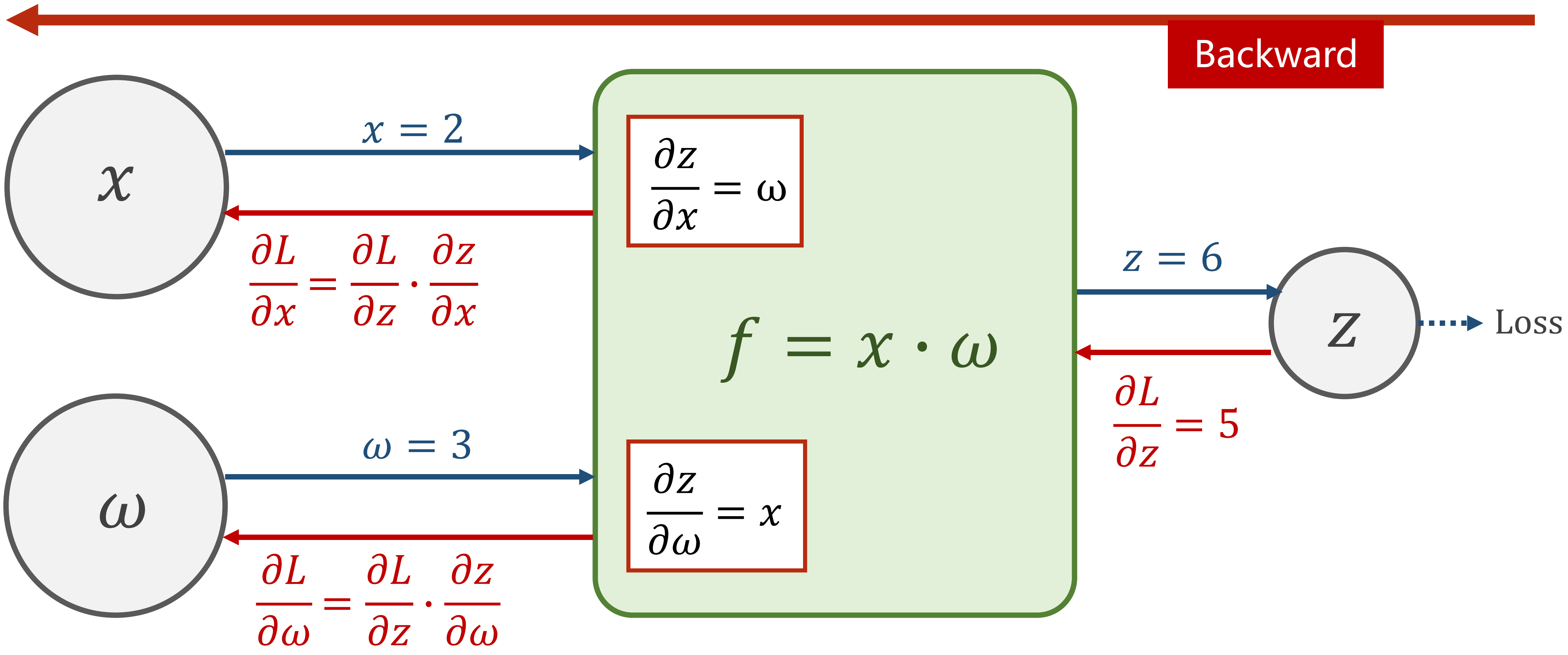
Example: $f = x \cdot \omega, x = 2, \omega = 3$



