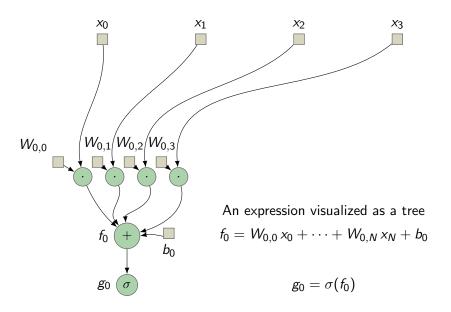
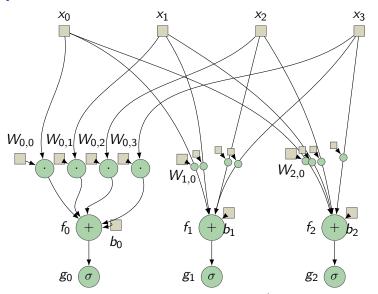
#### A neuron

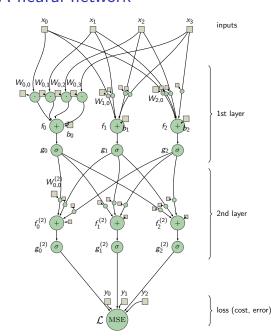


## A layer of neurons



An expression with re-used values gives a DAG (directed acyclic graph).

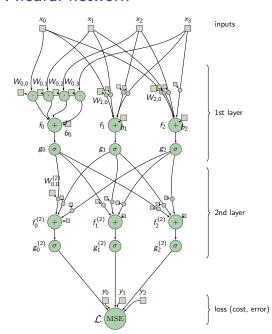
#### A neural network



For computing gradients, a neural network = one big expression for the loss.

a function of: inputs  $\bar{x}$ , weights and biases  $W, \bar{b}$ ,  $W^{(2)}, \bar{b}^{(2)}, \ldots$ , and targets  $\bar{y}$ .

#### A neural network



For computing gradients, a neural network = one big expression for the loss.

a function of: inputs  $\bar{x}$ , weights and biases  $W, \bar{b}$ ,  $W^{(2)}, \bar{b}^{(2)}, \ldots$ , and targets  $\bar{y}$ .

$$\mathcal{L} = \frac{1}{N^{(2)}} \sum_{i} (g_i^{(2)} - y_i)^2$$

where  $g_i^{(2)} = \sigma(f_i^{(2)})$ , ...

#### Gradients

Gradients = vectors to where a scalar function  $f(\bar{x})$  most increases.

Computationally, it's just a vector of partial derivatives,

$$\nabla_{\bar{x}} f = (\frac{\partial f}{\partial x_0}, \dots, \frac{\partial f}{\partial x_{N-1}})$$

#### Gradients

Gradients = vectors to where a scalar function  $f(\bar{x})$  most increases.

Computationally, it's just a vector of partial derivatives,

$$\nabla_{\bar{x}} f = (\frac{\partial f}{\partial x_0}, \dots, \frac{\partial f}{\partial x_{N-1}})$$

If  $\bar{x} \in \mathbb{R}^N$ , then  $\nabla_{\bar{x}} f \in \mathbb{R}^N$  (even if f has more inputs than just  $\bar{x}$ ).

Let f be a function of x, y, z.

$$\frac{\partial f}{\partial x} = a$$
 (at some fixed  $x, y, z$ ),

means increasing x by  $\varepsilon$  increases f by  $\approx a\varepsilon$  for small  $\varepsilon \in \mathbb{R}$ 

5 / 11

Let f be a function of x, y, z.

$$\frac{\partial f}{\partial x} = a$$
 (at some fixed  $x, y, z$ ),

means increasing x by  $\varepsilon$  increases f by  $\approx a\varepsilon$ 

for small  $\varepsilon \in \mathbb{R}$ 

Let 
$$g = g(f(x, y, z)), \frac{\partial f}{\partial x} = a_f, \frac{\partial g}{\partial f} = a_g.$$

Then increasing x by  $\varepsilon$  increases f by  $a_f \varepsilon$ ,

which increases g by  $a_g a_f \varepsilon$ . So  $\frac{\partial g}{\partial x} = a_g a_f$ .

