

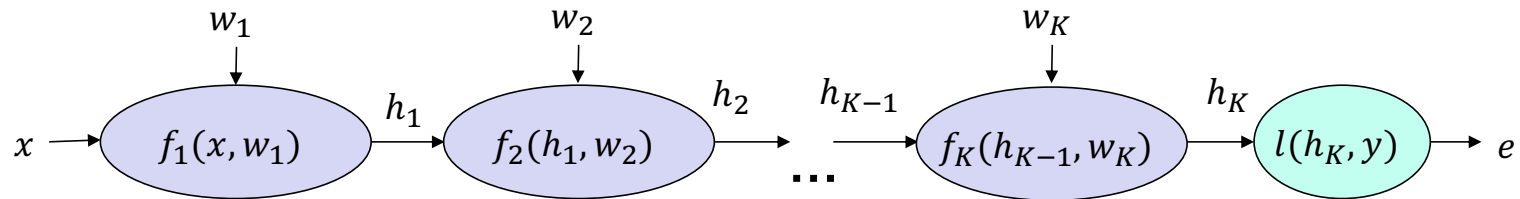
## Last time: Backpropagation

---

- Computation graphs
- Using the chain rule
- General backprop algorithm
- Toy examples of backward pass
- **Matrix-vector calculations: ReLU, linear layer**

# Review: Computation graph

---

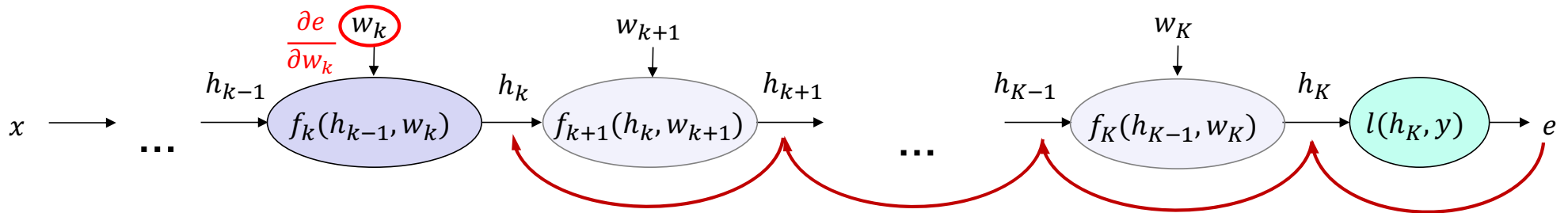


- Wanted: gradient of the error w.r.t. the parameters of each layer,  $\frac{\partial e}{\partial w_k}$
- Then we can train by applying SGD updates

$$w_k \leftarrow w_k - \eta \frac{\partial e}{\partial w_k}$$

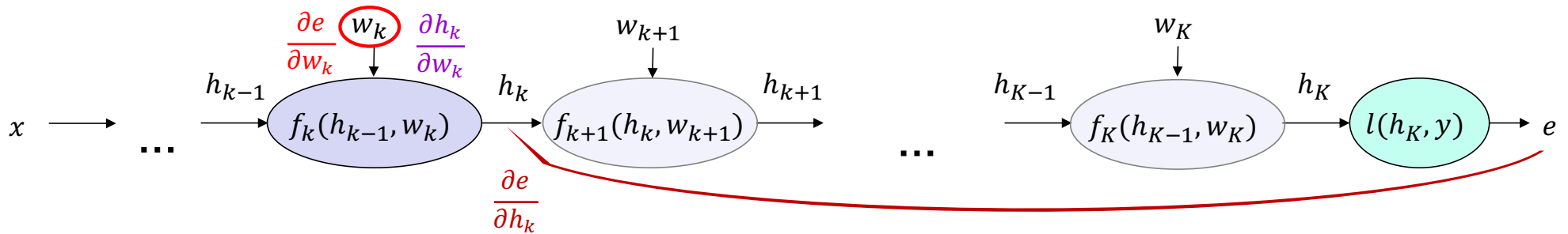
# The chain rule

---



$$\frac{\partial e}{\partial w_k} = \frac{\partial e}{\partial h_K} \frac{\partial h_K}{\partial h_{K-1}} \dots \frac{\partial h_{k+1}}{\partial h_k}$$

# The chain rule



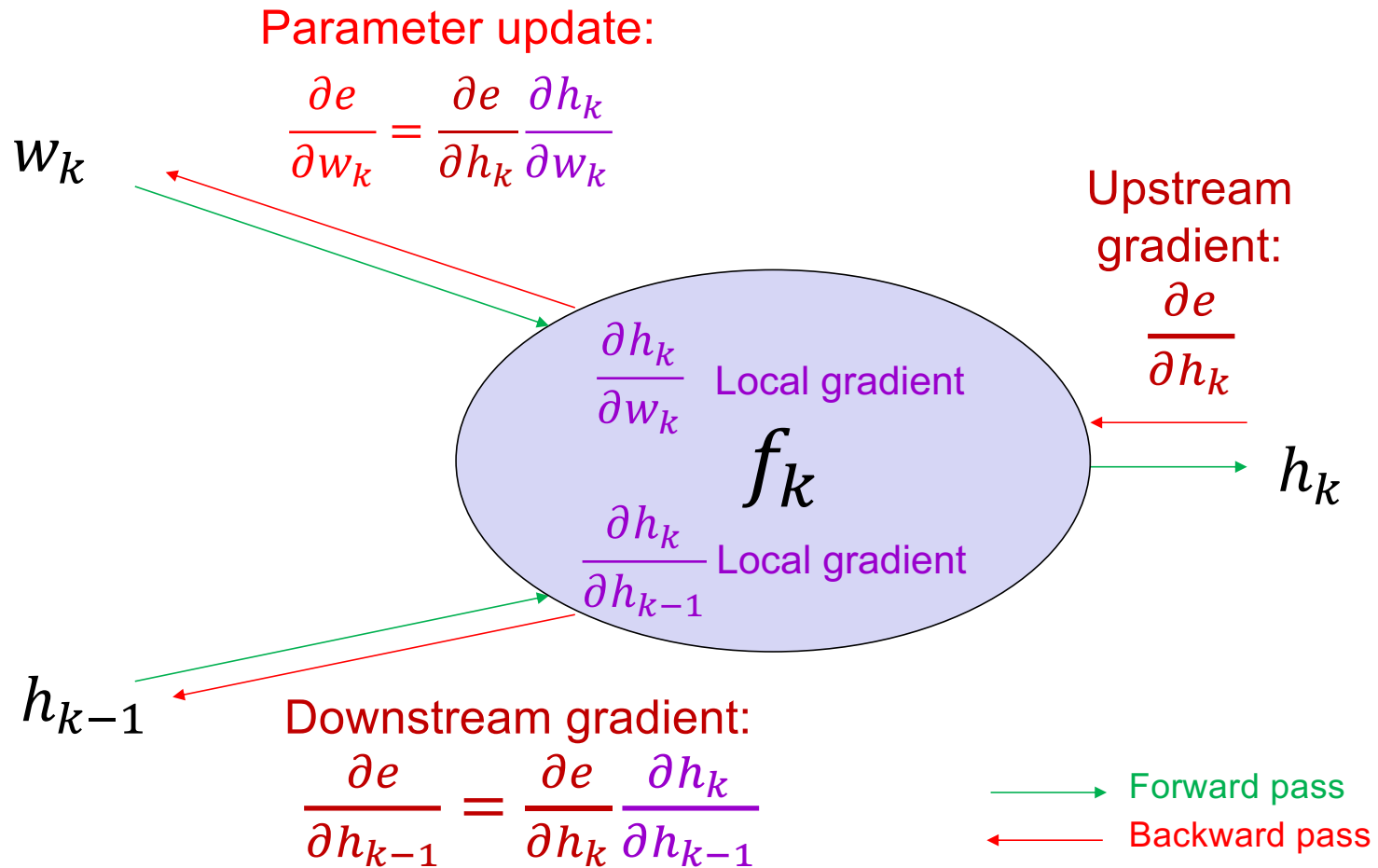
$$\frac{\partial e}{\partial w_k} = \frac{\partial e}{\partial h_K} \frac{\partial h_K}{\partial h_{K-1}} \cdots \frac{\partial h_{k+1}}{\partial h_k} \frac{\partial h_k}{\partial w_k}$$