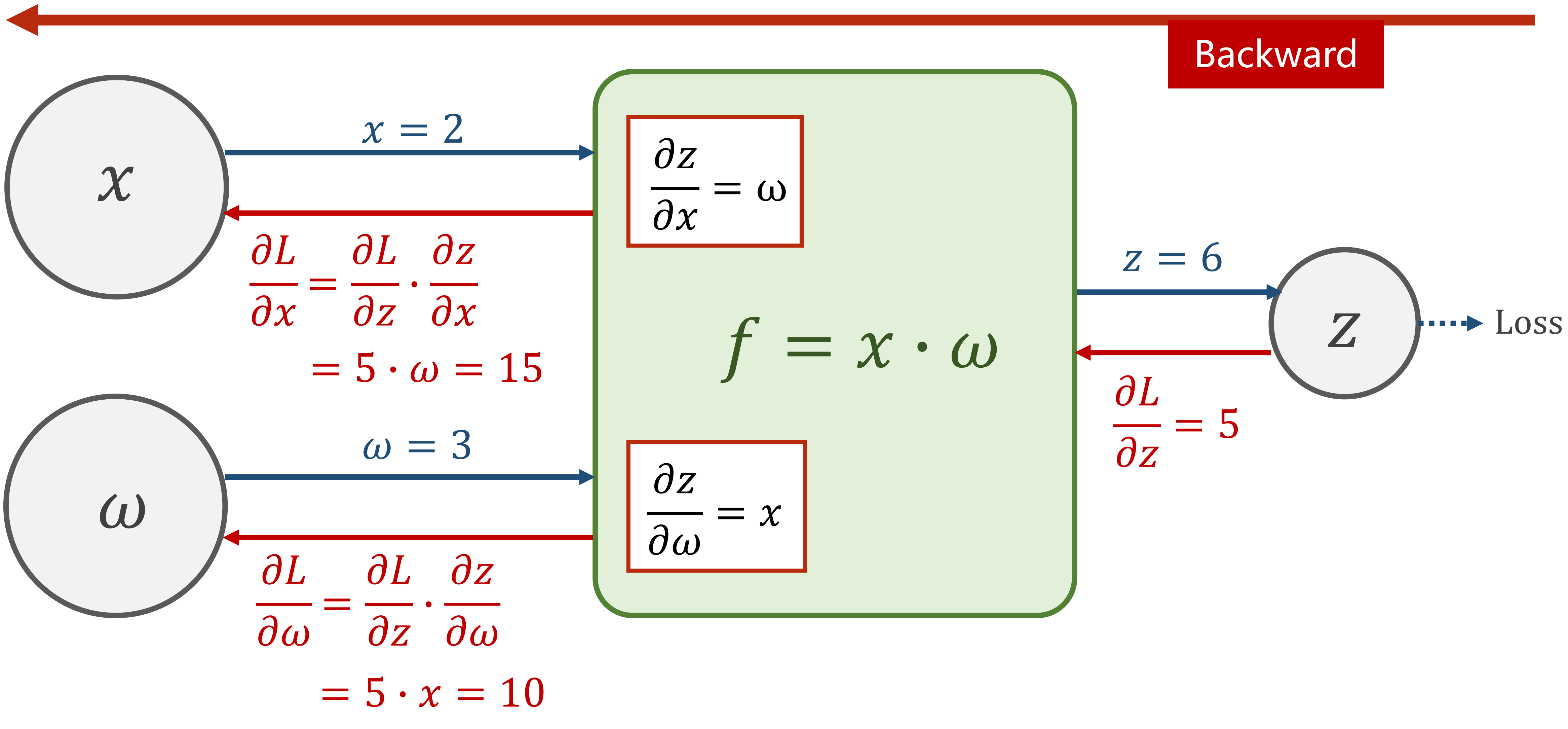
Example: Backward



Example: Backward



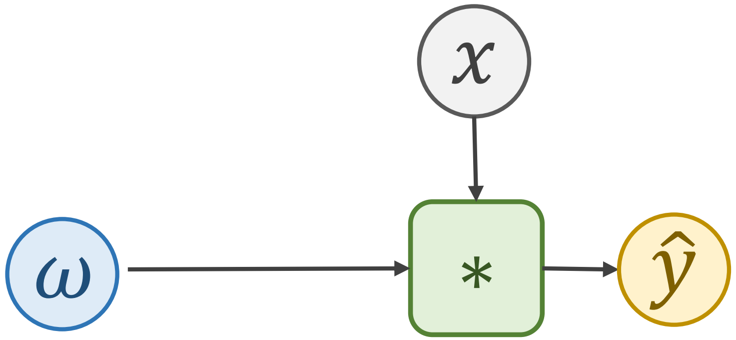
Example: Backward



Computational Graph of Linear Model

Linear Model

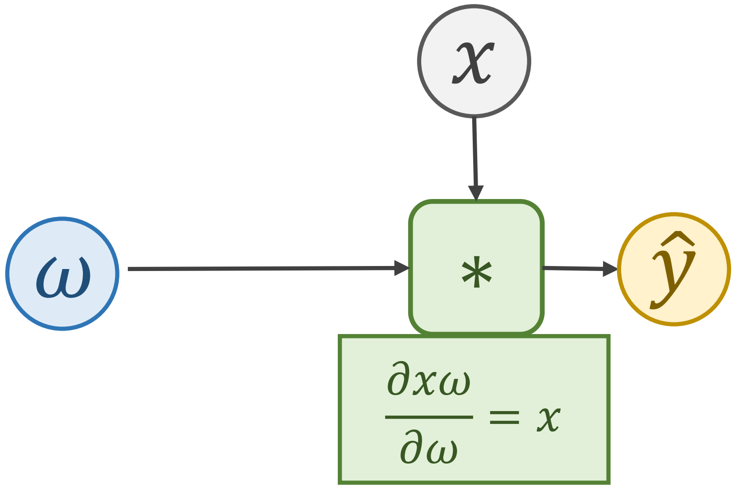
$$\hat{y} = x * \omega$$



Computational Graph of Linear Model

Linear Model

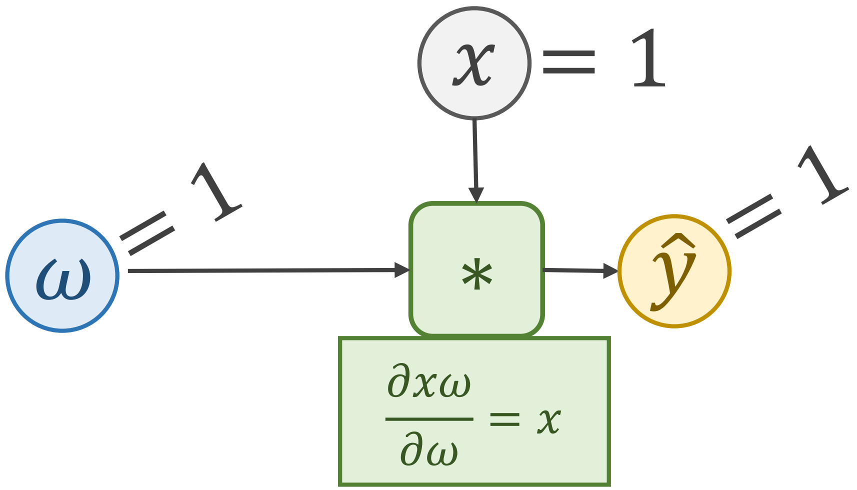
$$\hat{y} = x * \omega$$



Computational Graph of Linear Model

Linear Model