Let f be a function of x, y, z.

$$\frac{\partial f}{\partial x} = a$$
 (at some fixed  $x, y, z$ ),

means increasing x by  $\varepsilon$  increases f by  $\approx a\varepsilon$  for small  $\varepsilon \in \mathbb{R}$ 

Let 
$$g = g(f(x, y, z)), \frac{\partial f}{\partial x} = a_f, \frac{\partial g}{\partial f} = a_g.$$

Then increasing x by  $\varepsilon$  increases f by  $a_f \varepsilon$ ,

which increases g by  $a_g a_f \varepsilon$ . So  $\frac{\partial g}{\partial x} = a_g a_f$ . This proves  $\frac{\partial g}{\partial x} = \frac{\partial g}{\partial f} \frac{\partial f}{\partial x}$ .

Let f be a function of x, y, z.

$$\frac{\partial f}{\partial x} = a$$
 (at some fixed  $x, y, z$ ),

means increasing x by  $\varepsilon$  increases f by  $\approx a\varepsilon$  for small  $\varepsilon \in \mathbb{R}$ 

Let 
$$g = g(f(x, y, z)), \frac{\partial f}{\partial x} = a_f, \frac{\partial g}{\partial f} = a_g.$$

Then increasing x by  $\varepsilon$  increases f by  $a_f \varepsilon$ ,

which increases g by  $a_g a_f \varepsilon$ . So  $\frac{\partial g}{\partial x} = a_g a_f$ . This proves  $\frac{\partial g}{\partial x} = \frac{\partial g}{\partial f} \frac{\partial f}{\partial x}$ .

Note that  $\frac{\partial g}{\partial f}$  depends on the value of f, which depends on x, y, z.

Let f be a function of x, y, z.

$$\frac{\partial f}{\partial x} = a$$
 (at some fixed  $x, y, z$ ),

means increasing x by  $\varepsilon$  increases f by  $\approx a\varepsilon$ 

for small  $\varepsilon \in \mathbb{R}$ 

Let 
$$g = g(f(x, y, z)), \frac{\partial f}{\partial x} = a_f, \frac{\partial g}{\partial f} = a_g.$$

Then increasing x by  $\varepsilon$  increases f by  $a_f \varepsilon$ ,

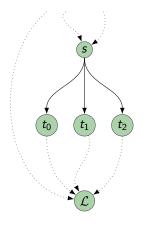
which increases g by  $a_g a_f \varepsilon$ . So  $\frac{\partial g}{\partial x} = a_g a_f$ . This proves  $\frac{\partial g}{\partial x} = \frac{\partial g}{\partial f} \frac{\partial f}{\partial x}$ .

Note that  $\frac{\partial g}{\partial f}$  depends on the value of f, which depends on x, y, z.

The expression for  $\frac{\partial g}{\partial f}$  often involves subexpressions equal to g.

Ex.: 
$$g = \sigma(f) \implies \frac{\partial g}{\partial x} = \sigma(f)(1 - \sigma(f))\frac{\partial f}{\partial x} = g(1 - g)\frac{\partial f}{\partial x}$$
.

## Derivates - in a graph

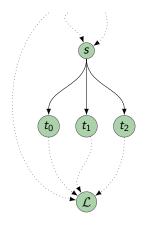


$$\frac{\partial \mathcal{L}}{\partial s} = \sum_{t \in \text{out}(s)} \frac{\partial \mathcal{L}}{\partial t} \cdot \frac{\partial t}{\partial s}$$

Because increasing s by  $\varepsilon$  increases each  $t_i$  by  $\frac{\partial t_i}{\partial s}\varepsilon$ .