Upstream gradient Local gradient

$\frac{\partial e}{\partial h_k}$

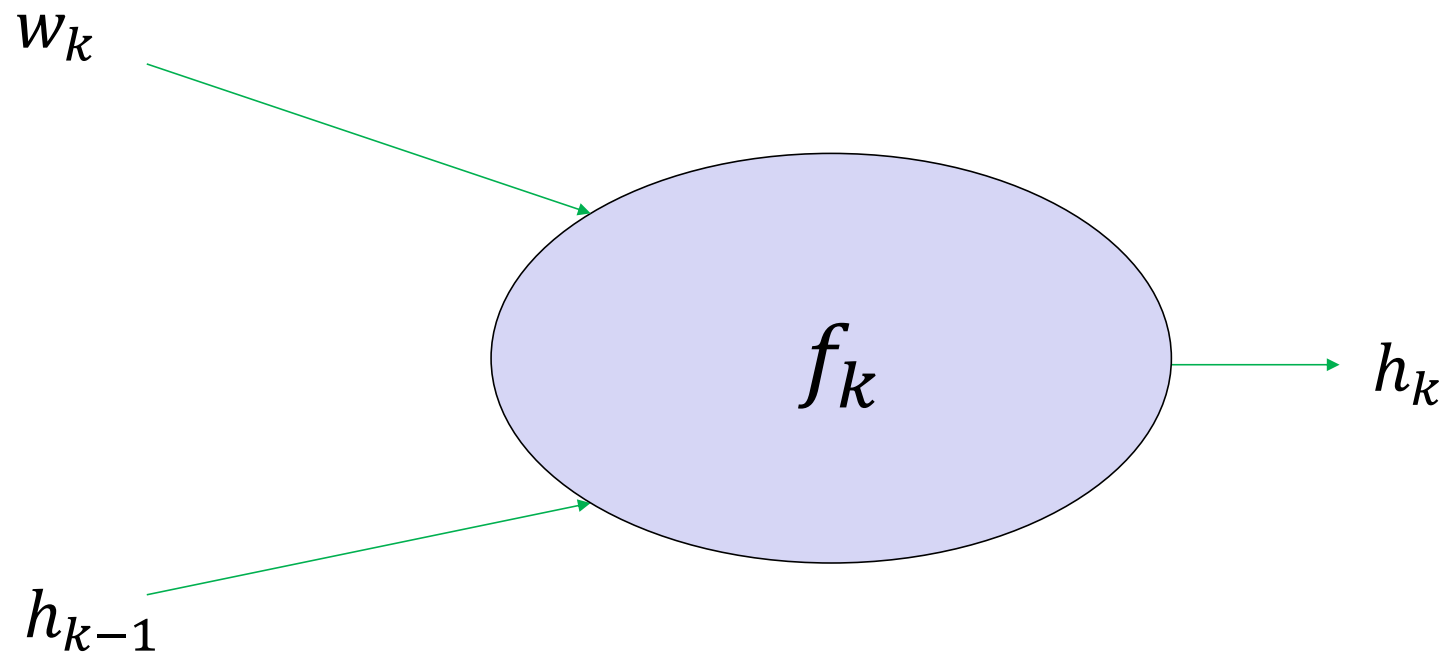
# Backpropagation summary

---



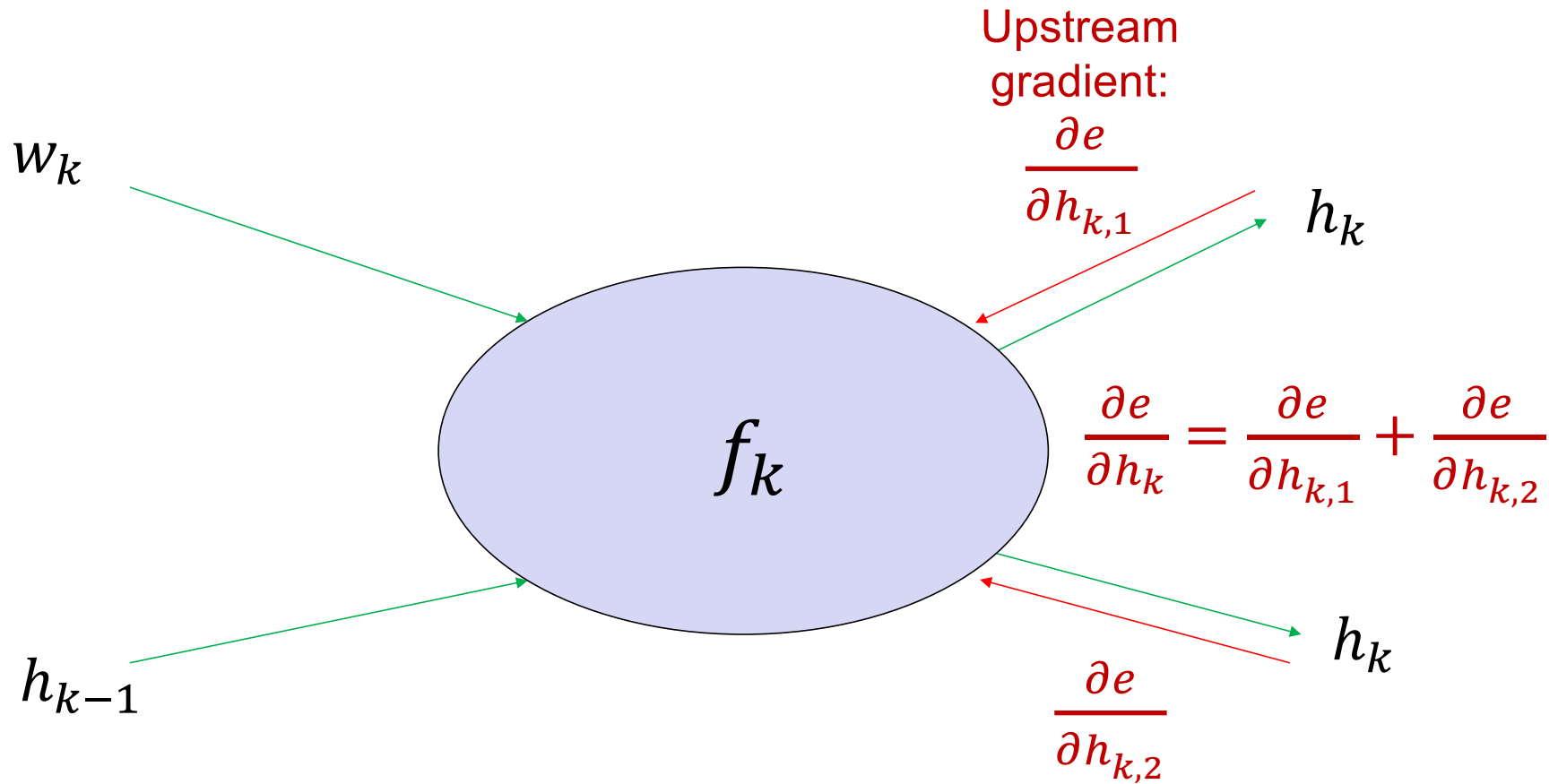
# What about more general computation graphs?