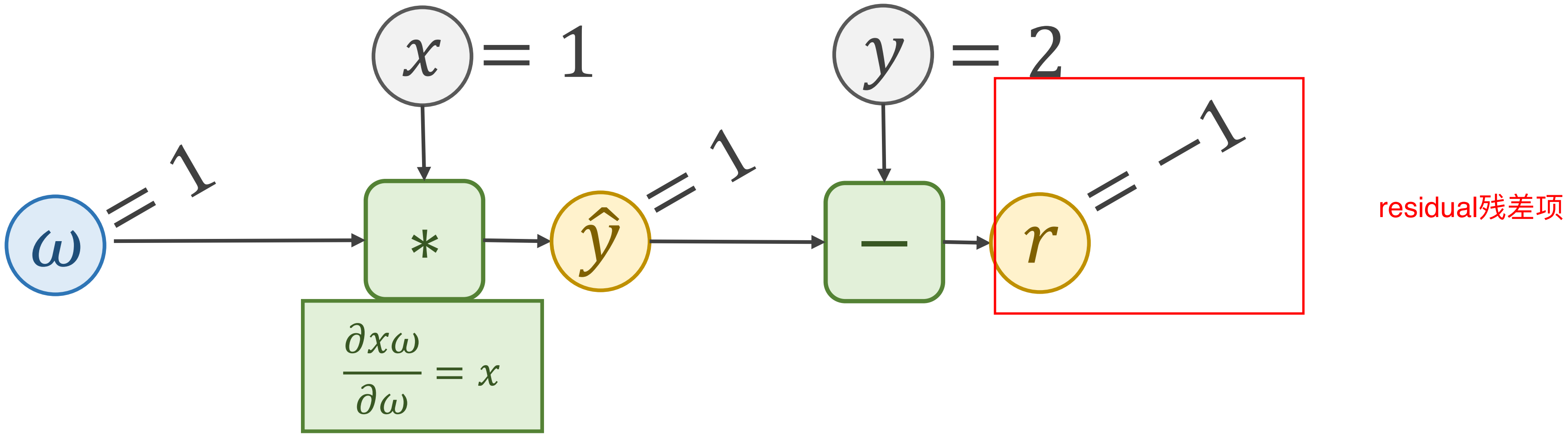
$$\hat{y} = x * \omega$$



Computational Graph of Linear Model

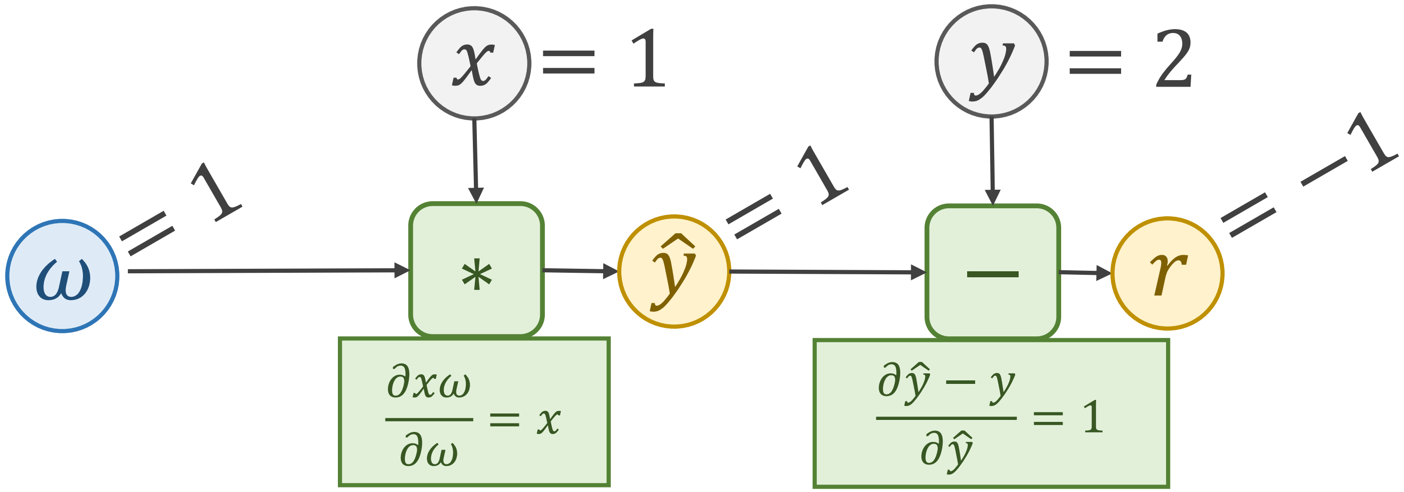
Linear Model
$\hat{y} = x * \omega$

Loss Function
$loss = (\hat{y} - y)^2 = (x \cdot \omega - y)^2$



Computational Graph of Linear Model

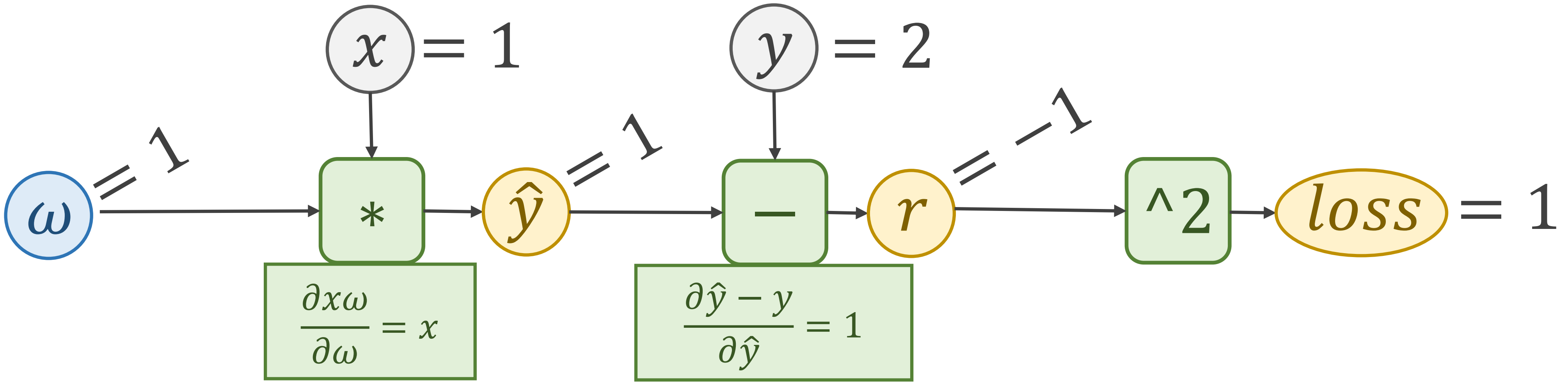
Linear Model	Loss Function
$\hat{y} = x * \omega$	$loss = (\hat{y} - y)^2 = (x \cdot \omega - y)^2$



