

인공지능 심화학습 프로그램
Convolution Neural Network
Computer vision

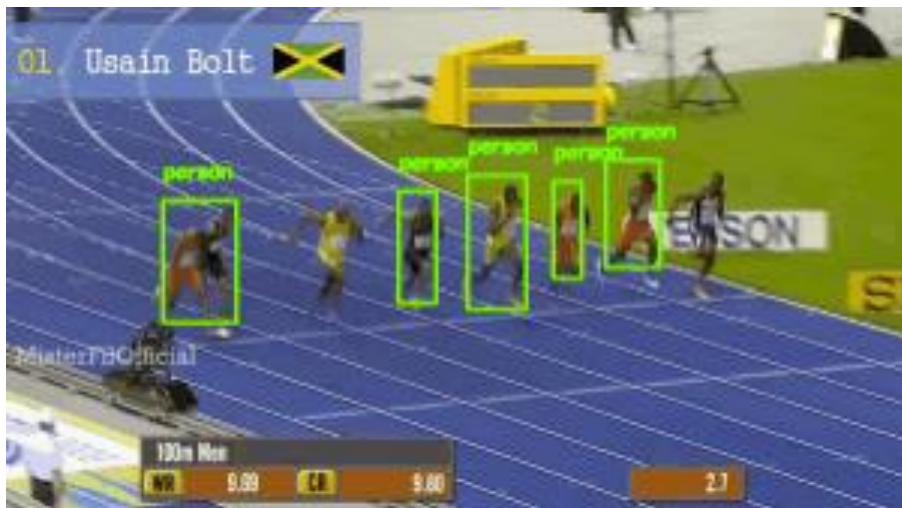
Vision and Learning laboratory
김민지
2023.11.06 ~ 2023.11.09

Index

- More Tasks
- Backpropagation
- Reminding: Convolution neural network
- Receptive Field
- Other Convolution layers
- QnA

More Tasks

- MOT (Multiple Object Tracking)



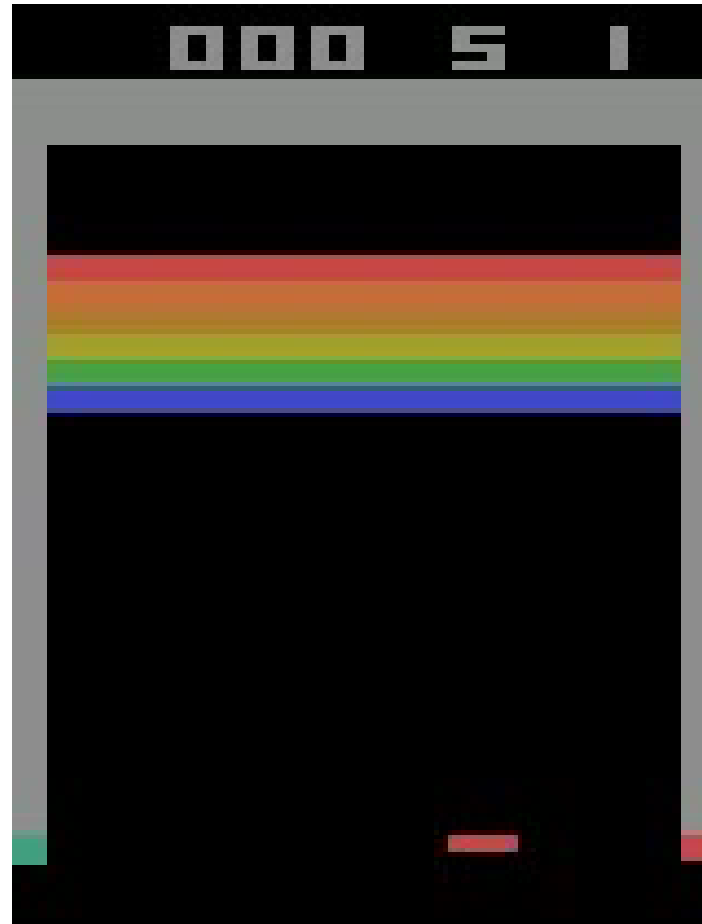
► More Tasks

- MOT (Multiple Object Tracking)

Tracking vs Detection vs Recognition

More Tasks

- Deep Reinforcement Learning



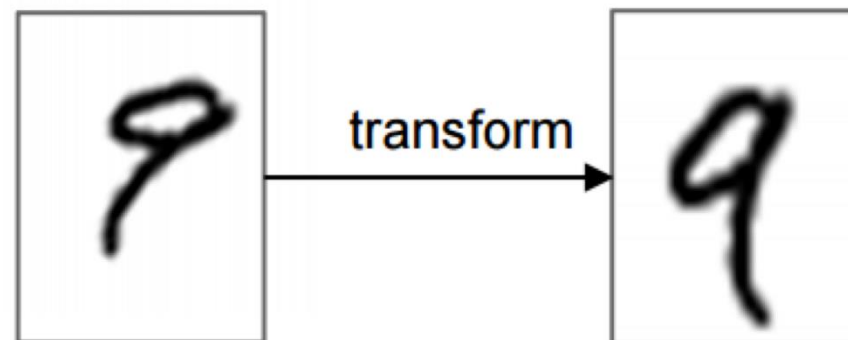
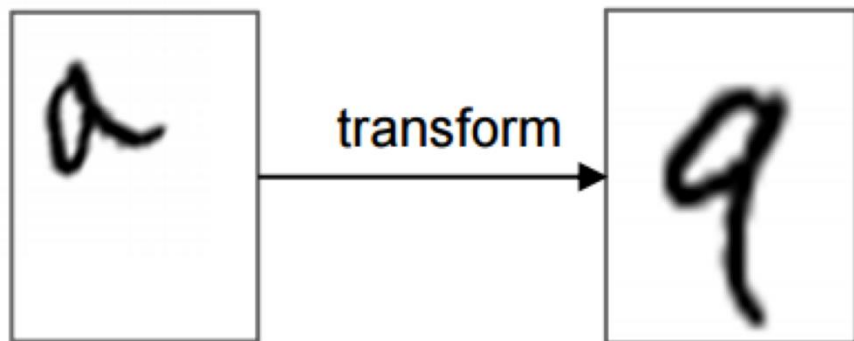
More Tasks

- Deep Reinforcement Learning



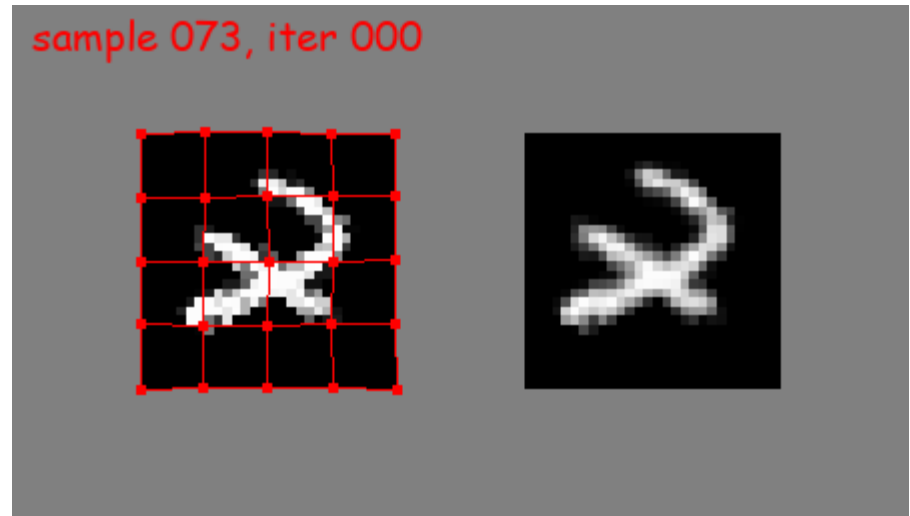
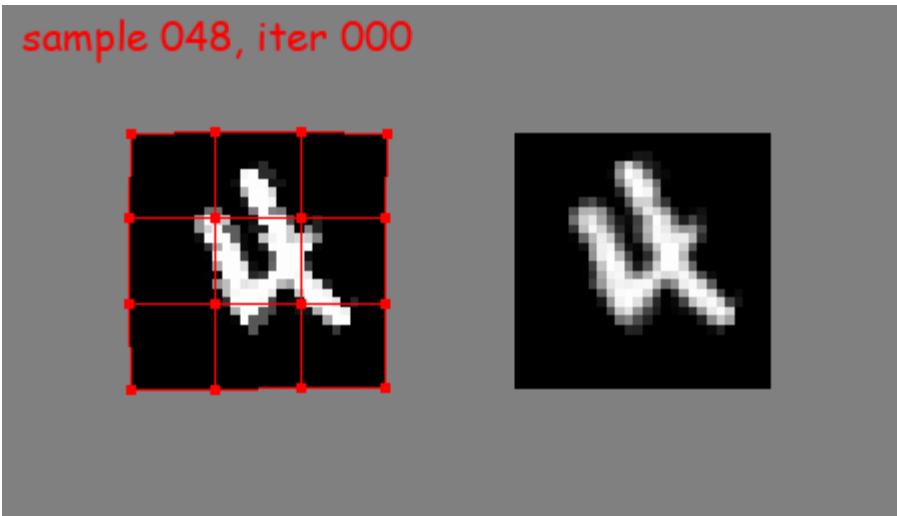
More Tasks

- Image Alignment



More Tasks

- Image Alignment



Source: https://github.com/WarBean/tps_stn_pytorch

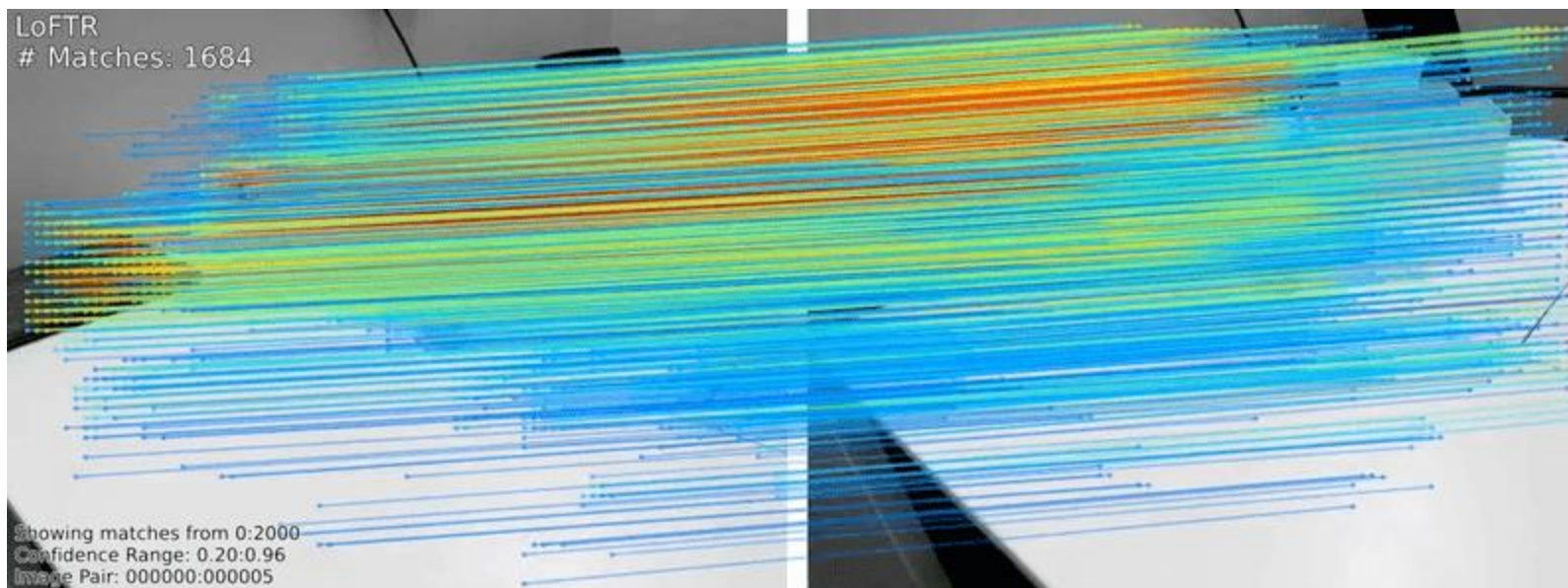
More Tasks

- Image Alignment (Keypoint matching)



More Tasks

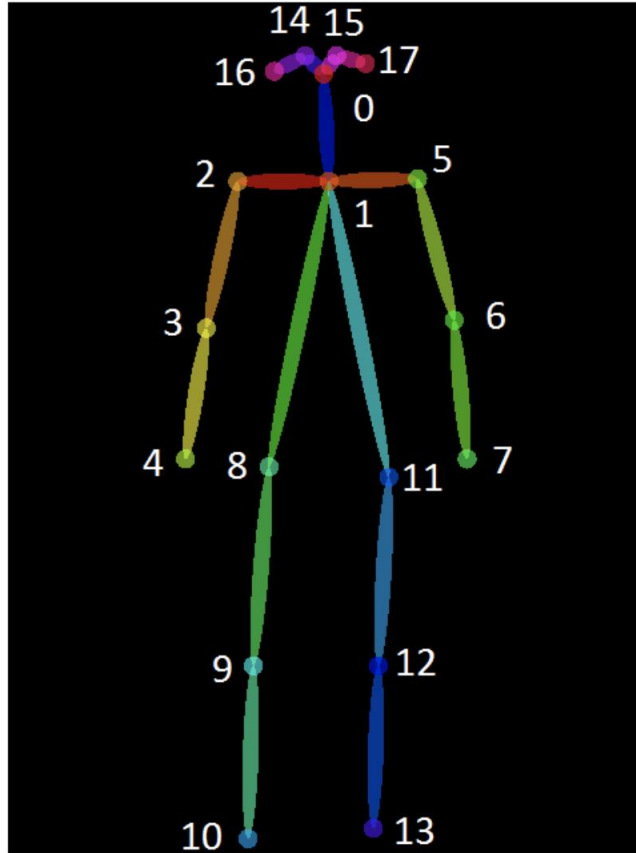
- Image Alignment (Keypoint matching)



Source: <https://github.com/zju3dv/LoFTR>

More Tasks

- Human Pose Estimation



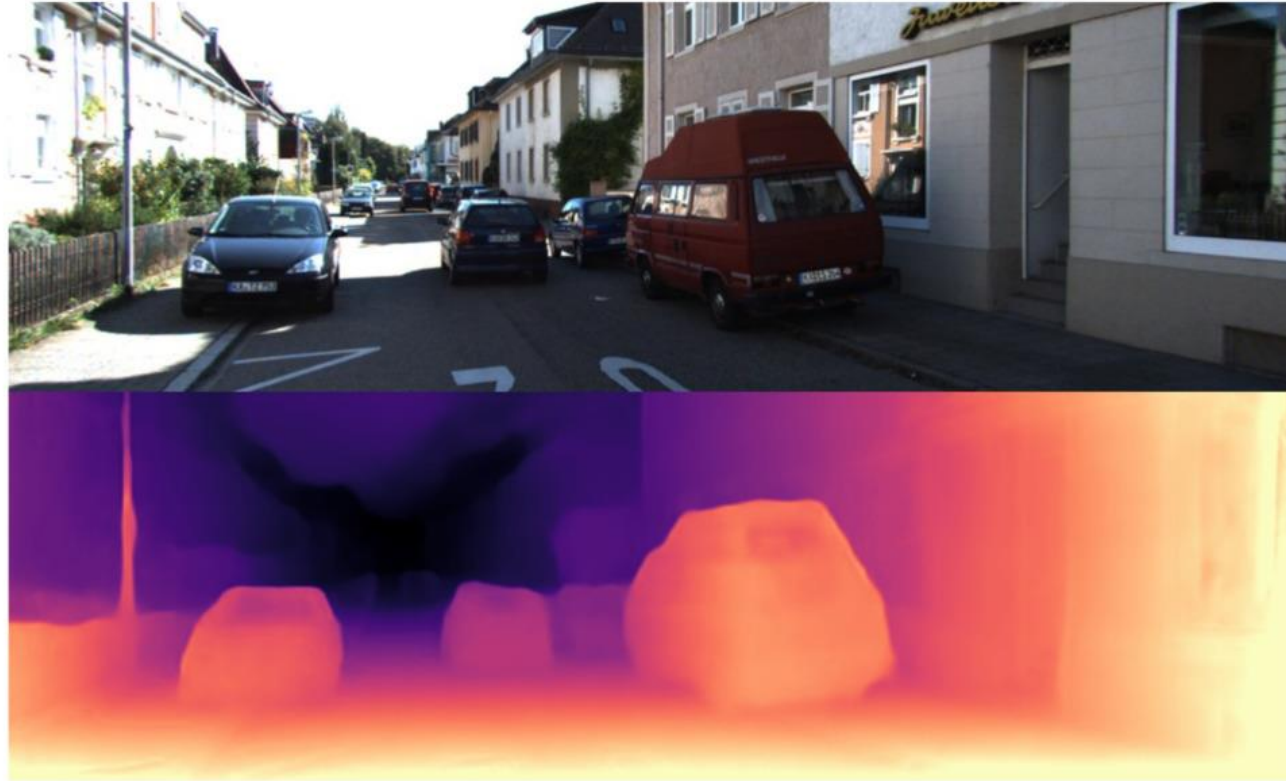
More Tasks

- Human Pose Estimation



More Tasks

- Depth Estimation



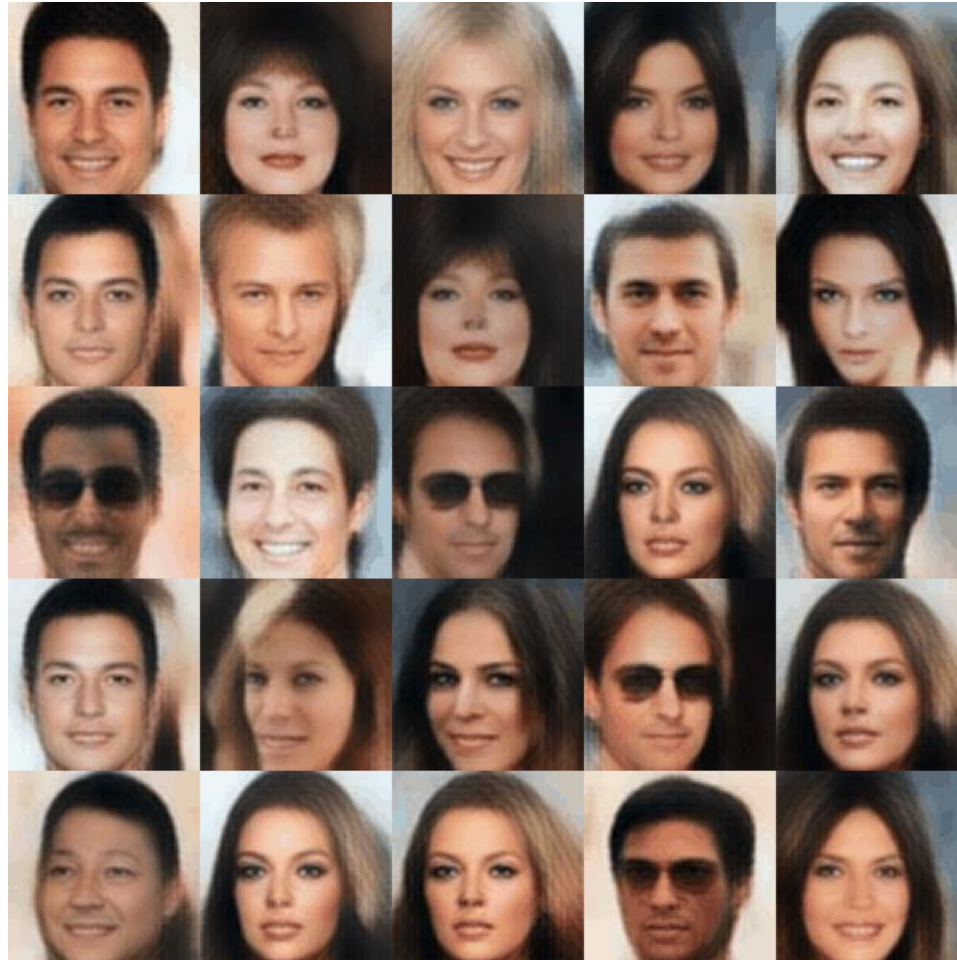
► More Tasks

- Depth Estimation



More Tasks

- Generative model



More Tasks

- Explainable AI

Grad-CAM for "Cat"

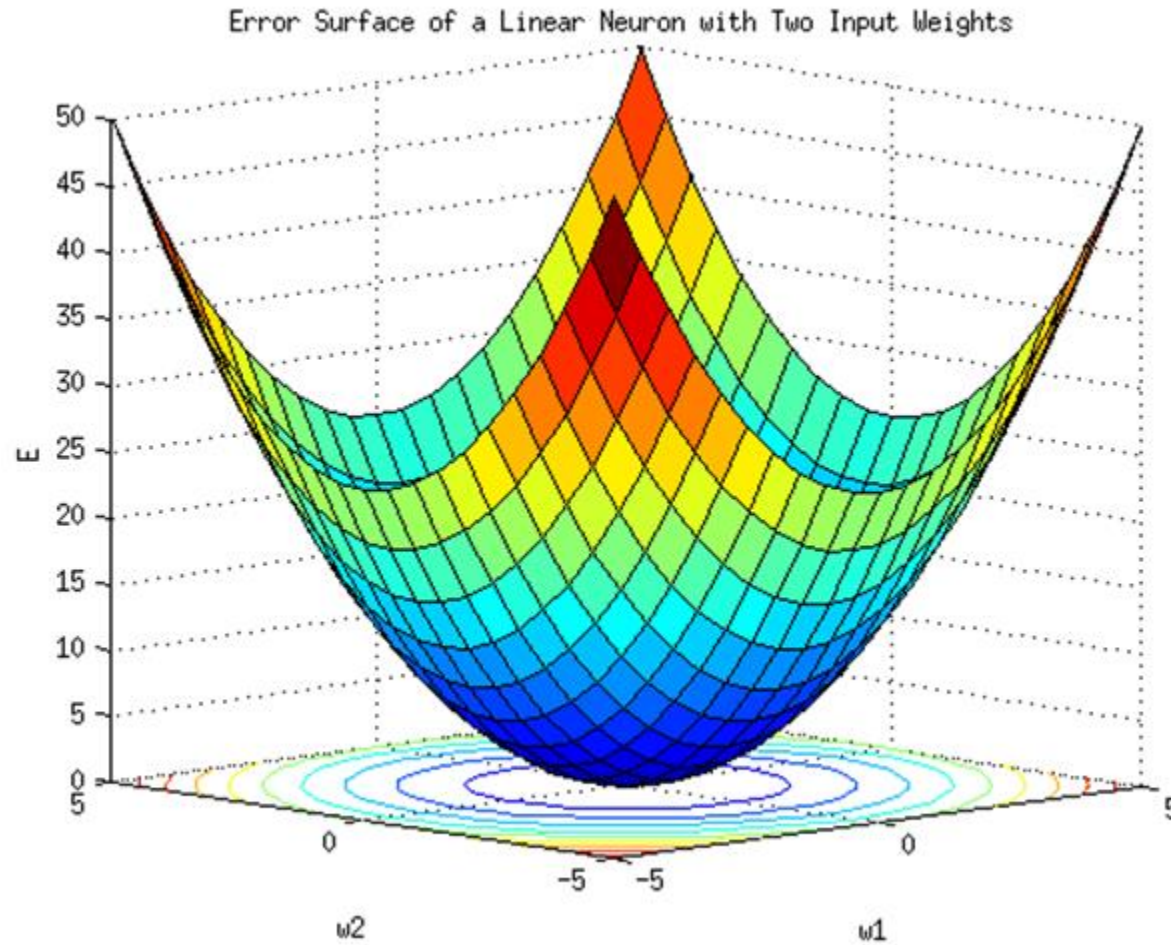


Grad-CAM for "Dog"



► Backpropagation

- What is backpropagation?



Backpropagation

■ What is Differentiation?

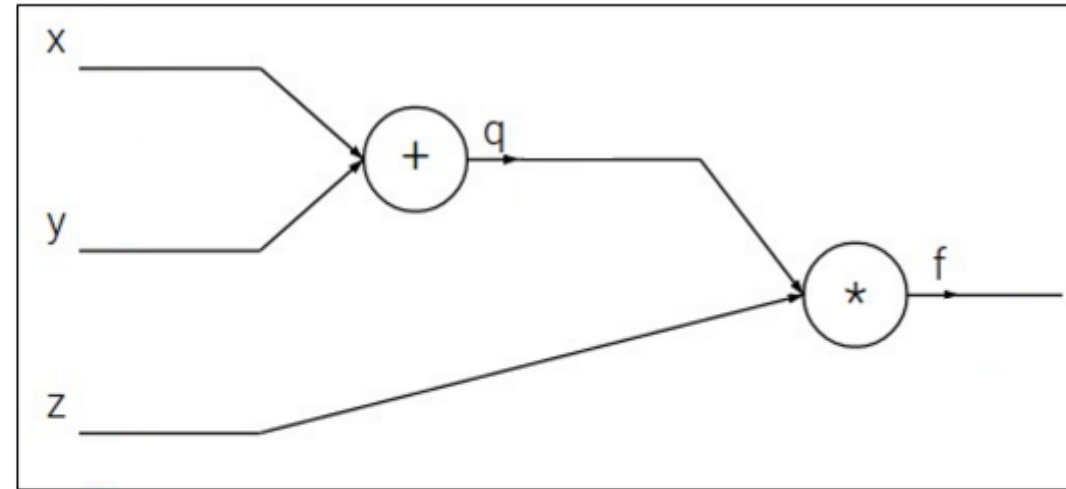
- 미분은 쉽게 말해 미세한 부분을 뜻함
- 미세한 변화를 연구하는 분야

$$\begin{aligned}f'(a) &= \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a} \\ &= \lim_{\Delta x \rightarrow 0} \frac{f(a + \Delta x) - f(a)}{\Delta x}\end{aligned}$$

► Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

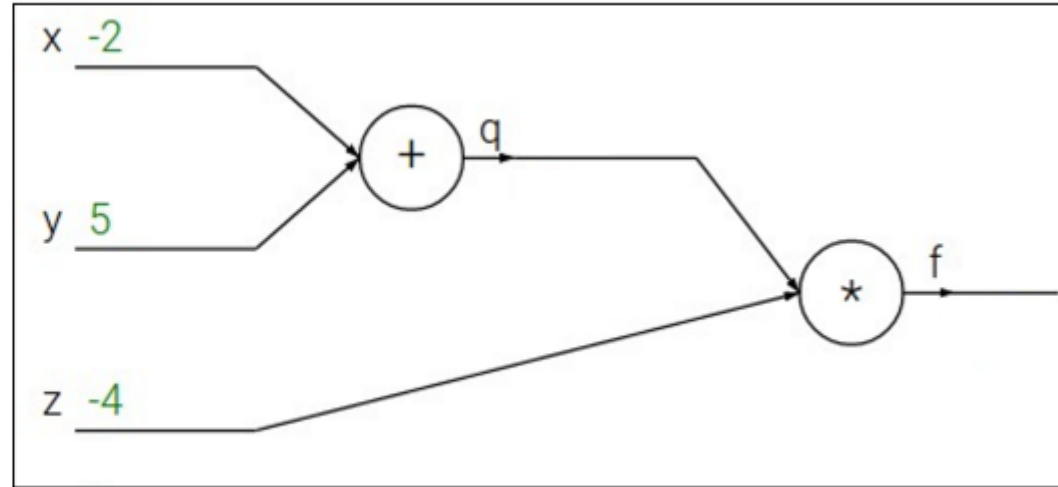


► Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. $x = -2, y = 5, z = -4$



Backpropagation

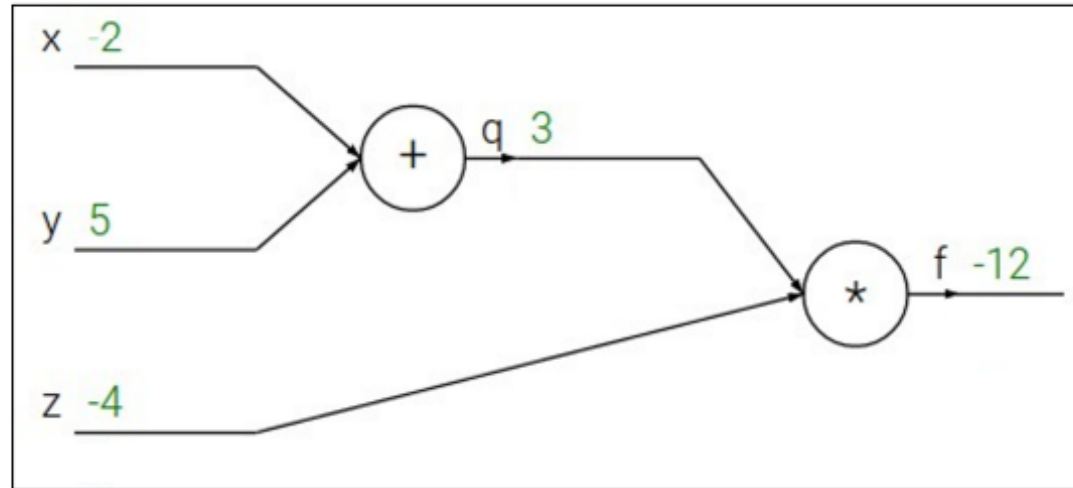
Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. $x = -2, y = 5, z = -4$

1. Forward pass: Compute outputs

$$q = x + y \quad f = q \cdot z$$

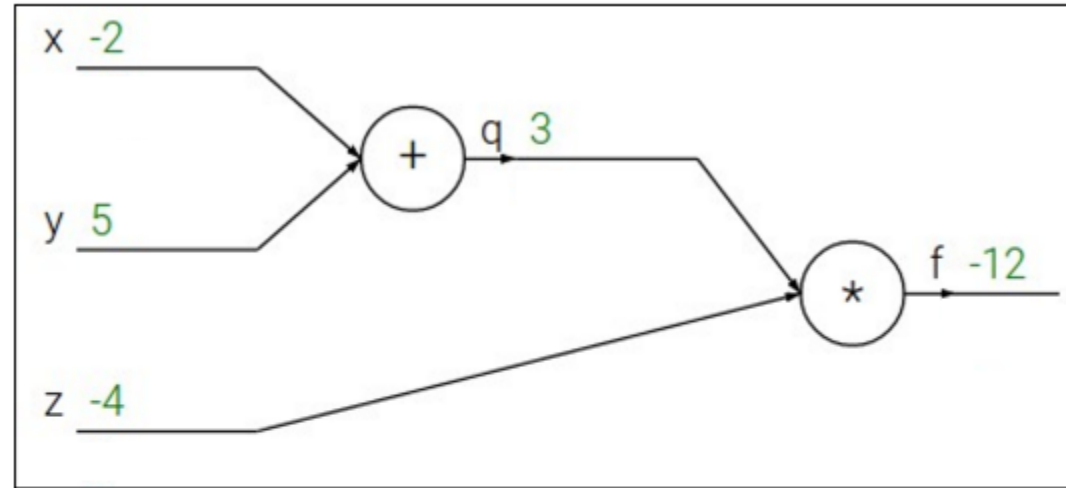


Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. $x = -2, y = 5, z = -4$



1. Forward pass: Compute outputs

$$q = x + y \quad f = q \cdot z$$

2. Backward pass: Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

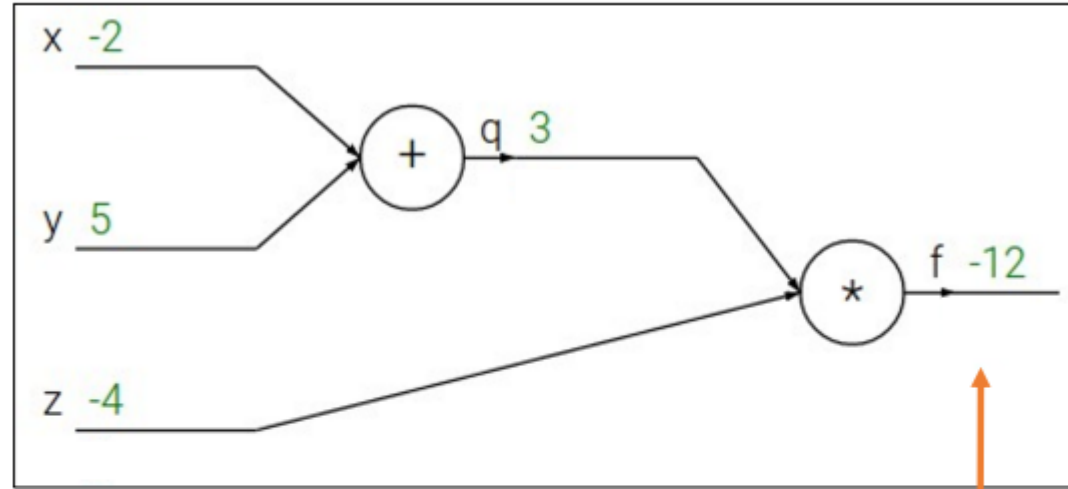
e.g. $x = -2, y = 5, z = -4$

1. **Forward pass:** Compute outputs

$$q = x + y \quad f = q \cdot z$$

2. **Backward pass:** Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial f}$$

Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

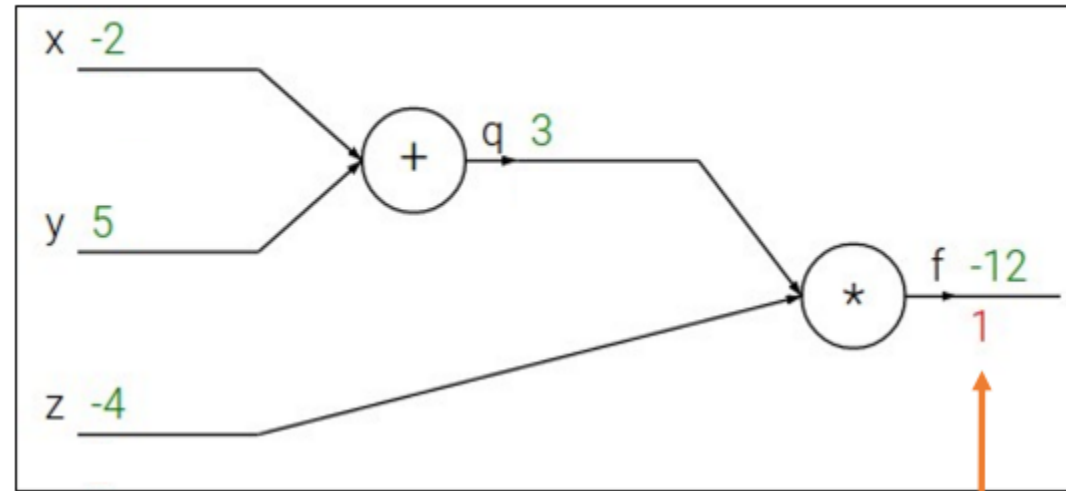
e.g. $x = -2, y = 5, z = -4$

1. **Forward pass:** Compute outputs

$$q = x + y \quad f = q \cdot z$$

2. **Backward pass:** Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial f}$$

Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

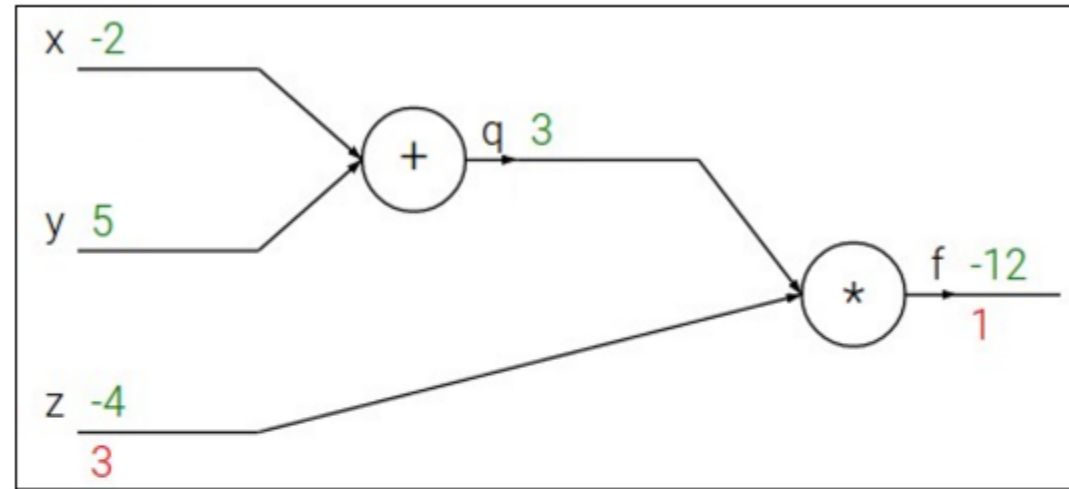
e.g. $x = -2, y = 5, z = -4$

1. **Forward pass:** Compute outputs

$$q = x + y \quad f = q \cdot z$$

2. **Backward pass:** Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial z}$$

Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

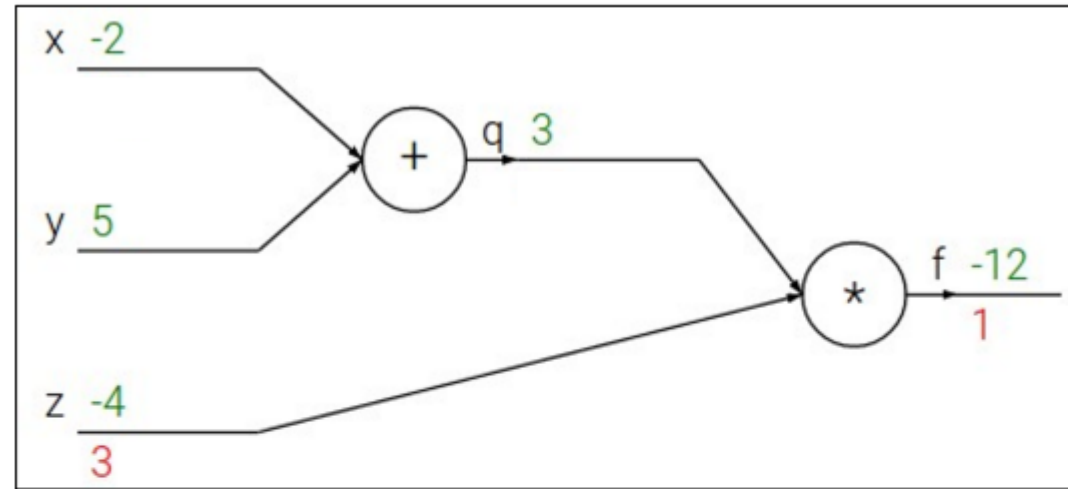
e.g. $x = -2, y = 5, z = -4$

1. **Forward pass:** Compute outputs

$$q = x + y \quad f = q \cdot z$$

2. **Backward pass:** Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial z} = q$$

Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

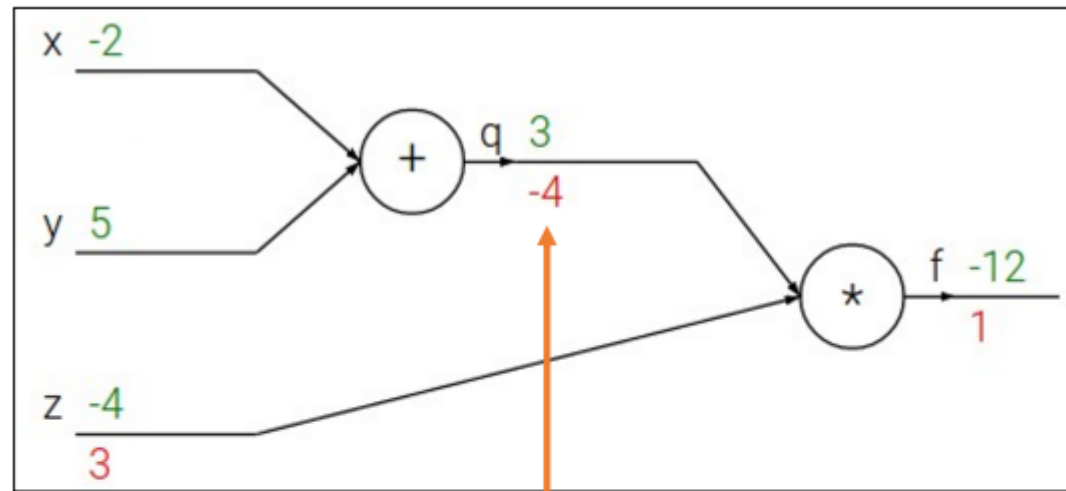
e.g. $x = -2, y = 5, z = -4$

1. **Forward pass:** Compute outputs

$$q = x + y \quad f = q \cdot z$$

2. **Backward pass:** Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial q}$$

Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

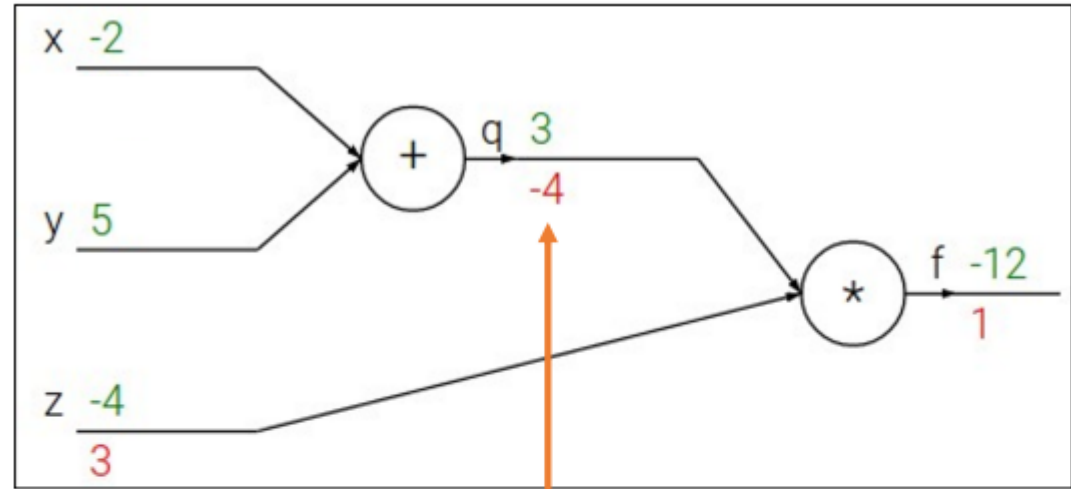
e.g. $x = -2, y = 5, z = -4$

1. **Forward pass:** Compute outputs

$$q = x + y \quad \boxed{f = q \cdot z}$$

2. **Backward pass:** Compute derivatives

$$\text{Want: } \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$$



$$\boxed{\frac{\partial f}{\partial q} = z}$$

Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

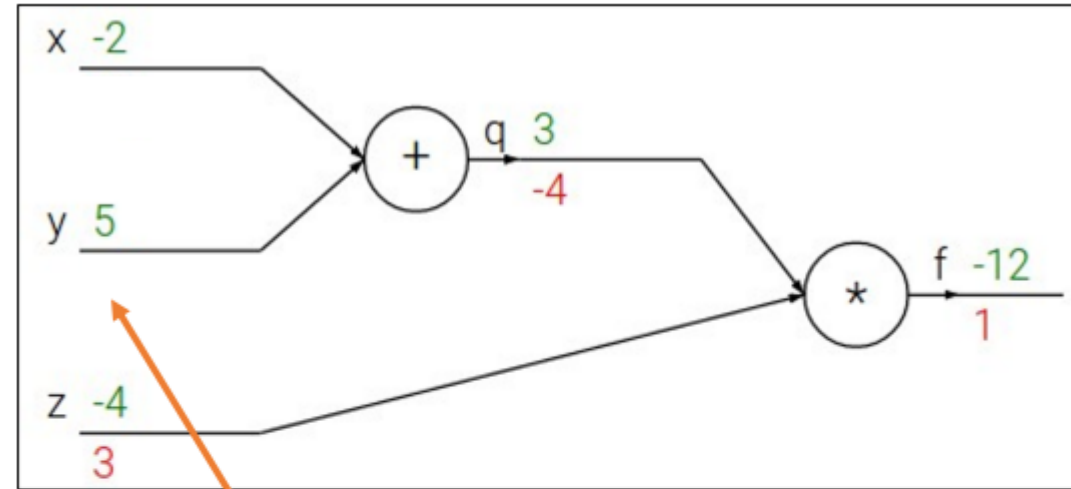
e.g. $x = -2, y = 5, z = -4$

1. **Forward pass:** Compute outputs

$$q = x + y \quad f = q \cdot z$$

2. **Backward pass:** Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial y}$$

Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

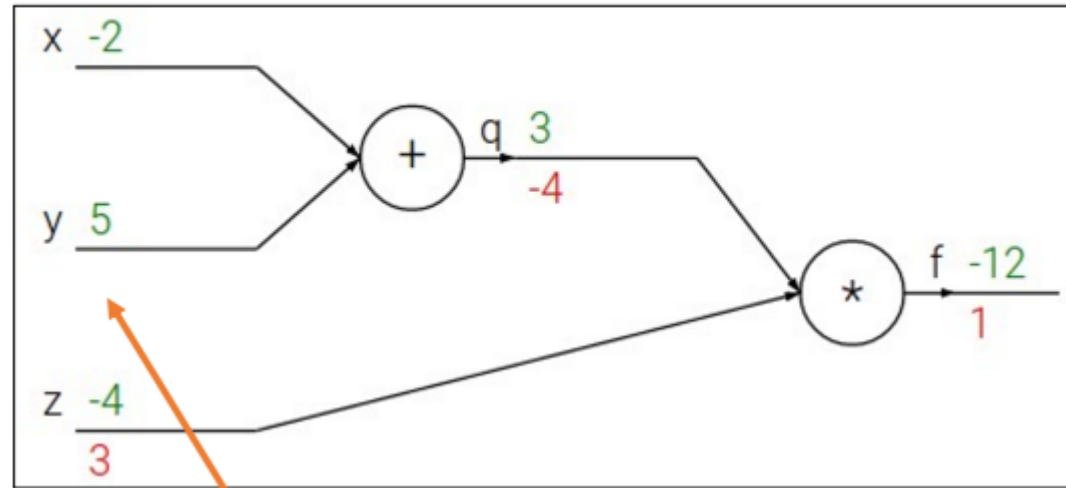
e.g. $x = -2, y = 5, z = -4$

1. **Forward pass:** Compute outputs

$$q = x + y \quad f = q \cdot z$$

2. **Backward pass:** Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain Rule

$$\frac{\partial f}{\partial y} = \frac{\partial q}{\partial y} \frac{\partial f}{\partial q}$$

Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

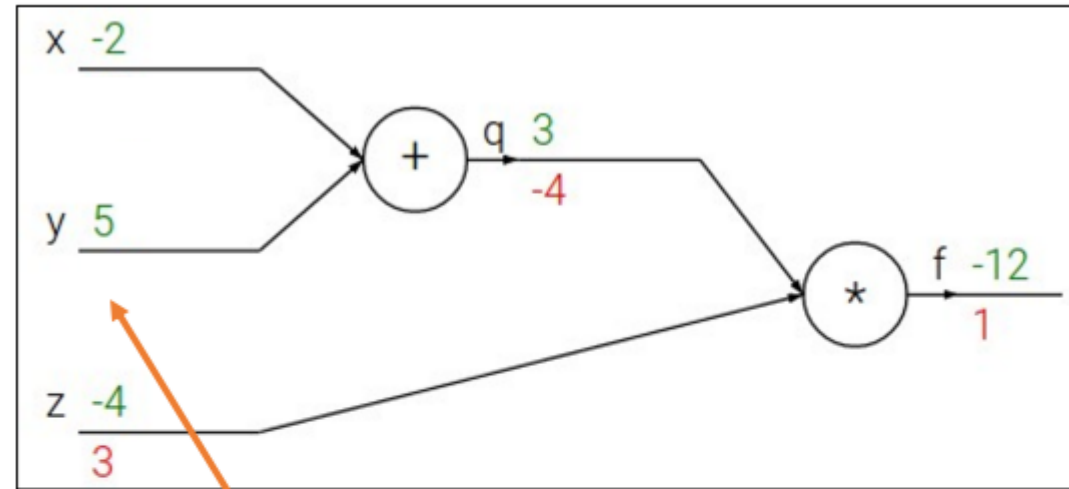
e.g. $x = -2, y = 5, z = -4$

1. **Forward pass:** Compute outputs

$$q = x + y \quad f = q \cdot z$$

2. **Backward pass:** Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain Rule

$$\frac{\partial f}{\partial y} = \frac{\partial q}{\partial y} \frac{\partial f}{\partial q}$$

Downstream
Gradient

Local
Gradient

Upstream
Gradient

Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

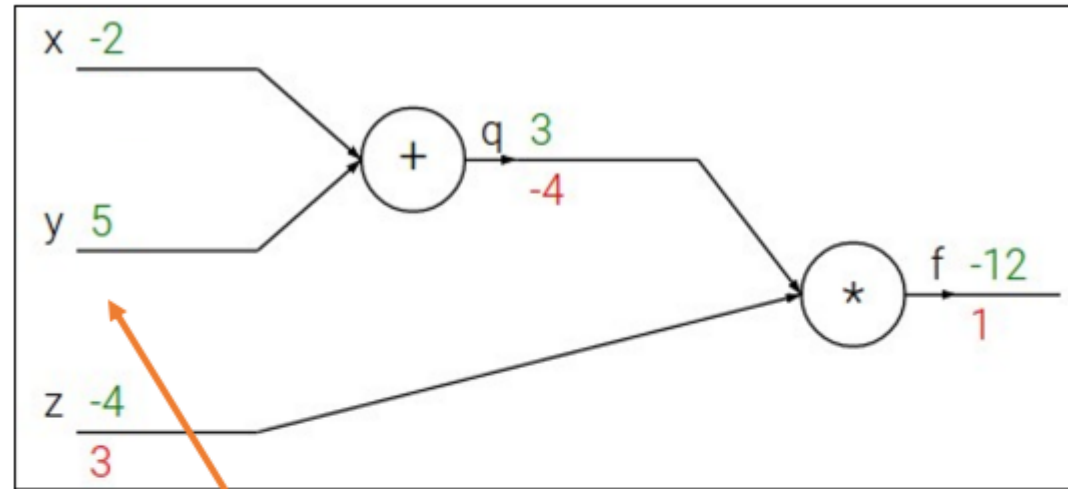
e.g. $x = -2, y = 5, z = -4$

1. **Forward pass:** Compute outputs

$$q = x + y \quad f = q \cdot z$$

2. **Backward pass:** Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain Rule

$$\frac{\partial f}{\partial y} = \frac{\partial q}{\partial y} \frac{\partial f}{\partial q}$$

$$\frac{\partial q}{\partial y} = 1$$

Downstream
Gradient

Local
Gradient

Upstream
Gradient

Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

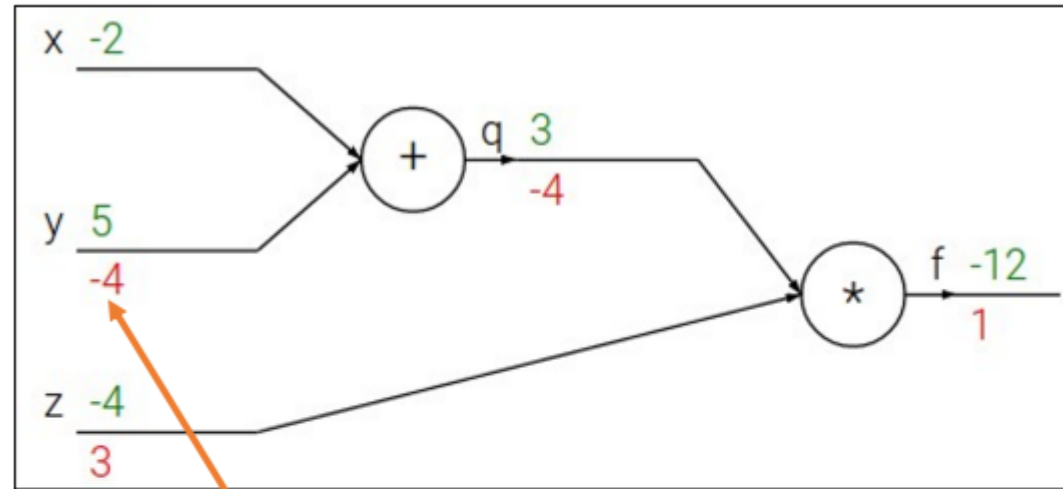
e.g. $x = -2, y = 5, z = -4$

1. **Forward pass:** Compute outputs

$$q = x + y \quad f = q \cdot z$$

2. **Backward pass:** Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain Rule

$$\frac{\partial f}{\partial y} = \frac{\partial q}{\partial y} \frac{\partial f}{\partial q}$$

$$\frac{\partial q}{\partial y} = 1$$

Downstream
Gradient

Local
Gradient

Upstream
Gradient

Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

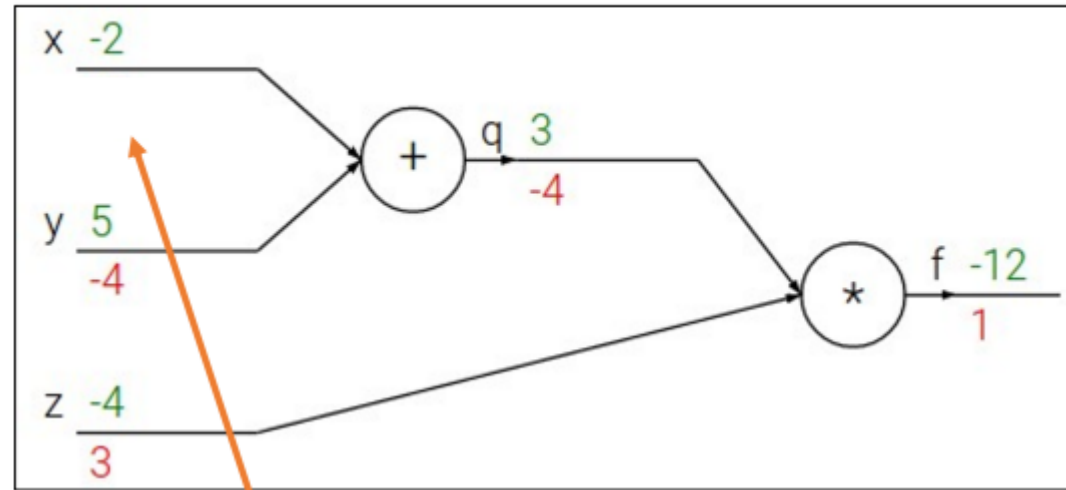
e.g. $x = -2, y = 5, z = -4$

1. **Forward pass:** Compute outputs

$$q = x + y \quad f = q \cdot z$$

2. **Backward pass:** Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain Rule

$$\frac{\partial f}{\partial x} = \frac{\partial q}{\partial x} \frac{\partial f}{\partial q}$$

$$\frac{\partial q}{\partial x} = 1$$

Downstream
Gradient

Local
Gradient

Upstream
Gradient

Backpropagation

Backpropagation:
Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

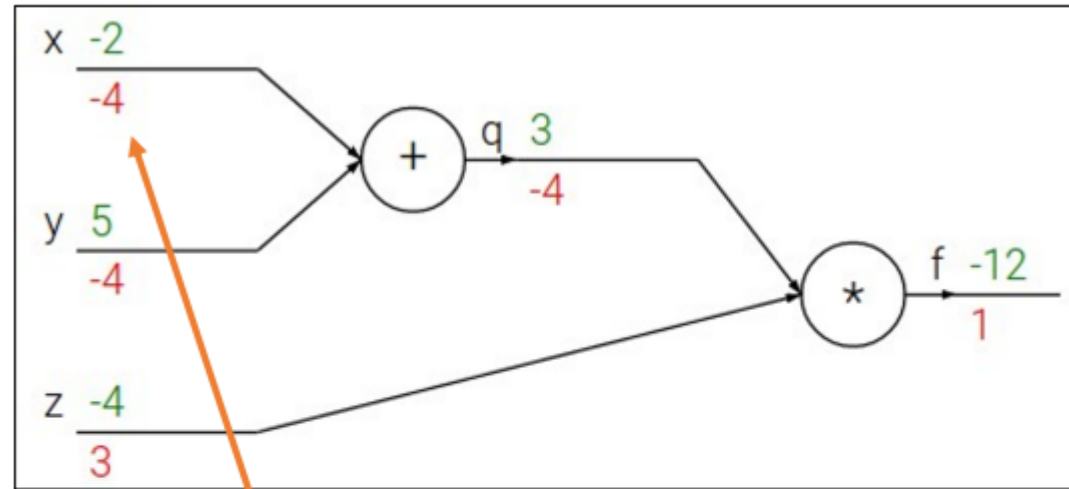
e.g. $x = -2, y = 5, z = -4$

1. **Forward pass:** Compute outputs

$$q = x + y \quad f = q \cdot z$$

2. **Backward pass:** Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain Rule

$$\frac{\partial f}{\partial x} = \frac{\partial q}{\partial x} \frac{\partial f}{\partial q}$$