---



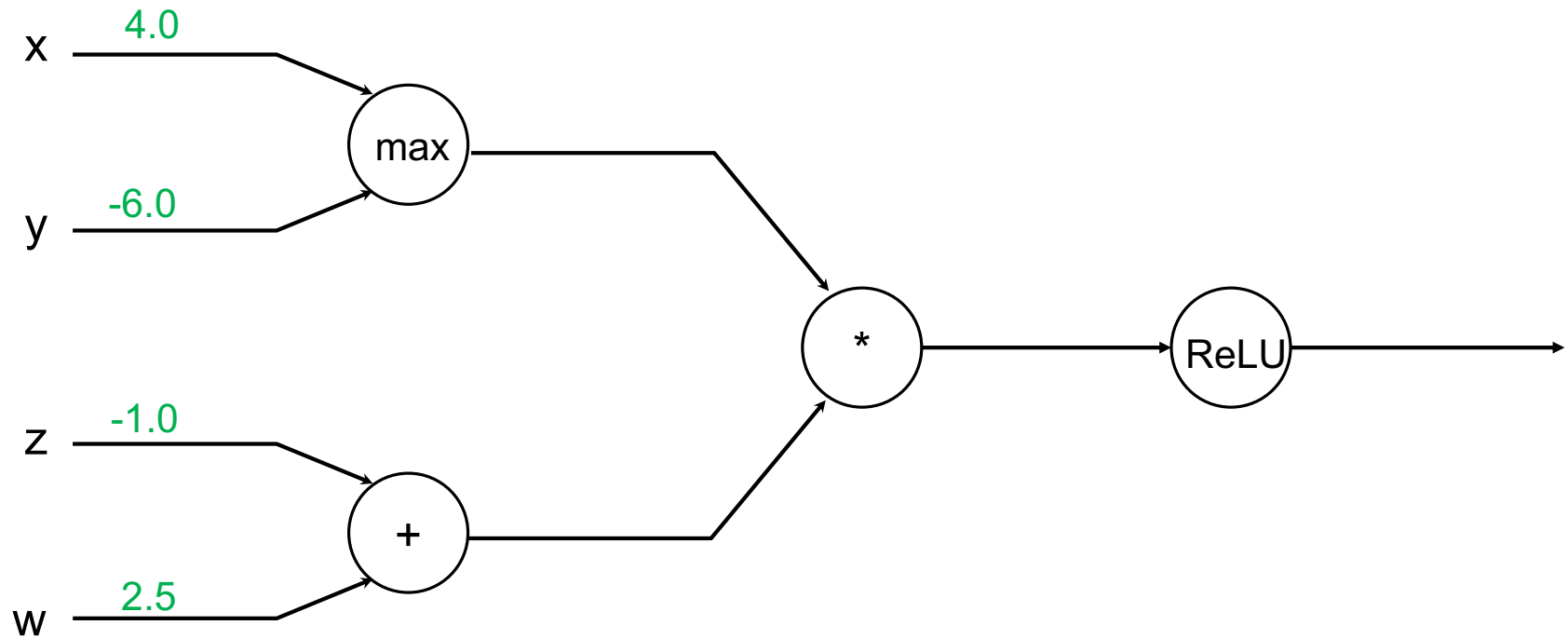
# What about more general computation graphs?

---



# Toy example of backward pass

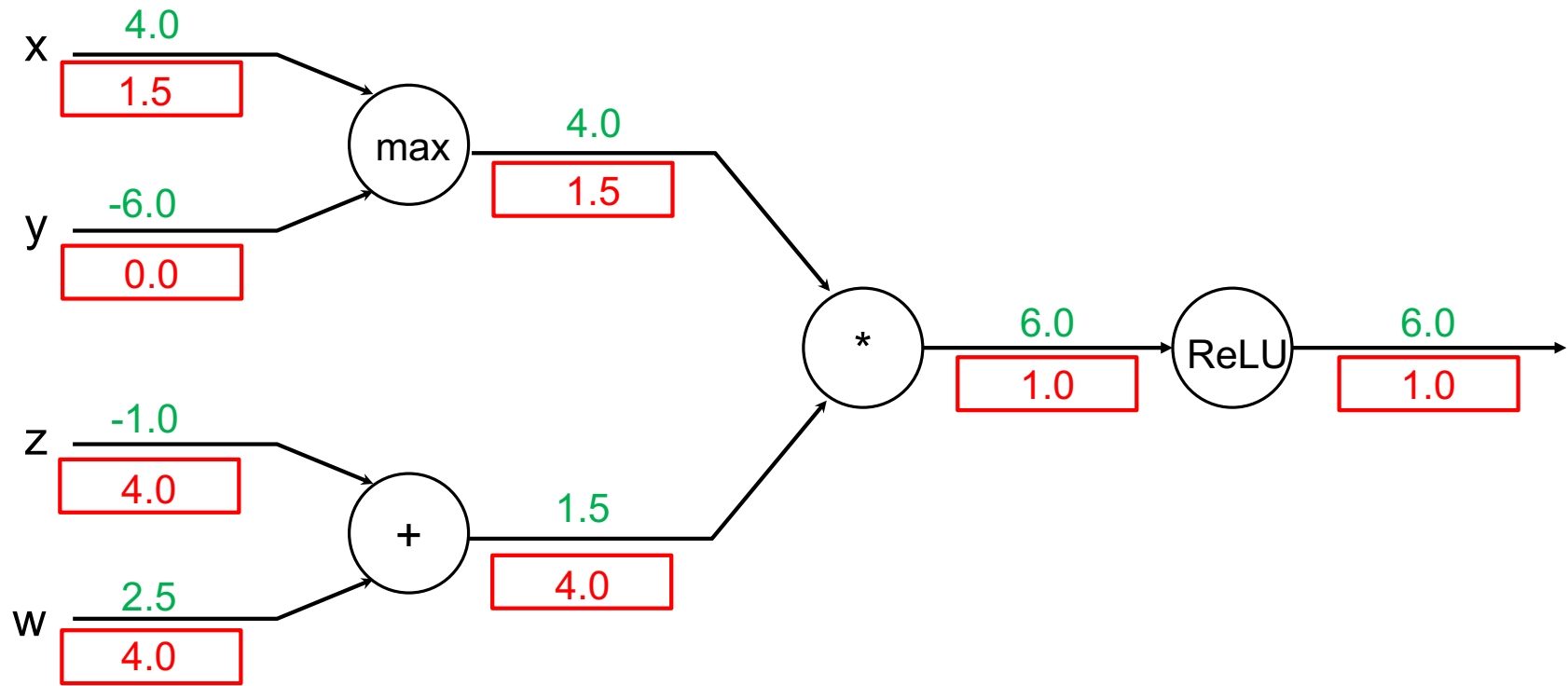
---



Source: [Stanford 231n](#)