Computational Graph of Linear Model

Linear Model
$\hat{y} = x * \omega$

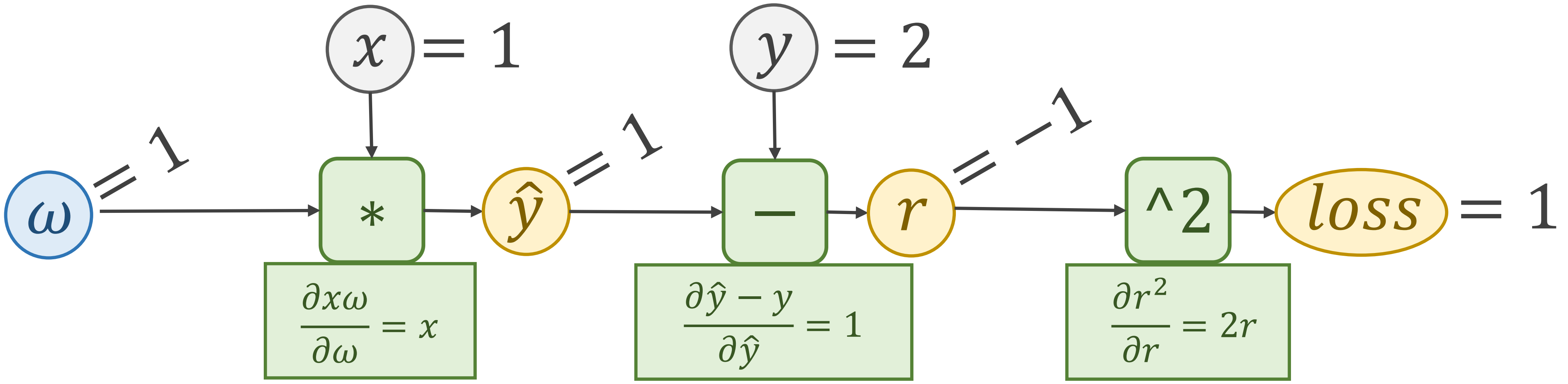
Loss Function
$loss = (\hat{y} - y)^2 = (x \cdot \omega - y)^2$



Computational Graph of Linear Model

Linear Model
$\hat{y} = x * \omega$

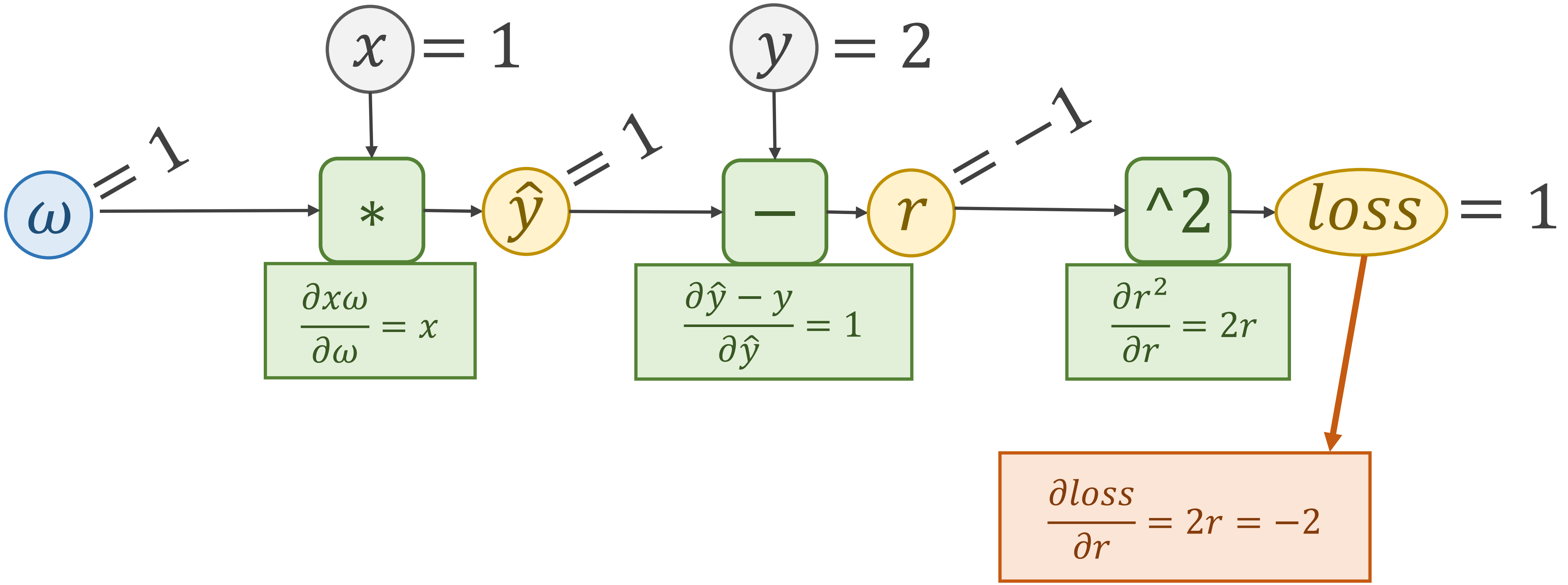
Loss Function
$loss = (\hat{y} - y)^2 = (x \cdot \omega - y)^2$



Computational Graph of Linear Model

Linear Model
$\hat{y} = x * \omega$

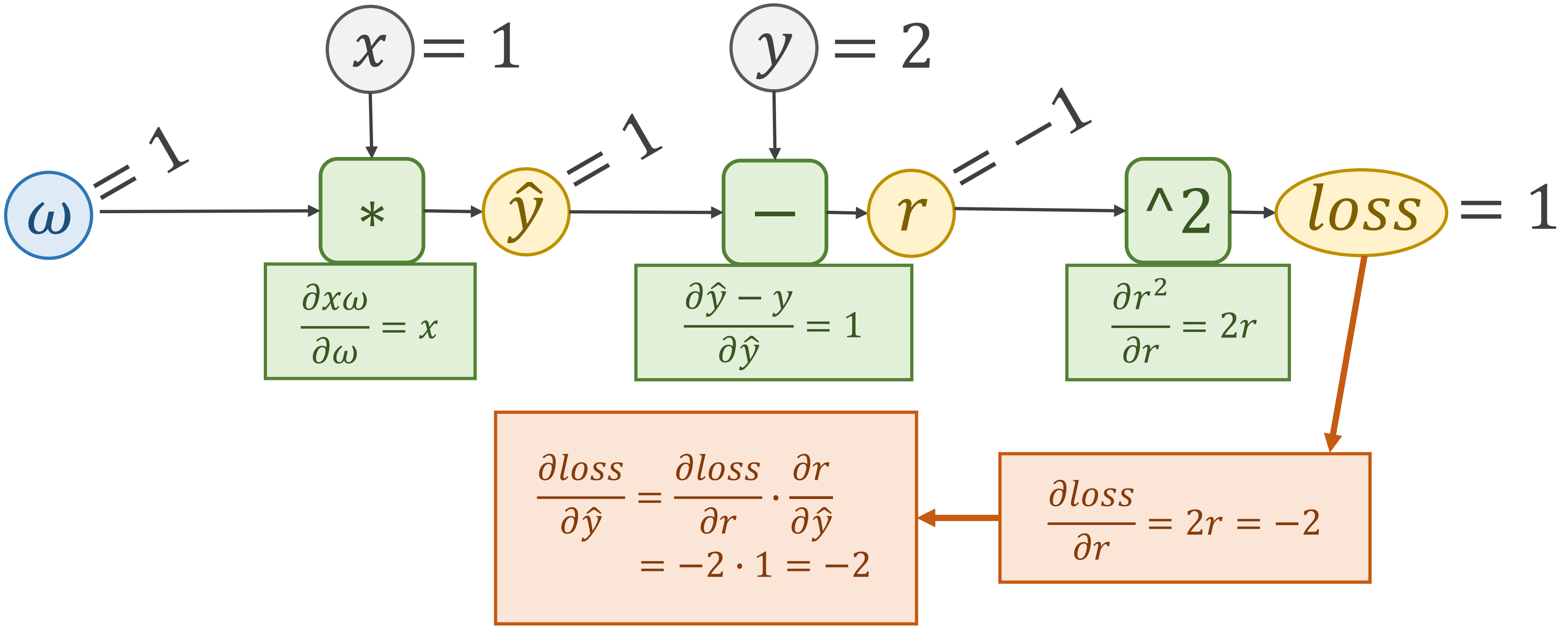
Loss Function
$loss = (\hat{y} - y)^2 = (x \cdot \omega - y)^2$