Each of these contribute to increasing  $\mathcal{L}$  by  $\frac{\partial \mathcal{L}}{\partial t_i} \left( \frac{\partial t_i}{\partial s} \varepsilon \right)$ .

## Derivates - in a graph

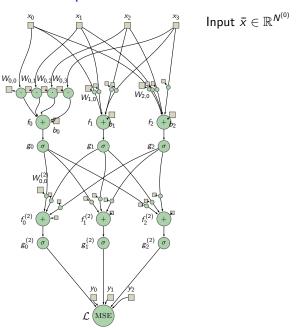


$$\frac{\partial \mathcal{L}}{\partial s} = \sum_{t \in \text{out}(s)} \frac{\partial \mathcal{L}}{\partial t} \cdot \frac{\partial t}{\partial s}$$

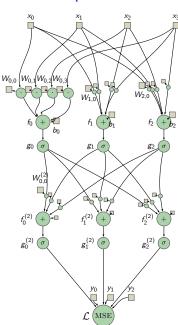
Because increasing s by  $\varepsilon$  increases each  $t_i$  by  $\frac{\partial t_i}{\partial s}\varepsilon$ .

Each of these contribute to increasing  $\mathcal{L}$  by  $\frac{\partial \mathcal{L}}{\partial t_i} \left( \frac{\partial t_i}{\partial s} \varepsilon \right)$ .

So we can compute  $\frac{\partial \mathcal{L}}{\partial s}$  for all nodes s, starting from  $\frac{\partial \mathcal{L}}{\partial \mathcal{L}} = 1$  and going back, as long as we can compute how each out-neighbor  $t_i$  depends on s  $(=\frac{\partial t_i}{\partial s})$ .



 $x_0,\dots,x_{N^{(0)}-1}$ 



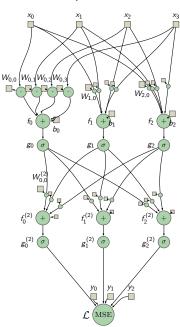
Input  $ar{x} \in \mathbb{R}^{N^{(0)}}$ 

 $x_0,\ldots,x_{N^{(0)}-1}$ 

Weights  $W^{\scriptscriptstyle (1)} \in \mathbb{R}^{N^{\scriptscriptstyle (1)} \times N^{\scriptscriptstyle (0)}}$  and biases  $b^{\scriptscriptstyle (1)} \in \mathbb{R}^{N^{\scriptscriptstyle (1)}}$ 

Pre-activations

$$\bar{f}^{(1)} = W^{(1)}\bar{x} + \bar{b}^{(1)} \in \mathbb{R}^{N^{(1)}}$$
  $f_i^{(1)} = \sum_j W_{i,j}^{(1)} x_j + b_i^{(1)}$ 



Input  $\bar{x} \in \mathbb{R}^{N^{(0)}}$ 

 $X_0, \ldots, X_{N^{(0)}-1}$ 

Weights  $W^{(1)} \in \mathbb{R}^{N^{(1)} \times N^{(0)}}$  and biases  $b^{(1)} \in \mathbb{R}^{N^{(1)}}$ 

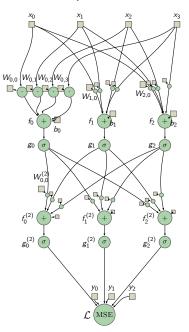
Pre-activations

$$\bar{f}^{(1)} = W^{(1)}\bar{x} + \bar{b}^{(1)} \in \mathbb{R}^{N^{(1)}}$$
  $f_i^{(1)} = \sum_j W_{i,j}^{(1)} x_j + b_i^{(1)}$ 