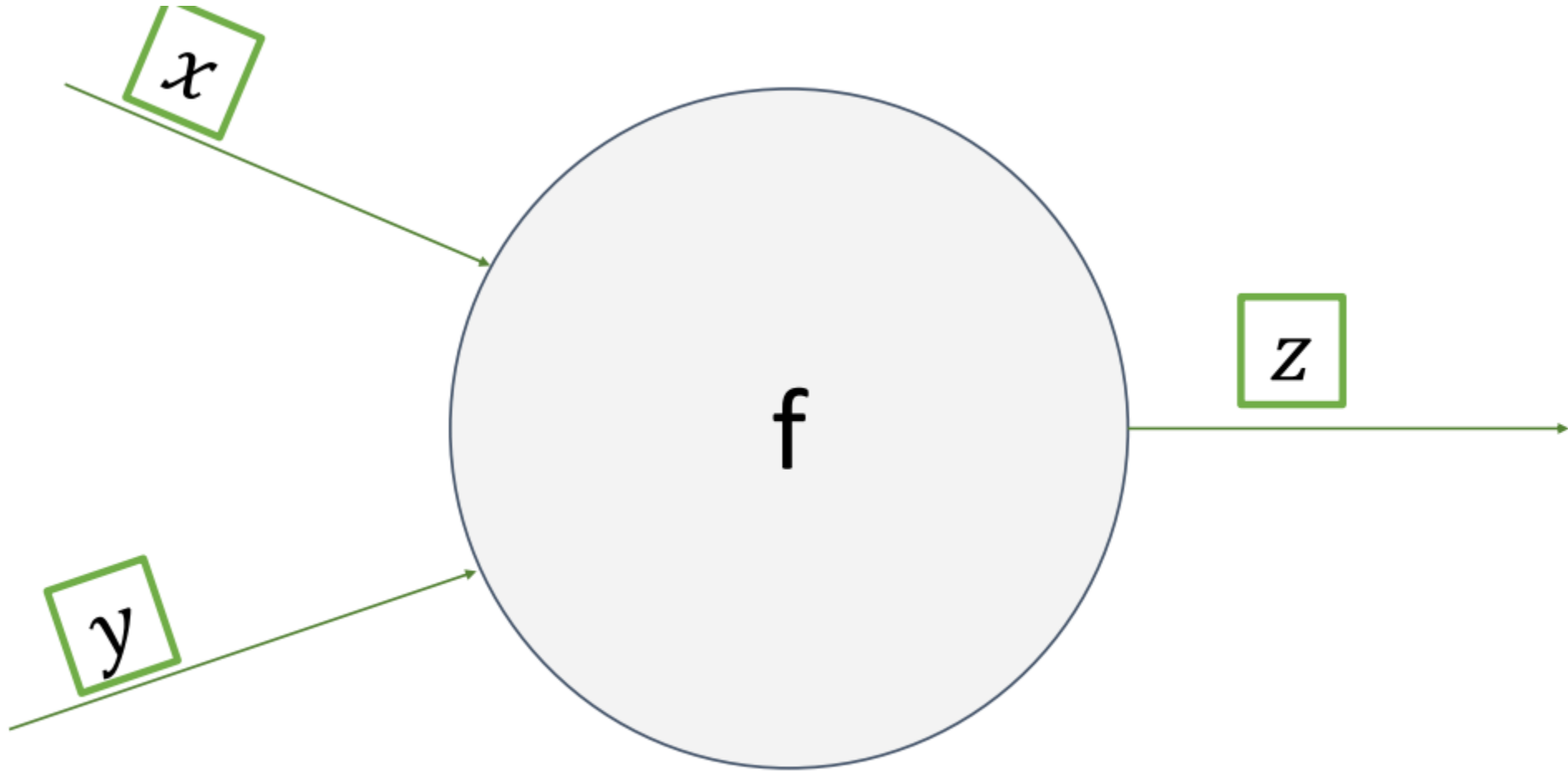
$$\frac{\partial q}{\partial x} = 1$$

Downstream
Gradient

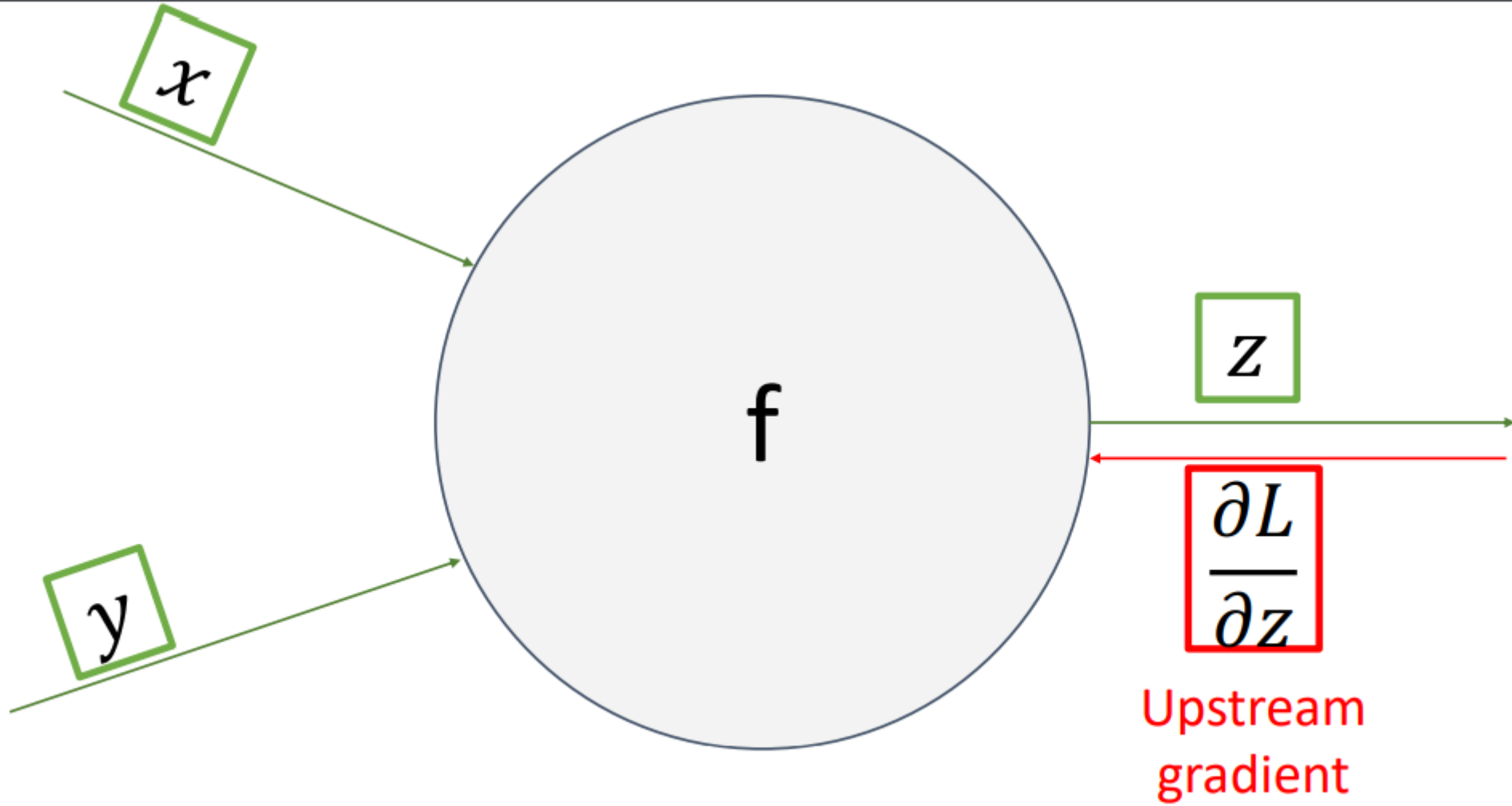
Local
Gradient

Upstream
Gradient

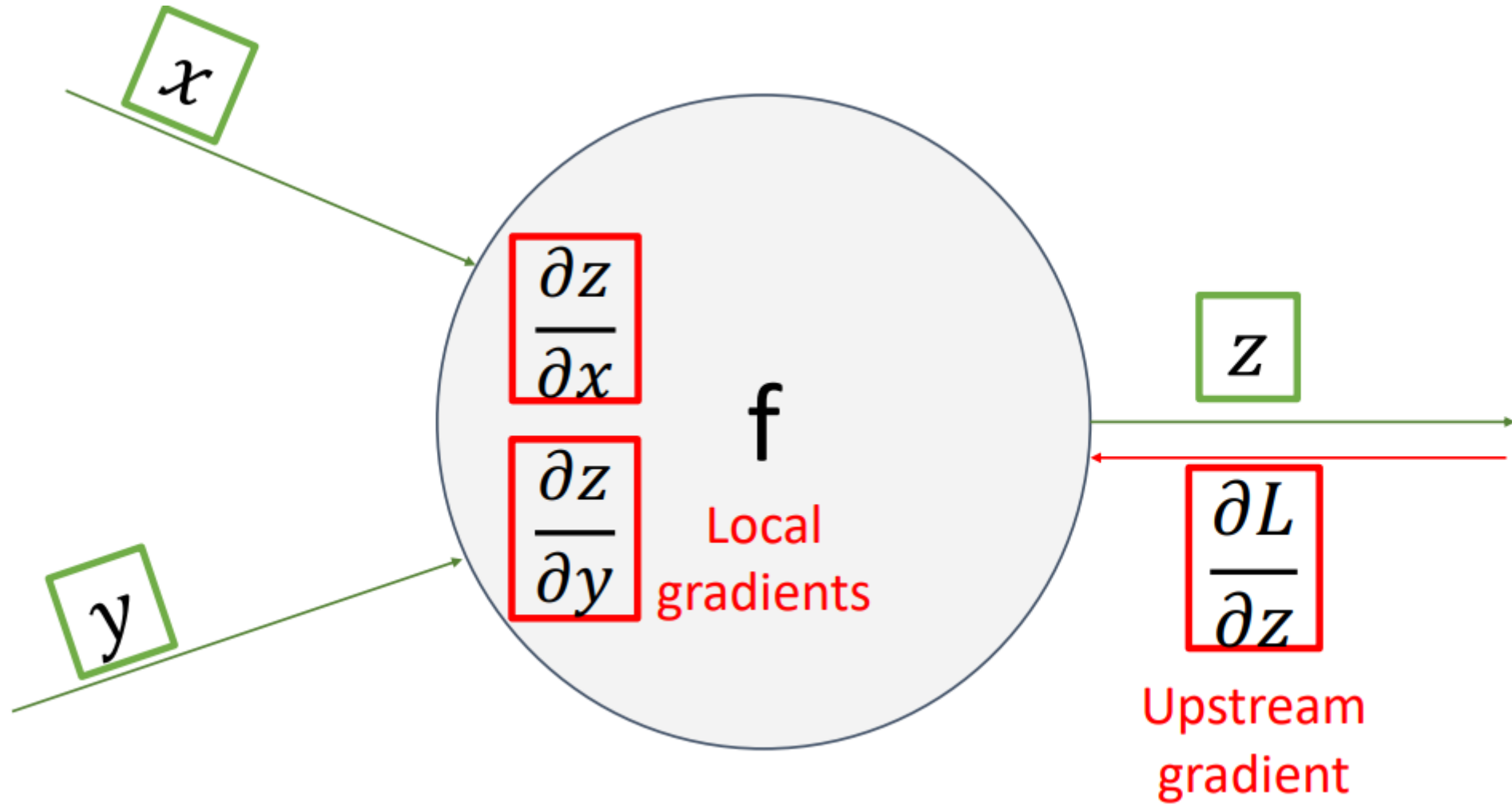
► Backpropagation



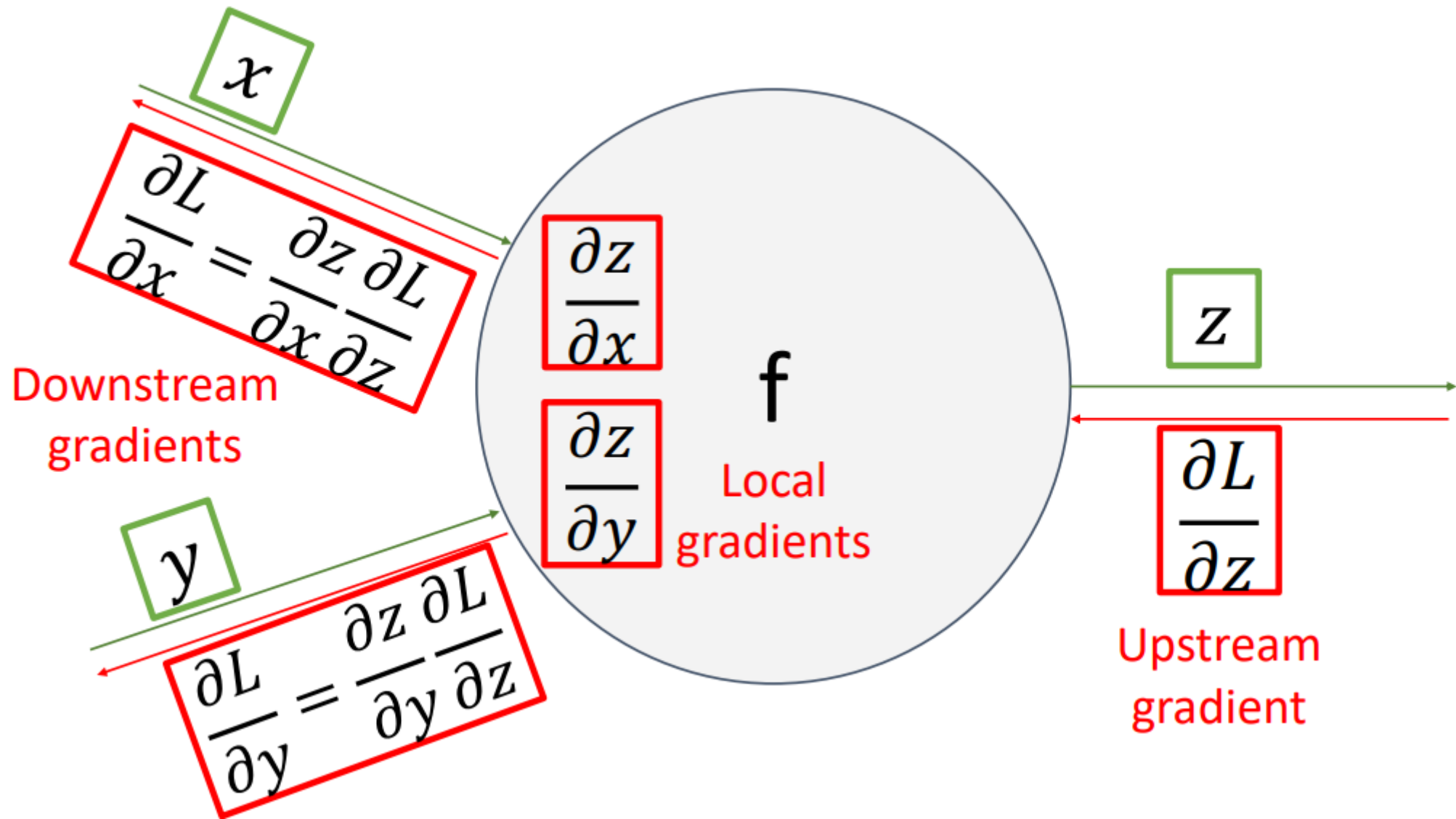
Backpropagation



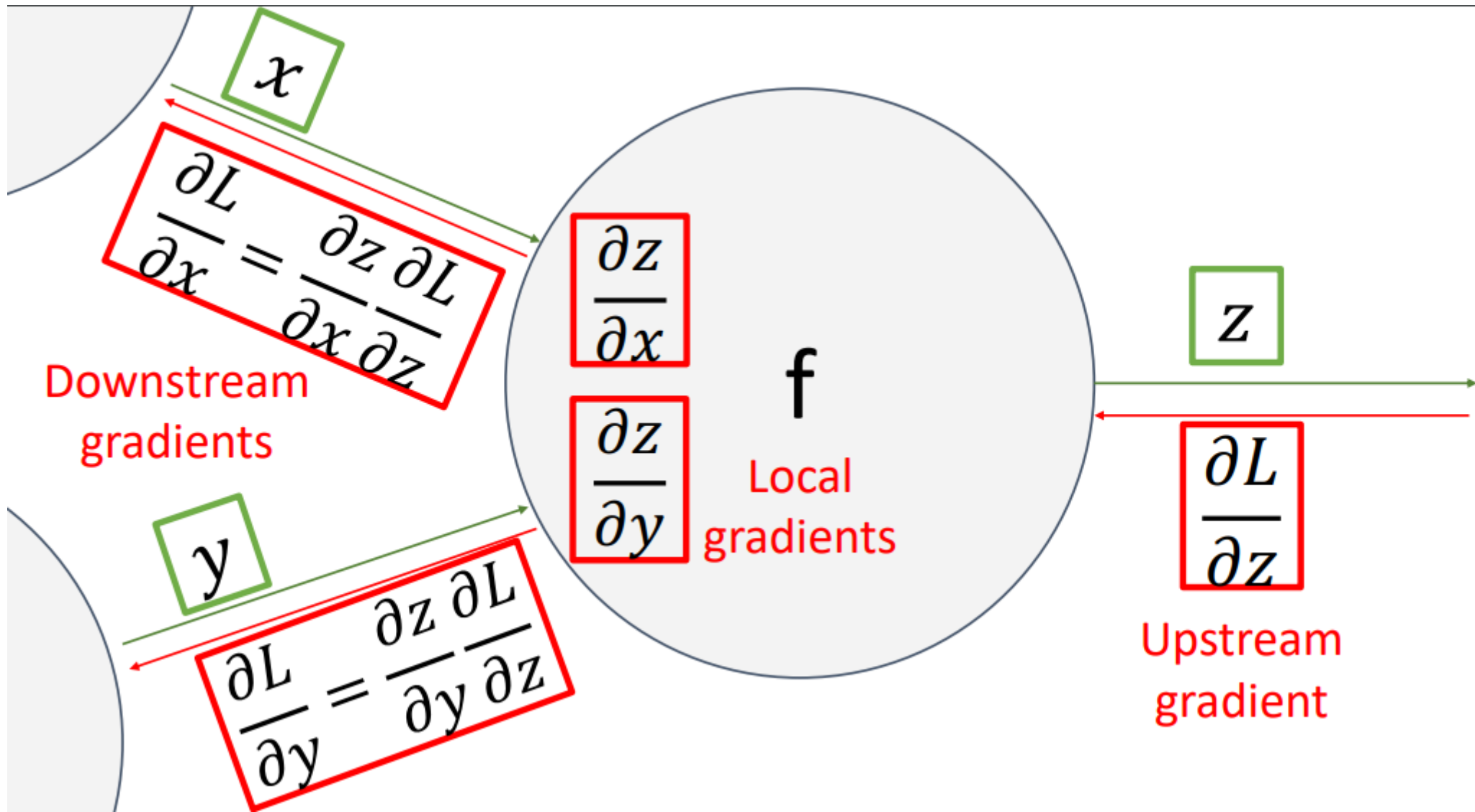
Backpropagation



Backpropagation

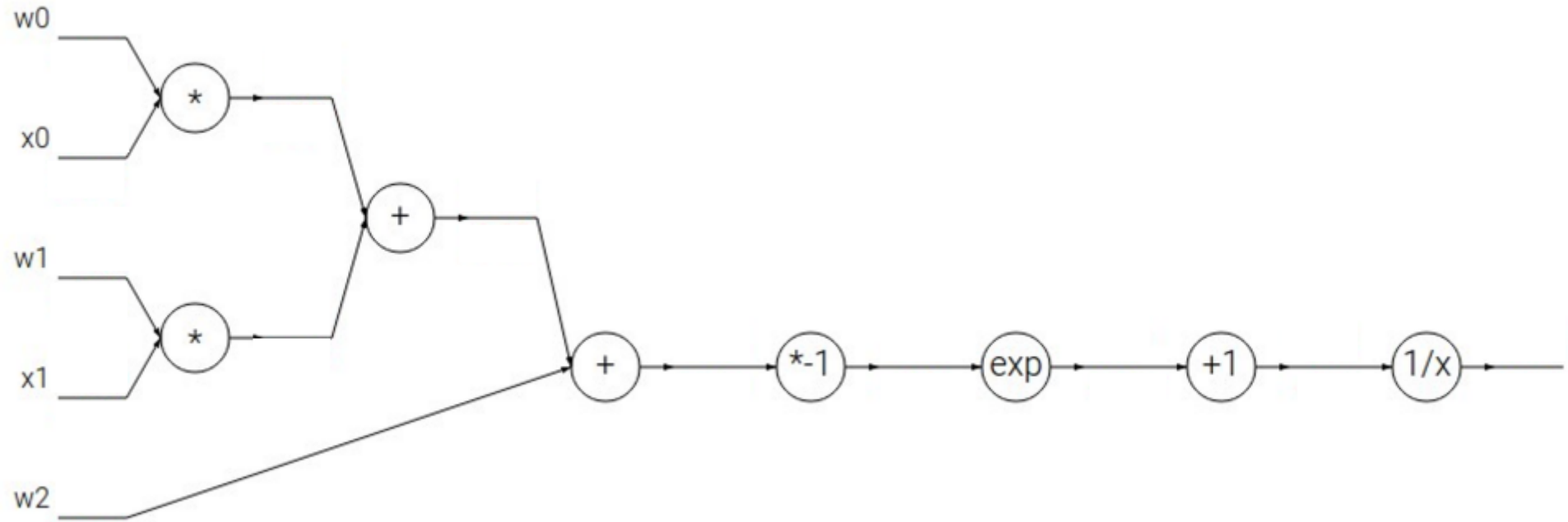


Backpropagation



Backpropagation

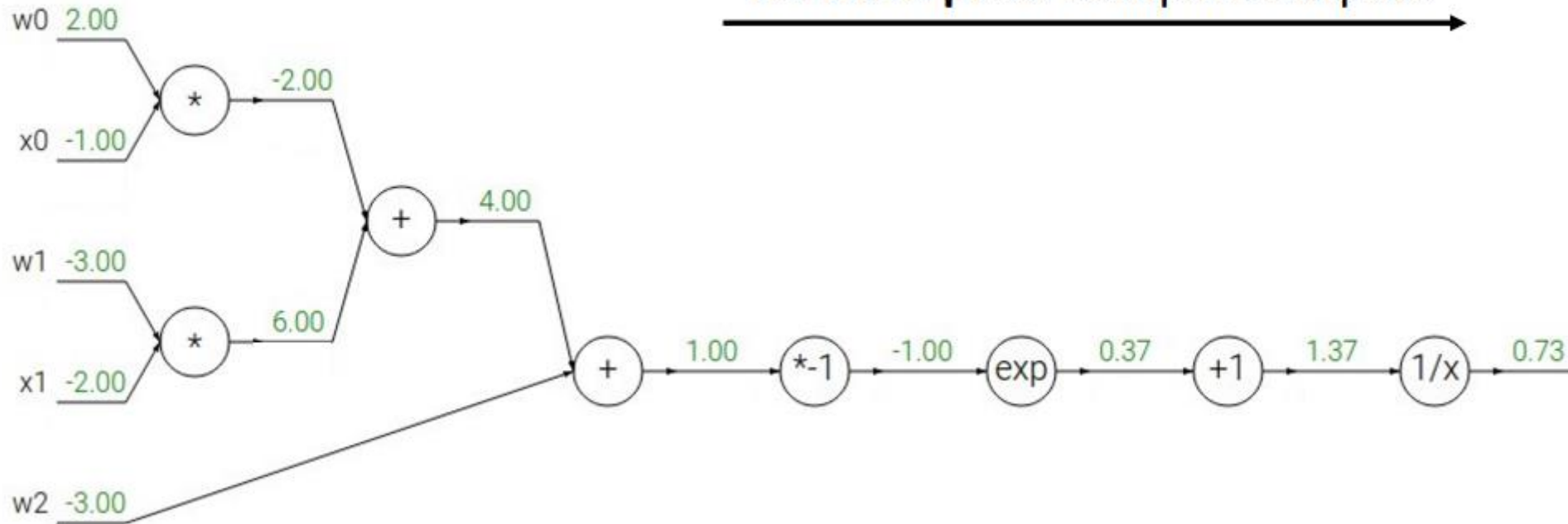
Another Example $f(x, w) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$