Computational Graph of Linear Model

Linear Model
$\hat{y} = x * \omega$

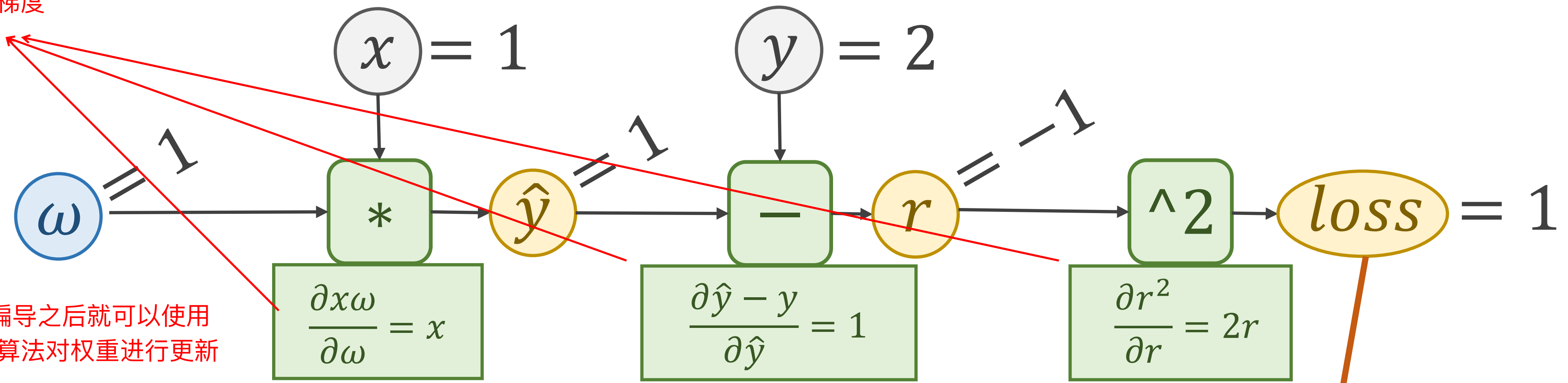
Loss Function
$loss = (\hat{y} - y)^2 = (x \cdot \omega - y)^2$



Computational Graph of Linear Model

Linear Model	Loss Function
$\hat{y} = x * \omega$	$loss = (\hat{y} - y)^2 = (x \cdot \omega - y)^2$

局部梯度



得到对权重的偏导之后就可以使用梯度下降算法对权重进行更新

$$\frac{\partial loss}{\partial \omega} = \frac{\partial loss}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \omega} = -2 \cdot x = -2 \cdot 1 = -2$$

$$\frac{\partial loss}{\partial \hat{y}} = \frac{\partial loss}{\partial r} \cdot \frac{\partial r}{\partial \hat{y}} = -2 \cdot 1 = -2$$

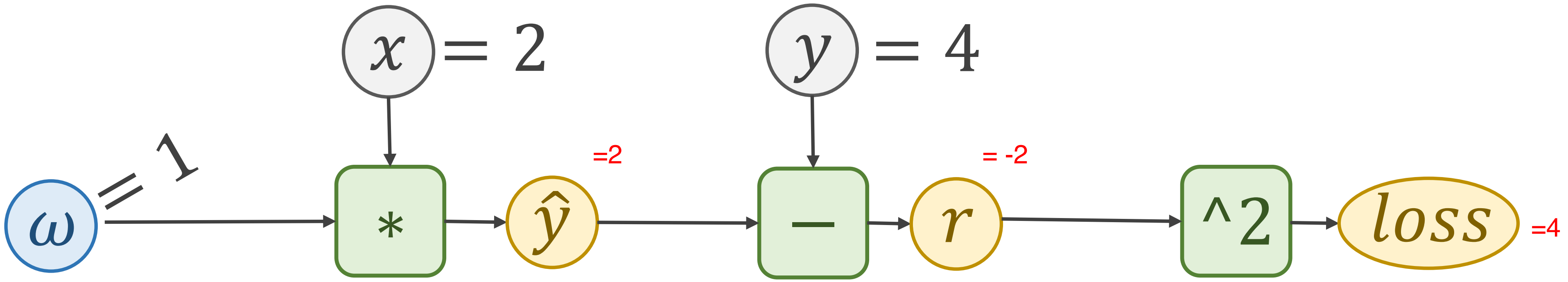
$$\frac{\partial loss}{\partial r} = 2r = -2$$