# Toy example of backward pass

---



Source: [Stanford 231n](#)

# Overview: Backpropagation

---

- Computation graphs
- Using the chain rule
- General backprop algorithm
- Toy examples of backward pass
- **Matrix-vector calculations: ReLU, linear layer**

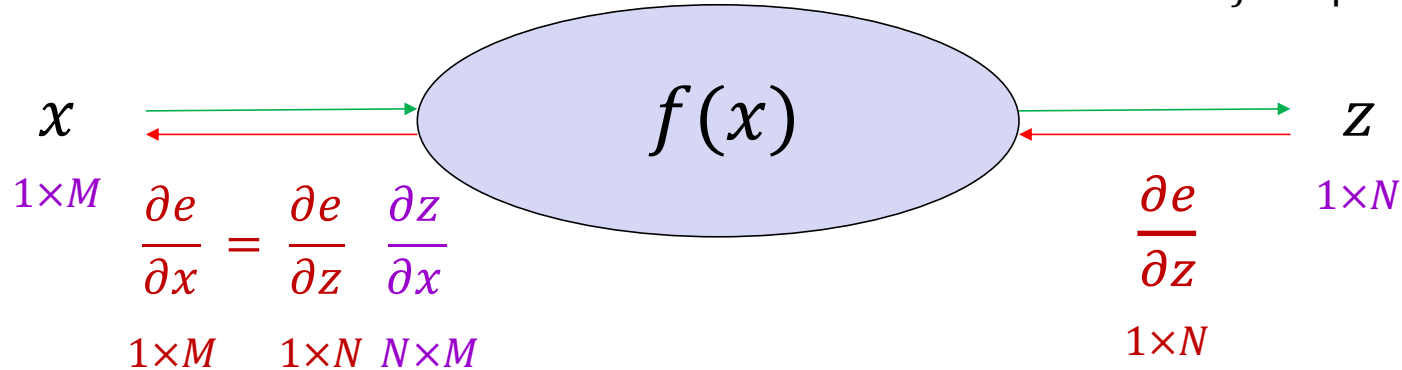
# Dealing with vectors

---

$$\frac{\partial z}{\partial x} = \begin{pmatrix} \frac{\partial z^{(1)}}{\partial x^{(1)}} & \cdots & \frac{\partial z^{(1)}}{\partial x^{(M)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial z^{(N)}}{\partial x^{(1)}} & \cdots & \frac{\partial z^{(N)}}{\partial x^{(M)}} \end{pmatrix}$$

$N \times M$   
Jacobian

**Jacobian:** row indices correspond to outputs, column indices correspond to inputs. The  $i, j$ th element of the Jacobian is the partial derivative of the  $i$ th output w.r.t.  $j$ th input



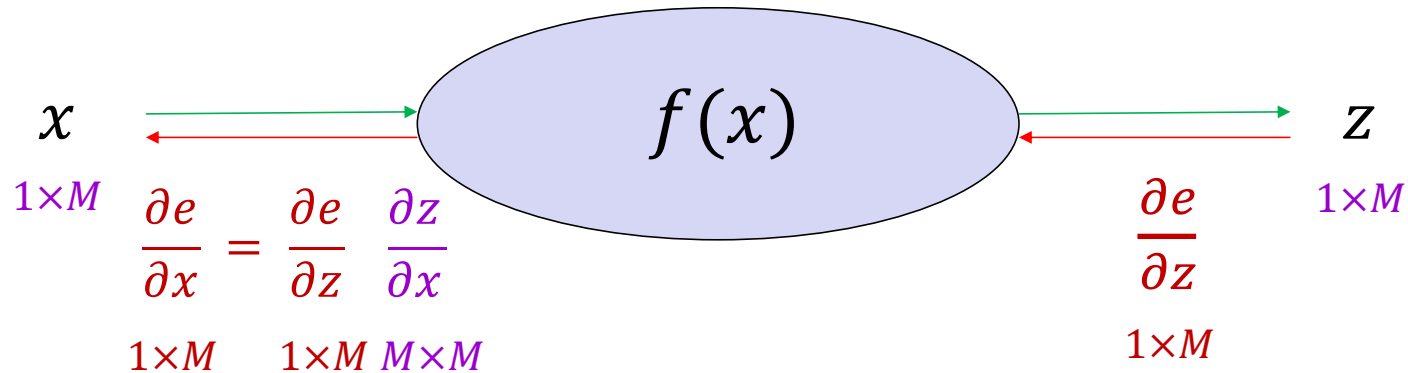
# Simple case: Elementwise operation

---

$$\frac{\partial z}{\partial x} = \begin{pmatrix} \frac{\partial z^{(1)}}{\partial x^{(1)}} & \cdots & \frac{\partial z^{(1)}}{\partial x^{(M)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial z^{(M)}}{\partial x^{(1)}} & \cdots & \frac{\partial z^{(M)}}{\partial x^{(M)}} \end{pmatrix}$$

$M \times M$   
Jacobian

What does the Jacobian for an elementwise function look like?



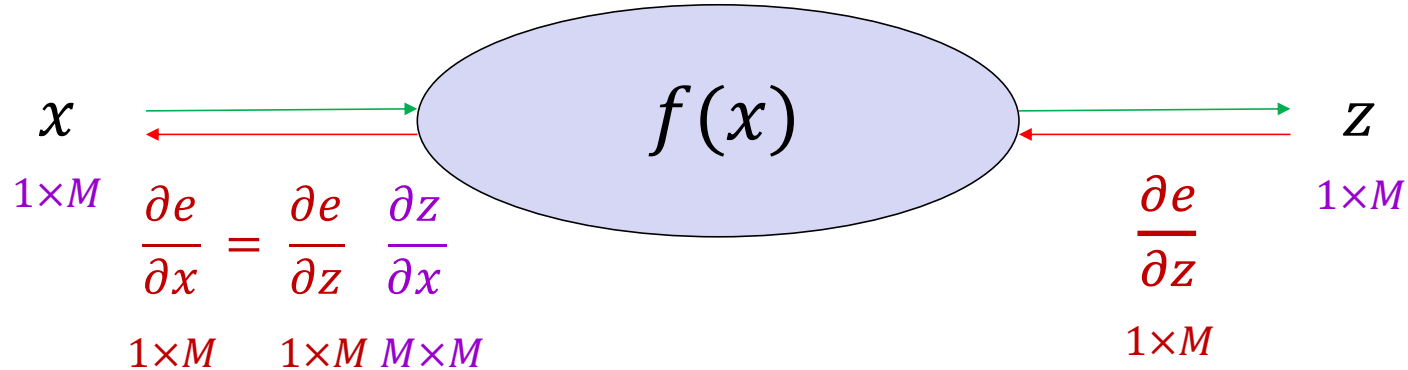
# Simple case: Elementwise operation

---

$$\frac{\partial z}{\partial x} = \begin{pmatrix} \frac{\partial z^{(1)}}{\partial x^{(1)}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{\partial z^{(M)}}{\partial x^{(M)}} \end{pmatrix}$$

$M \times M$  Jacobian

What does the Jacobian for an elementwise function look like?