Backpropagation

Another Example $f(x, w) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$

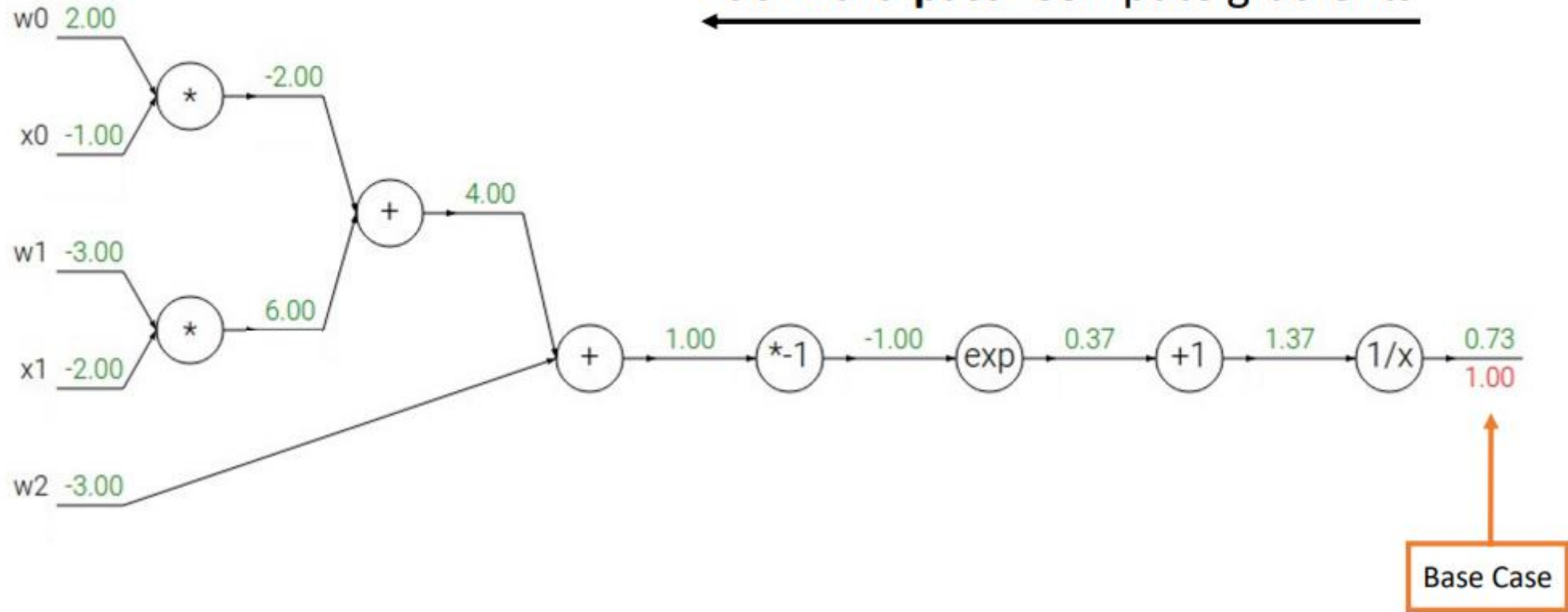
Forward pass: Compute outputs



Backpropagation

Another Example $f(x, w) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$

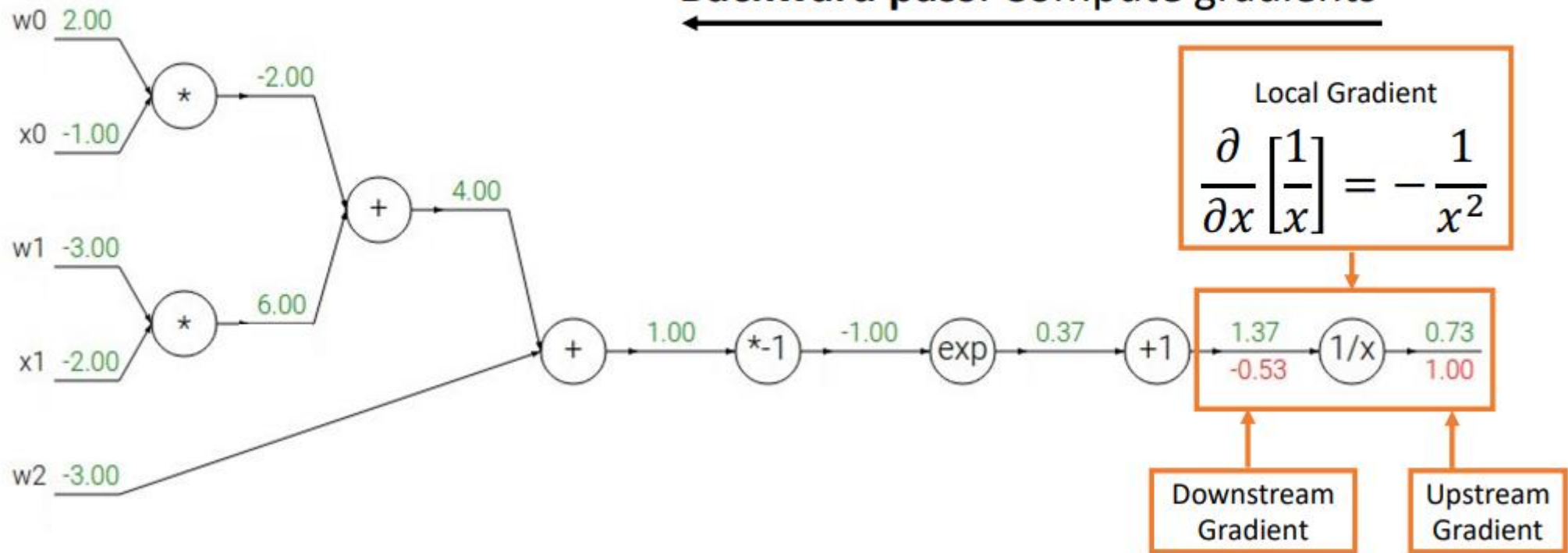
Backward pass: Compute gradients



Backpropagation

Another Example $f(x, w) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$

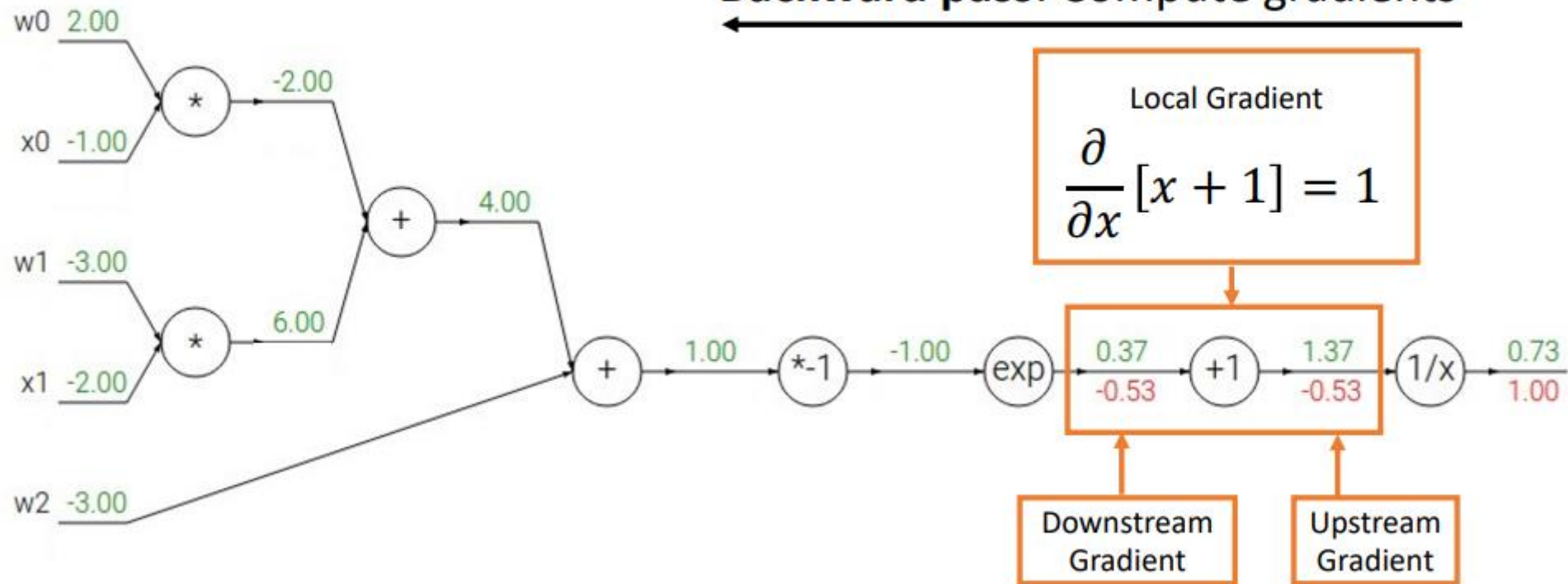
Backward pass: Compute gradients



Backpropagation

Another Example $f(x, w) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$

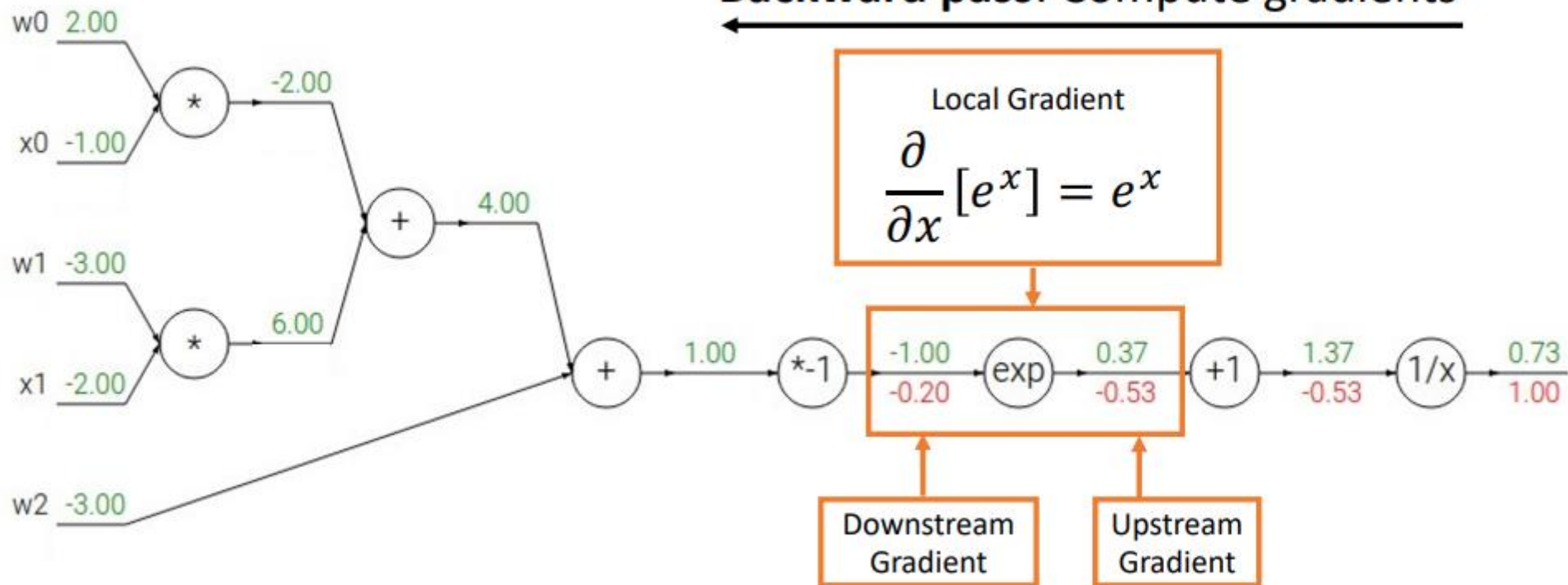
Backward pass: Compute gradients



Backpropagation

Another Example $f(x, w) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$

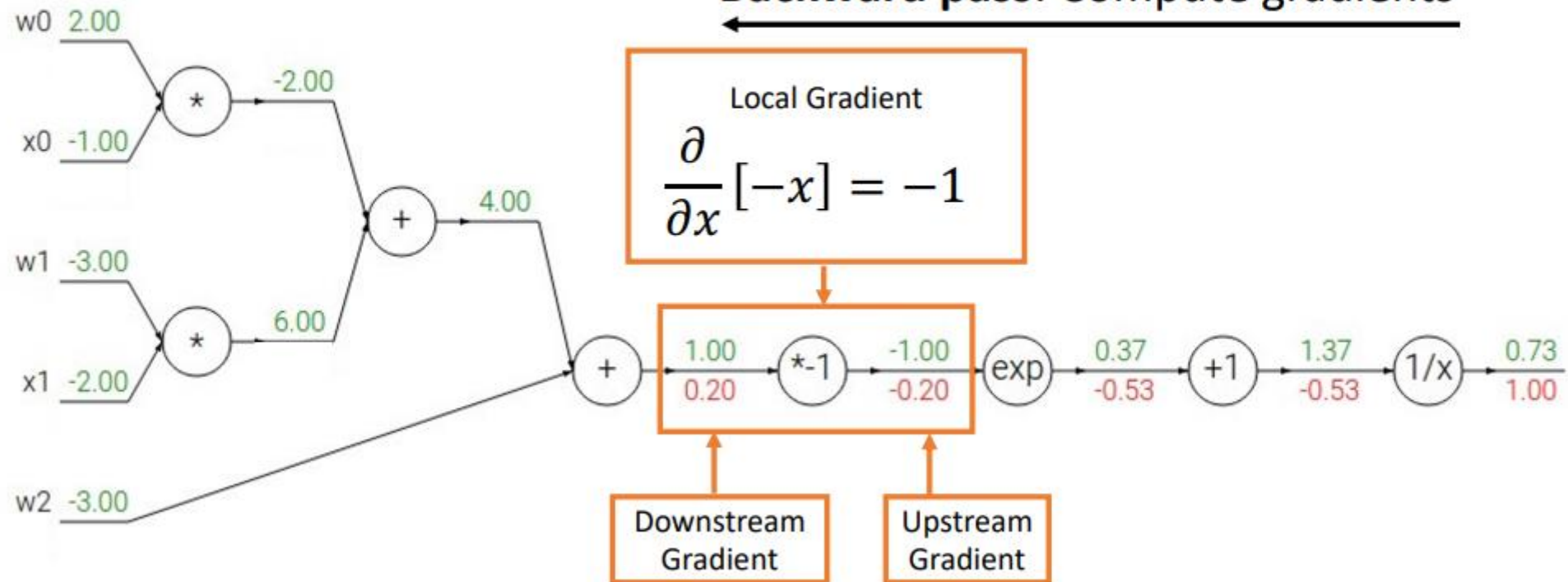
Backward pass: Compute gradients



Backpropagation

Another Example $f(x, w) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$

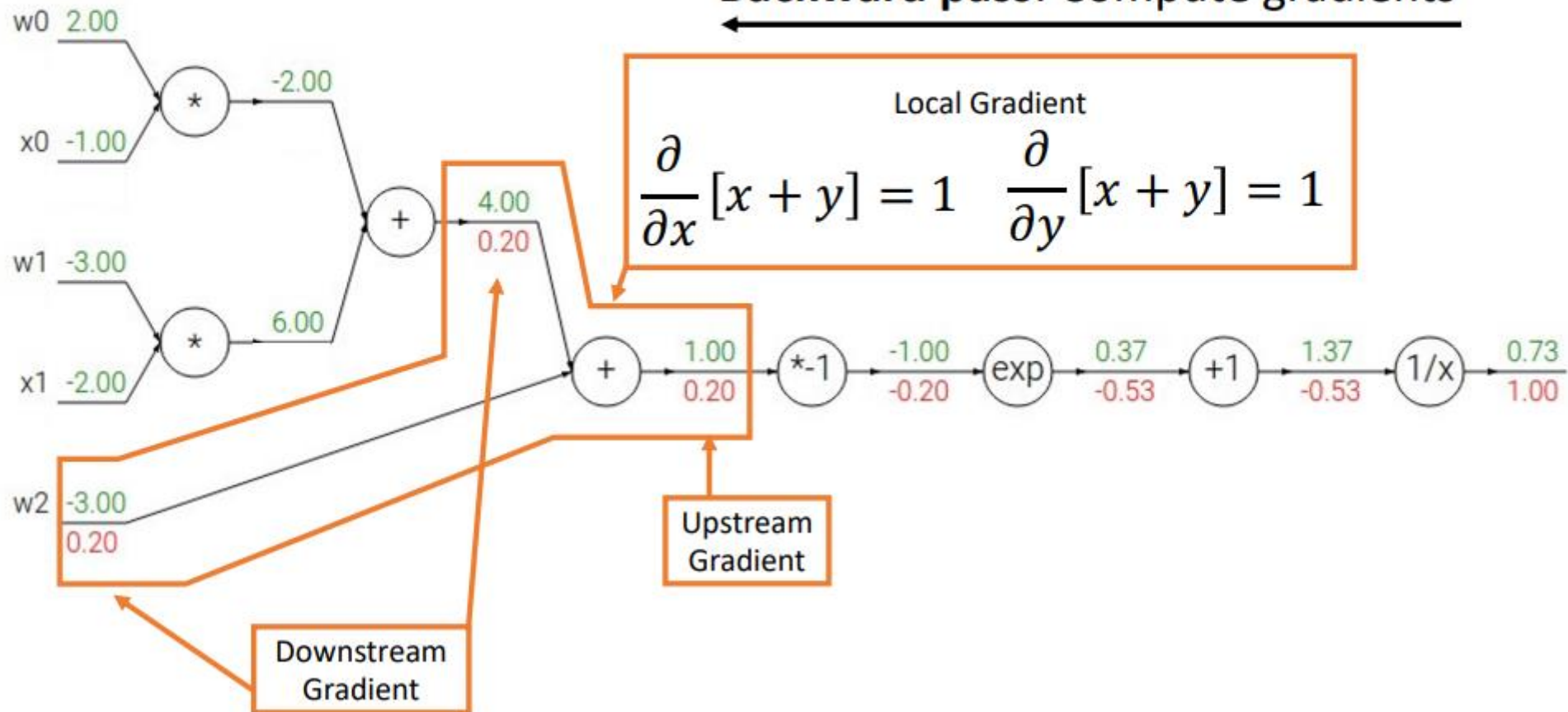
Backward pass: Compute gradients



Backpropagation

Another Example $f(x, w) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$

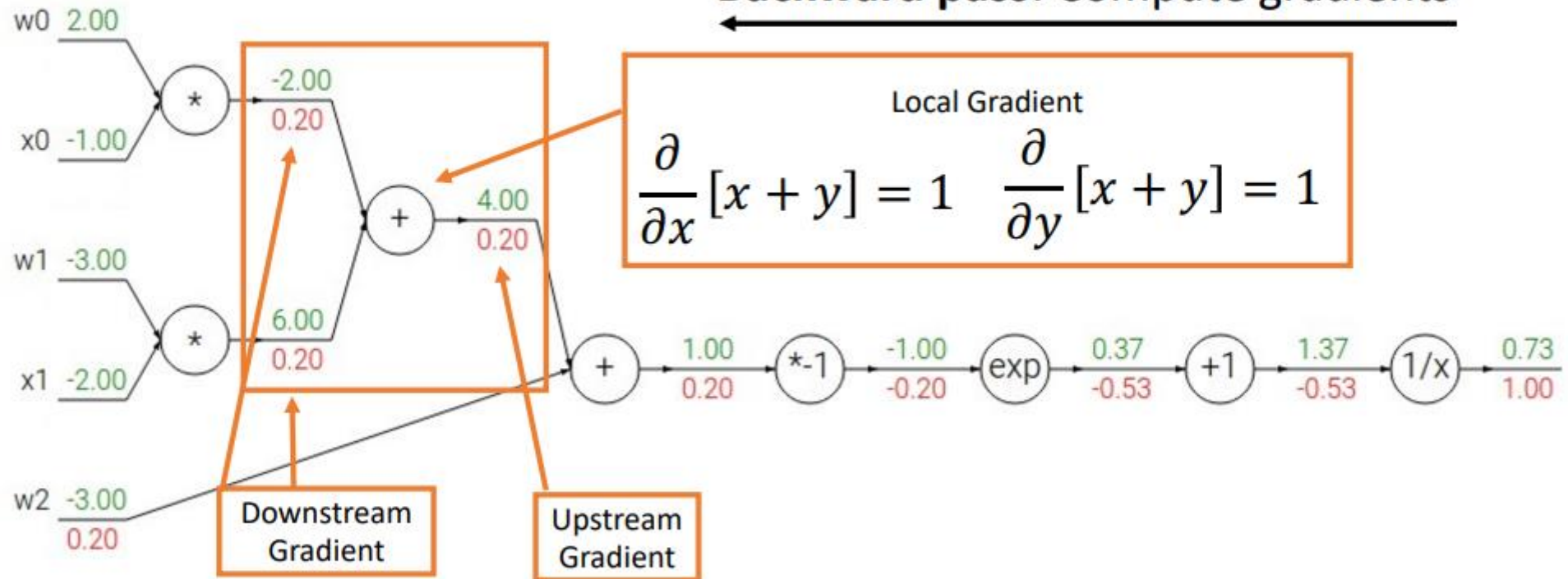
Backward pass: Compute gradients



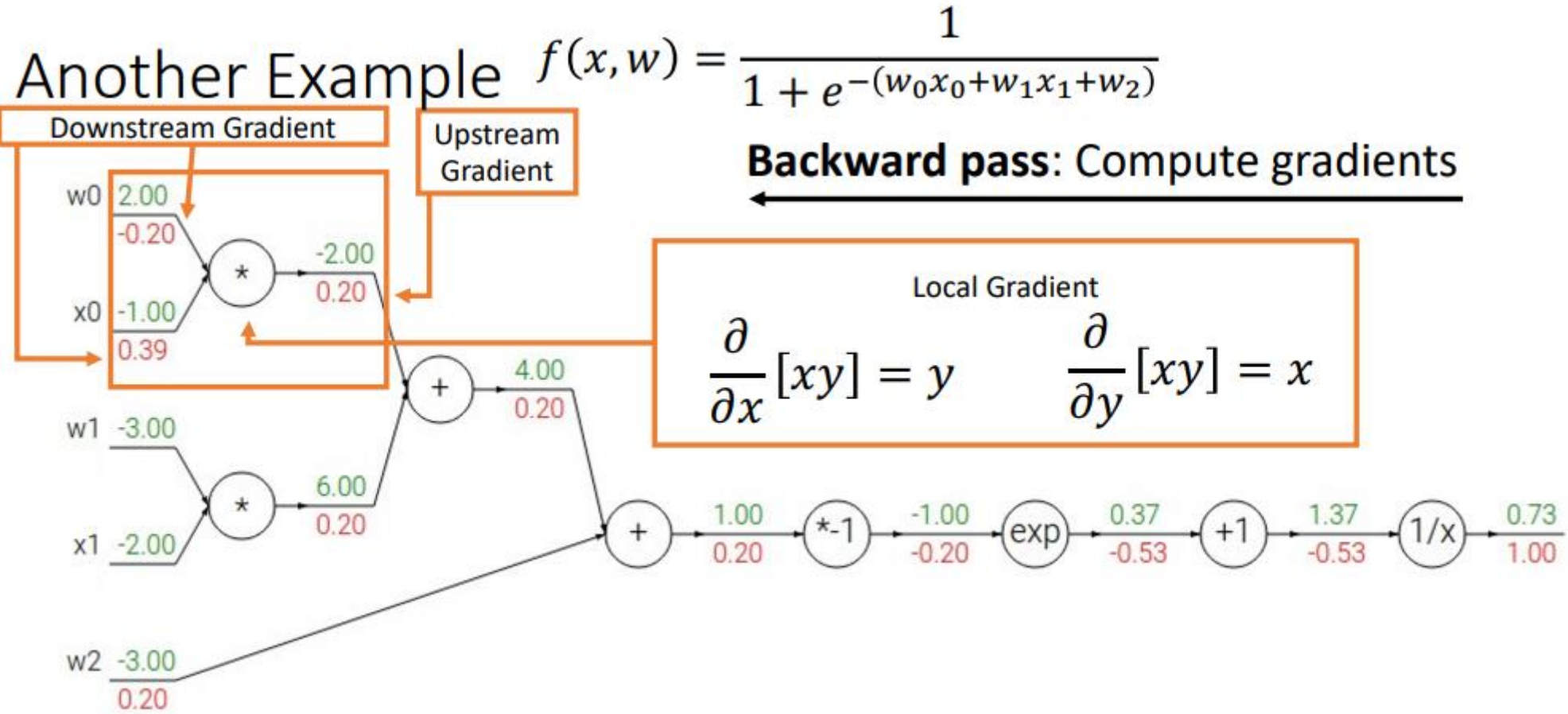
Backpropagation

Another Example $f(x, w) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$

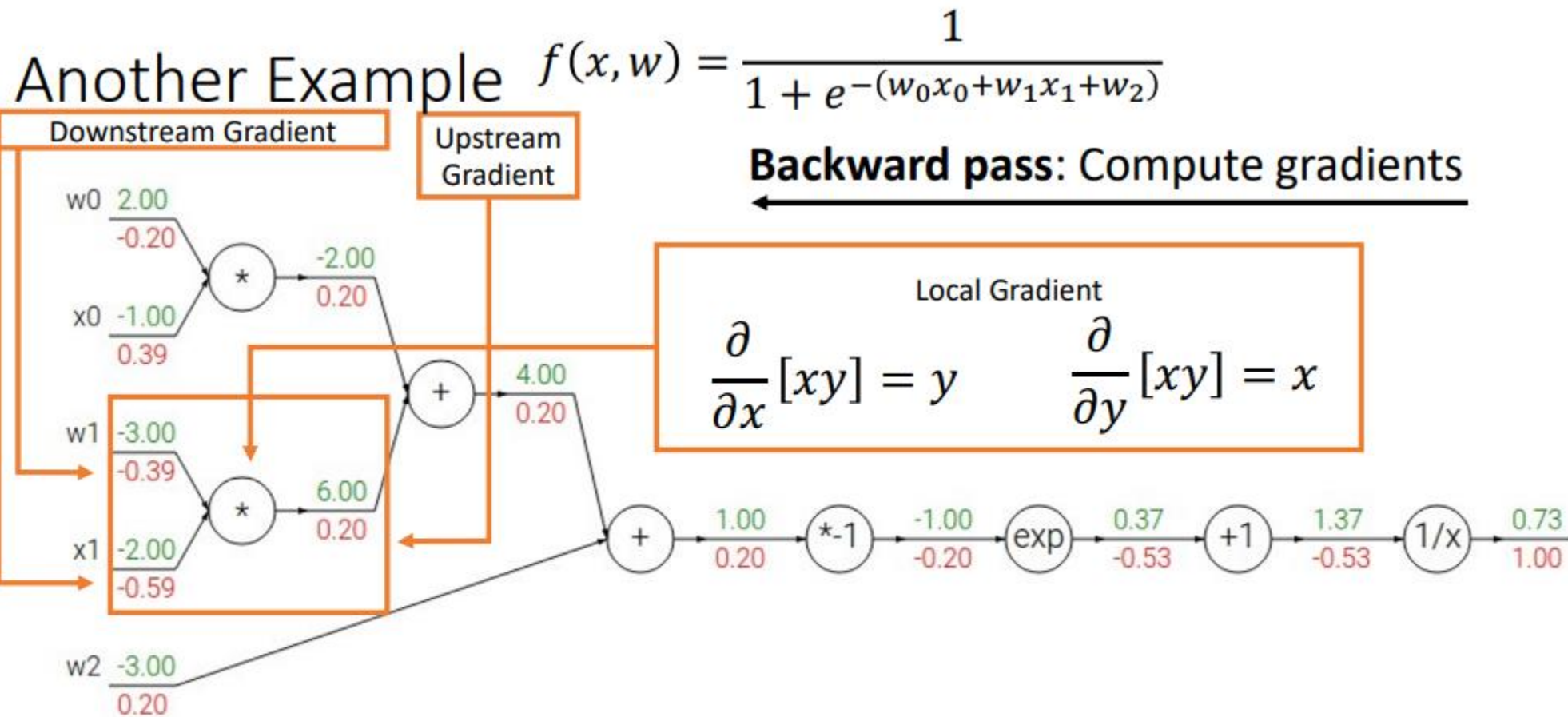
Backward pass: Compute gradients



Backpropagation



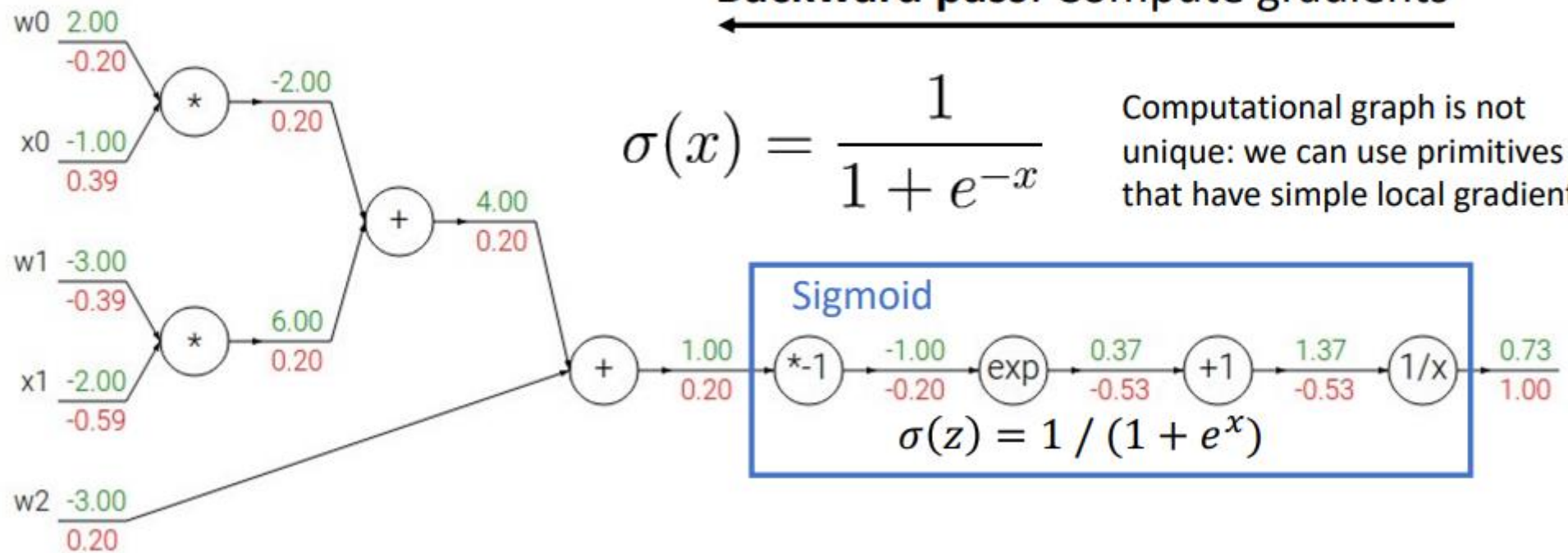
Backpropagation



Backpropagation

Another Example $f(x, w) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}} = \sigma(w_0x_0 + w_1x_1 + w_2)$

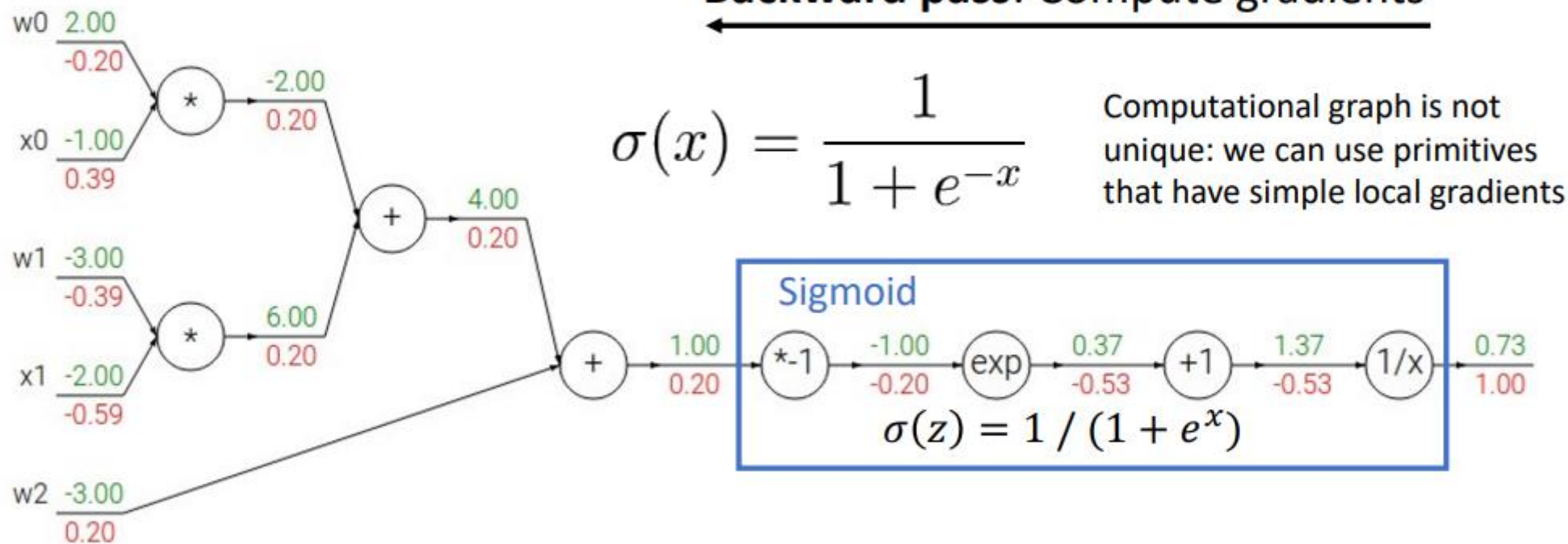
Backward pass: Compute gradients



Backpropagation

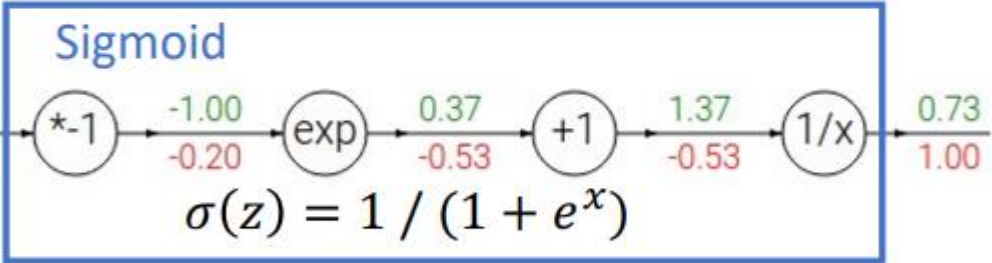
Another Example $f(x, w) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}} = \sigma(w_0x_0 + w_1x_1 + w_2)$

Backward pass: Compute gradients



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Computational graph is not unique: we can use primitives that have simple local gradients

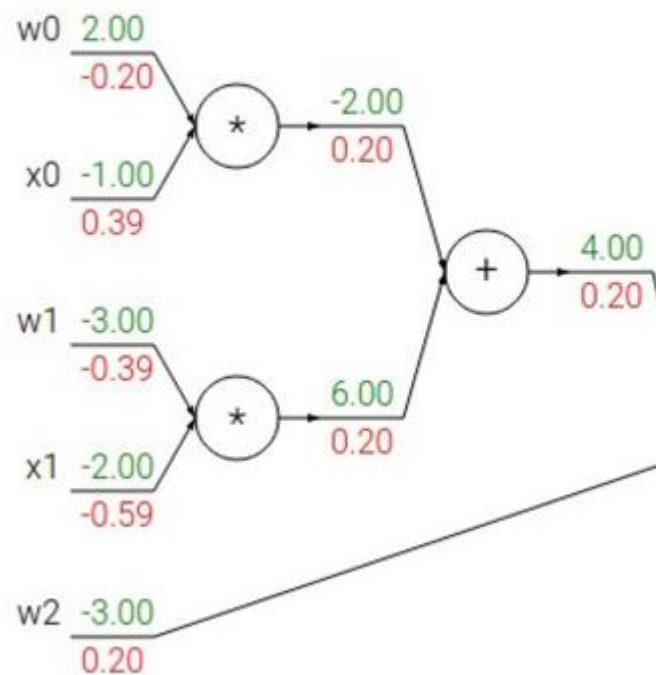


Sigmoid local gradient: $\frac{\partial}{\partial x} [\sigma(x)] = \frac{e^{-x}}{(1 + e^{-x})^2} = \left(\frac{1 + e^{-x} - 1}{1 + e^{-x}} \right) \left(\frac{1}{1 + e^{-x}} \right) = (1 - \sigma(x))\sigma(x)$

Backpropagation

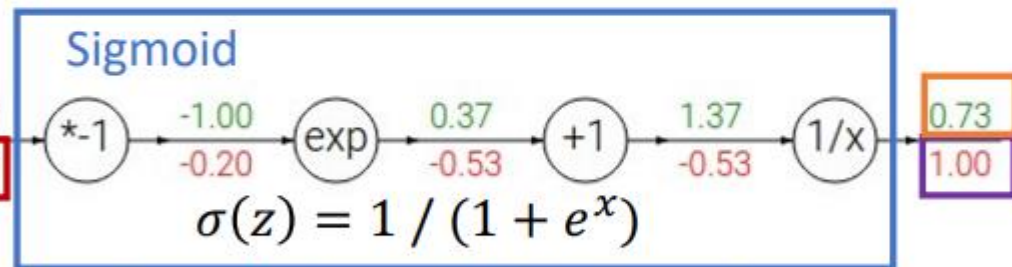
Another Example $f(x, w) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}} = \sigma(w_0x_0 + w_1x_1 + w_2)$

Backward pass: Compute gradients



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Computational graph is not unique: we can use primitives that have simple local gradients



$$[\text{Downstream}] = [\text{Local}] * [\text{Upstream}]$$

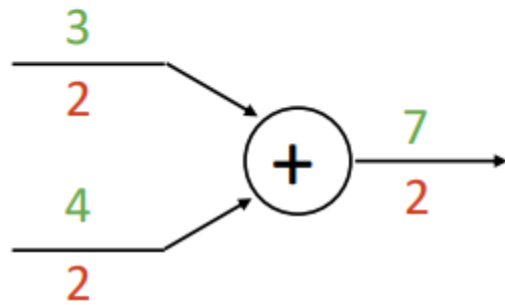
$$= (1 - 0.73) * 0.73 * 1.0 = 0.2$$

Sigmoid local gradient: $\frac{\partial}{\partial x} [\sigma(x)] = \frac{e^{-x}}{(1 + e^{-x})^2} = \left(\frac{1 + e^{-x} - 1}{1 + e^{-x}} \right) \left(\frac{1}{1 + e^{-x}} \right) = (1 - \sigma(x))\sigma(x)$

► Backpropagation

Patterns in Gradient Flow

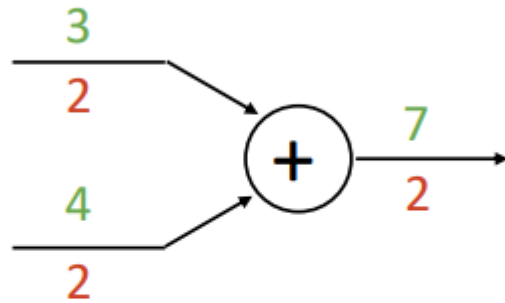
add gate: gradient distributor



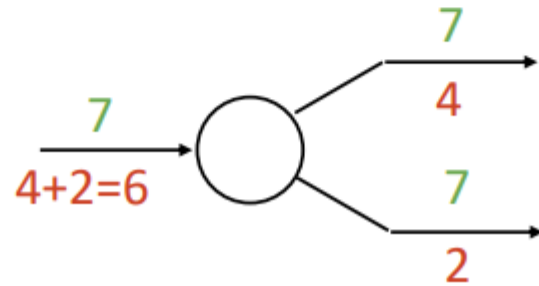
► Backpropagation

Patterns in Gradient Flow

add gate: gradient distributor



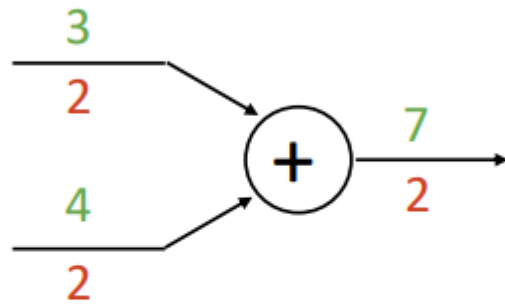
copy gate: gradient adder



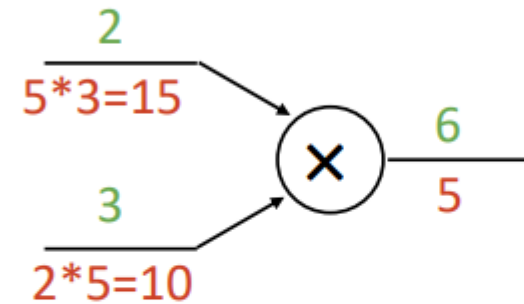
Backpropagation

Patterns in Gradient Flow

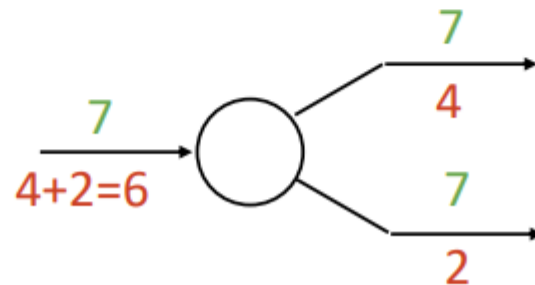
add gate: gradient distributor



mul gate: "swap multiplier"



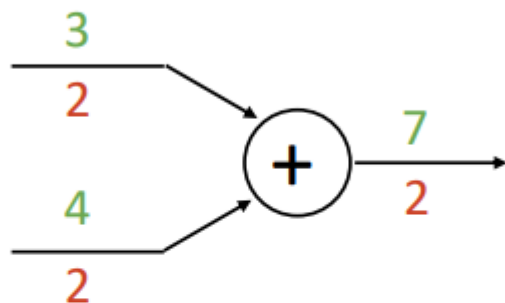
copy gate: gradient adder



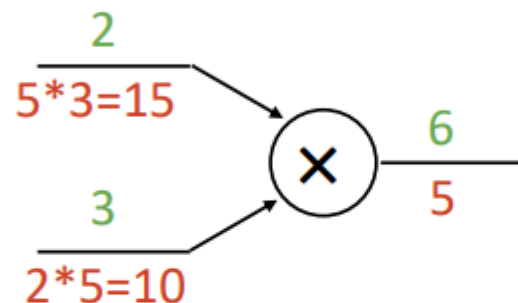
Backpropagation

Patterns in Gradient Flow

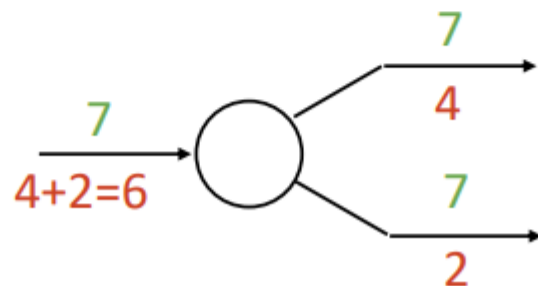
add gate: gradient distributor



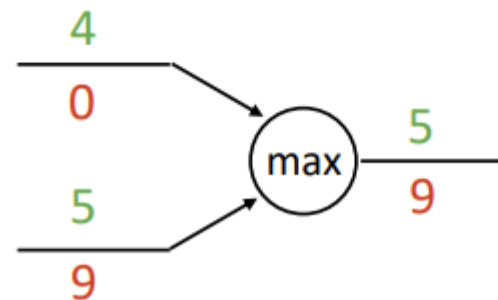
mul gate: "swap multiplier"



copy gate: gradient adder



max gate: gradient router



► Backpropagation

So far: backprop with scalars

What about vector-valued functions?

► Backpropagation

Recap: Vector derivatives

Scalar to Scalar

$$x \in \mathbb{R}, y \in \mathbb{R}$$

Regular derivative:

$$\frac{\partial y}{\partial x} \in \mathbb{R}$$

If x changes by a small amount, how much will y change?

Backpropagation

Recap: Vector derivatives

Scalar to Scalar

$$x \in \mathbb{R}, y \in \mathbb{R}$$

Regular derivative:

$$\frac{\partial y}{\partial x} \in \mathbb{R}$$

If x changes by a small amount, how much will y change?

Vector to Scalar

$$x \in \mathbb{R}^N, y \in \mathbb{R}$$

Derivative is **Gradient**:

$$\frac{\partial y}{\partial x} \in \mathbb{R}^N \quad \left(\frac{\partial y}{\partial x} \right)_n = \frac{\partial y}{\partial x_n}$$

For each element of x , if it changes by a small amount then how much will y change?

Backpropagation

Recap: Vector derivatives

Scalar to Scalar

$$x \in \mathbb{R}, y \in \mathbb{R}$$

Regular derivative:

$$\frac{\partial y}{\partial x} \in \mathbb{R}$$

If x changes by a small amount, how much will y change?

Vector to Scalar

$$x \in \mathbb{R}^N, y \in \mathbb{R}$$

Derivative is **Gradient**:

$$\frac{\partial y}{\partial x} \in \mathbb{R}^N \quad \left(\frac{\partial y}{\partial x} \right)_n = \frac{\partial y}{\partial x_n}$$

For each element of x , if it changes by a small amount then how much will y change?

Vector to Vector

$$x \in \mathbb{R}^N, y \in \mathbb{R}^M$$

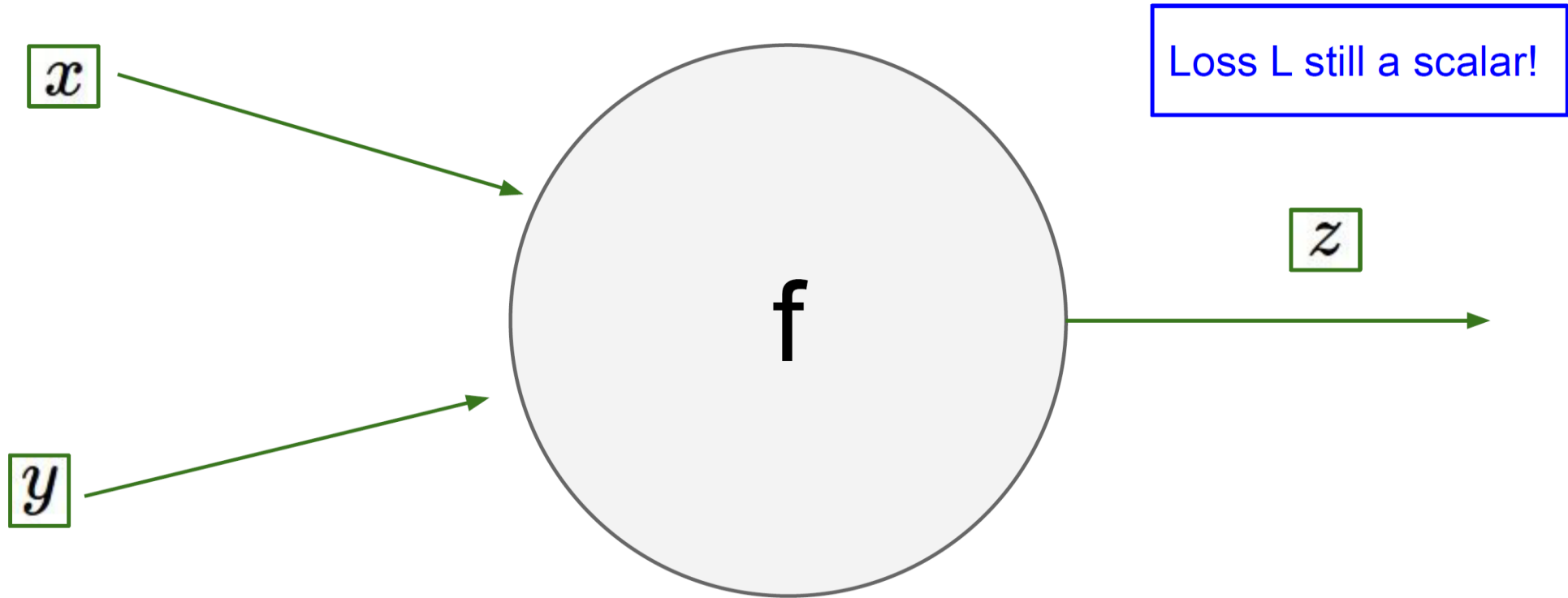
Derivative is **Jacobian**:

$$\frac{\partial y}{\partial x} \in \mathbb{R}^{N \times M} \quad \left(\frac{\partial y}{\partial x} \right)_{n,m} = \frac{\partial y_m}{\partial x_n}$$

For each element of x , if it changes by a small amount then how much will each element of y change?