开始反向传播

Exercise 4-1: Compute the gradient with Computational Graph

Linear Model
$\hat{y} = x * \omega$

Loss Function
$loss = (\hat{y} - y)^2 = (x \cdot \omega - y)^2$



d_loss/d_w= d_loss/d_y' * d_y'/d_w
= -4 * 2 = -8

d_loss / d_y' = d_loss/ dr * d_r/d_y'
= -4 * 1 = -4

d_loss/ d_r = 2r = -4

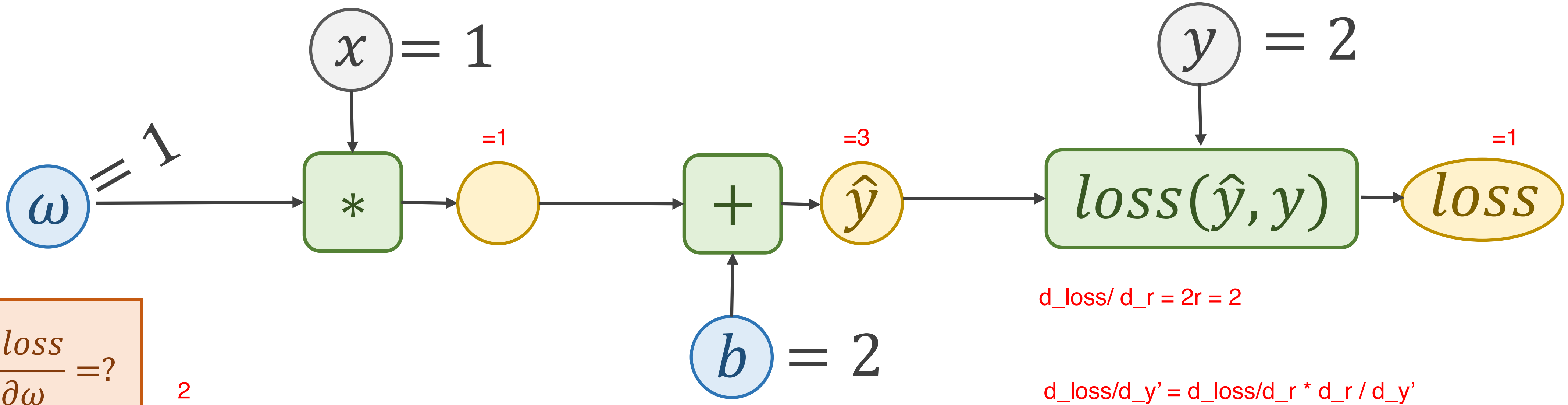
$$\frac{\partial loss}{\partial \omega} = ?$$

-8

Exercise 4-2: Compute gradient of Affine model

Affine Model
$\hat{y} = x * \omega + b$

Loss Function
$loss = (\hat{y} - y)^2 = (x \cdot \omega - y)^2$



d_loss/ d_r = 2r = 2

$\frac{\partial loss}{\partial \omega} = ?$

2

d_loss/d_y' = d_loss/d_r * d_r / d_y'
= 2 * 1 = 2

d_loss / db = d_loss / d_y' * d_y'/d_b
= 2 * 1 = 2

$\frac{\partial loss}{\partial b} = ?$

Tensor in PyTorch

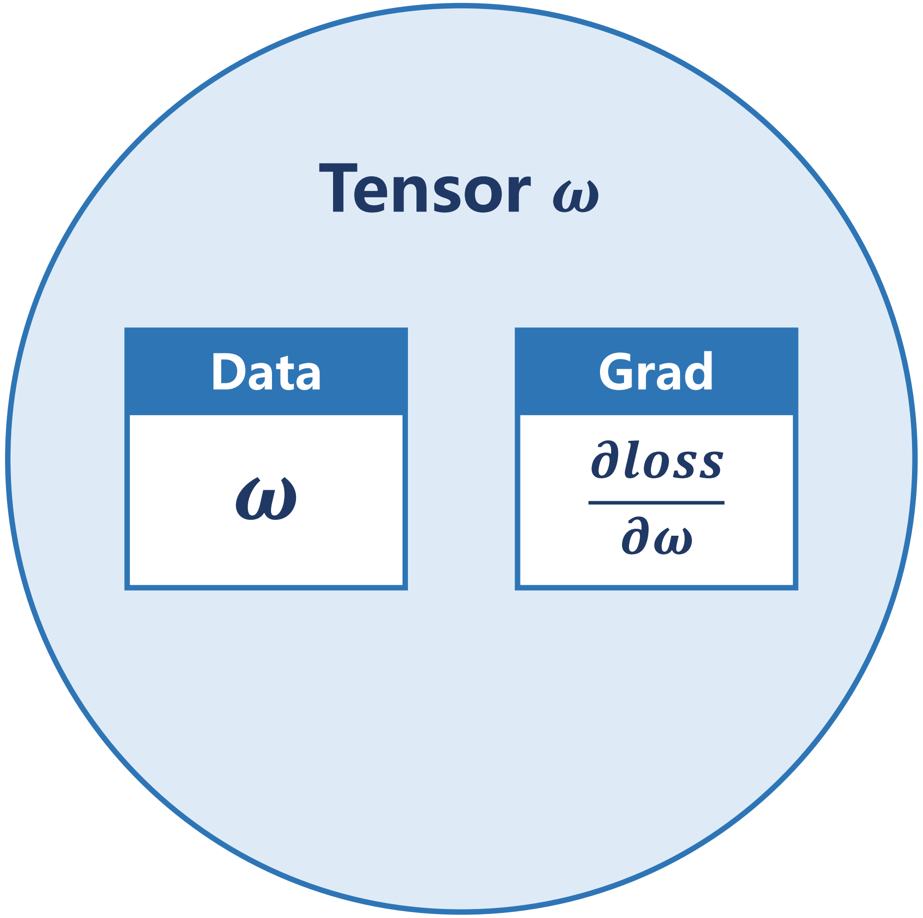
可以存标量、向量、矩阵、高维tensor...

In PyTorch, **Tensor** is the important component in constructing dynamic computational graph.

保存权重　　　　保存损失函数对权重的导数

It contains **data** and **grad**, which storage the value of node and gradient w.r.t loss respectively.