Activations

Activations 
$$g^{(1)} = \sigma(\bar{f}^{(1)})$$

 $g_i^{\scriptscriptstyle (1)} = \sigma(f_i^{\scriptscriptstyle (1)})$ 



Input 
$$\bar{x} \in \mathbb{R}^{N^{(0)}}$$

Weights  $W^{(1)} \in \mathbb{R}^{N^{(1)} \times N^{(0)}}$  and biases  $b^{(1)} \in \mathbb{R}^{N^{(1)}}$ 

Pre-activations

Fre-activations 
$$\bar{f}^{(1)} = W^{(1)}\bar{x} + \bar{b}^{(1)} \in \mathbb{R}^{N^{(1)}} \quad f_i^{(1)} = \sum_j W_{i,j}^{(1)} x_j + b_i^{(1)}$$

 $\bar{g}^{(2)} = \sigma(\bar{f}^{(2)})$ 

Activations  $\bar{g^{(1)}} = \sigma(\bar{f}^{(1)})$ 

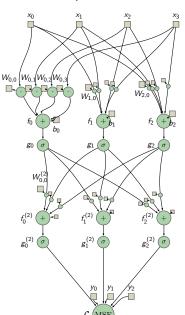
$$\sigma = \sigma(\Gamma^*)$$

$$g_i^{\scriptscriptstyle (1)} = \sigma(f_i^{\scriptscriptstyle (1)})$$

 $x_0, \ldots, x_{N^{(0)}-1}$ 

$$ar{f}^{(2)} = W^{(2)} ar{g}^{(1)} + ar{b}^{(2)} \in \mathbb{R}^{N^{(2)}}$$
 $ar{g}^{(2)} = \sigma(ar{f}^{(2)})$ 

7 / 11



Input 
$$\bar{x} \in \mathbb{R}^{N^{(0)}}$$

Weights  $W^{(1)} \in \mathbb{R}^{N^{(1)} \times N^{(0)}}$  and biases  $b^{(1)} \in \mathbb{R}^{N^{(1)}}$ 

Pre-activations

 $\bar{f}^{(1)} = W^{(1)}\bar{x} + \bar{b}^{(1)} \in \mathbb{R}^{N^{(1)}} \quad f_i^{(1)} = \sum_j W_{i,j}^{(1)} x_j + b_i^{(1)}$ 

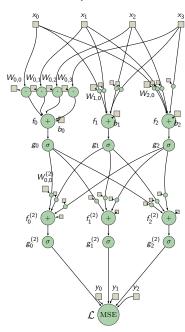
Activations 
$$g^{(1)} = \sigma(\bar{f}^{(1)})$$

 $g_i^{(1)} = \sigma(f_i^{(1)})$ 

 $X_0, \ldots, X_{N^{(0)}-1}$ 

$$ar{f}^{(2)} = \mathcal{W}^{(2)} ar{g}^{(1)} + ar{b}^{(2)} \in \mathbb{R}^{\mathcal{N}^{(2)}} \ ar{g}^{(2)} = \sigma(ar{f}^{(2)})$$

$$\mathcal{L} = \mathrm{mean} \left[ (\bar{g}^{(2)} - \bar{y})^2 \right] = \frac{1}{N^{(2)}} \sum_i (g_i^{(2)} - y_i)^2$$



Input  $ar{x} \in \mathbb{R}^{N^{(0)}}$ 

Weights  $W^{\scriptscriptstyle (1)} \in \mathbb{R}^{N^{\scriptscriptstyle (1)} \times N^{\scriptscriptstyle (0)}}$  and biases  $b^{\scriptscriptstyle (1)} \in \mathbb{R}^{N^{\scriptscriptstyle (1)}}$ 

Pre-activations

$$\bar{f}^{(1)} = W^{(1)}\bar{x} + \bar{b}^{(1)} \in \mathbb{R}^{N^{(1)}}$$

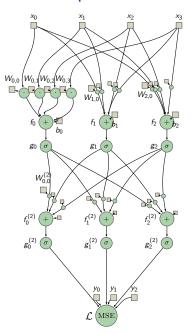
Activations

$$ar{g^{\scriptscriptstyle{(1)}}} = \sigma(ar{f}^{\scriptscriptstyle{(1)}})$$

$$\bar{f}^{(2)} = W^{(2)}\bar{g}^{(1)} + \bar{b}^{(2)} \in \mathbb{R}^{N^{(2)}}$$