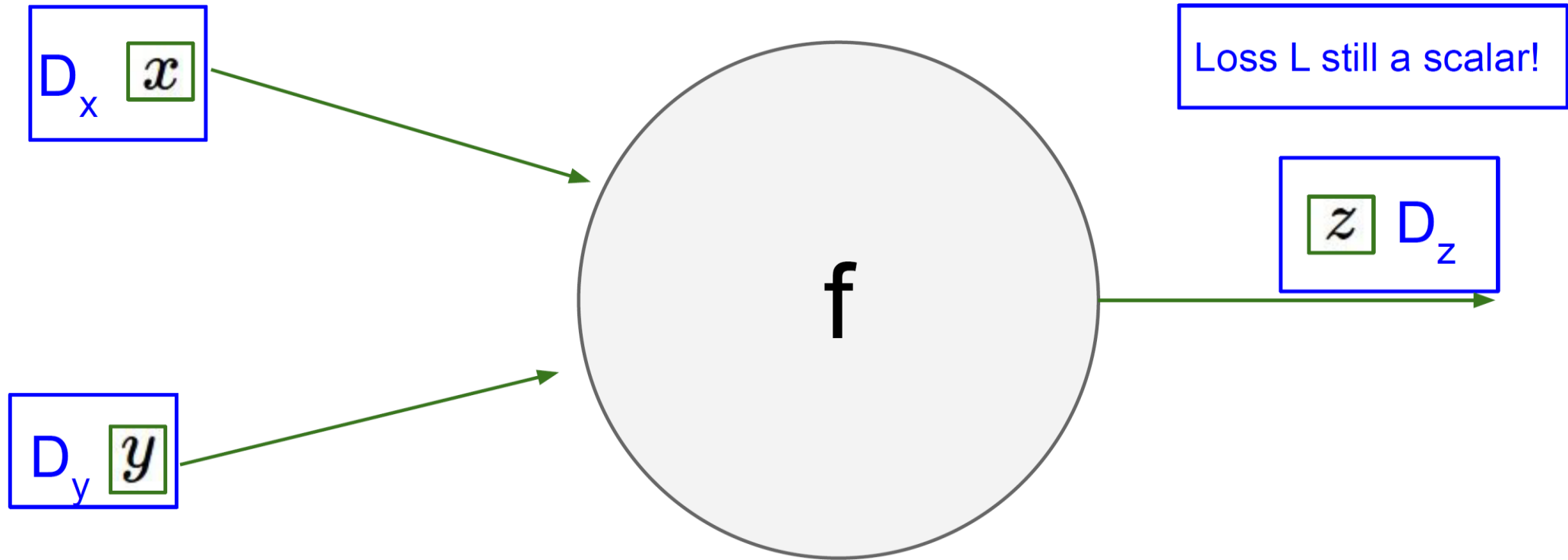
► Backpropagation

Backprop with Vectors



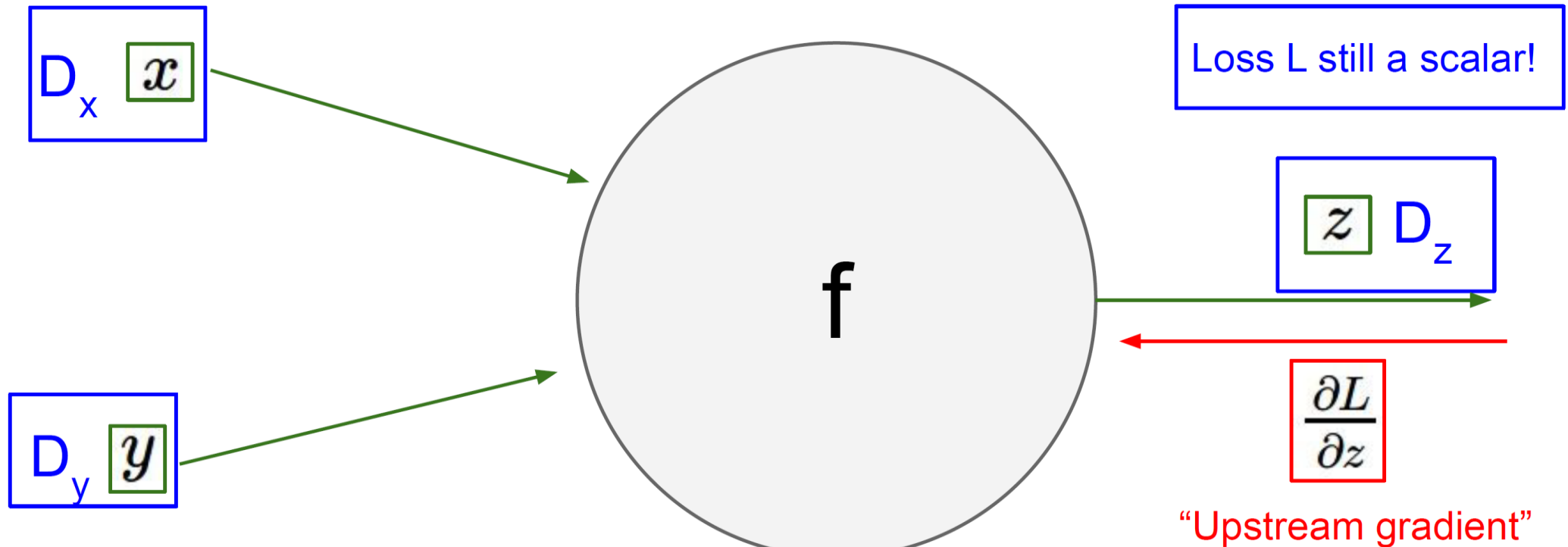
Backpropagation

Backprop with Vectors



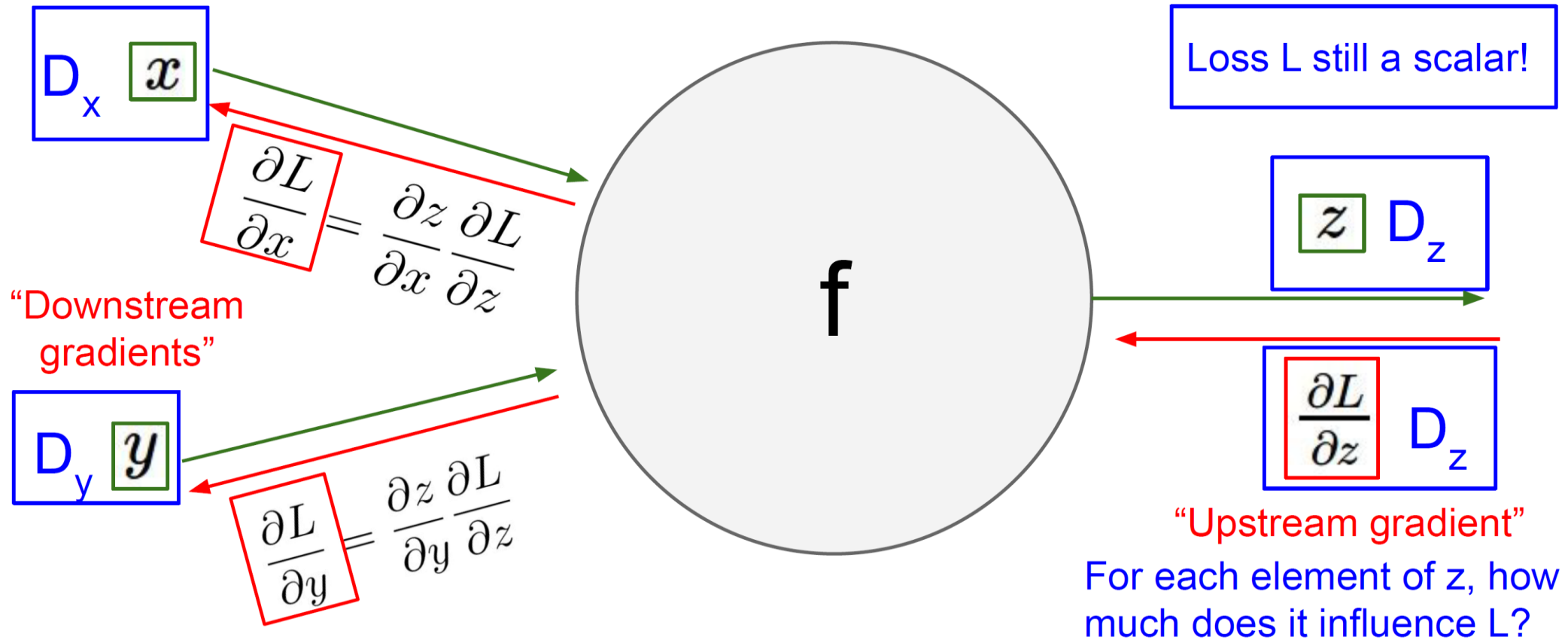
Backpropagation

Backprop with Vectors



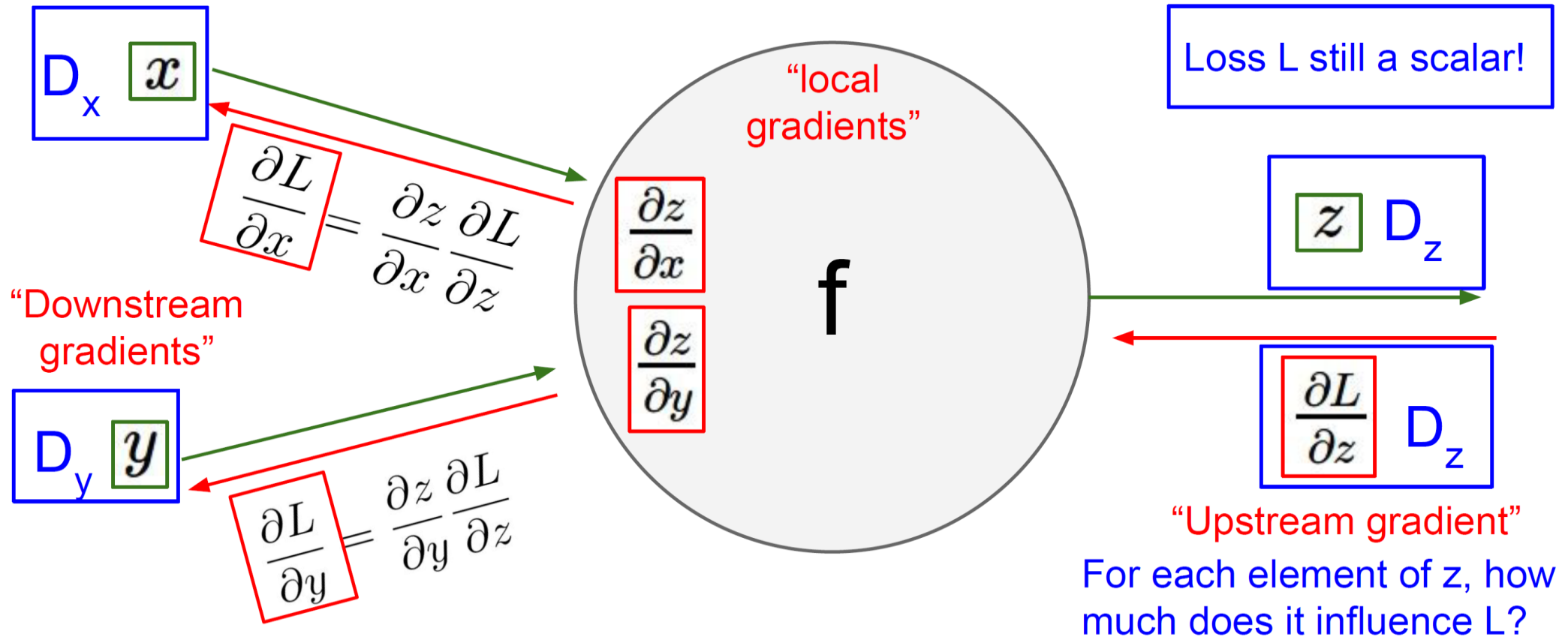
Backpropagation

Backprop with Vectors



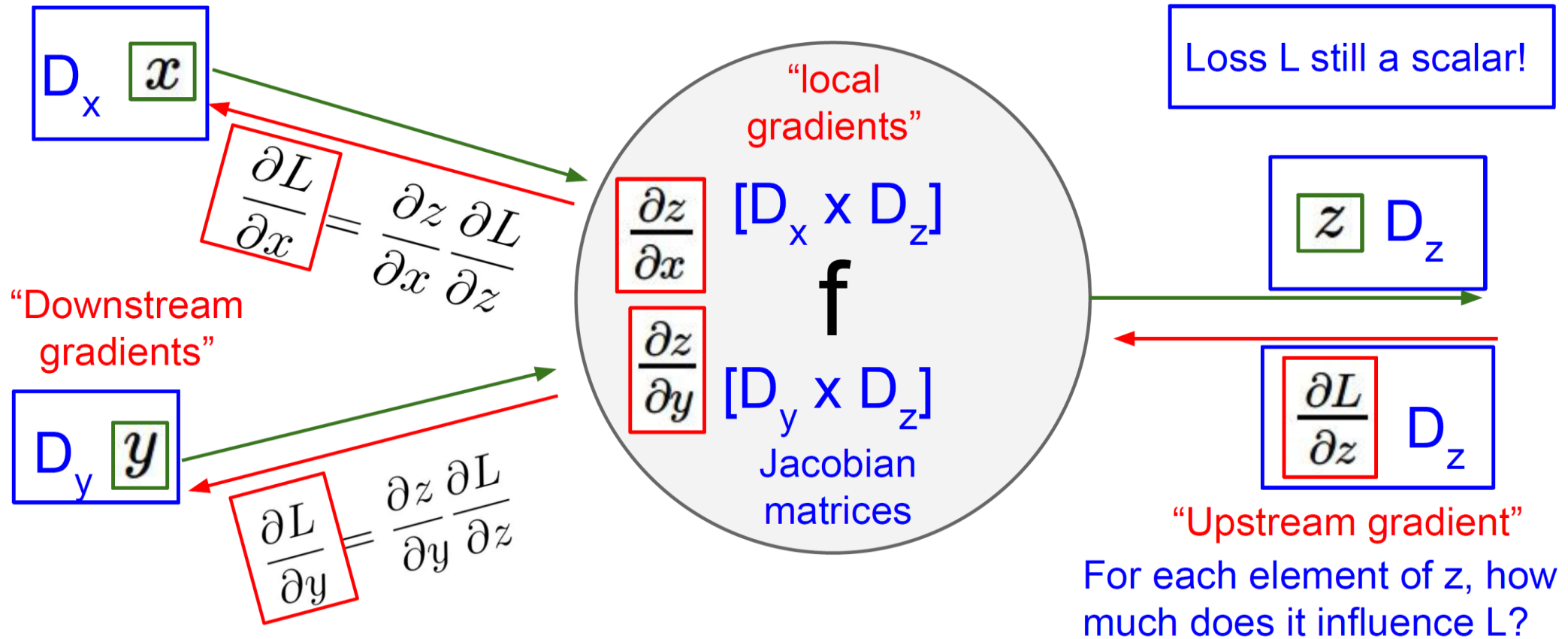
Backpropagation

Backprop with Vectors



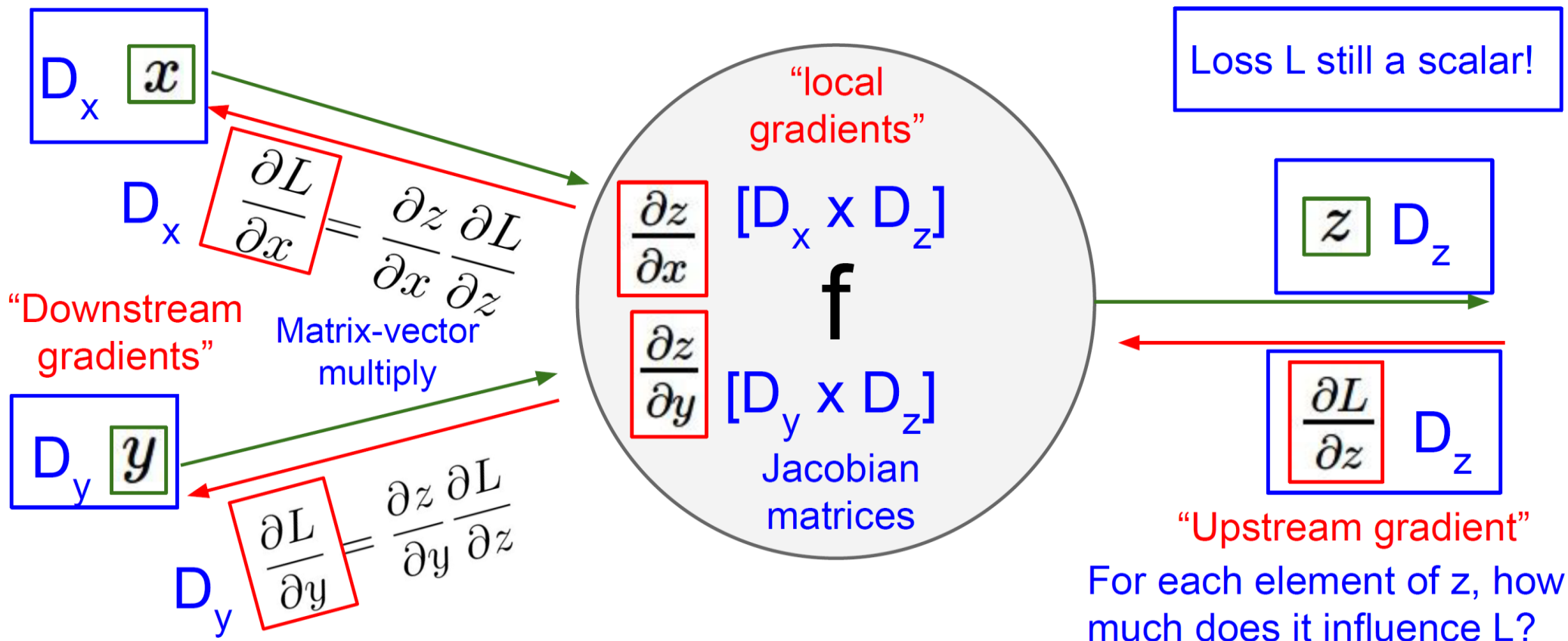
Backpropagation

Backprop with Vectors



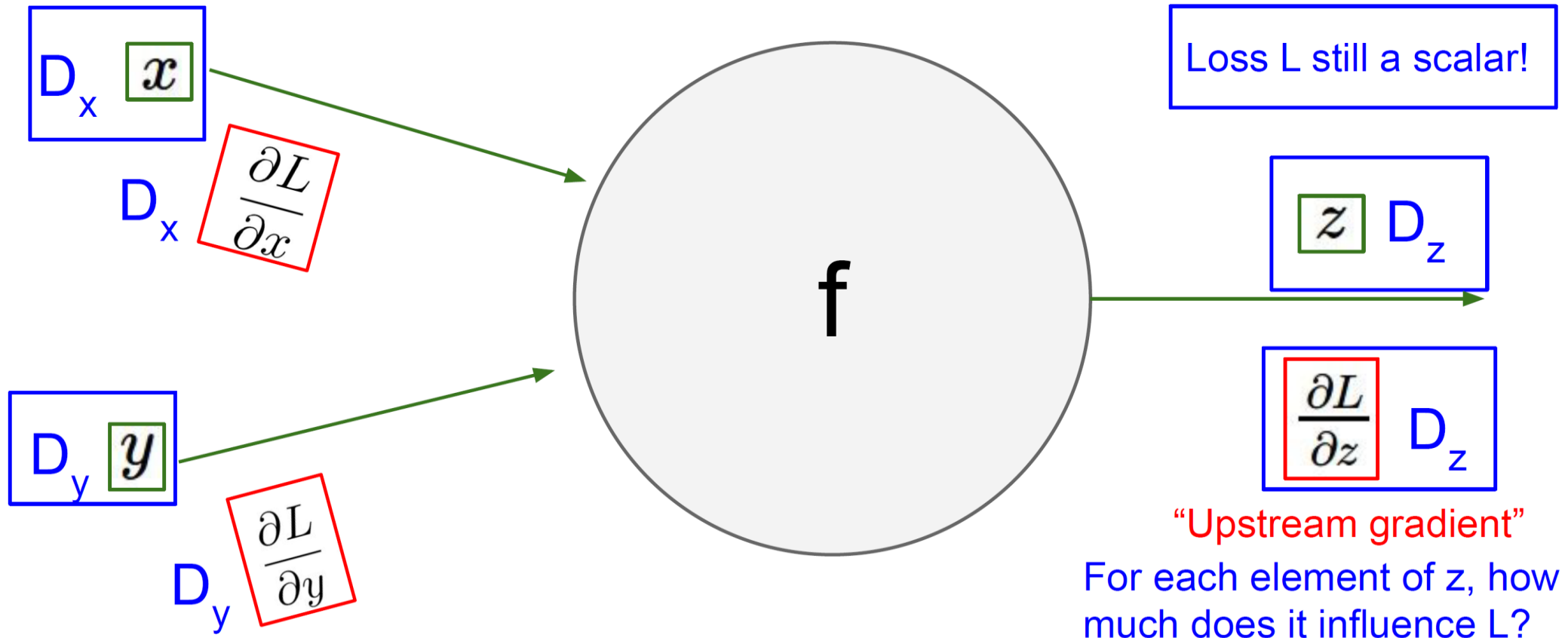
Backpropagation

Backprop with Vectors



Backpropagation

Gradients of variables wrt loss have same dims as the original variable



▶ Backpropagation

Backprop with Vectors

4D input x:

$\begin{bmatrix} 1 \\ -2 \\ 3 \\ -1 \end{bmatrix}$

$f(x) = \max(0, x)$
(*elementwise*)

4D output z:

$\begin{bmatrix} 1 \\ 0 \\ 3 \\ 0 \end{bmatrix}$

▶ Backpropagation

Backprop with Vectors

4D input x:

$\begin{bmatrix} 1 \\ -2 \\ 3 \\ -1 \end{bmatrix}$

$f(x) = \max(0, x)$
(*elementwise*)

4D output z:

$\begin{bmatrix} 1 \\ 0 \\ 3 \\ 0 \end{bmatrix}$

4D dL/dz:

$\begin{bmatrix} 4 \\ -1 \\ 5 \\ 9 \end{bmatrix}$

Upstream
gradient

Backpropagation

Backprop with Vectors

4D input x:

$\begin{bmatrix} 1 \\ -2 \\ 3 \\ -1 \end{bmatrix}$

$f(x) = \max(0, x)$
(*elementwise*)

4D output z:

$\begin{bmatrix} 1 \\ 0 \\ 3 \\ 0 \end{bmatrix}$

Jacobian dz/dx

$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

4D dL/dz :

$\begin{bmatrix} 4 \\ -1 \\ 5 \\ 9 \end{bmatrix}$

Upstream
gradient

Backpropagation

Backprop with Vectors

4D input x:

$\begin{bmatrix} 1 \\ -2 \\ 3 \\ -1 \end{bmatrix}$

$f(x) = \max(0, x)$
(*elementwise*)

4D output z:

$\begin{bmatrix} 1 \\ 0 \\ 3 \\ 0 \end{bmatrix}$

$\begin{bmatrix} dz/dx & dL/dz \end{bmatrix}$

$\begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 4 \end{bmatrix}$

$\begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -1 \end{bmatrix}$

$\begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 5 \end{bmatrix}$

$\begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 9 \end{bmatrix}$

4D dL/dz:

$\begin{bmatrix} 4 \end{bmatrix}$

$\begin{bmatrix} -1 \end{bmatrix}$

$\begin{bmatrix} 5 \end{bmatrix}$

$\begin{bmatrix} 9 \end{bmatrix}$

Upstream
gradient

Backpropagation

Backprop with Vectors

4D input x:

$\begin{bmatrix} 1 \\ -2 \\ 3 \\ -1 \end{bmatrix}$

$f(x) = \max(0, x)$
(*elementwise*)

4D output z:

$\begin{bmatrix} 1 \\ 0 \\ 3 \\ 0 \end{bmatrix}$

4D dL/dx:

$\begin{bmatrix} 4 \\ 0 \\ 5 \\ 0 \end{bmatrix}$

$\begin{bmatrix} dz/dx & dL/dz \end{bmatrix}$

$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ -1 \\ 5 \\ 9 \end{bmatrix}$

4D dL/dz:

$\begin{bmatrix} 4 \\ -1 \\ 5 \\ 9 \end{bmatrix}$

Upstream
gradient

Backpropagation

Backprop with Vectors

4D input x:

$$\begin{bmatrix} 1 \\ -2 \\ 3 \\ -1 \end{bmatrix}$$

$$f(x) = \max(0, x)$$

(*elementwise*)

4D output z:

$$\begin{bmatrix} 1 \\ 0 \\ 3 \\ 0 \end{bmatrix}$$

Jacobian is **sparse**:
off-diagonal entries
always zero! Never
explicitly form
Jacobian -- instead
use **implicit**
multiplication

4D dL/dx:

$$\begin{bmatrix} 4 \\ 0 \\ 5 \\ 0 \end{bmatrix}$$

$[dz/dx]$ $[dL/dz]$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ -1 \\ 5 \\ 9 \end{bmatrix}$$

4D dL/dz:

$$\begin{bmatrix} 4 \\ -1 \\ 5 \\ 9 \end{bmatrix}$$

Upstream
gradient

Backpropagation

Backprop with Vectors

4D input x:

$\begin{bmatrix} 1 \\ -2 \\ 3 \\ -1 \end{bmatrix}$

$f(x) = \max(0, x)$
(elementwise)

4D output z:

$\begin{bmatrix} 1 \\ 0 \\ 3 \\ 0 \end{bmatrix}$

Jacobian is **sparse**:
 off-diagonal entries
 always zero! Never
explicitly form
 Jacobian -- instead
 use **implicit**
 multiplication

4D dL/dx:

$\begin{bmatrix} 4 \\ 0 \\ 5 \\ 0 \end{bmatrix}$

$[dz/dx] [dL/dz]$

$$\left(\frac{\partial L}{\partial x}\right)_i = \begin{cases} \left(\frac{\partial L}{\partial z}\right)_i & \text{if } x_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

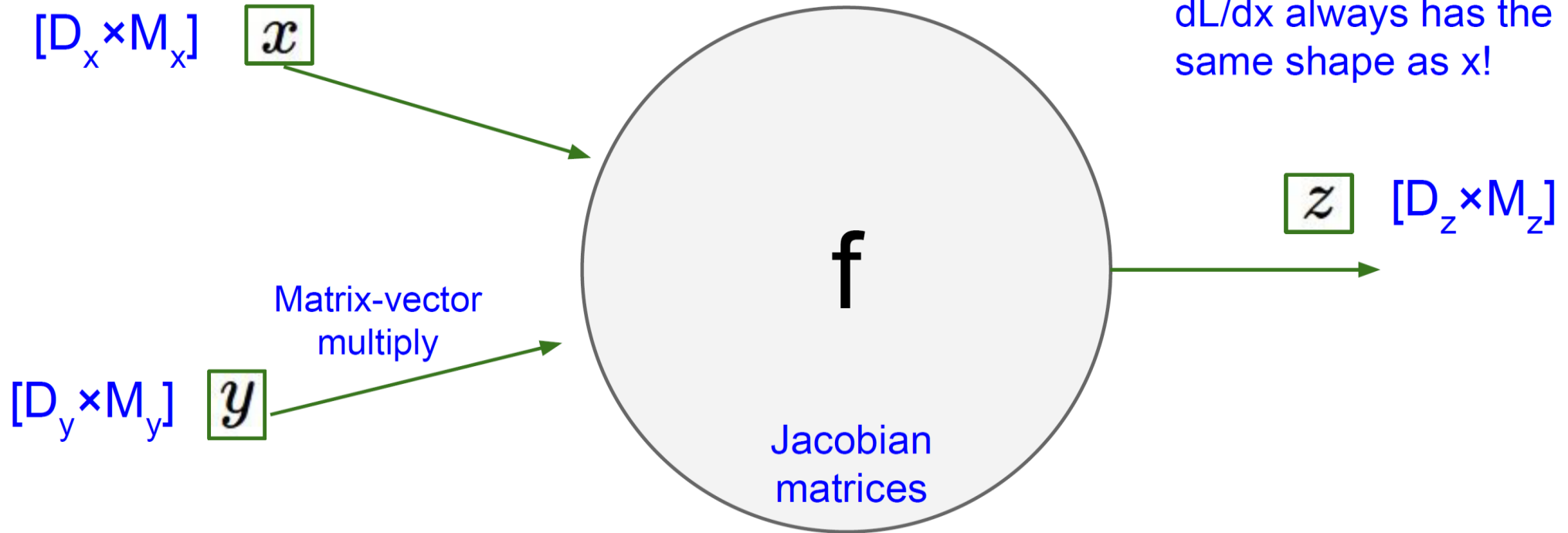
4D dL/dz:

$\begin{bmatrix} 4 \\ -1 \\ 5 \\ 9 \end{bmatrix}$

Upstream
 gradient

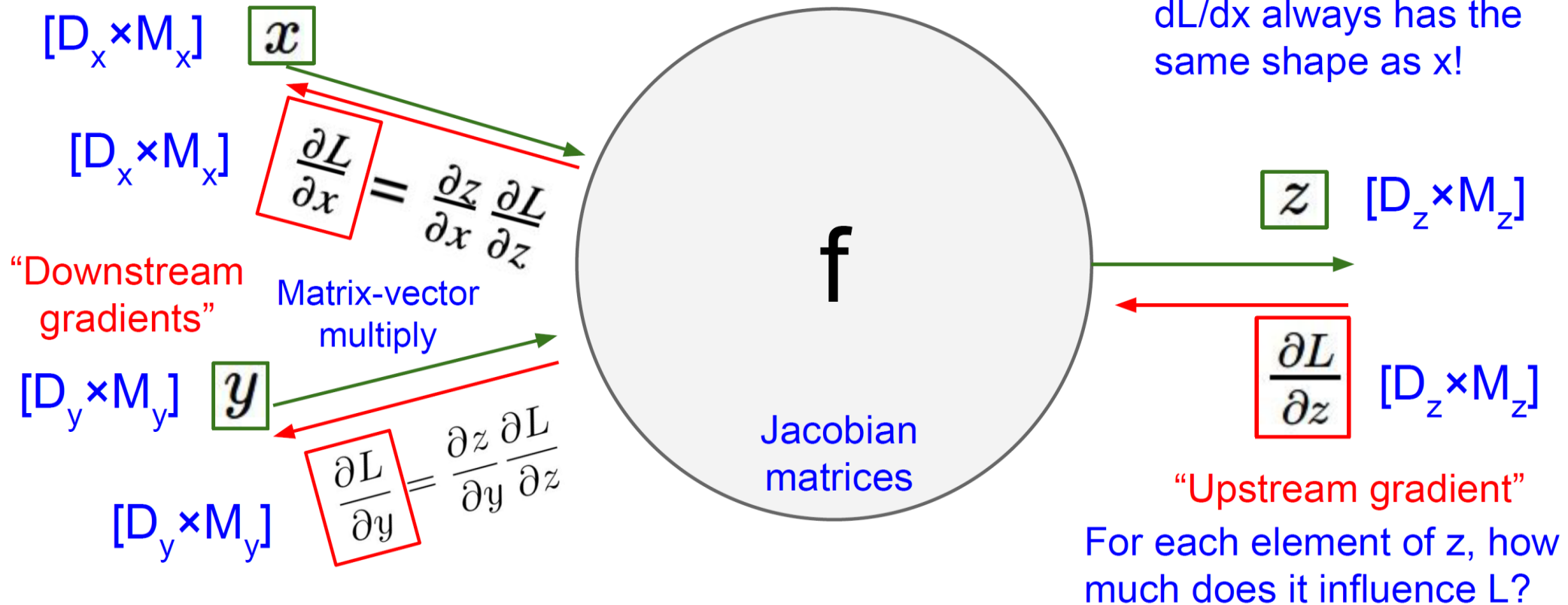
Backpropagation

Backprop with Matrices (or Tensors)



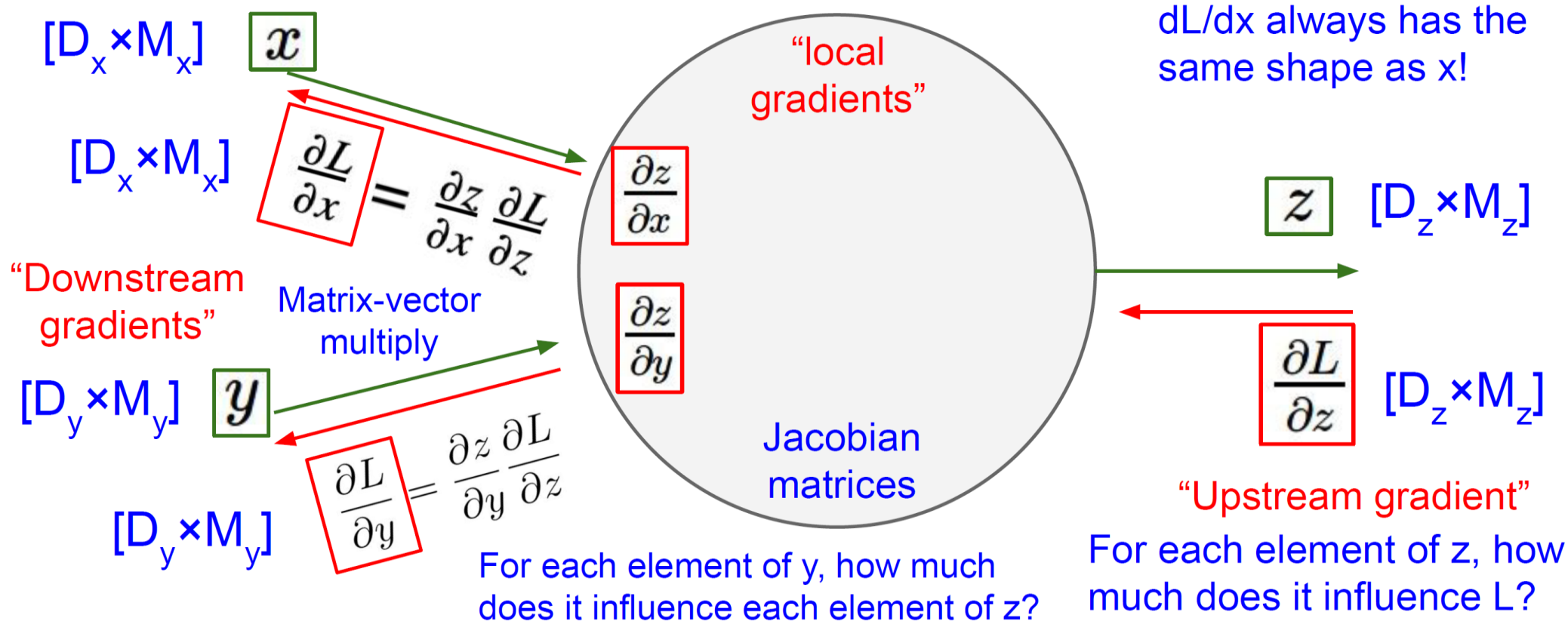
Backpropagation

Backprop with Matrices (or Tensors)



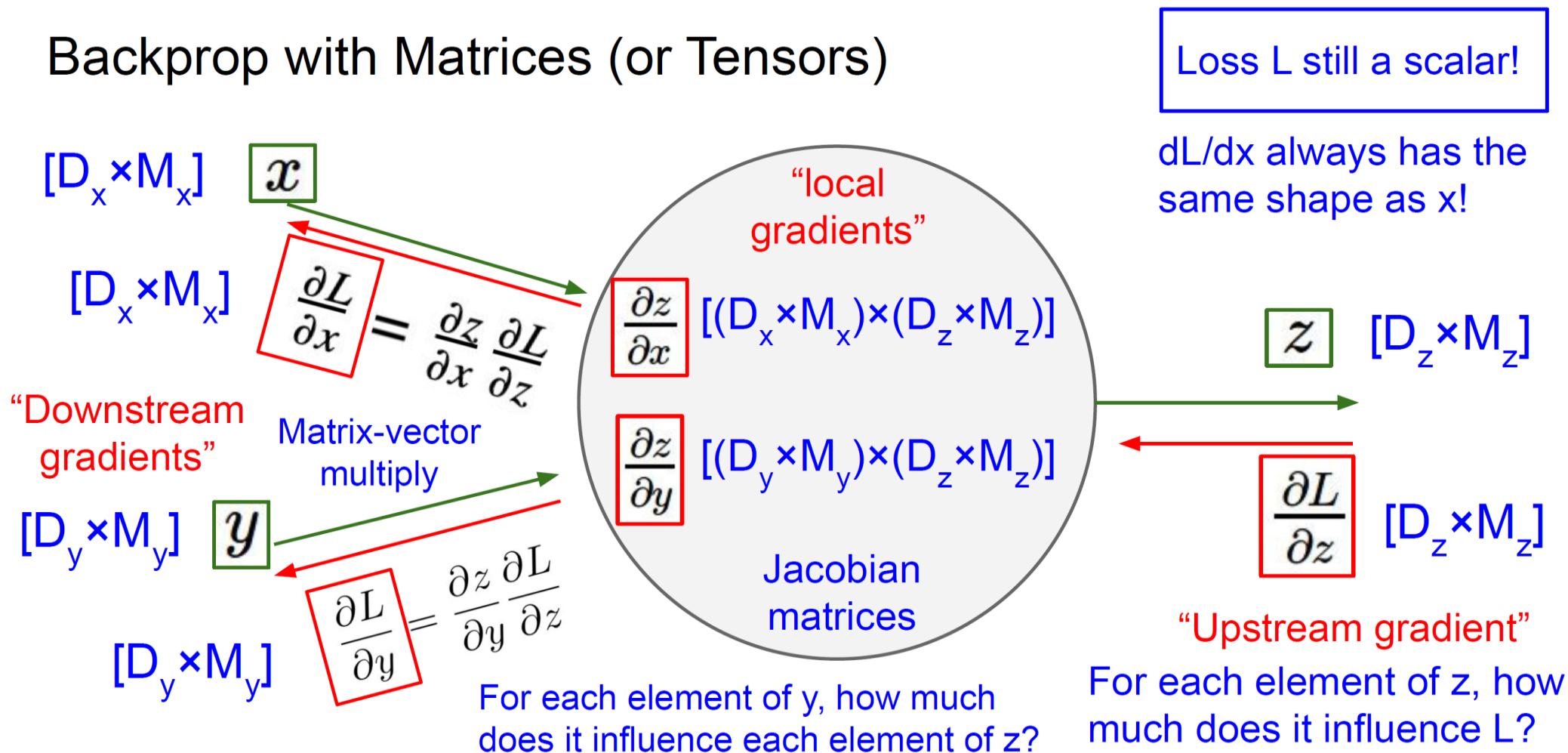
Backpropagation

Backprop with Matrices (or Tensors)



Backpropagation

Backprop with Matrices (or Tensors)



► Backpropagation

Example: Matrix Multiplication

$x: [N \times D]$
[2 1 -3]
[-3 4 2]

$w: [D \times M]$
[3 2 1 -1]
[2 1 3 2]
[3 2 1 -2]



Matrix Multiply $y = xw$

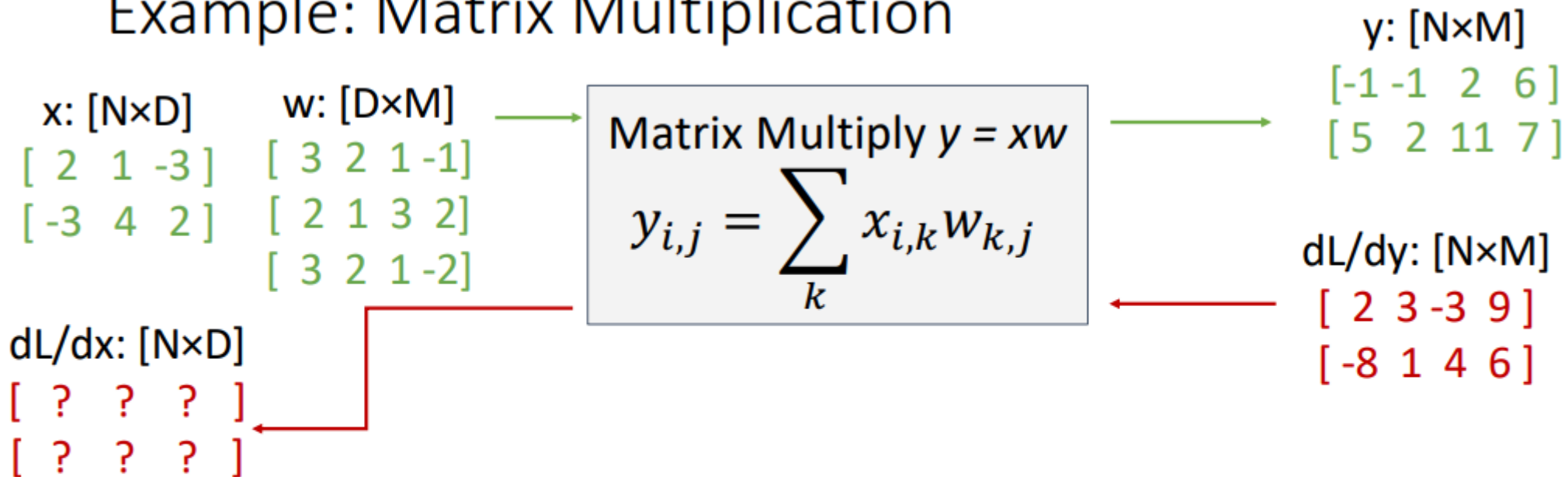
$$y_{i,j} = \sum_k x_{i,k} w_{k,j}$$



$y: [N \times M]$
[-1 -1 2 6]
[5 2 11 7]