ReLU layer

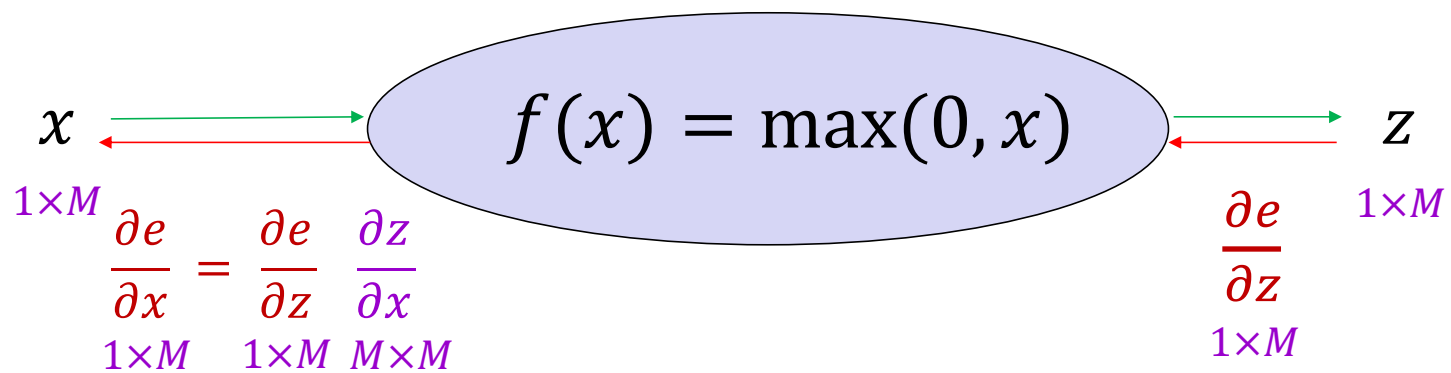
---

# ReLU layer

---

$$\frac{\partial z}{\partial x} = \begin{pmatrix} \frac{\partial z^{(1)}}{\partial x^{(1)}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial z^{(M)}}{\partial x^{(M)}} \end{pmatrix}$$

$M \times M$   
Jacobian

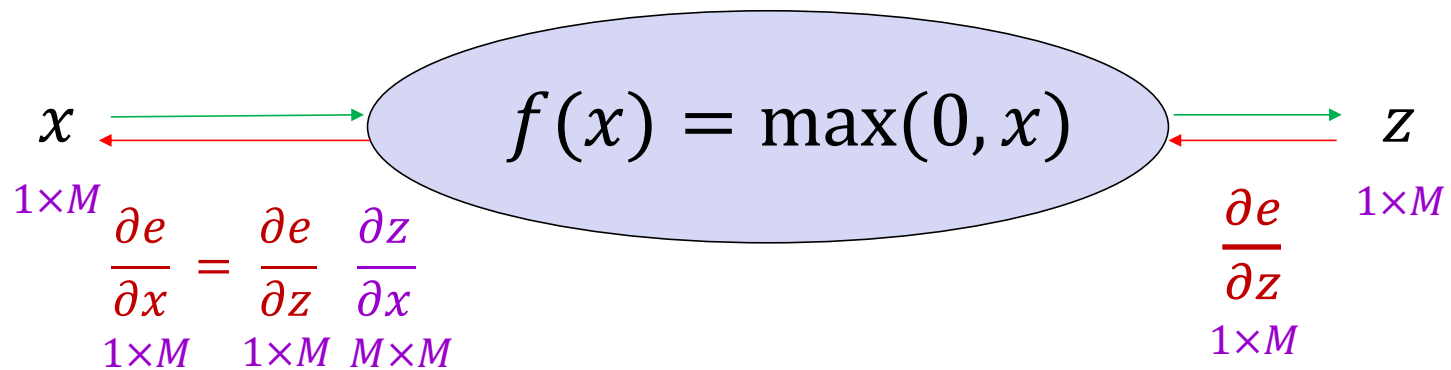


# ReLU layer

---

$$\frac{\partial z}{\partial x} = \begin{pmatrix} \mathbb{I}[x^{(1)} > 0] & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbb{I}[x^{(M)} > 0] \end{pmatrix}$$

$M \times M$   
Jacobian



$$\frac{\partial e}{\partial x^{(i)}} = \frac{\partial e}{\partial z^{(i)}} \mathbb{I}[x^{(i)} > 0]$$

What happens if some  $x^{(i)}$  is always negative?

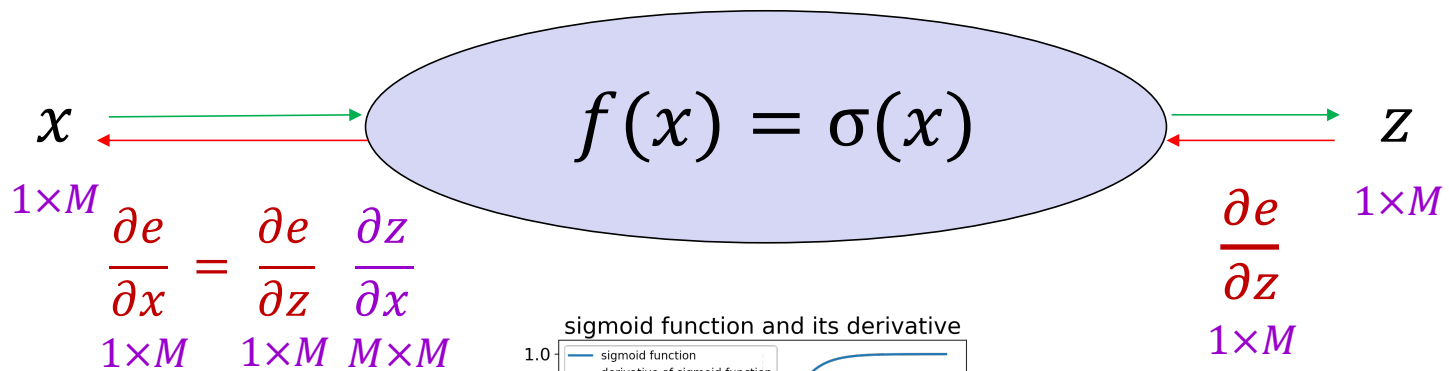
$$\frac{\partial e}{\partial x} = \frac{\partial e}{\partial z} \odot \mathbb{I}[x > 0]$$

This is known as the “dead ReLU” problem

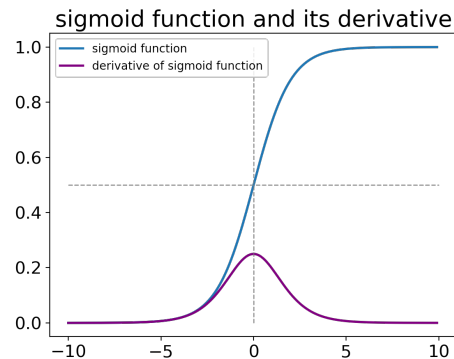
# What about sigmoid?