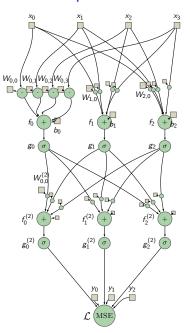
$$\bar{g}^{(2)} = \sigma(\bar{f}^{(2)})$$

$$\mathcal{L} = \operatorname{mean}\left[(\bar{g}^{(2)} - \bar{y})^2\right]$$



```
import numpy as np

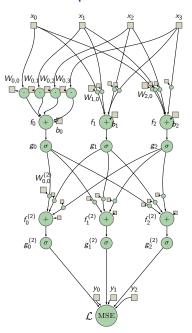
g = x
for b, w in zip(biases, weights):
    f = w @ g + b
    g = sigmoid(f)
```



```
import numpy as np

g = x
for b, w in zip(biases, weights):
    f = w @ g + b
    g = sigmoid(f)

When batched:
x has shape (B, N<sup>(0)</sup>)
g has shape before iteration i
```

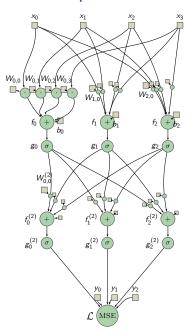


```
import numpy as np
```

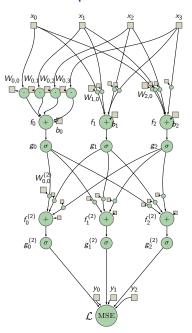
```
g = x
for b, w in zip(biases, weights):
f = w @ g + b
g = sigmoid(f)
```

When batched:

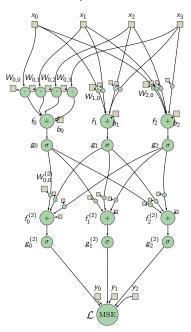
- x has shape  $(B, N^{(0)})$
- g has shape  $(B, N^{(i)})$  before iteration i w has shape



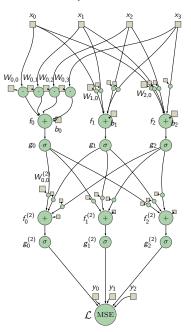
```
import numpy as np
for b, w in zip(biases, weights):
           f = w @ g + b
           g = sigmoid(f)
When batched:
x has shape (B, N^{(0)})
g has shape (B, N^{(i)}) before iteration i w has shape (N^{(i+1)}, N^{(i)})
w @ g.T has shape
```



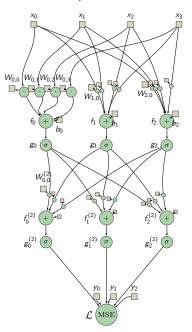
```
import numpy as np
for b, w in zip(biases, weights):
         f = w @ g + b
          g = sigmoid(f)
When batched:
x has shape (B, N^{(0)})
g has shape (B, N^{(i)}) before iteration i
w has shape (N^{(i+1)}, N^{(i)})
w @ g.T has shape (N^{(i+1)}, B)
```



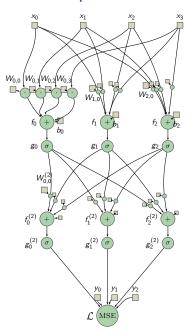
```
import numpy as np
for b, w in zip(biases, weights):
           f = w @ g + b
           g = sigmoid(f)
When batched:
x has shape (B, N^{(0)})
g has shape (B, N^{(i)}) before iteration i
w has shape (N^{(i+1)}, N^{(i)})
w @ g.T has shape (N^{(i+1)}, B)
(w @ g.T).T has shape (B, N^{(i+1)})
```