Tensor之间发生运算时会产生计算图，每次backward之后之前的计算图会释放。



Implementation of linear model with PyTorch

```python
import torch

x_data = [1.0, 2.0, 3.0]
y_data = [2.0, 4.0, 6.0]

w = torch.Tensor([1.0])
w.requires_grad = True
```

创建一个权重变量w

默认创建的tensor不会计算梯度，这里将w设置为需要计算梯度

If **autograd mechanics** are required, the element variable **requires_grad** of **Tensor** has to be set to **True**.

Implementation of linear model with PyTorch

实际哈桑进行的是tensor之间的乘法

```
def forward(x):
    return x * w

def loss(x, y):
    y_pred = forward(x)
    return (y_pred - y) ** 2
```

Define the linear model:

Linear Model

$$\hat{y} = x * \omega$$

之后由w计算出的中间变量都是默认计算梯度的

Define the loss function:

```
def forward(x):
    return x * w

def loss(x, y):
    y_pred = forward(x)
    return (y_pred - y) ** 2
```

Loss Function

$$loss = (\hat{y} - y)^2 = (x \cdot \omega - y)^2$$

Implementation of linear model with PyTorch

```
print("predict (before training)",  4, forward(4).item())

for epoch in range(100):
    for x, y in zip(x_data, y_data):
        l = loss(x, y)
        l.backward()
        print('\tgrad:', x, y, w.grad.item())
        w.data = w.data - 0.01 * w.grad.data

        w.grad.data.zero_()

    print("progress:", epoch, l.item())

print("predict (after training)", 4, forward(4).item())
```

Forward, compute the loss.

Implementation of linear model with PyTorch

l是表示loss的张量，调用其的backward()函数，可以自动反向传播计算梯度。

```
print("predict (before training)",  4, forward(4).item())

for epoch in range(100):
    for x, y in zip(x_data, y_data):
        l = loss(x, y)
        l.backward()
        print('\tgrad:', x, y, w.grad.item())
        w.data = w.data - 0.01 * w.grad.data

        w.grad.data.zero_()

    print("progress:", epoch, l.item())

print("predict (after training)", 4, forward(4).item())
```

权重更新

每次进行完反向传播之后，pytorch会将计算图释放

w.grad也是一个tensor，要用.item()方法取出这个数值

Backward, compute grad for Tensor whose **requires_grad** set to True

W.grad.item()：直接取出梯度值，返回一个Python的标量
W.gard.data，因为w.grad也是一个tensor，要用.item()取出其值来运算

显示将权重变量当中的梯度值清零。

Implementation of linear model with PyTorch

```
print("predict (before training)",  4, forward(4).item())

for epoch in range(100):
    for x, y in zip(x_data, y_data):
        l = loss(x, y)
        l.backward()
        print('\tgrad:', x, y, w.grad.item())
        w.data = w.data - 0.01 * w.grad.data

        w.grad.data.zero_()

    print("progress:", epoch, l.item())

print("predict (after training)", 4, forward(4).item())
```

The **grad** is utilized to update weight.

Implementation of linear model with PyTorch

```
print("predict (before training)",  4, forward(4).item())

for epoch in range(100):
    for x, y in zip(x_data, y_data):
        l = loss(x, y)
        l.backward()
        print('\tgrad:', x, y, w.grad.item())
        w.data = w.data - 0.01 * w.grad.data

        w.grad.data.zero_()

    print("progress:", epoch, l.item())

print("predict (after training)", 4, forward(4).i
```

NOTICE:

The grad computed by *.backward()* will be **accumulated**.

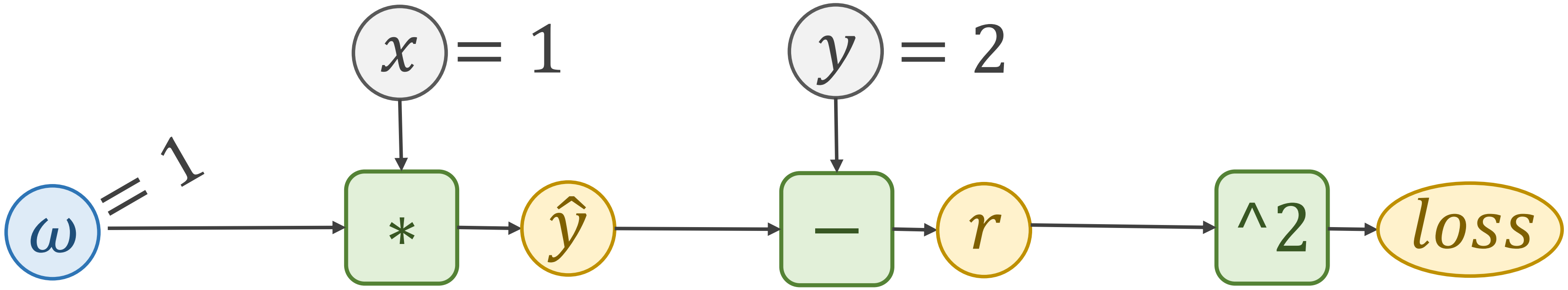
So after update, remember set the grad to **ZERO**!!!