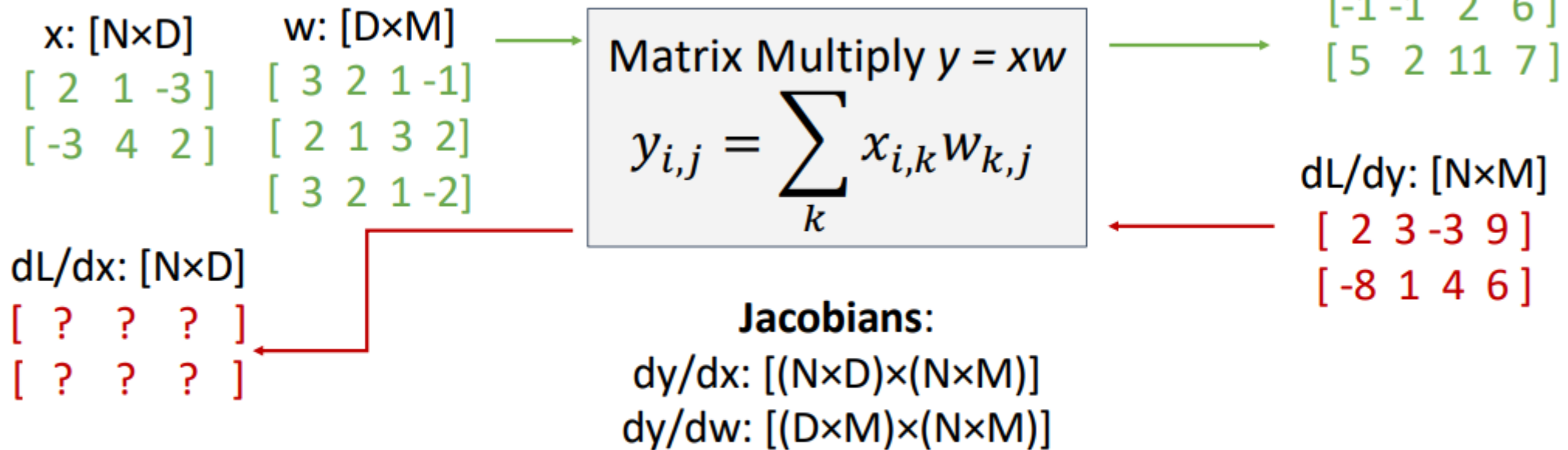
Backpropagation

Example: Matrix Multiplication



Backpropagation

Example: Matrix Multiplication



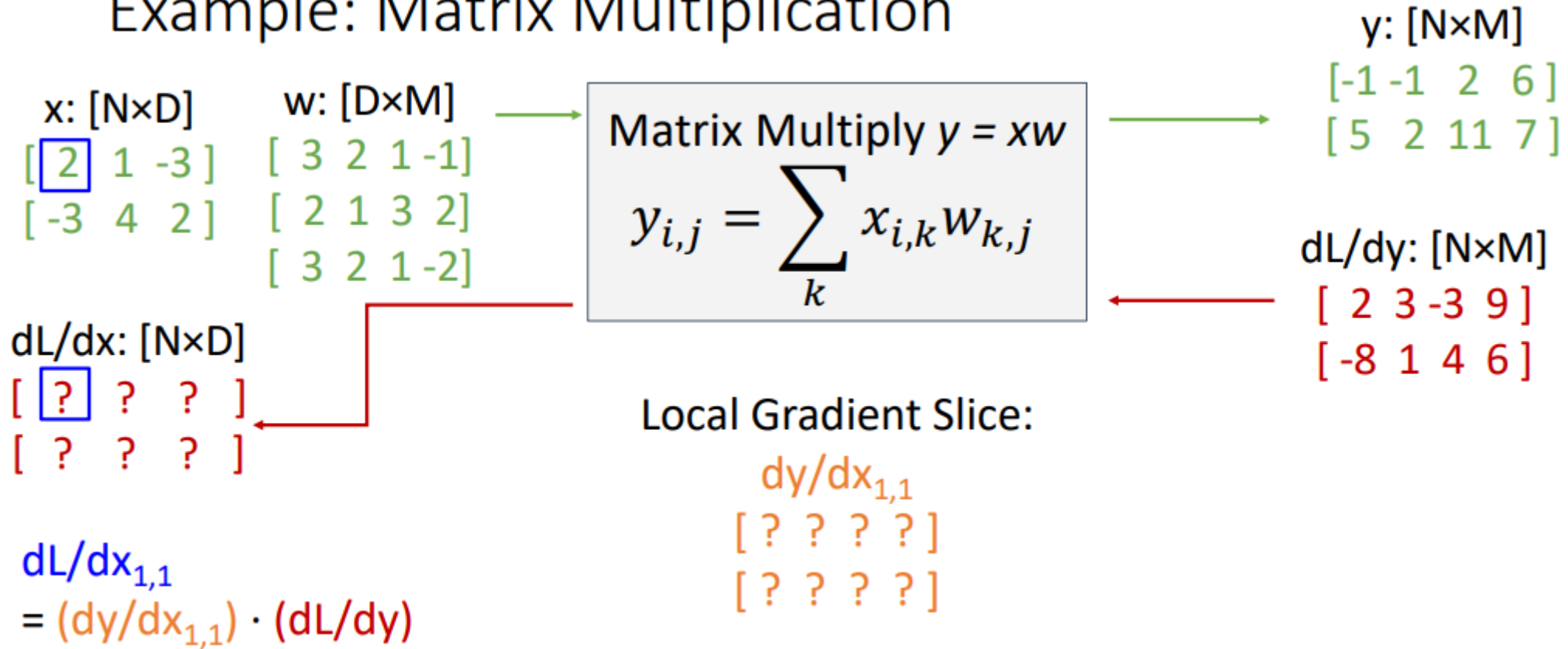
For a neural net we may have

$N=64, D=M=4096$

Each Jacobian takes 256 GB of memory! Must work with them implicitly!

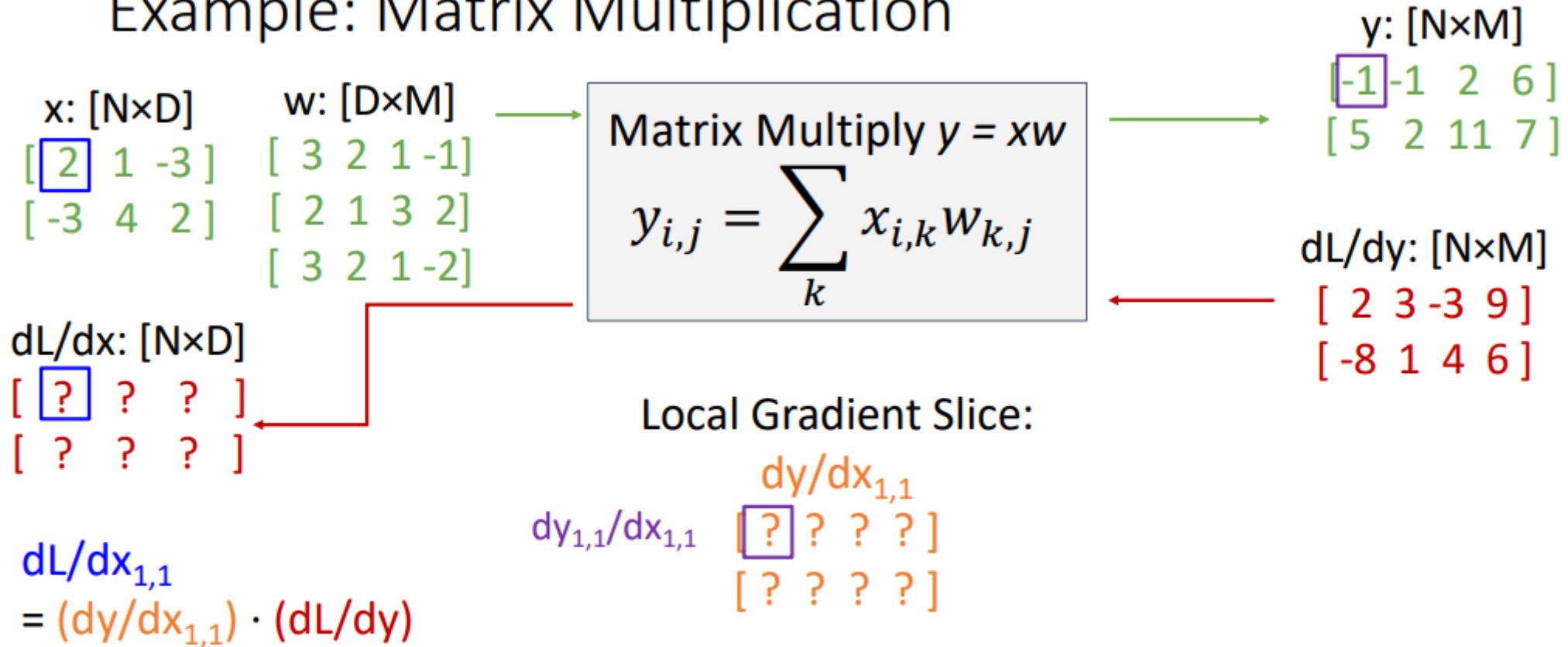
Backpropagation

Example: Matrix Multiplication



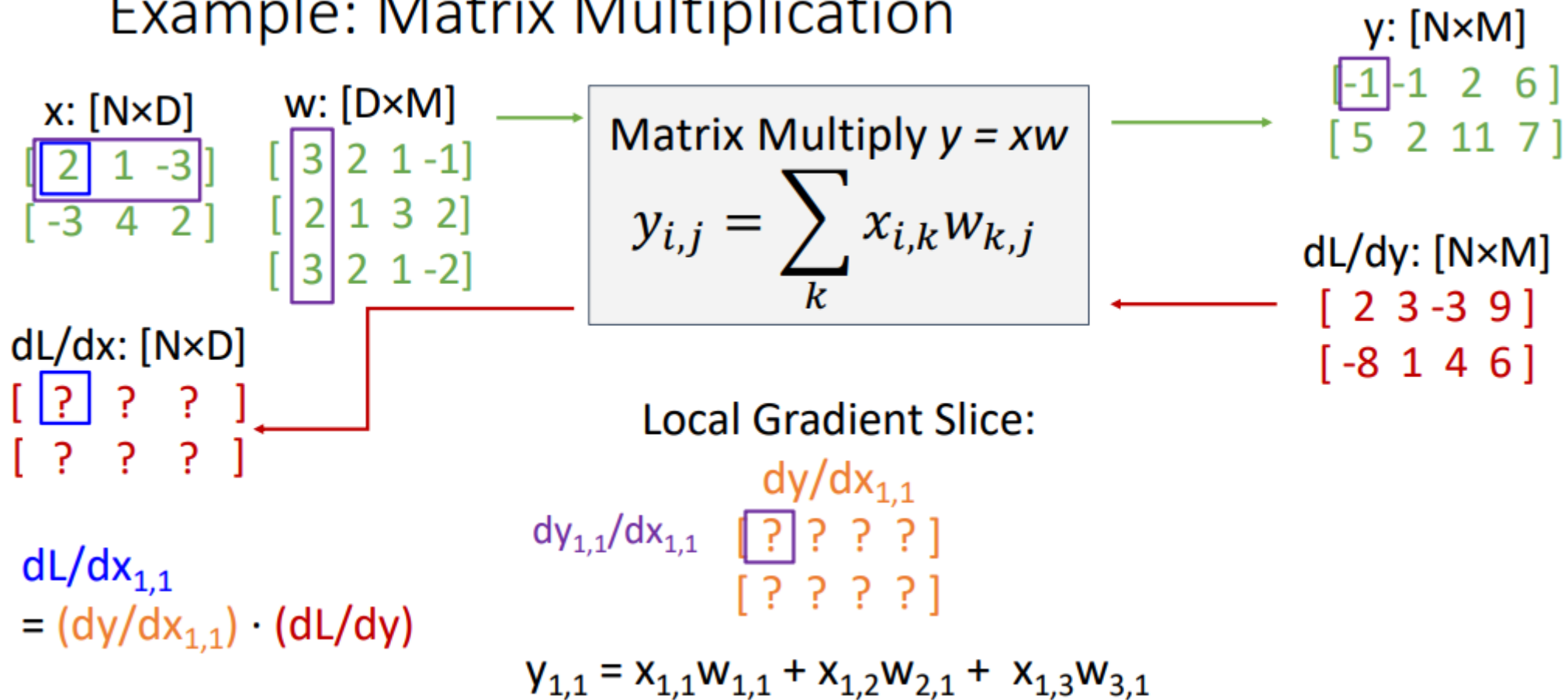
Backpropagation

Example: Matrix Multiplication



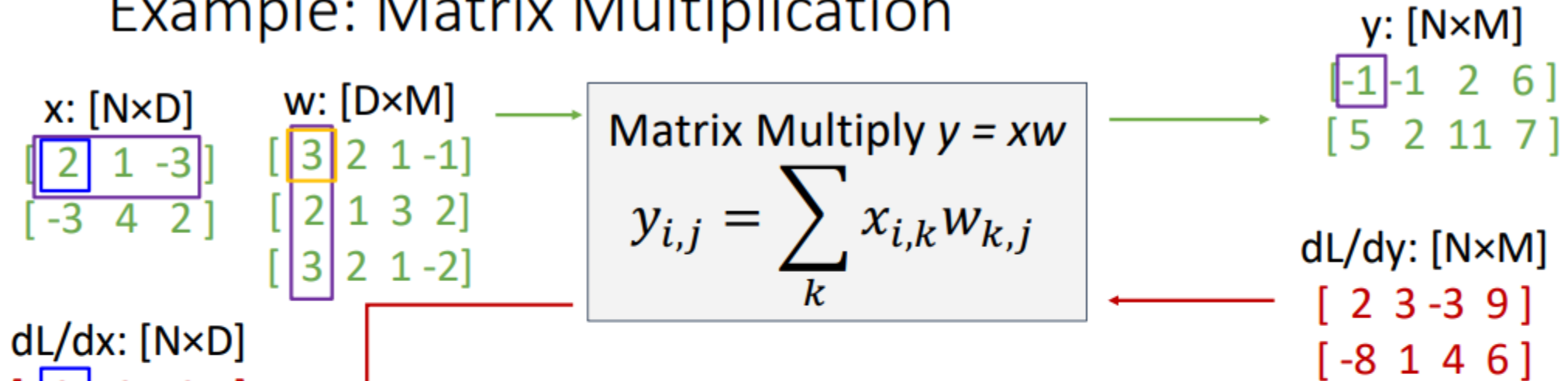
Backpropagation

Example: Matrix Multiplication



Backpropagation

Example: Matrix Multiplication



Local Gradient Slice:

$dy/dx_{1,1}$

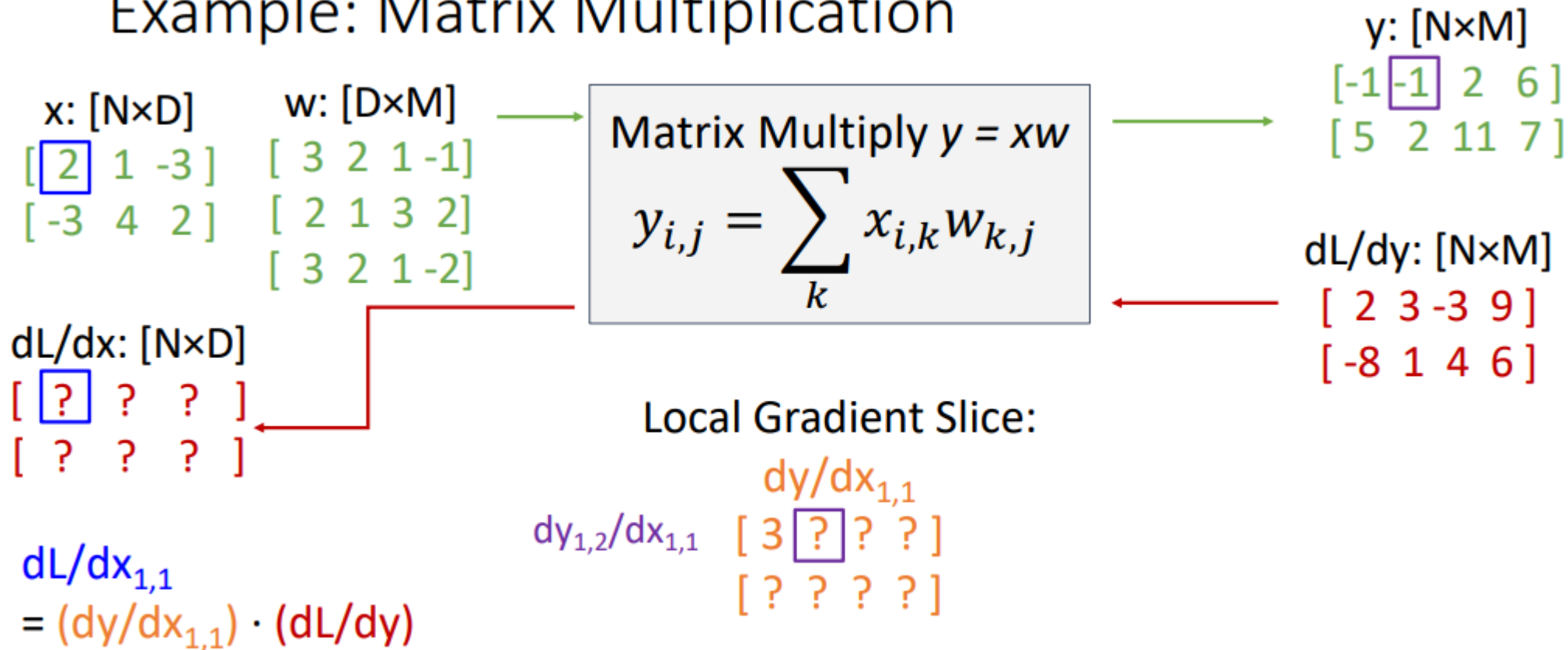
$$dy_{1,1}/dx_{1,1} \begin{bmatrix} 3 & ? & ? & ? \\ ? & ? & ? & ? \end{bmatrix}$$

$$y_{1,1} = x_{1,1} w_{1,1} + x_{1,2} w_{2,1} + x_{1,3} w_{3,1}$$
$$\Rightarrow dy_{1,1}/dx_{1,1} = w_{1,1}$$

$$dL/dx_{1,1} = (dy/dx_{1,1}) \cdot (dL/dy)$$

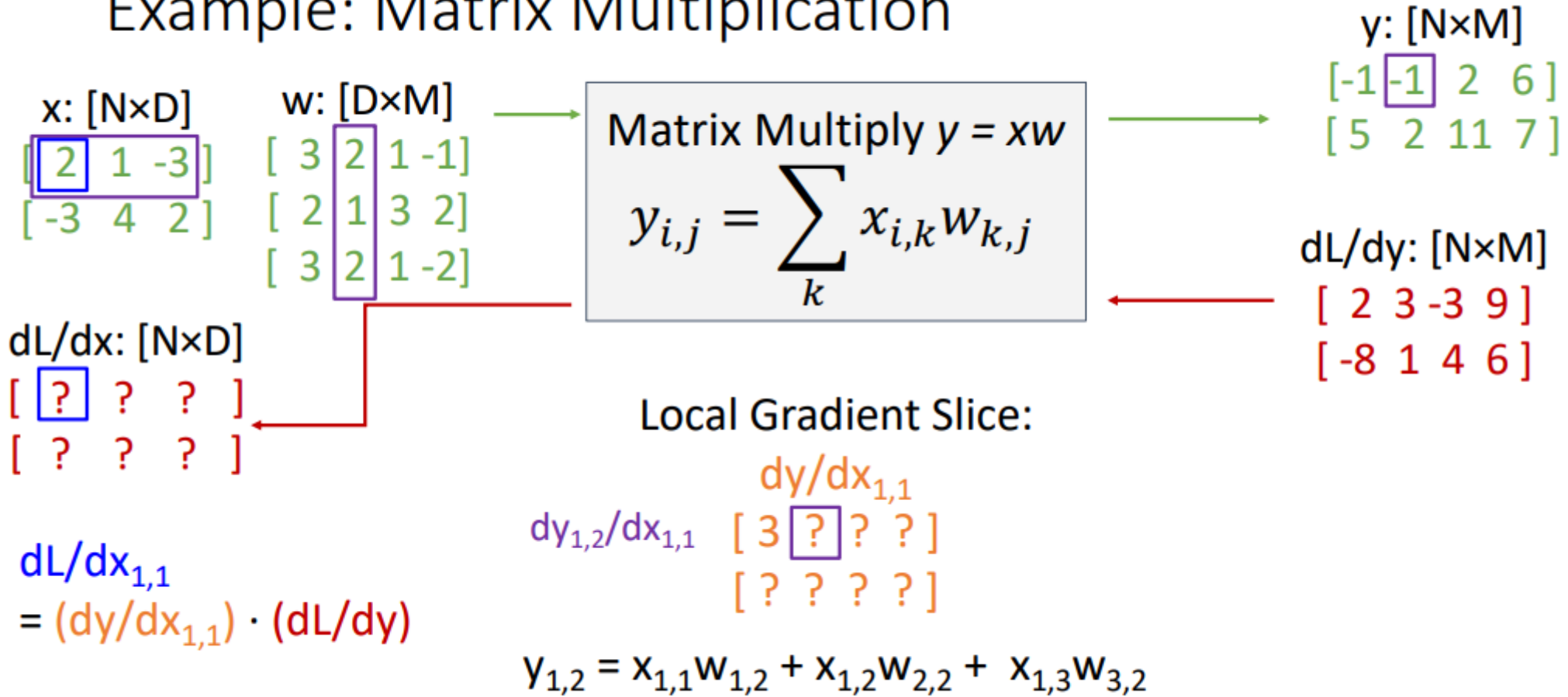
Backpropagation

Example: Matrix Multiplication



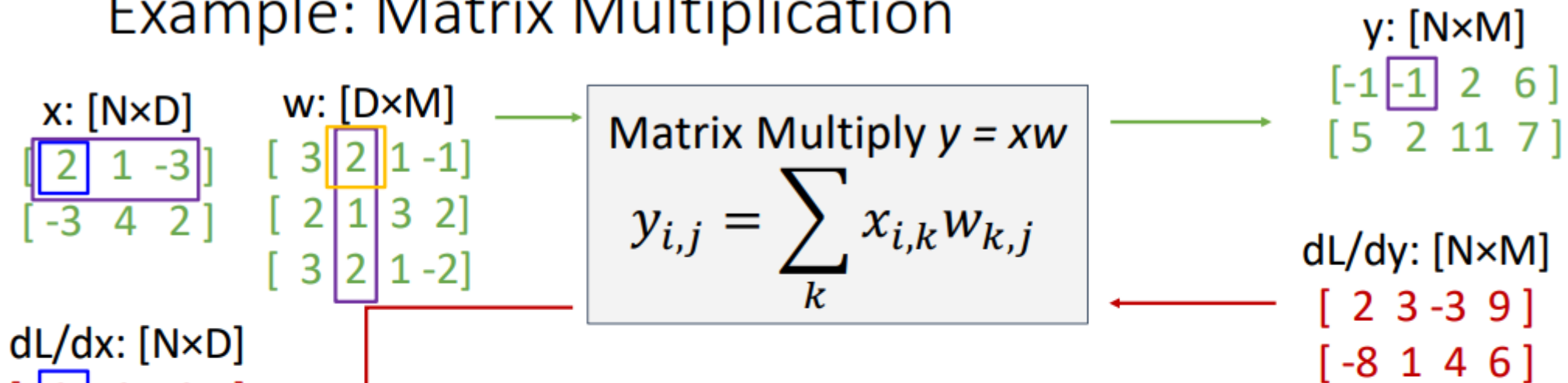
Backpropagation

Example: Matrix Multiplication



Backpropagation

Example: Matrix Multiplication



Local Gradient Slice:

$$dy/dx_{1,1}$$

$$dy_{1,2}/dx_{1,1} \begin{bmatrix} 3 & 2 & ? & ? \\ ? & ? & ? & ? \end{bmatrix}$$

$$dL/dx_{1,1}$$

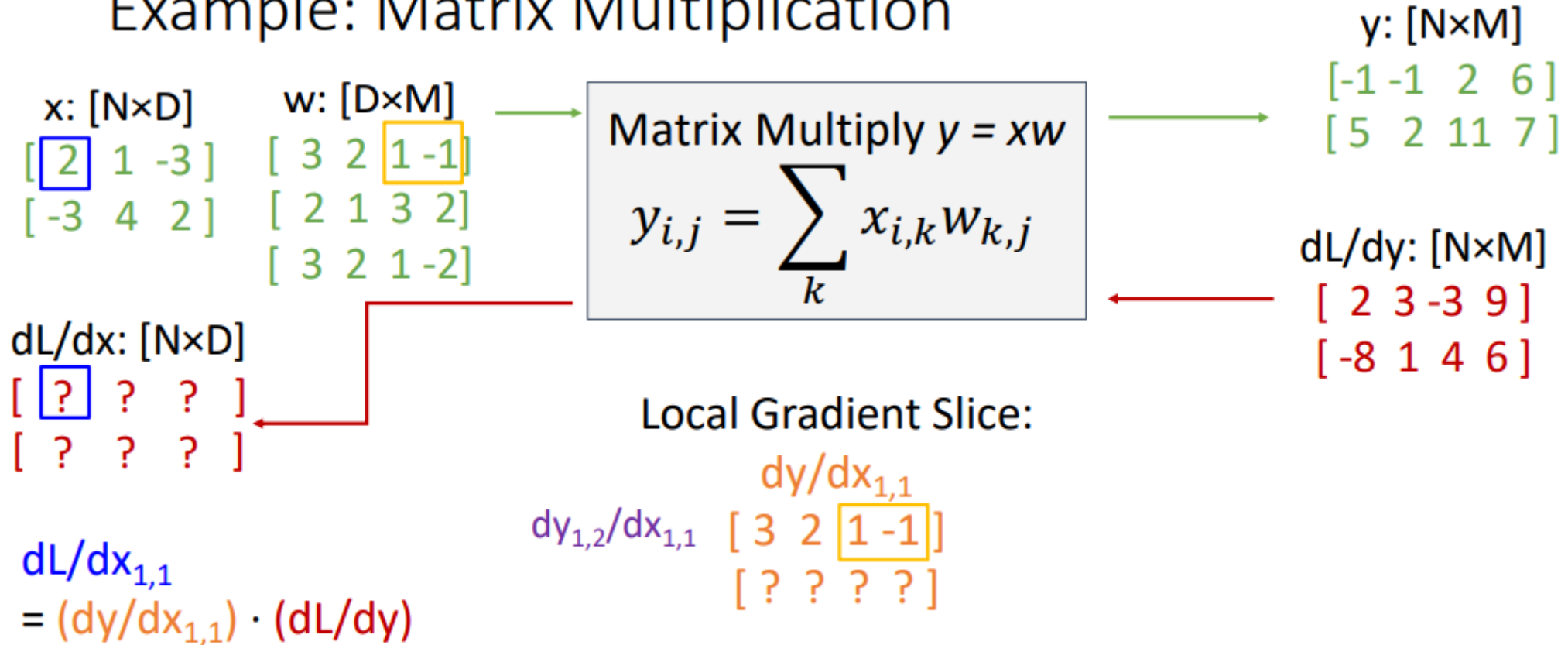
$$= (dy/dx_{1,1}) \cdot (dL/dy)$$

$$y_{1,2} = x_{1,1} w_{1,2} + x_{1,2} w_{2,2} + x_{1,3} w_{3,2}$$

$$\Rightarrow dy_{1,2}/dx_{1,1} = w_{1,2}$$

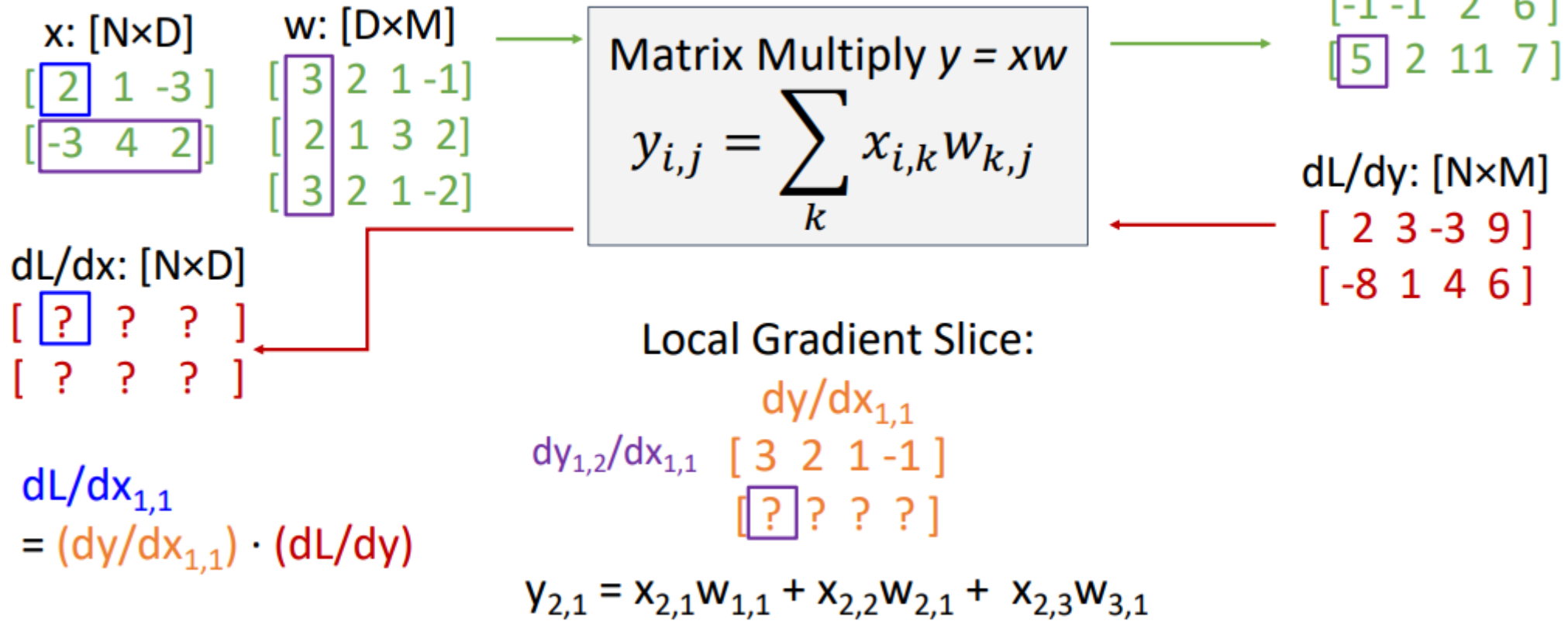
Backpropagation

Example: Matrix Multiplication



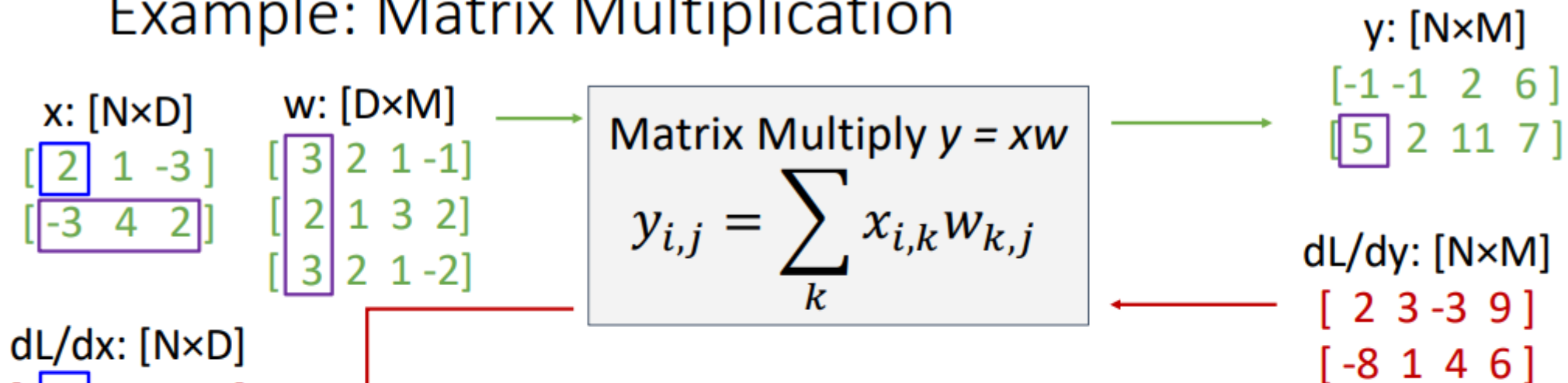
Backpropagation

Example: Matrix Multiplication



Backpropagation

Example: Matrix Multiplication



Local Gradient Slice:

$$dy/dx_{1,1}$$

$$dy_{1,2}/dx_{1,1} \begin{bmatrix} 3 & 2 & 1 & -1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & ? & ? & ? \end{bmatrix}$$