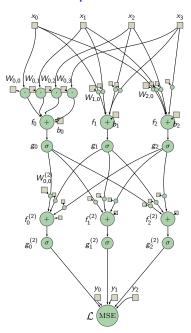
```
import numpy as np
for b, w in zip(biases, weights):
         f = w @ g + b
          g = sigmoid(f)
When batched:
x has shape (B, N^{(0)})
g has shape (B, N^{(i)}) before iteration i
w has shape (N^{(i+1)}, N^{(i)})
w @ g.T has shape (N^{(i+1)}, B)
(w @ g.T).T has shape (B, N^{(i+1)})
b has shape
```



```
import numpy as np
for b, w in zip(biases, weights):
          f = w @ g + b
          g = sigmoid(f)
When batched:
x has shape (B, N^{(0)})
g has shape (B, N^{(i)}) before iteration i
w has shape (N^{(i+1)}, N^{(i)})
w @ g.T has shape (N^{(i+1)}, B)
(w @ g.T).T has shape (B, N^{(i+1)})
b has shape (N^{(i+1)})
```



```
import numpy as np
for b, w in zip(biases, weights):
         f = w @ g + b
          g = sigmoid(f)
When batched:
x has shape (B, N^{(0)})
g has shape (B, N^{(i)}) before iteration i
w has shape (N^{(i+1)}, N^{(i)})
w @ g.T has shape (N^{(i+1)}, B)
(w @ g.T).T has shape (B, N^{(i+1)})
b has shape (N^{(i+1)})
f = (w @ g.T).T + b
(with b broadcasted along the batch)
```



```
import numpy as np
for b, w in zip(biases, weights):
         f = w @ g + b
          g = sigmoid(f)
When batched:
x has shape (B, N^{(0)})
g has shape (B, N^{(i)}) before iteration i
w has shape (N^{(i+1)}, N^{(i)})
w @ g.T has shape (N^{(i+1)}, B)
(w @ g.T).T has shape (B, N^{(i+1)})
b has shape (N^{(i+1)})
f = (w @ g.T).T + b
(with b broadcasted along the batch)
@ is the same as np.matmul
```

$$f_i^{(\ell)} = \sum_j W_{i,j}^{(\ell)} g_j^{(\ell-1)} + b_i^{(\ell)}$$

$$g_i^{(\ell)} = \sigma(f_i^{(\ell)})$$

$$\mathcal{L} = \frac{1}{N^{(L)}} \sum_{i} (g_i^{(L)} - y_i)^2$$

$$f_{i}^{(\ell)} = \sum_{j} W_{i,j}^{(\ell)} g_{j}^{(\ell-1)} + b_{i}^{(\ell)}$$

$$g_i^{(\ell)} = \sigma(f_i^{(\ell)})$$

$$\frac{\partial \mathcal{L}}{\partial g_i^{(L)}} =$$

$$\mathcal{L} = \frac{1}{N^{(L)}} \sum_{i} (g_i^{(L)} - y_i)^2$$

$$f_{i}^{(\ell)} = \sum_{j} W_{i,j}^{(\ell)} g_{j}^{(\ell-1)} + b_{i}^{(\ell)}$$

$$g_i^{(\ell)} = \sigma(f_i^{(\ell)})$$

$$\mathcal{L} = \frac{1}{M^{(L)}} \sum_{i} (g_i^{(L)} - y_i)^2$$

$$\frac{\partial \mathcal{L}}{\partial g_{i}^{(L)}} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right) \frac{\partial \left( g_{i}^{(L)} - y_{i} \right)}{\partial g_{i}^{(L)}}$$