$$\frac{\partial z}{\partial x} = \begin{pmatrix} \sigma'(x^{(1)}) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma'(x^{(M)}) \end{pmatrix}$$

$M \times M$   
Jacobian



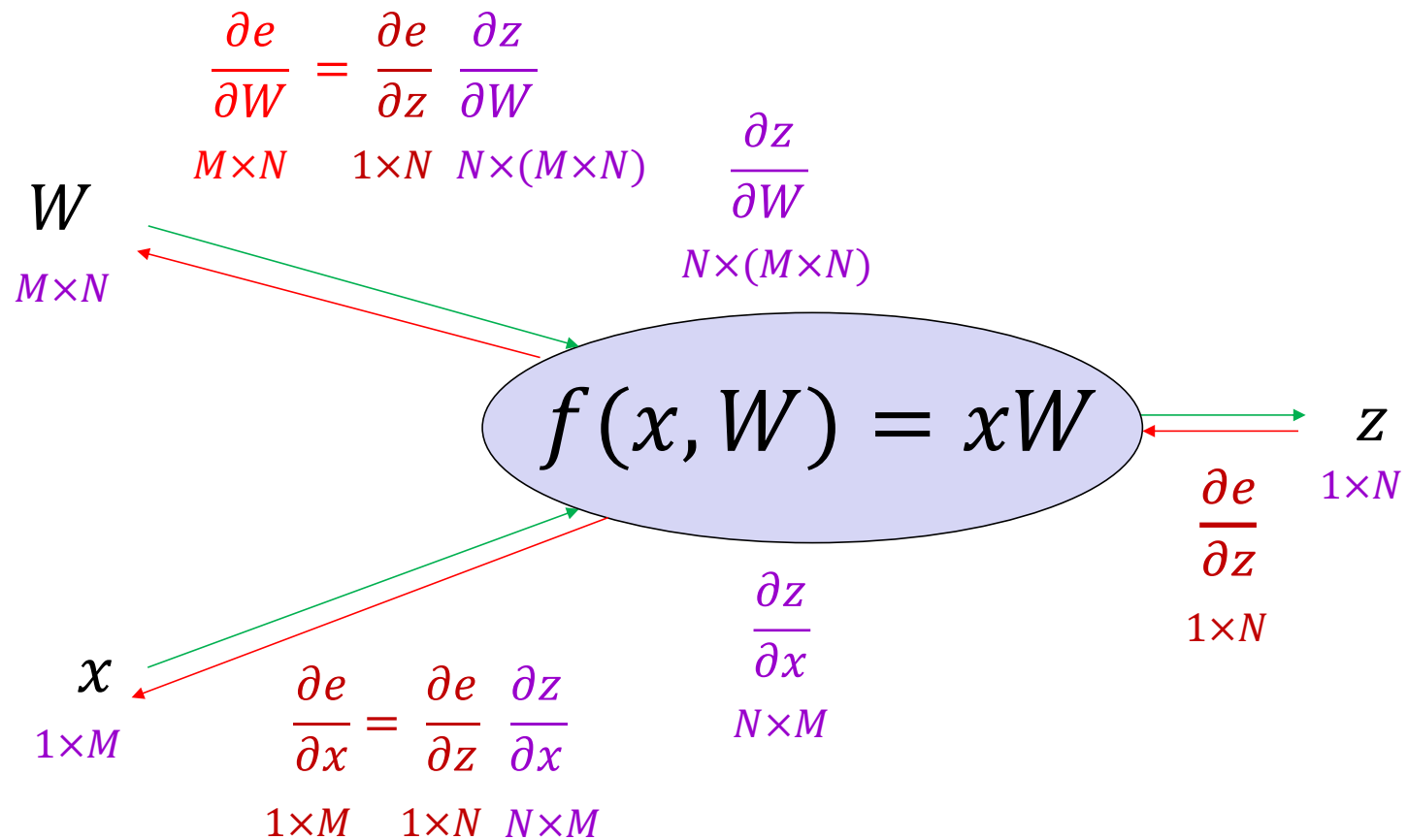
$$\frac{\partial e}{\partial x^{(i)}} = \frac{\partial e}{\partial z^{(i)}} \sigma'(x^{(i)})$$



Derivative vanishes unless the input value is close to zero!

# Matrix-vector multiplication (linear layer)

---

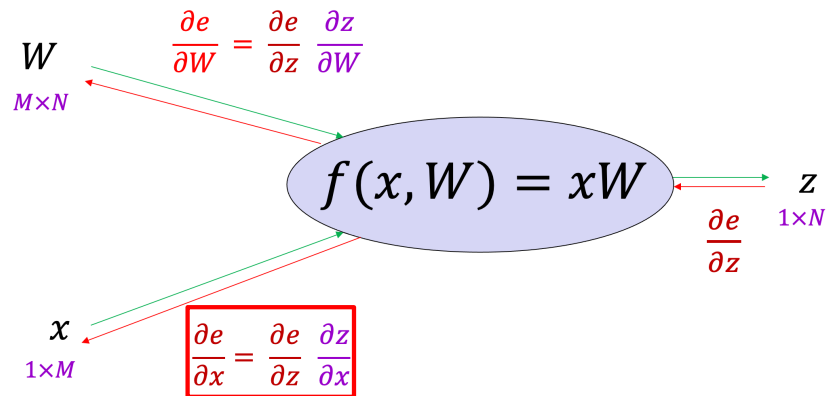


# Matrix-vector multiplication (linear layer)

$$(z^{(1)} \quad \dots \quad z^{(N)}) = (x^{(1)} \quad \dots \quad x^{(M)}) \begin{pmatrix} W^{(11)} & \dots & W^{(1N)} \\ \vdots & \ddots & \vdots \\ W^{(M1)} & \dots & W^{(MN)} \end{pmatrix} \quad z^{(j)} = \sum_{i=1}^M x^{(i)} W^{(ij)}$$

Want:  $\frac{\partial e}{\partial x} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial x}$

$1 \times M$        $1 \times N$     $N \times M$



# Matrix-vector multiplication (linear layer)

---

$$(z^{(1)} \quad \dots \quad z^{(N)}) = (x^{(1)} \quad \dots \quad x^{(M)}) \begin{pmatrix} W^{(11)} & \dots & W^{(1N)} \\ \vdots & \ddots & \vdots \\ W^{(M1)} & \dots & W^{(MN)} \end{pmatrix} \quad z^{(j)} = \sum_{i=1}^M x^{(i)} W^{(ij)}$$

Want:  $\frac{\partial e}{\partial x} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial x}$

$1 \times M$        $1 \times N$     $N \times M$

$$\frac{\partial z^{(j)}}{\partial x^{(i)}} =$$

$j$ th row,  $i$ th column  
of Jacobian

$$\frac{\partial z}{\partial x} = W^T$$

# Matrix-vector multiplication (linear layer)

---

$$(z^{(1)} \quad \dots \quad z^{(N)}) = (x^{(1)} \quad \dots \quad x^{(M)}) \begin{pmatrix} W^{(11)} & \dots & W^{(1N)} \\ \vdots & \ddots & \vdots \\ W^{(M1)} & \dots & W^{(MN)} \end{pmatrix} \quad z^{(j)} = \sum_{i=1}^M x^{(i)} W^{(ij)}$$