$$dL/dx_{1,1}$$

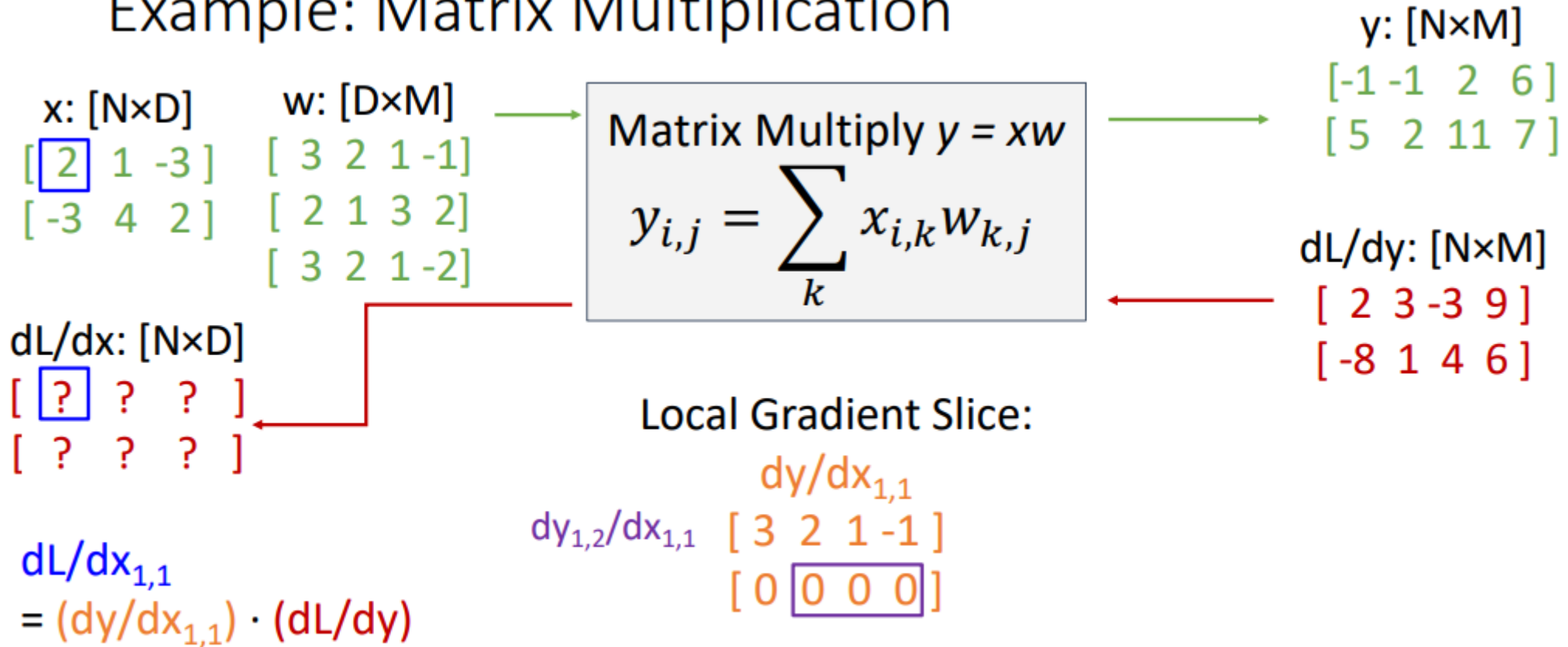
$$= (dy/dx_{1,1}) \cdot (dL/dy)$$

$$y_{2,1} = x_{2,1}w_{1,1} + x_{2,2}w_{2,1} + x_{2,3}w_{3,1}$$

$$\Rightarrow dy_{2,1}/dx_{1,1} = 0$$

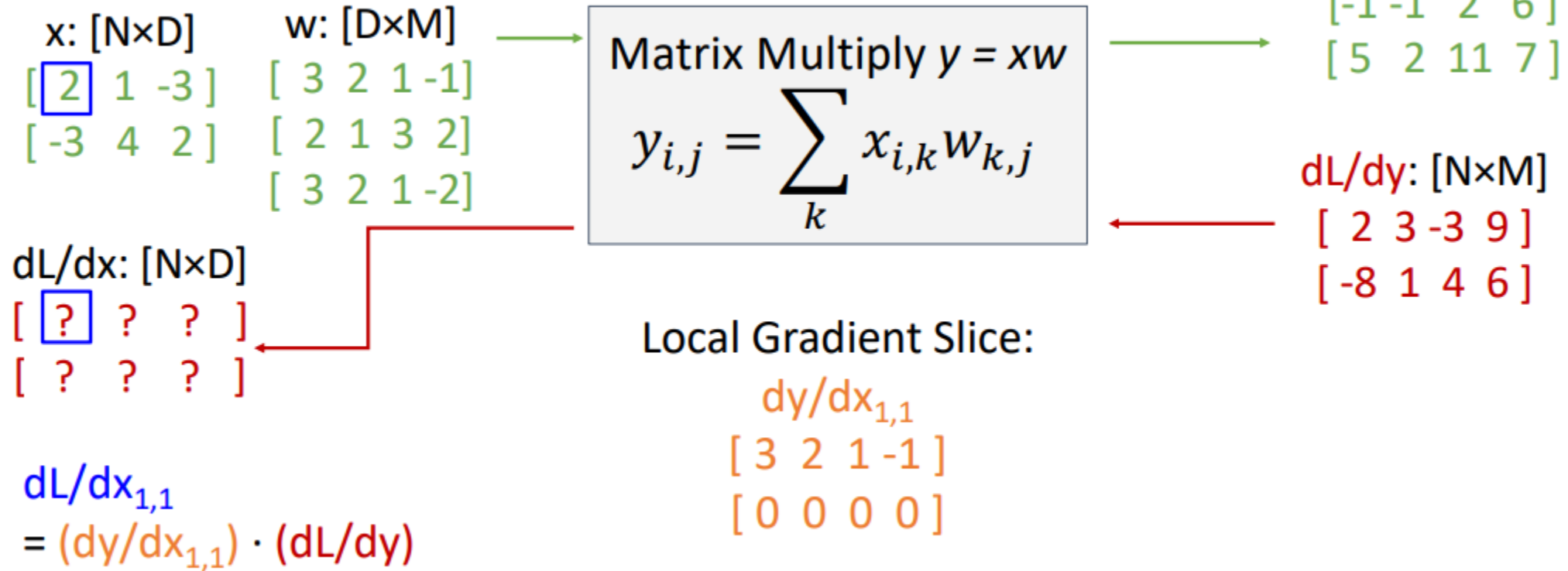
Backpropagation

Example: Matrix Multiplication



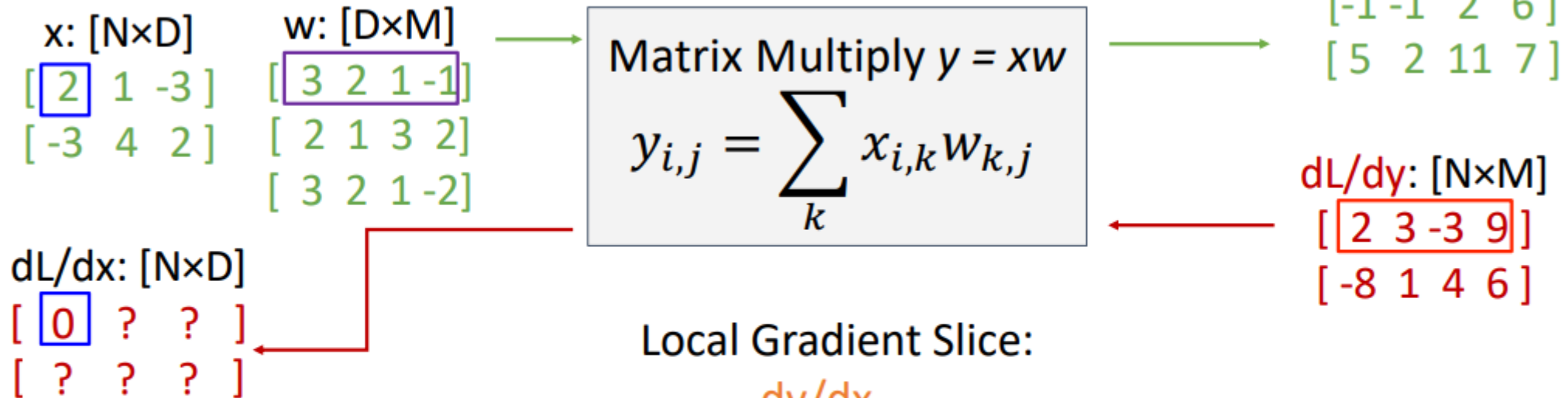
Backpropagation

Example: Matrix Multiplication



Backpropagation

Example: Matrix Multiplication



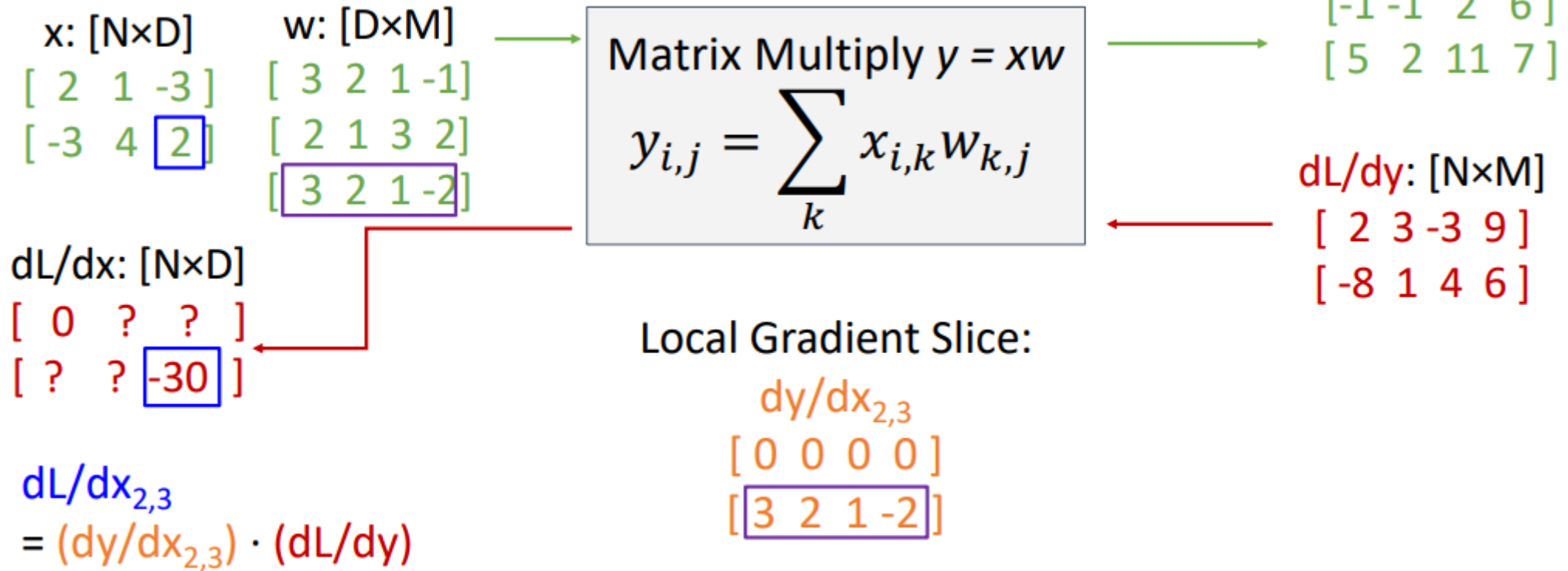
Local Gradient Slice:

$$\frac{dy}{dx_{1,1}}$$
$$\begin{bmatrix} 3 & 2 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{aligned} \frac{dL}{dx_{1,1}} &= \left(\frac{dy}{dx_{1,1}}\right) \cdot \left(\frac{dL}{dy}\right) \\ &= (w_{1,:}) \cdot (dL/dy_{1,:}) \\ &= 3*2 + 2*3 + 1*(-3) + (-1)*9 = 0 \end{aligned}$$

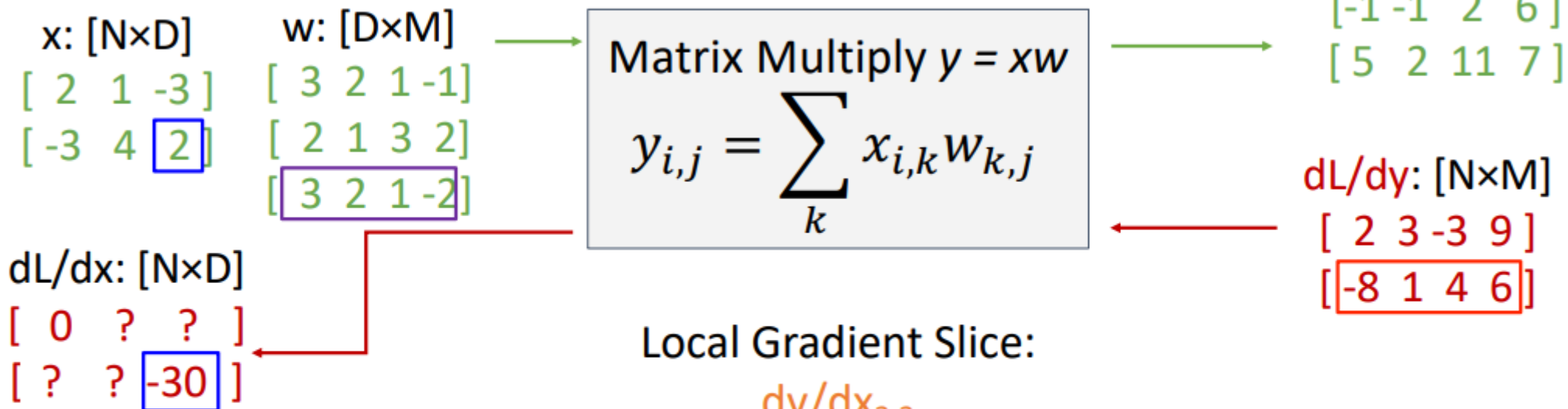
Backpropagation

Example: Matrix Multiplication



Backpropagation

Example: Matrix Multiplication



Local Gradient Slice:

$$\frac{dy}{dx_{2,3}}$$

0	0	0	0
3	2	1	-2

$$\frac{dL}{dx_{2,3}}$$

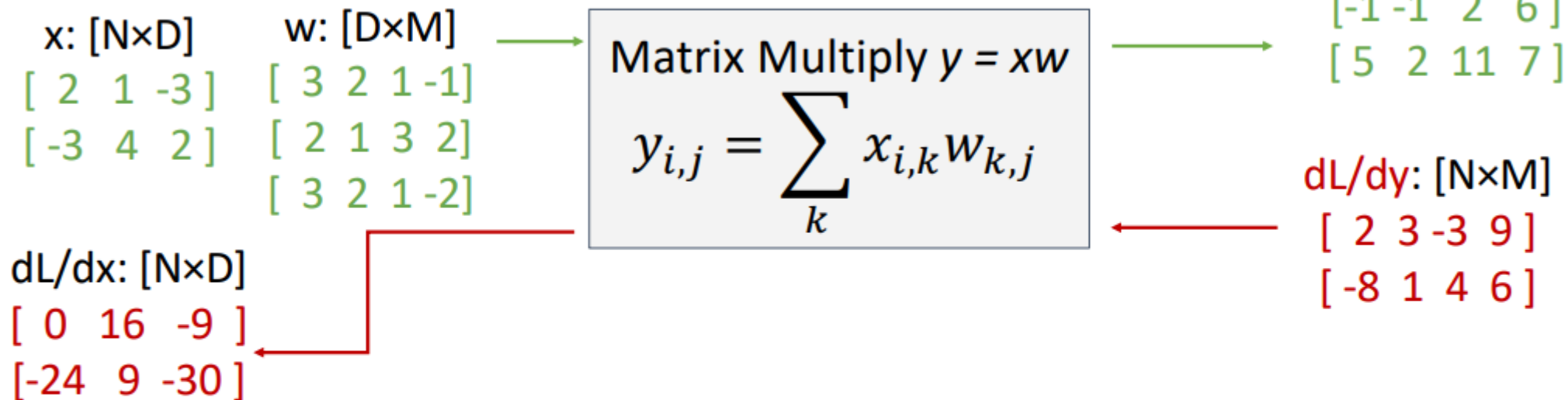
$$= \left(\frac{dy}{dx_{2,3}}\right) \cdot \left(\frac{dL}{dy}\right)$$

$$= (w_{3,:}) \cdot \left(\frac{dL}{dy_{2,:}}\right)$$

$$= 3*(-8) + 2*1 + 1*4 + (-2)*6 = -30$$

Backpropagation

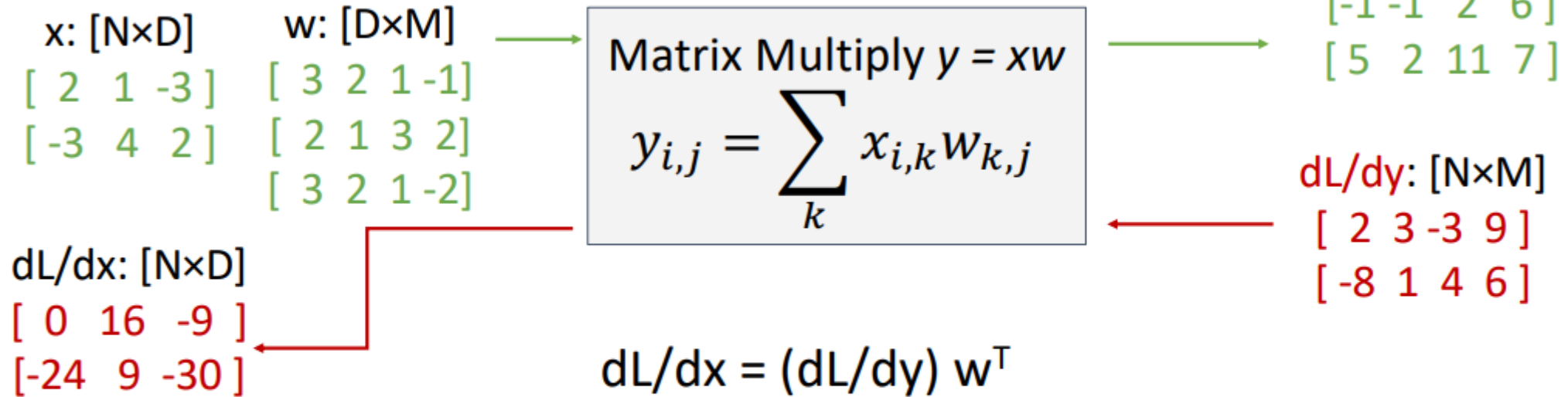
Example: Matrix Multiplication



$$\begin{aligned} dL/dx_{i,j} &= (dy/dx_{i,j}) \cdot (dL/dy) \\ &= (w_{j,:}) \cdot (dL/dy_{i,:}) \end{aligned}$$

Backpropagation

Example: Matrix Multiplication



$$dL/dx = (dL/dy) w^T$$

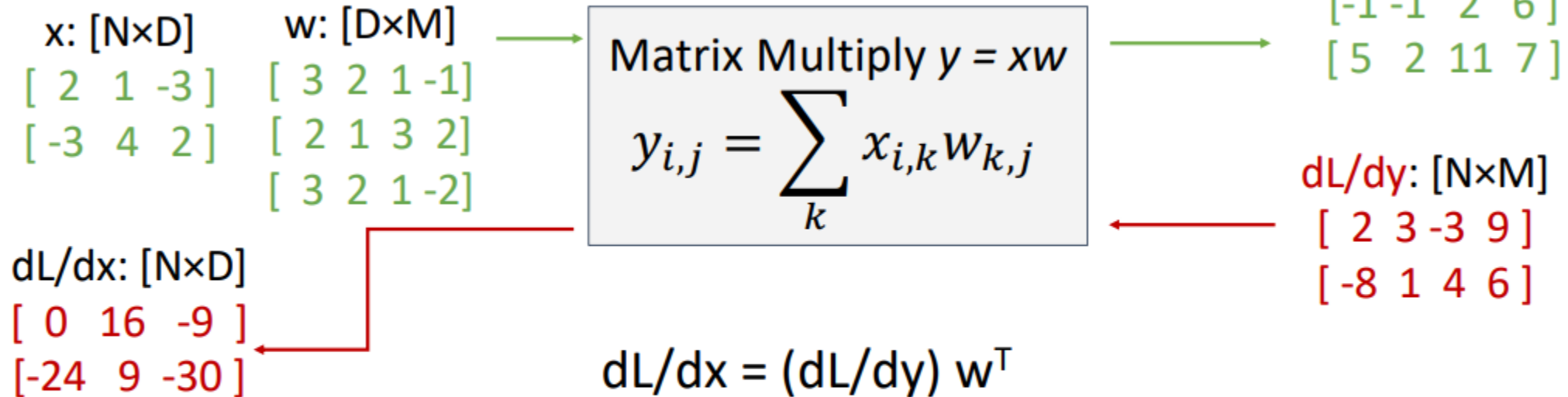
$[N \times D] \quad [N \times M] \quad [M \times D]$

$$\begin{aligned} dL/dx_{i,j} &= (dy/dx_{i,j}) \cdot (dL/dy) \\ &= (w_{j,:}) \cdot (dL/dy_{i,:}) \end{aligned}$$

Easy way to remember:
It's the only way the
shapes work out!

Backpropagation

Example: Matrix Multiplication



$$dL/dx = (dL/dy) w^T$$

$[N \times D] \quad [N \times M] \quad [M \times D]$

$$dL/dw = x^T (dL/dy)$$

$[D \times M] \quad [D \times N] \quad [N \times M]$

Easy way to remember:
It's the only way the
shapes work out!

► Choosing hyperparameters

Choosing Hyperparameters: Grid Search

Choose several values for each hyperparameter
(Often space choices log-linearly)

Example:

Weight decay: $[1 \times 10^{-4}, 1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}]$

Learning rate: $[1 \times 10^{-4}, 1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}]$

Evaluate all possible choices on this
hyperparameter grid

► Choosing hyperparameters

Choosing Hyperparameters: Random Search

Choose several values for each hyperparameter
(Often space choices log-linearly)

Example:

Weight decay: log-uniform on $[1 \times 10^{-4}, 1 \times 10^{-1}]$

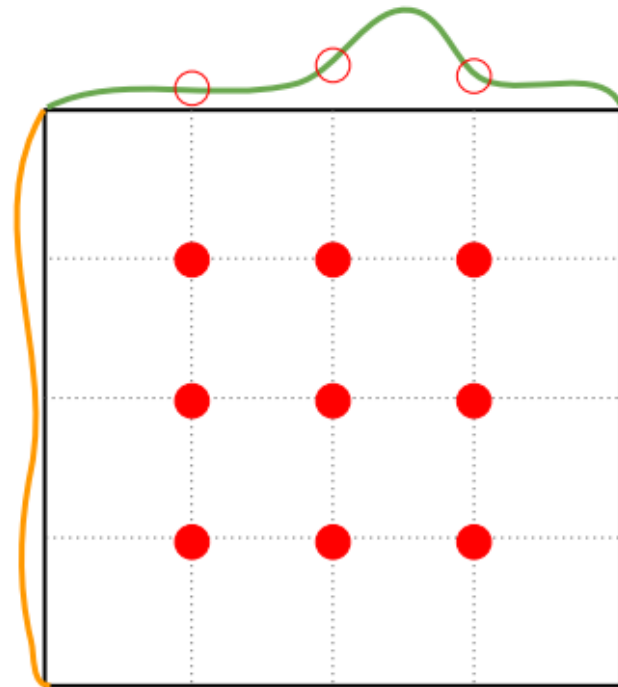
Learning rate: log-uniform on $[1 \times 10^{-4}, 1 \times 10^{-1}]$

Run many different trials

▶ Choosing hyperparameters

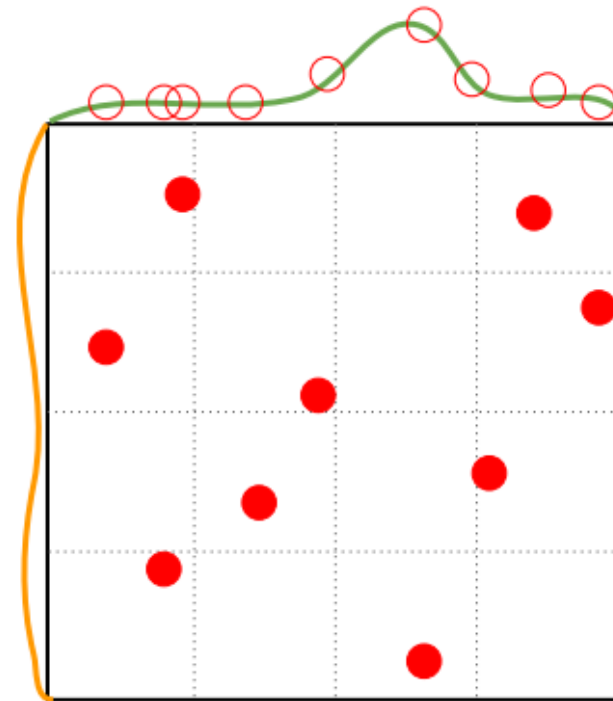
Hyperparameters: Random vs Grid Search

Grid Layout



Important
Parameter

Random Layout



Important
Parameter

► Choosing hyperparameters

Choosing Hyperparameters

Step 1: Check initial loss

Turn off weight decay, sanity check loss at initialization
e.g. $\log(C)$ for softmax with C classes

► Choosing hyperparameters

Choosing Hyperparameters

Step 1: Check initial loss

Step 2: Overfit a small sample

Try to train to 100% training accuracy on a small sample of training data (~5-10 minibatches); fiddle with architecture, learning rate, weight initialization. Turn off regularization.

Loss not going down? LR too low, bad initialization

Loss explodes to Inf or NaN? LR too high, bad initialization

► Choosing hyperparameters

Choosing Hyperparameters

Step 1: Check initial loss

Step 2: Overfit a small sample

Step 3: Find LR that makes loss go down

Use the architecture from the previous step, use all training data, turn on small weight decay, find a learning rate that makes the loss drop significantly within ~ 100 iterations

Good learning rates to try: $1e-1$, $1e-2$, $1e-3$, $1e-4$

► Choosing hyperparameters

Choosing Hyperparameters

Step 1: Check initial loss

Step 2: Overfit a small sample

Step 3: Find LR that makes loss go down

Step 4: Coarse grid, train for ~1-5 epochs

Choose a few values of learning rate and weight decay around what worked from Step 3, train a few models for ~1-5 epochs.

Good weight decay to try: $1e-4$, $1e-5$, 0

► Choosing hyperparameters

Choosing Hyperparameters

Step 1: Check initial loss

Step 2: Overfit a small sample

Step 3: Find LR that makes loss go down

Step 4: Coarse grid, train for ~1-5 epochs

Step 5: Refine grid, train longer

Pick best models from Step 4, train them for longer (~10-20 epochs) without learning rate decay

► Choosing hyperparameters

Choosing Hyperparameters

Step 1: Check initial loss

Step 2: Overfit a small sample

Step 3: Find LR that makes loss go down

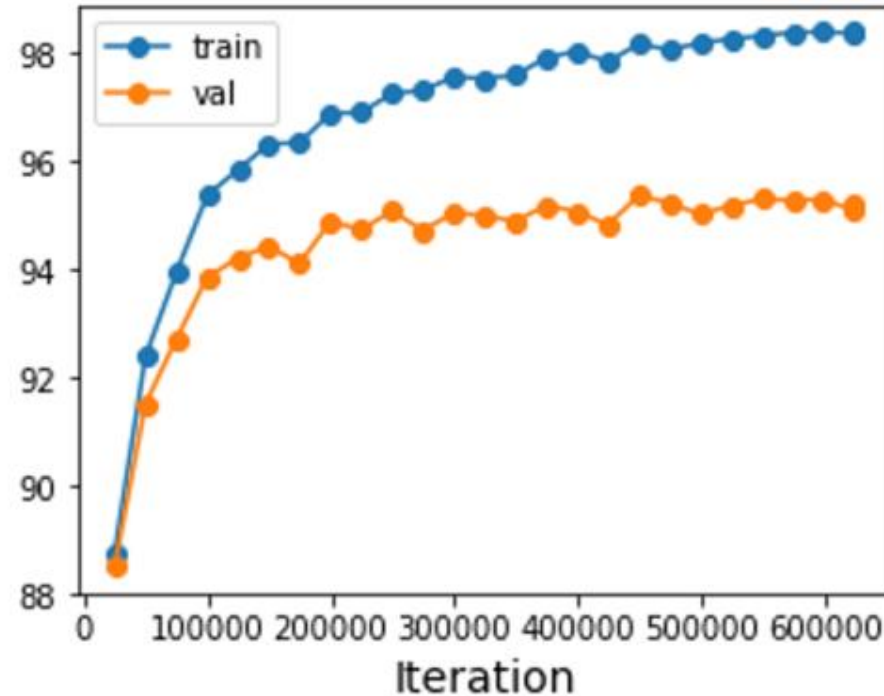
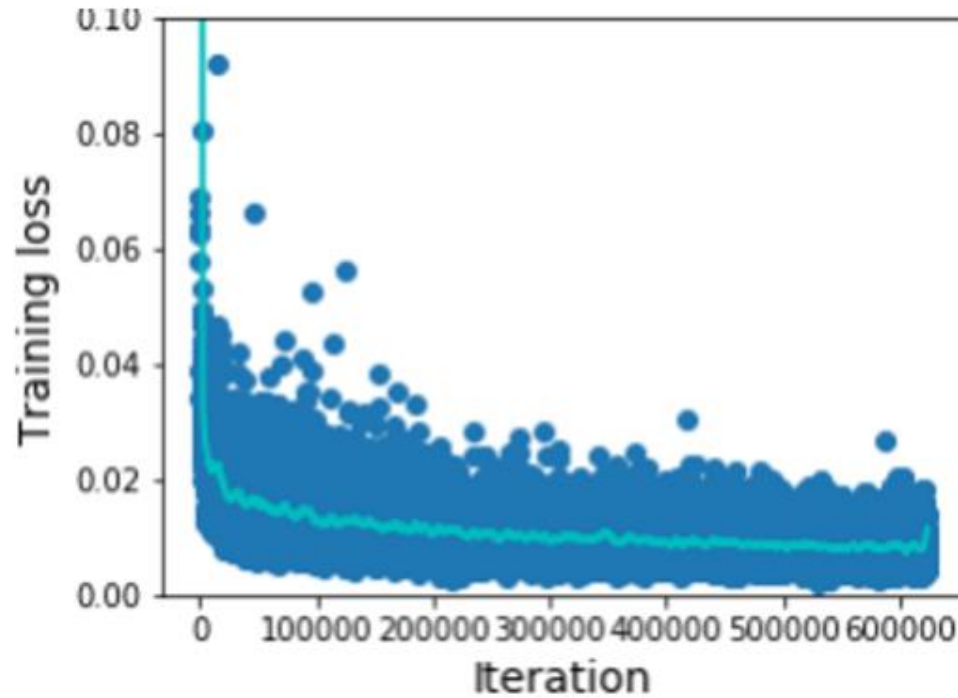
Step 4: Coarse grid, train for ~1-5 epochs

Step 5: Refine grid, train longer

Step 6: Look at learning curves

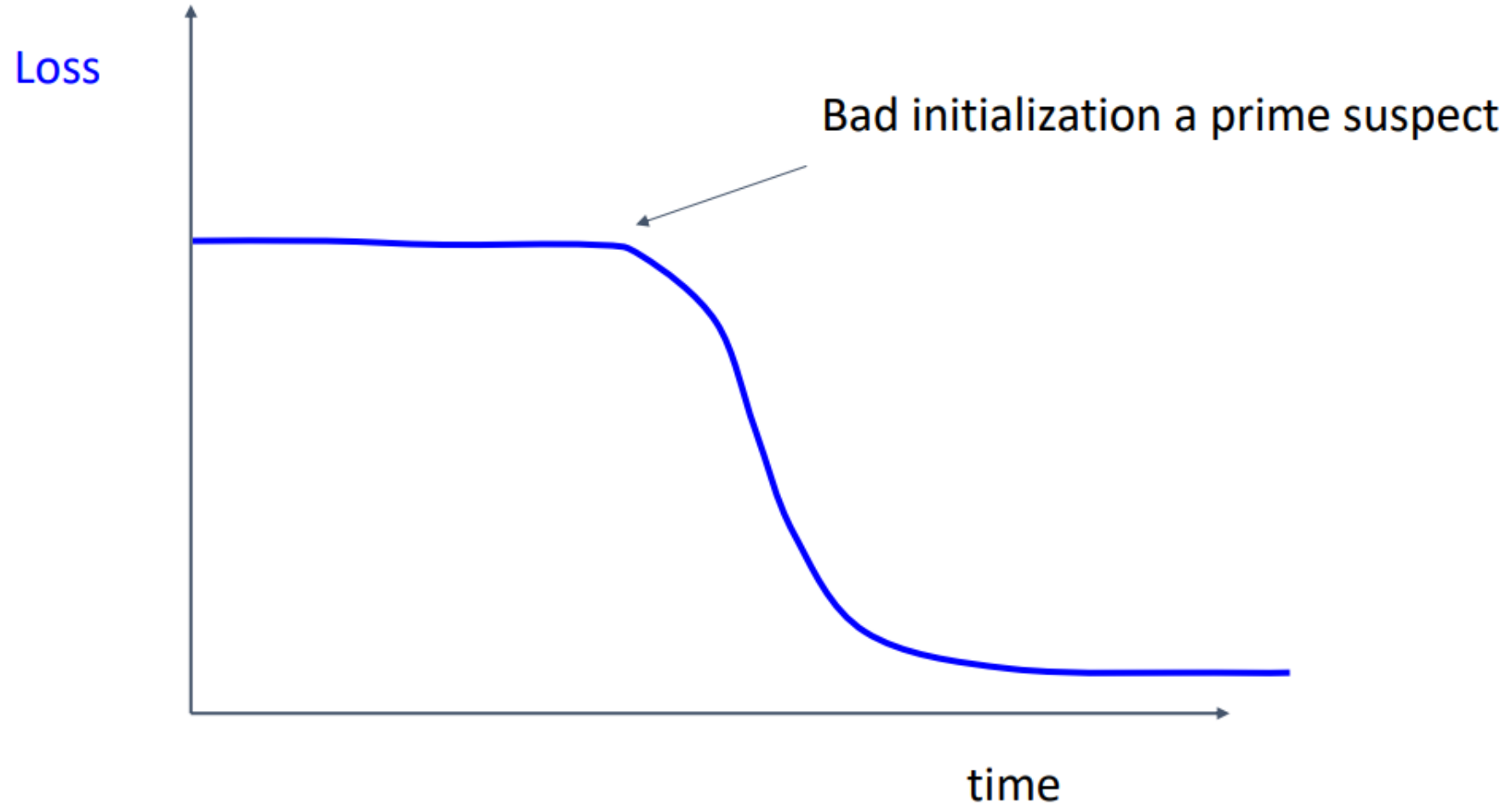
► Choosing hyperparameters

Look at Learning Curves!