$$\begin{split} f_i^{(\ell)} &= \sum_j W_{i,j}^{(\ell)} g_j^{(\ell-1)} + b_i^{(\ell)} \\ &\qquad \qquad \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} = \\ g_i^{(\ell)} &= \sigma(f_i^{(\ell)}) \\ &\qquad \qquad \frac{\partial \mathcal{L}}{\partial g_i^{(L)}} = \frac{2}{N^{(L)}} \left( g_i^{(L)} - y_i \right) \\ \mathcal{L} &= \frac{1}{N^{(L)}} \sum_i (g_i^{(L)} - y_i)^2 \end{split}$$

$$\begin{split} f_{i}^{(\ell)} &= \sum_{j} W_{i,j}^{(\ell)} g_{j}^{(\ell-1)} + b_{i}^{(\ell)} \\ & \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial g_{i}^{(\ell)}} \frac{\partial g_{i}^{(\ell)}}{\partial f_{i}^{(\ell)}} = \\ g_{i}^{(\ell)} &= \sigma(f_{i}^{(\ell)}) \\ & \frac{\partial \mathcal{L}}{\partial g_{i}^{(L)}} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right) \\ \mathcal{L} &= \frac{1}{N^{(L)}} \sum_{i} (g_{i}^{(L)} - y_{i})^{2} \end{split}$$

$$\begin{split} f_i^{(\ell)} &= \sum_j W_{i,j}^{(\ell)} g_j^{(\ell-1)} + b_i^{(\ell)} \\ &\qquad \qquad \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} \frac{\partial g_i^{(\ell)}}{\partial f_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} \sigma'(f_i^{(\ell)}) = \\ g_i^{(\ell)} &= \sigma(f_i^{(\ell)}) \\ &\qquad \qquad \frac{\partial \mathcal{L}}{\partial g_i^{(L)}} = \frac{2}{N^{(L)}} \left( g_i^{(L)} - y_i \right) \\ \mathcal{L} &= \frac{1}{N^{(L)}} \sum_i (g_i^{(L)} - y_i)^2 \end{split}$$

$$f_i^{(\ell)} = \sum_j W_{i,j}^{(\ell)} g_j^{(\ell-1)} + b_i^{(\ell)}$$

$$\mathcal{C}' = \sum_{j} W_{i,j} \mathcal{C}' \mathcal{G}_{j} \mathcal{C}^{-j} + \mathcal{D}_{i}^{k}$$

$$g_i^{(\ell)} = \sigma(f_i^{(\ell)})$$

$$\mathcal{L} = \frac{1}{N^{(L)}} \sum_{i} (g_i^{(L)} - y_i)^2$$

$$egin{aligned} rac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} &= rac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} \sigma'(f_i^{(\ell)}) = rac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} \, g_i^{(\ell)} \, (1 - g_i^{(\ell)}) \ & rac{\partial \mathcal{L}}{\partial g_i^{(L)}} &= rac{2}{N^{(L)}} \left( g_i^{(L)} - y_i 
ight) \end{aligned}$$

$$\begin{split} \frac{\partial \mathcal{L}}{\partial g_j^{(\ell-1)}} = \\ f_i^{(\ell)} = \sum_j W_{i,j}^{(\ell)} g_j^{(\ell-1)} + b_i^{(\ell)} \\ \\ \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} \sigma'(f_i^{(\ell)}) = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} g_i^{(\ell)} (1 - g_i^{(\ell)}) \\ g_i^{(\ell)} = \sigma(f_i^{(\ell)}) \end{split}$$

 $rac{\partial \mathcal{L}}{\partial g_i^{(L)}} = rac{2}{N^{(L)}} \left( g_i^{(L)} - y_i 
ight)$   $\mathcal{L} = rac{1}{N^{(L)}} \sum_i (g_i^{(L)} - y_i)^2$ 