Want:  $\frac{\partial e}{\partial x} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial x}$

$1 \times M \quad 1 \times N \quad N \times M$

$$\frac{\partial z^{(j)}}{\partial x^{(i)}} = W^{(ij)} \quad \text{jth row, } i\text{th column of Jacobian}$$

$$\frac{\partial z}{\partial x} = W^T$$

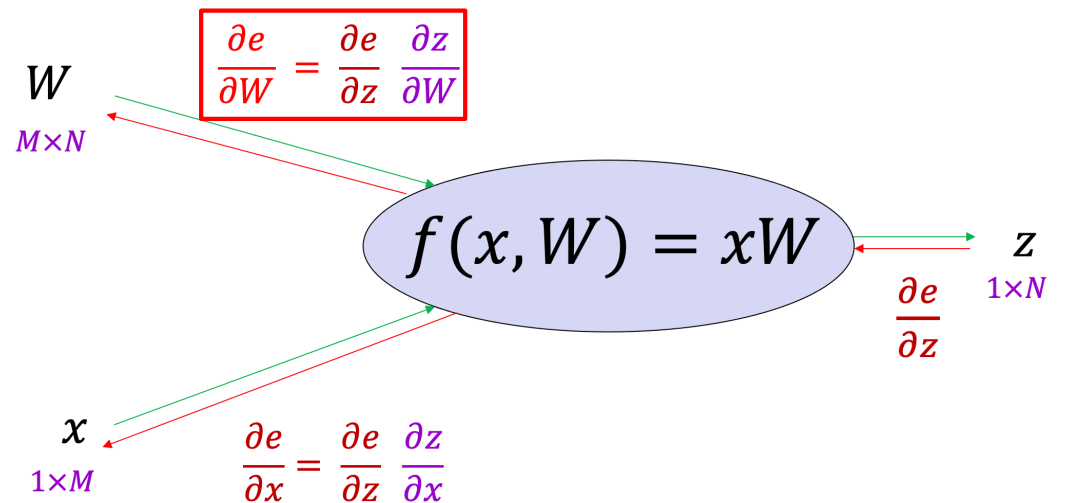
$$\frac{\partial e}{\partial x} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial x} = \frac{\partial e}{\partial z} W^T$$

# Matrix-vector multiplication (linear layer)

$$(z^{(1)} \quad \dots \quad z^{(N)}) = (x^{(1)} \quad \dots \quad x^{(M)}) \begin{pmatrix} W^{(11)} & \dots & W^{(1N)} \\ \vdots & \ddots & \vdots \\ W^{(M1)} & \dots & W^{(MN)} \end{pmatrix} \quad z^{(j)} = \sum_{i=1}^M x^{(i)} W^{(ij)}$$

Want:  $\frac{\partial e}{\partial W} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial W}$

$M \times N$        $1 \times N$      $N \times (M \times N)$



# Matrix-vector multiplication (linear layer)

---

$$(z^{(1)} \quad \dots \quad z^{(N)}) = (x^{(1)} \quad \dots \quad x^{(M)}) \begin{pmatrix} W^{(11)} & \dots & W^{(1N)} \\ \vdots & \ddots & \vdots \\ W^{(M1)} & \dots & W^{(MN)} \end{pmatrix}$$

$$z^{(j)} = \sum_{i=1}^M x^{(i)} W^{(ij)}$$

Want:  $\frac{\partial e}{\partial W} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial W}$

$M \times N$        $1 \times N$        $N \times (M \times N)$

$$\frac{\partial z^{(k)}}{\partial W^{(ij)}}$$

$z^{(k)}$  depends only  
on  $k$ th column of  $W$

# Matrix-vector multiplication (linear layer)

---

$$(z^{(1)} \quad \dots \quad z^{(N)}) = (x^{(1)} \quad \dots \quad x^{(M)}) \begin{pmatrix} W^{(11)} & \dots & W^{(1N)} \\ \vdots & \ddots & \vdots \\ W^{(M1)} & \dots & W^{(MN)} \end{pmatrix}$$

$$z^{(j)} = \sum_{i=1}^M x^{(i)} W^{(ij)}$$

Want:  $\frac{\partial e}{\partial W} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial W}$

$M \times N$        $1 \times N$        $N \times (M \times N)$

$$\frac{\partial z^{(k)}}{\partial W^{(ij)}} = \mathbb{I}[k = j] x^{(i)}$$

$z^{(k)}$  depends only on  $k$ th column of  $W$

$$\frac{\partial e}{\partial W^{(ij)}} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial W^{(ij)}}$$

# Matrix-vector multiplication (linear layer)

---

$$(z^{(1)} \quad \dots \quad z^{(N)}) = (x^{(1)} \quad \dots \quad x^{(M)}) \begin{pmatrix} W^{(11)} & \dots & W^{(1N)} \\ \vdots & \ddots & \vdots \\ W^{(M1)} & \dots & W^{(MN)} \end{pmatrix}$$

$$z^{(j)} = \sum_{i=1}^M x^{(i)} W^{(ij)}$$

Want:  $\frac{\partial e}{\partial W} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial W}$

$M \times N$        $1 \times N$      $N \times (M \times N)$

$$\frac{\partial z^{(k)}}{\partial W^{(ij)}} = \mathbb{I}[k = j] x^{(i)}$$

$z^{(k)}$  depends only on  $k$ th column of  $W$

$$\frac{\partial e}{\partial W^{(ij)}} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial W^{(ij)}} = \sum_{k=1}^N \frac{\partial e}{\partial z^{(k)}} \frac{\partial z^{(k)}}{\partial W^{(ij)}}$$

# Matrix-vector multiplication (linear layer)

---

$$(z^{(1)} \quad \dots \quad z^{(N)}) = (x^{(1)} \quad \dots \quad x^{(M)}) \begin{pmatrix} W^{(11)} & \dots & W^{(1N)} \\ \vdots & \ddots & \vdots \\ W^{(M1)} & \dots & W^{(MN)} \end{pmatrix} \quad z^{(j)} = \sum_{i=1}^M x^{(i)} W^{(ij)}$$

Want:  $\frac{\partial e}{\partial W} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial W}$

$M \times N$        $1 \times N$      $N \times (M \times N)$

$$\frac{\partial z^{(k)}}{\partial W^{(ij)}} = \mathbb{I}[k = j] x^{(i)}$$

$z^{(k)}$  depends only on  $k$ th column of  $W$

$$\frac{\partial e}{\partial W^{(ij)}} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial W^{(ij)}} = \sum_{k=1}^N \frac{\partial e}{\partial z^{(k)}} \frac{\partial z^{(k)}}{\partial W^{(ij)}} = \frac{\partial e}{\partial z^{(j)}} \frac{\partial z^{(j)}}{\partial W^{(ij)}}$$