Implementation of linear model with PyTorch

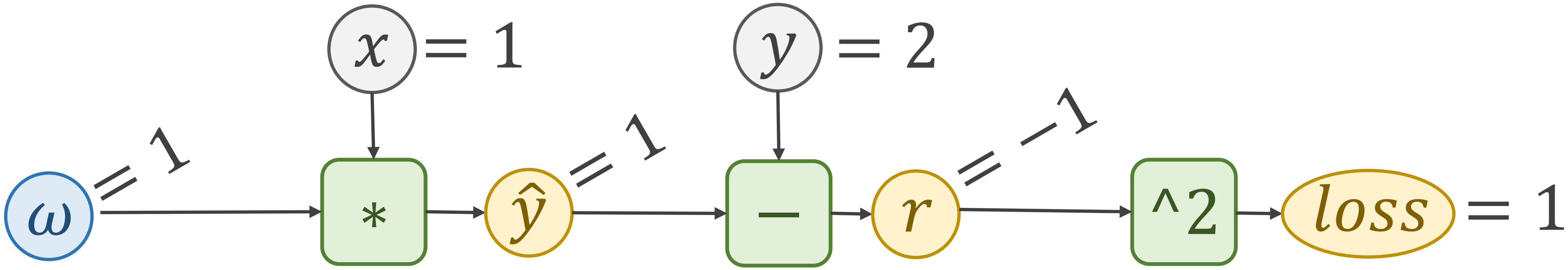
```python
print("predict (before training)",  4, forward(4).item())

for epoch in range(100):
    for x, y in zip(x_data, y_data):
        l = loss(x, y)
        l.backward()
        print('\tgrad:', x, y, w.grad.item())
        w.data = w.data - 0.01 * w.grad.data

        w.grad.data.zero_()

    print("progress:", epoch, l.item())

print("predict (after training)", 4, forward(4).item())
```

```
predict (before training) 4 4.0
	grad: 1.0 2.0 -2.0
	grad: 2.0 4.0 -7.840000152587891
	grad: 3.0 6.0 -16.228801727294922
progress: 0 7.315943717956543
	grad: 1.0 2.0 -1.478623867034912
	grad: 2.0 4.0 -5.796205520629883
	grad: 3.0 6.0 -11.998146057128906
progress: 1 3.9987640380859375
	grad: 1.0 2.0 -1.0931644439697266
	grad: 2.0 4.0 -4.285204887390137
	grad: 3.0 6.0 -8.870372772216797
progress: 2 2.1856532096862793
	grad: 1.0 2.0 -0.8081896305084229
	grad: 2.0 4.0 -3.1681032180786133
	grad: 3.0 6.0 -6.557973861694336
progress: 3 1.1946394443511963
	grad: 1.0 2.0 -0.5975041389465332
	grad: 2.0 4.0 -2.3422164916992188
	grad: 3.0 6.0 -4.848389625549316
progress: 4 0.6529689431190491
	grad: 1.0 2.0 -0.4417421817779541
	grad: 2.0 4.0 -1.7316293716430664
	grad: 3.0 6.0 -3.58447265625
progress: 5 0.35690122842788696
	grad: 1.0 2.0 -0.3265852928161621
	grad: 2.0 4.0 -1.2802143096923828
	grad: 3.0 6.0 -2.650045394897461
```

Forward/Backward in PyTorch

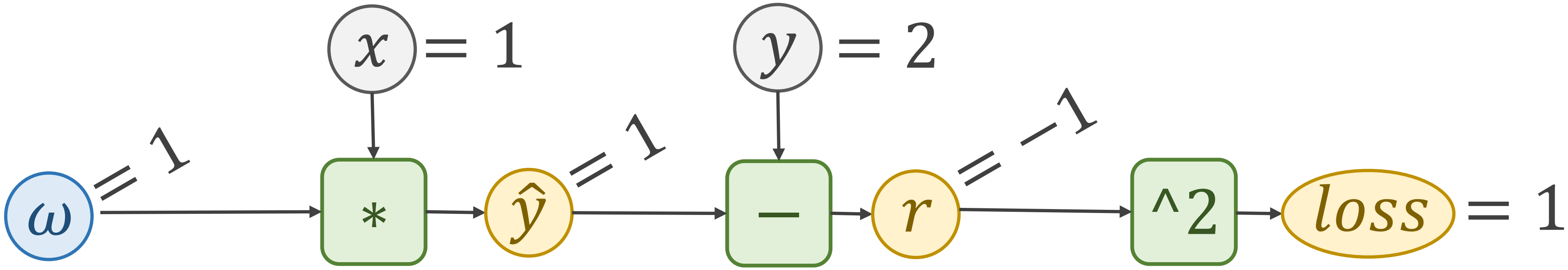


Forward in PyTorch



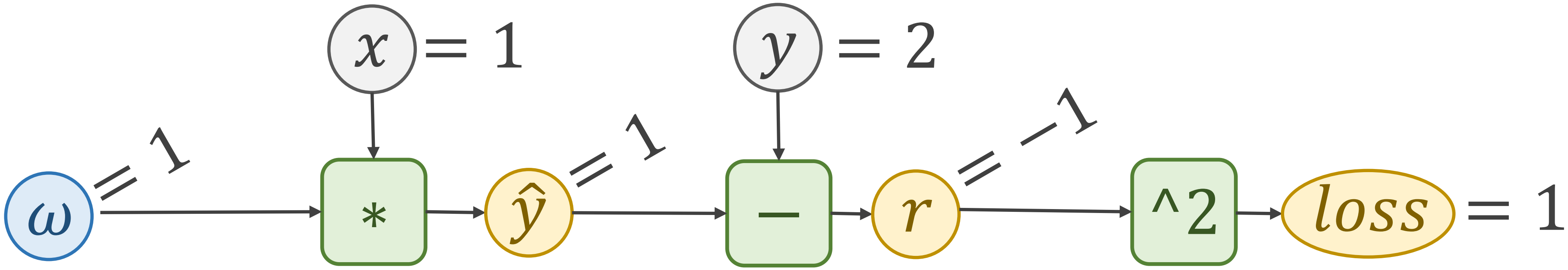

```python
w = torch.Tensor([1.0])
w.requires_grad = True

l = loss(x, y)
```

Backward in PyTorch



$$1.\ \mathrm{backward}() \rightarrow \frac{\partial loss}{\partial \omega} = w.grad$$

Update weight in PyTorch



1. backward()

$$\frac{\partial loss}{\partial \omega} = w.grad$$

```
w.data = w.data - 0.01 * w.grad.data
```

Exercise 4-3: Compute gradients using computational graph

Quadratic Model

$$\hat{y} = \omega_1 x^2 + \omega_2 x + b$$

Loss Function

$$loss = (\hat{y} - y)^2 = (x \cdot \omega - y)^2$$

$$\frac{\partial loss}{\partial \omega_1} = ?$$

$$\frac{\partial loss}{\partial \omega_2} = ?$$

$$\frac{\partial loss}{\partial b} = ?$$

Exercise 4-4: Compute gradients using PyTorch

Quadratic Model

$$\hat{y} = \omega_1 x^2 + \omega_2 x + b$$

Loss Function

$$loss = (\hat{y} - y)^2 = (x \cdot \omega - y)^2$$

$$\frac{\partial loss}{\partial \omega_1} = ?$$

$$\frac{\partial loss}{\partial \omega_2} = ?$$

$$\frac{\partial loss}{\partial b} = ?$$

PyTorch Tutorial

04. Back Propagation