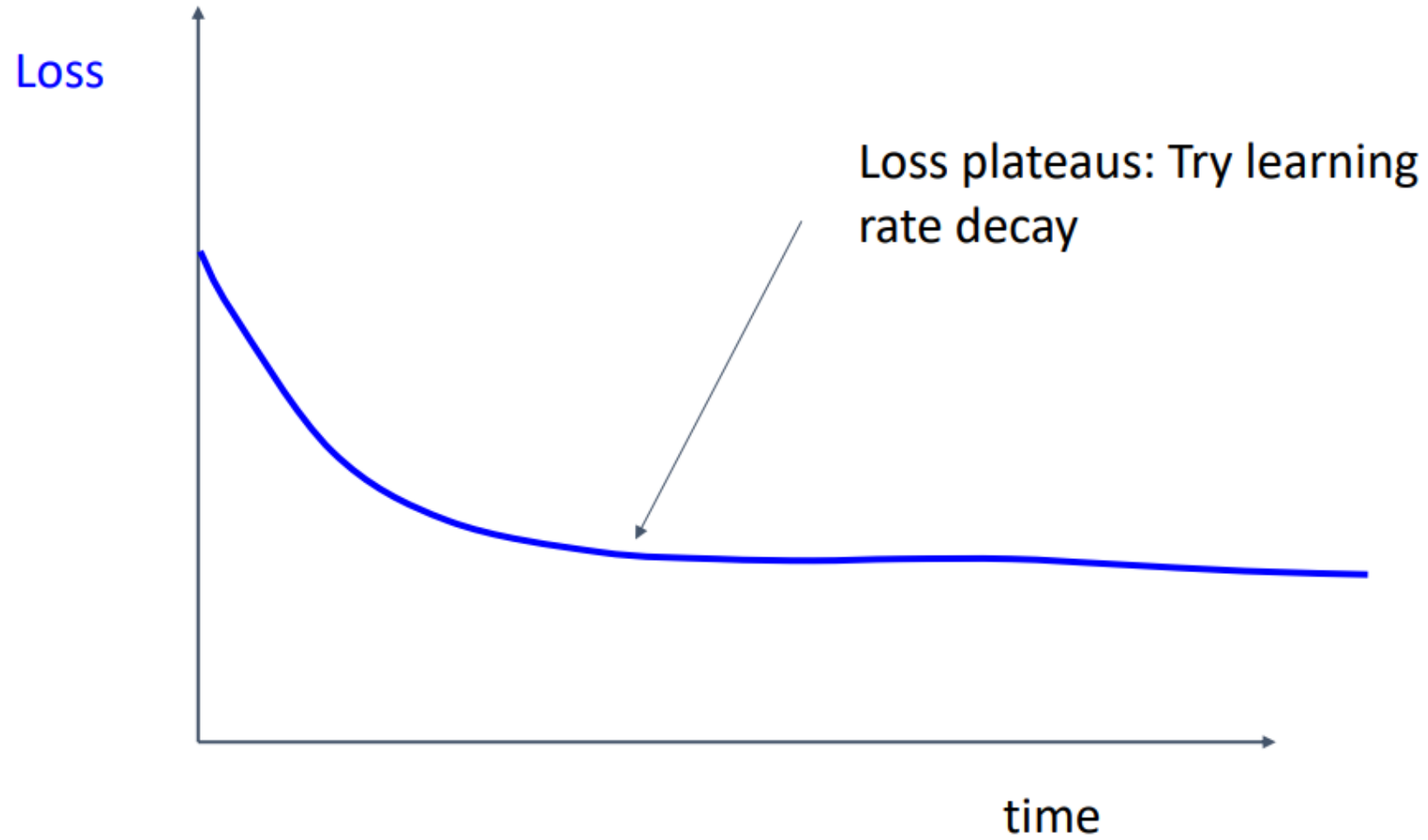


Losses may be noisy, use a scatter plot and also plot moving average to see trends better

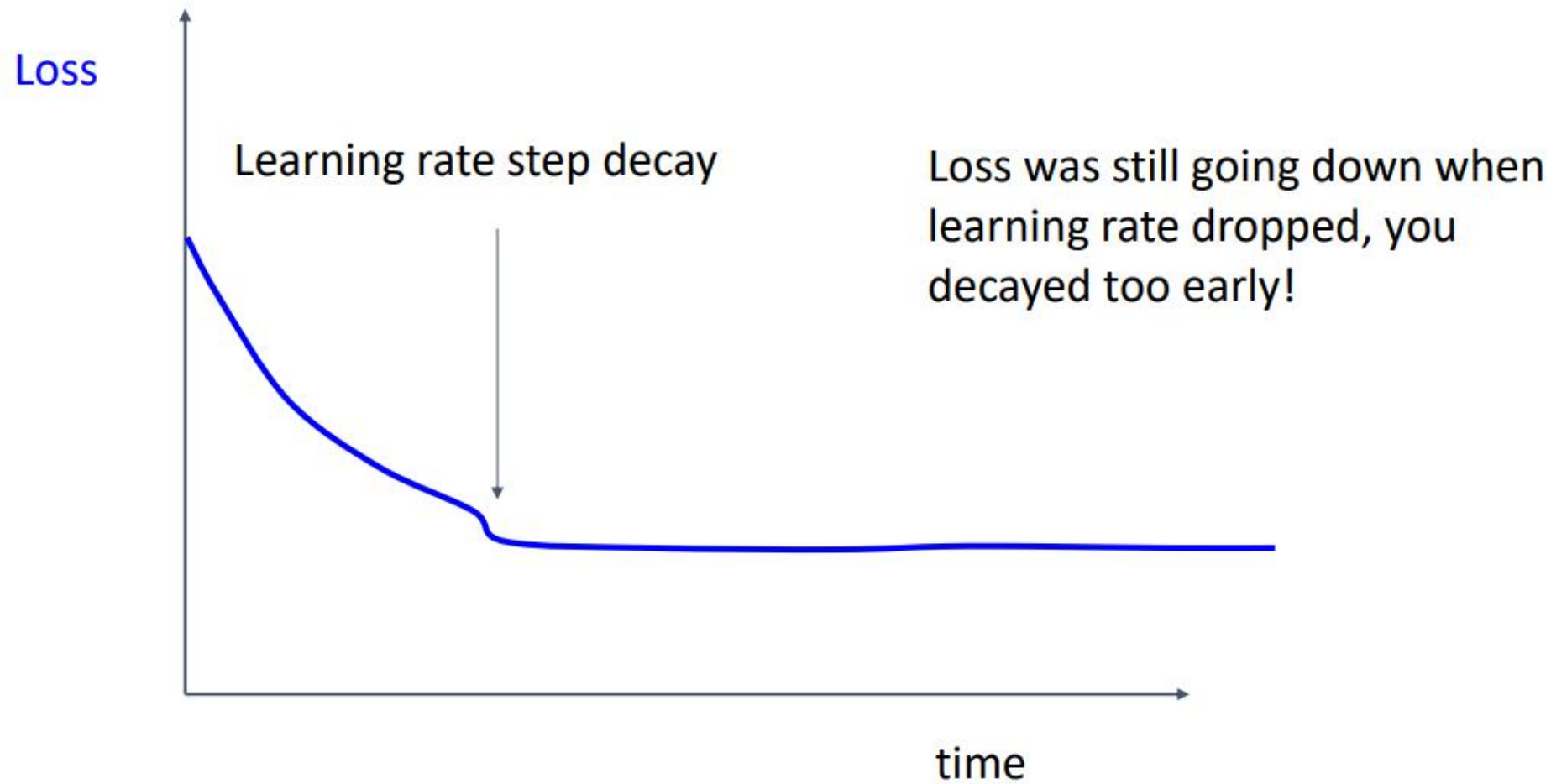
▶ Choosing hyperparameters



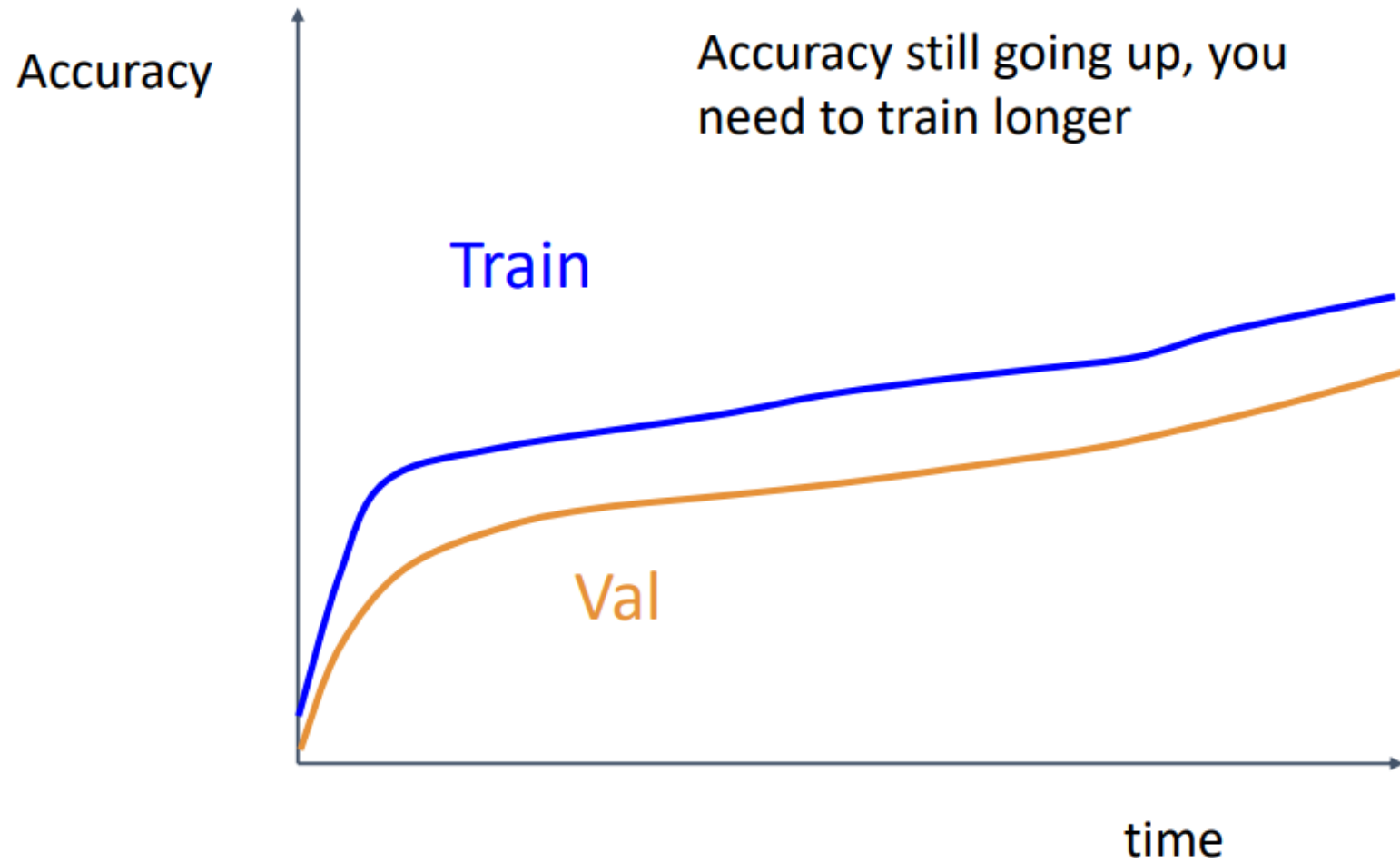
► Choosing hyperparameters



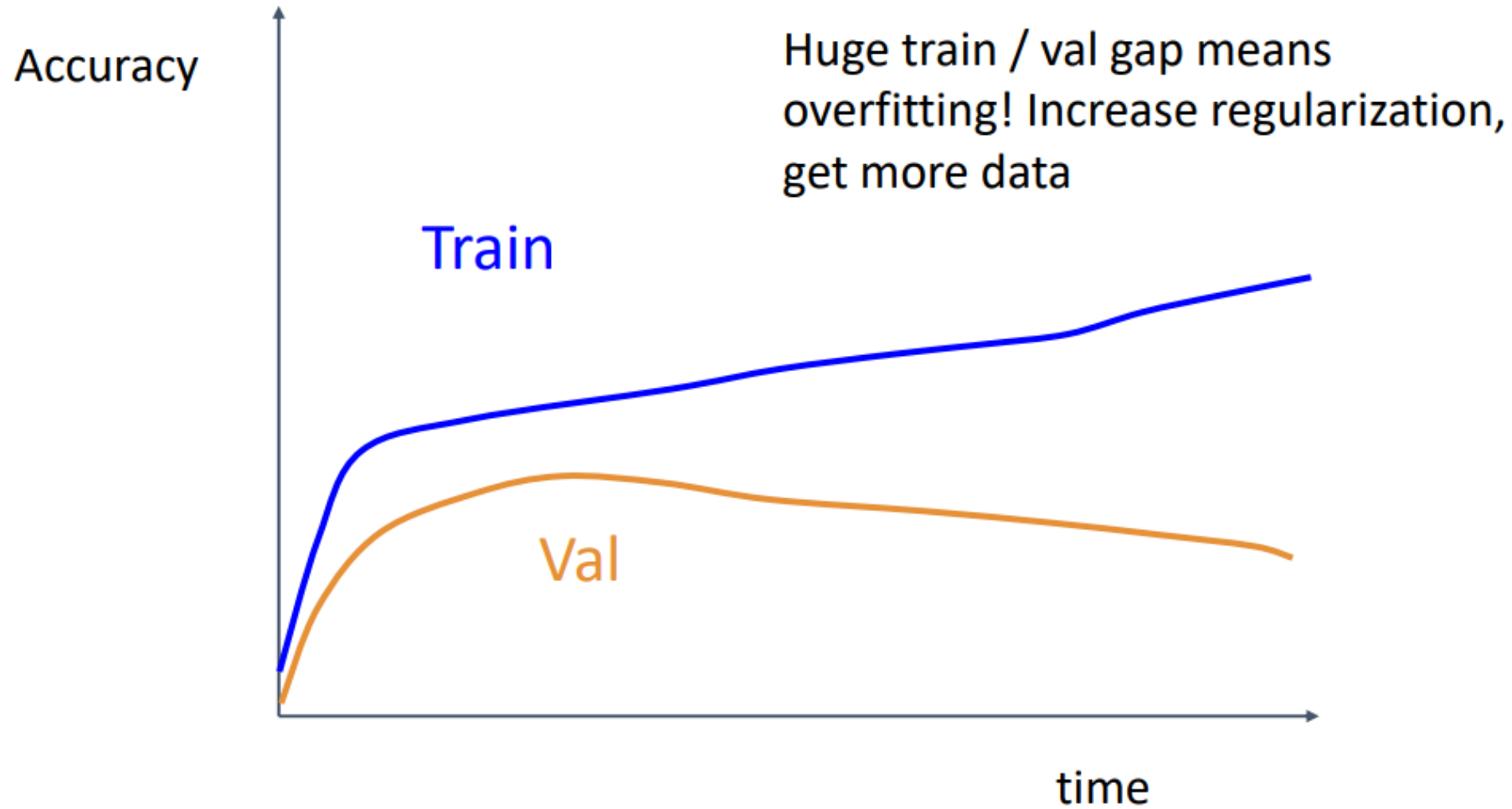
▶ Choosing hyperparameters



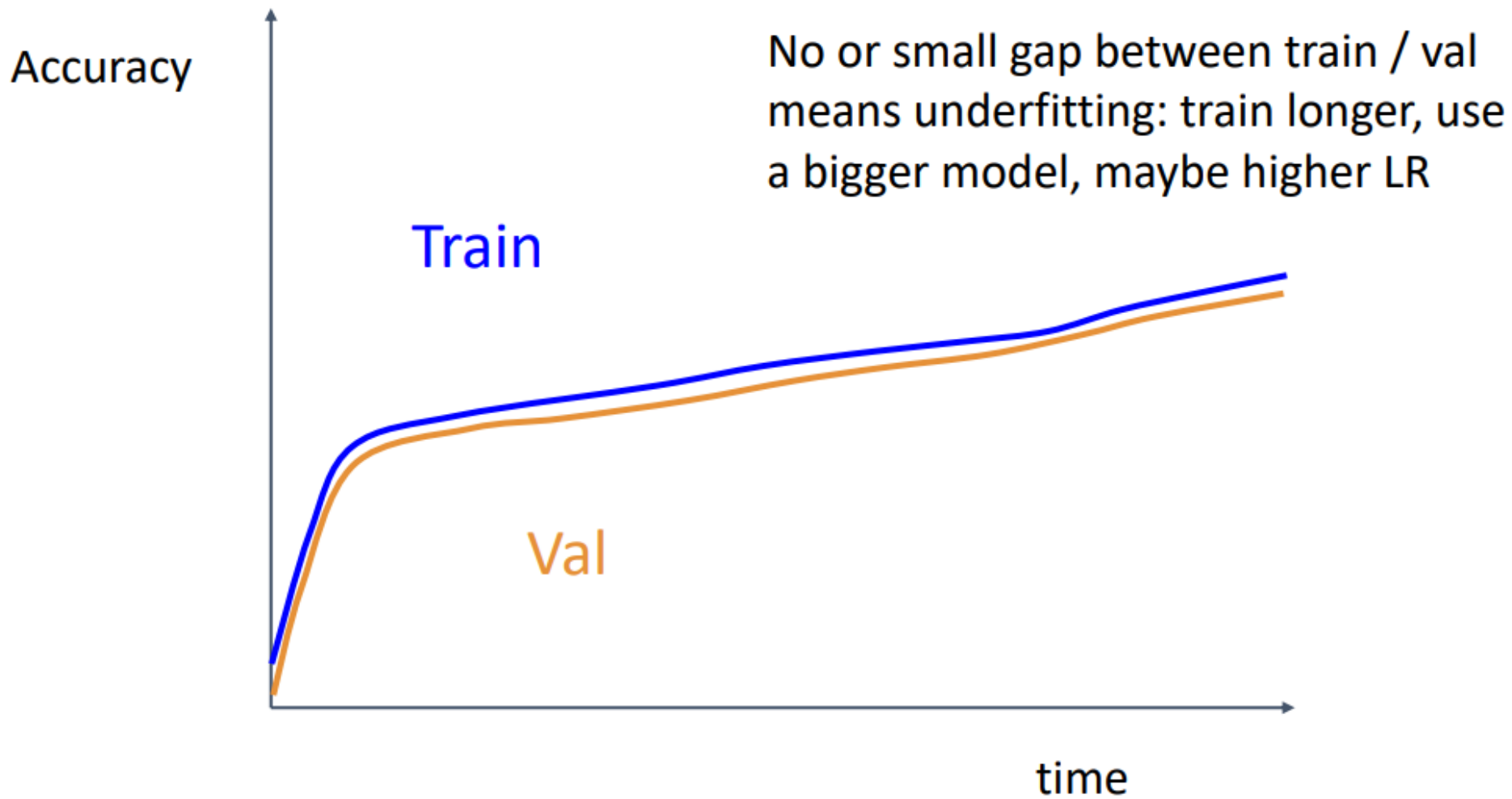
▶ Choosing hyperparameters



▶ Choosing hyperparameters



▶ Choosing hyperparameters



► Choosing hyperparameters

Choosing Hyperparameters

Step 1: Check initial loss

Step 2: Overfit a small sample

Step 3: Find LR that makes loss go down

Step 4: Coarse grid, train for ~1-5 epochs

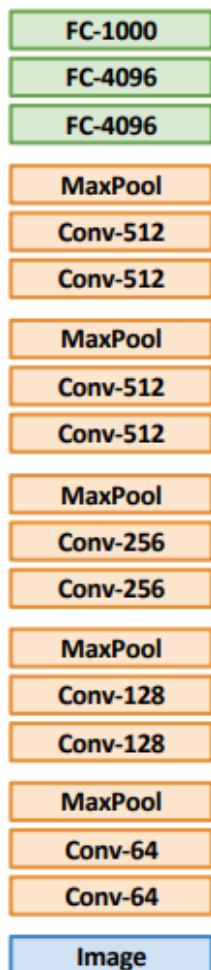
Step 5: Refine grid, train longer

Step 6: Look at loss curves

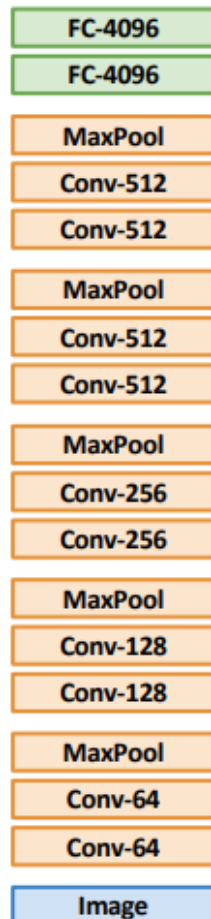
Step 7: GOTO step 5

Transfer learning

1. Train on ImageNet



2. Use CNN as a feature extractor

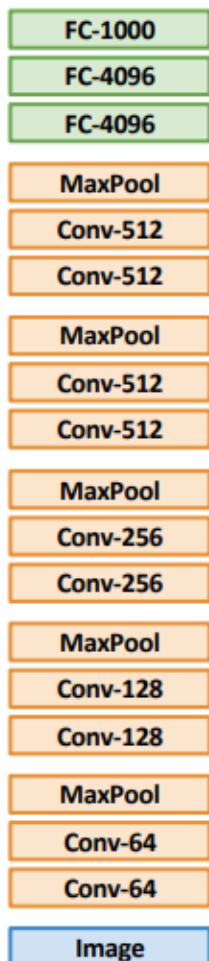


Remove last layer

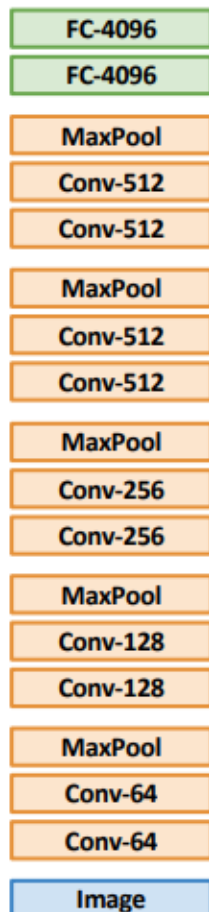
Freeze these

Transfer learning

1. Train on ImageNet



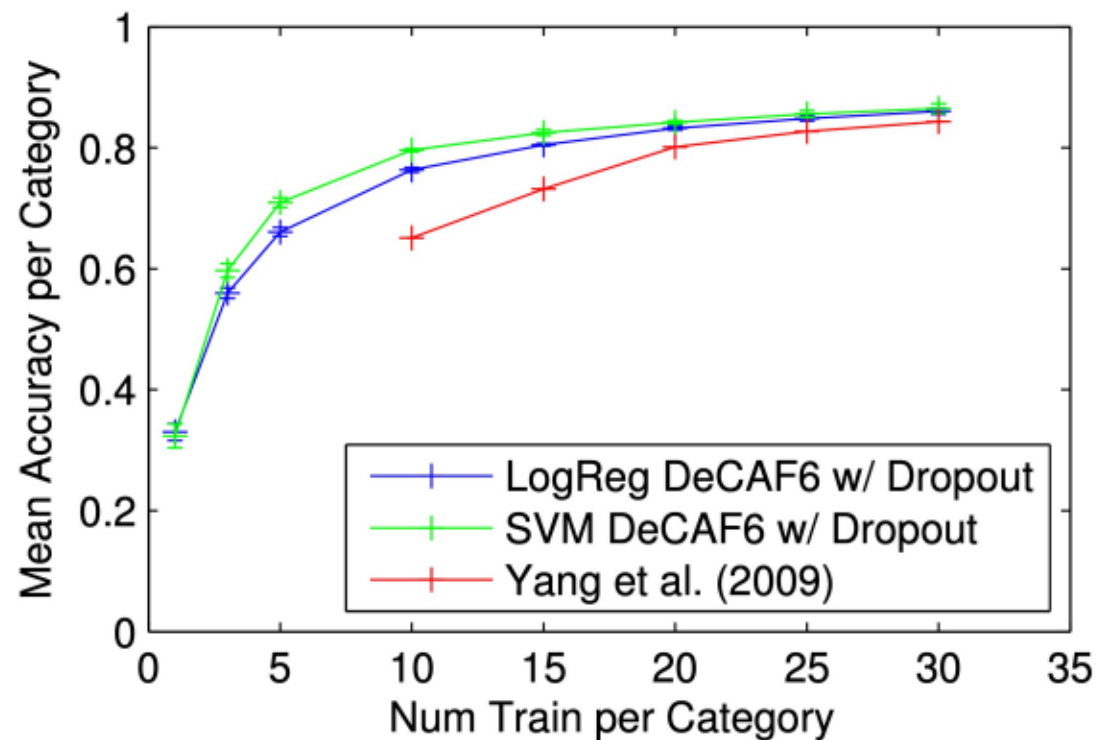
2. Use CNN as a feature extractor



Remove last layer

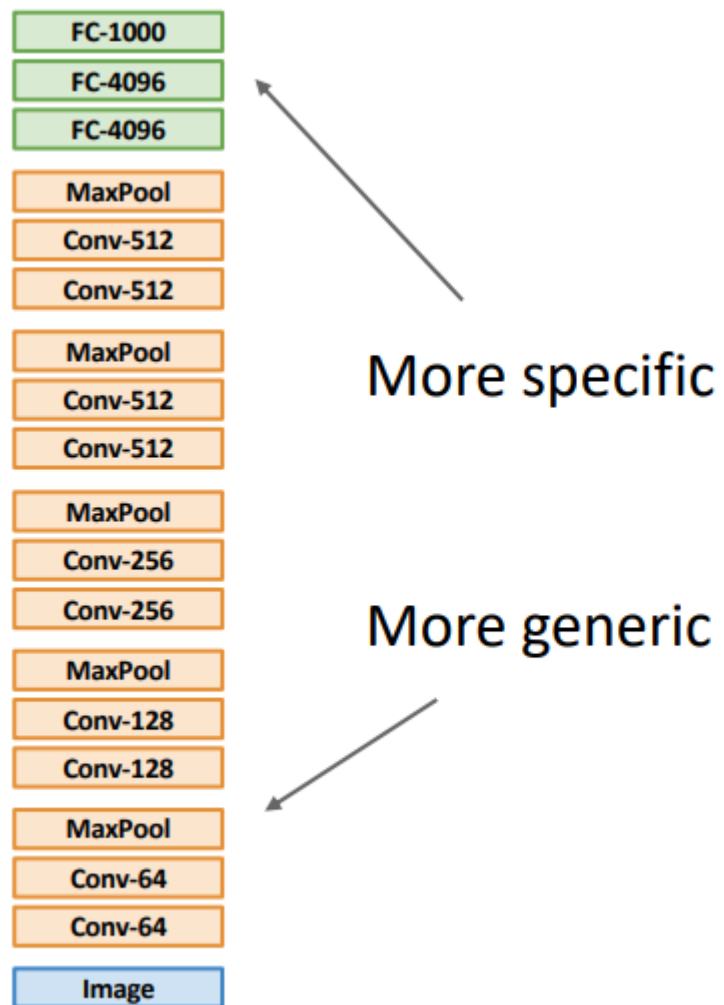
Freeze these

Classification on Caltech-101



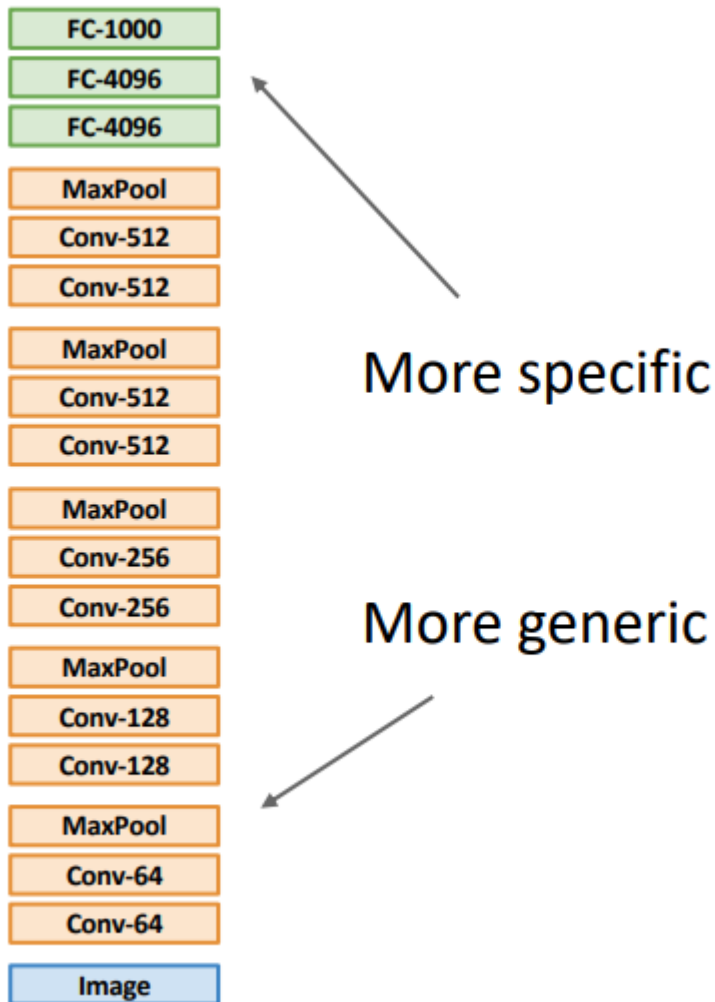
Donahue et al, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ICML 2014

Transfer learning



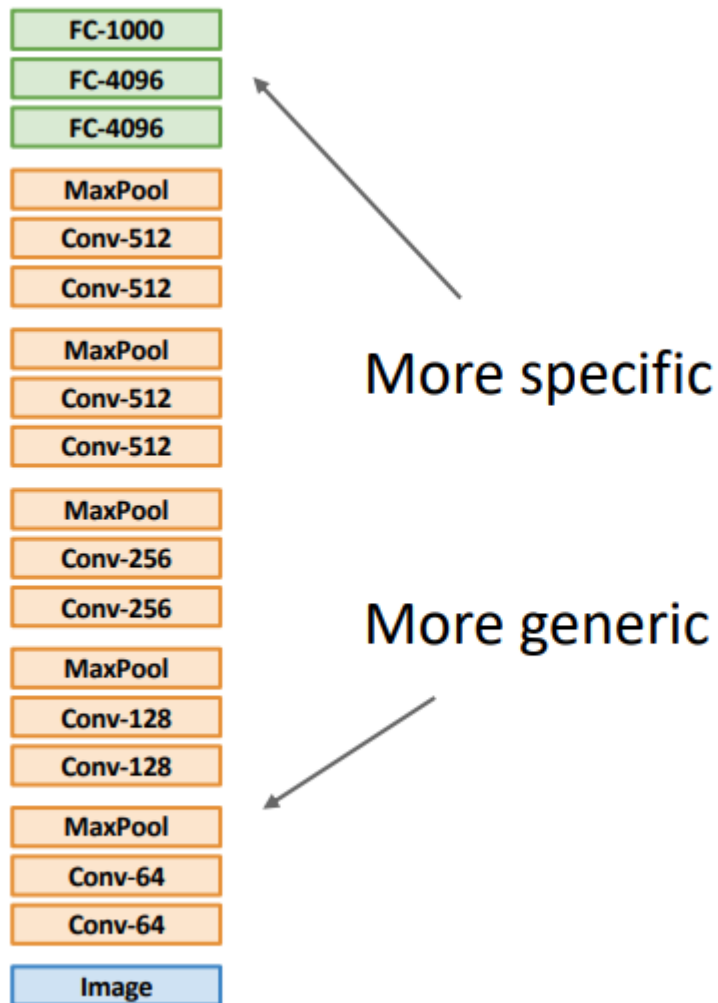
	Dataset similar to ImageNet	Dataset very different from ImageNet
very little data (10s to 100s)	?	?
quite a lot of data (100s to 1000s)	?	?

Transfer learning



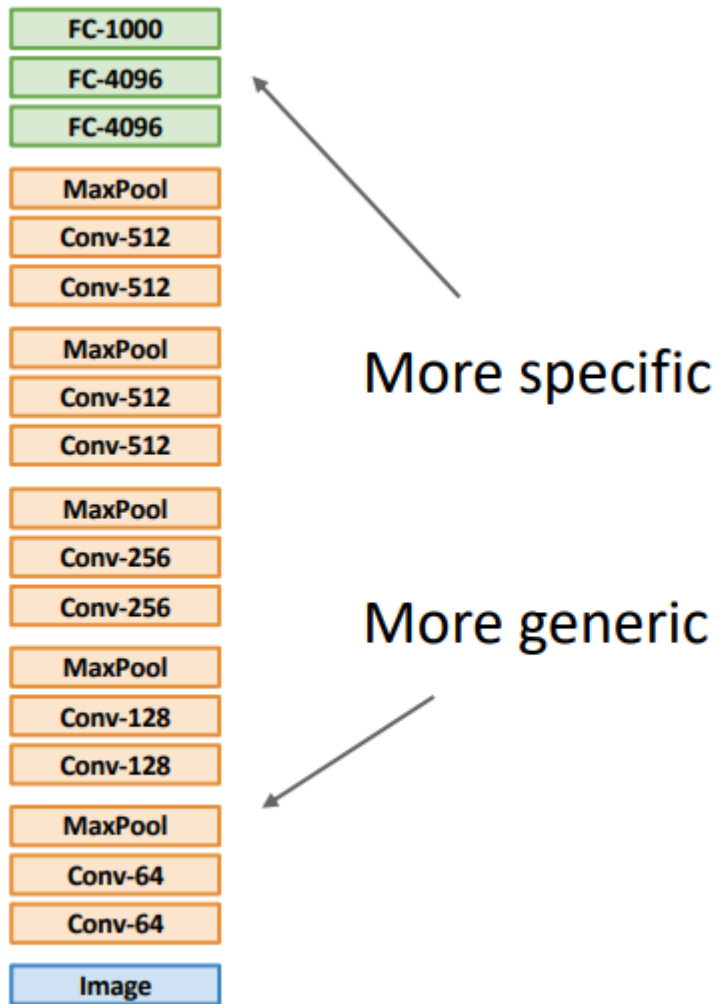
	Dataset similar to ImageNet	Dataset very different from ImageNet
very little data (10s to 100s)	Use Linear Classifier on top layer	?
quite a lot of data (100s to 1000s)	Finetune a few layers	?

Transfer learning



	Dataset similar to ImageNet	Dataset very different from ImageNet
very little data (10s to 100s)	Use Linear Classifier on top layer	?
quite a lot of data (100s to 1000s)	Finetune a few layers	Finetune a larger number of layers

Transfer learning



	Dataset similar to ImageNet	Dataset very different from ImageNet
very little data (10s to 100s)	Use Linear Classifier on top layer	You're in trouble... Try linear classifier from different stages
quite a lot of data (100s to 1000s)	Finetune a few layers	Finetune a larger number of layers

▶ Reference and Readings

Reference

- <http://cs231n.stanford.edu>

Readings

- “Computing Receptive Fields of Convolutional Neural Networks” In Distill 2019.
- <https://naokishibuya.medium.com/up-sampling-with-transposed-convolution-9ae4f2df52d0>

Thank You!!