# Matrix-vector multiplication (linear layer)

---

$$(z^{(1)} \quad \dots \quad z^{(N)}) = (x^{(1)} \quad \dots \quad x^{(M)}) \begin{pmatrix} W^{(11)} & \dots & W^{(1N)} \\ \vdots & \ddots & \vdots \\ W^{(M1)} & \dots & W^{(MN)} \end{pmatrix} \quad z^{(j)} = \sum_{i=1}^M x^{(i)} W^{(ij)}$$

Want:  $\frac{\partial e}{\partial W} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial W}$

$M \times N$        $1 \times N$      $N \times (M \times N)$

$$\frac{\partial z^{(k)}}{\partial W^{(ij)}} = \mathbb{I}[k = j] x^{(i)}$$

$z^{(k)}$  depends only on  $k$ th column of  $W$

$$\frac{\partial e}{\partial W^{(ij)}} = \frac{\partial e}{\partial z} \frac{\partial z}{\partial W^{(ij)}} = \sum_{k=1}^N \frac{\partial e}{\partial z^{(k)}} \frac{\partial z^{(k)}}{\partial W^{(ij)}} = \frac{\partial e}{\partial z^{(j)}} \frac{\partial z^{(j)}}{\partial W^{(ij)}} = \frac{\partial e}{\partial z^{(j)}} x^{(i)}$$

## Matrix-vector multiplication (linear layer)

---

$$(z^{(1)} \quad \dots \quad z^{(N)}) = (x^{(1)} \quad \dots \quad x^{(M)}) \begin{pmatrix} W^{(11)} & \dots & W^{(1N)} \\ \vdots & \ddots & \vdots \\ W^{(M1)} & \dots & W^{(MN)} \end{pmatrix}$$

$$z^{(j)} = \sum_{i=1}^M x^{(i)} W^{(ij)}$$

Want:  $\frac{\partial e}{\partial W}$   $M \times N$  =  $\frac{\partial e}{\partial z}$   $1 \times N$   $\frac{\partial z}{\partial W}$   $N \times (M \times N)$

$$\frac{\partial e}{\partial W^{(ij)}} = \frac{\partial e}{\partial z^{(j)}} x^{(i)}$$

$$\frac{\partial e}{\partial W} = \begin{pmatrix} \frac{\partial e}{\partial z^{(1)}} x^{(1)} & \dots & \frac{\partial e}{\partial z^{(N)}} x^{(1)} \\ \vdots & \ddots & \vdots \\ \frac{\partial e}{\partial z^{(1)}} x^{(M)} & \dots & \frac{\partial e}{\partial z^{(N)}} x^{(M)} \end{pmatrix}$$

# Matrix-vector multiplication (linear layer)

---

$$(z^{(1)} \quad \dots \quad z^{(N)}) = (x^{(1)} \quad \dots \quad x^{(M)}) \begin{pmatrix} W^{(11)} & \dots & W^{(1N)} \\ \vdots & \ddots & \vdots \\ W^{(M1)} & \dots & W^{(MN)} \end{pmatrix}$$

$$z^{(j)} = \sum_{i=1}^M x^{(i)} W^{(ij)}$$

Want:  $\frac{\partial e}{\partial W}$   $=$   $\frac{\partial e}{\partial z}$   $\frac{\partial z}{\partial W}$

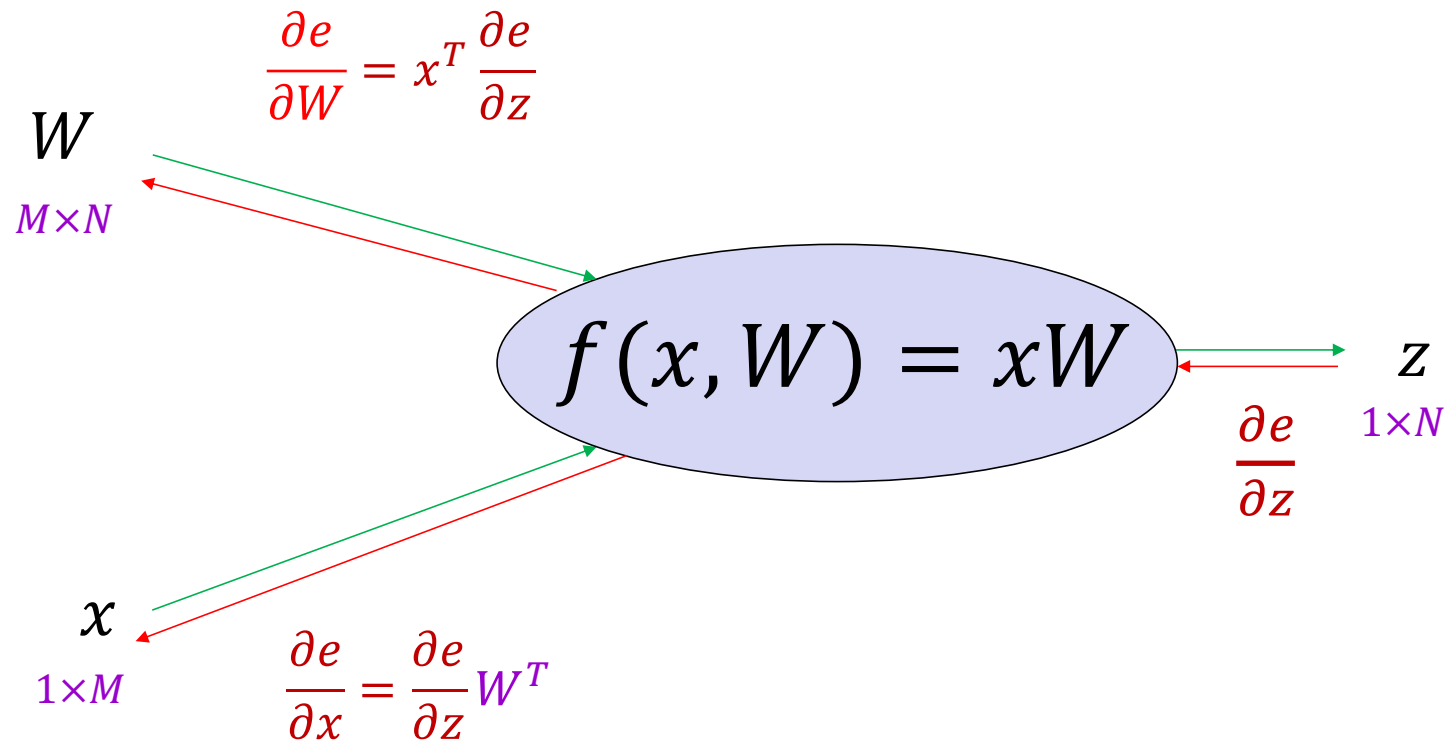
$M \times N$        $1 \times N$      $N \times (M \times N)$

$$\frac{\partial e}{\partial W} = x^T \frac{\partial e}{\partial z}$$

# Matrix-vector multiplication (linear layer)

---

- Summary of backward pass:



## General tips

---

- Derive error signal (upstream gradient) directly, avoid explicit computation of huge local derivatives
- Write out expression for a single element of the Jacobian, then deduce the overall formula
- Keep consistent indexing conventions, order of operations
- Use dimension analysis
- **For further reading:**
  - Lecture 4 of [Stanford 231n](#) and associated links in the syllabus
  - [Yes you should understand backprop](#) by Andrej Karpathy