 $\mathcal{L} = \frac{1}{N(L)} \sum_{i} (g_i^{(L)} - y_i)^2$ 

$$\begin{split} \frac{\partial \mathcal{L}}{\partial g_j^{(\ell-1)}} &= \sum_i \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \frac{\partial f_i^{(\ell)}}{\partial g_j^{(\ell-1)}} \\ f_i^{(\ell)} &= \sum_j W_{i,j}^{(\ell)} g_j^{(\ell-1)} + b_i^{(\ell)} \\ &\qquad \qquad \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} \sigma'(f_i^{(\ell)}) = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} g_i^{(\ell)} (1 - g_i^{(\ell)}) \\ g_i^{(\ell)} &= \sigma(f_i^{(\ell)}) \end{split}$$

 $\frac{\partial \mathcal{L}}{\partial g_{i}(L)} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right)$ 

 $\mathcal{L} = \frac{1}{N(L)} \sum_{i} (g_i^{(L)} - y_i)^2$ 

$$\begin{split} \frac{\partial \mathcal{L}}{\partial g_j^{(\ell-1)}} &= \sum_i \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} W_{i,j}^{(\ell)} \\ f_i^{(\ell)} &= \sum_j W_{i,j}^{(\ell)} g_j^{(\ell-1)} + b_i^{(\ell)} \\ \\ \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} &= \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} \sigma'(f_i^{(\ell)}) = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} g_i^{(\ell)} (1 - g_i^{(\ell)}) \\ g_i^{(\ell)} &= \sigma(f_i^{(\ell)}) \end{split}$$

 $\frac{\partial \mathcal{L}}{\partial g_{i}^{(L)}} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right)$ 

 $\mathcal{L} = \frac{1}{N(L)} \sum_{i} (g_i^{(L)} - y_i)^2$ 

$$\begin{split} \frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} &= \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \\ \frac{\partial \mathcal{L}}{\partial g_j^{(\ell-1)}} &= \sum_i \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} W_{i,j}^{(\ell)} \\ f_i^{(\ell)} &= \sum_j W_{i,j}^{(\ell)} g_j^{(\ell-1)} + b_i^{(\ell)} \\ \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} &= \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} \sigma'(f_i^{(\ell)}) = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} g_i^{(\ell)} (1 - g_i^{(\ell)}) \\ g_i^{(\ell)} &= \sigma(f_i^{(\ell)}) \end{split}$$

 $\frac{\partial \mathcal{L}}{\partial \sigma_{i}(L)} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right)$ 

$$\begin{split} \frac{\partial \mathcal{L}}{\partial b_{i}^{(\ell)}} &= \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} \frac{\partial f_{i}^{(\ell)}}{\partial b_{i}^{(\ell)}} \\ \frac{\partial \mathcal{L}}{\partial g_{j}^{(\ell-1)}} &= \sum_{i} \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} W_{i,j}^{(\ell)} \\ f_{i}^{(\ell)} &= \sum_{j} W_{i,j}^{(\ell)} g_{j}^{(\ell-1)} + b_{i}^{(\ell)} \\ \\ \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} &= \frac{\partial \mathcal{L}}{\partial g_{i}^{(\ell)}} \sigma'(f_{i}^{(\ell)}) = \frac{\partial \mathcal{L}}{\partial g_{i}^{(\ell)}} g_{i}^{(\ell)} (1 - g_{i}^{(\ell)}) \\ g_{i}^{(\ell)} &= \sigma(f_{i}^{(\ell)}) \end{split}$$

 $\frac{\partial \mathcal{L}}{\partial \sigma_{i}(L)} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right)$ 

 $\mathcal{L} = \frac{1}{N^{(L)}} \sum_{i} (g_i^{(L)} - y_i)^2$ 

 $\mathcal{L} = \frac{1}{N(L)} \sum_{i} (g_i^{(L)} - y_i)^2$ 

$$\begin{split} \frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} &= \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \\ \frac{\partial \mathcal{L}}{\partial g_j^{(\ell-1)}} &= \sum_i \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} W_{i,j}^{(\ell)} \\ f_i^{(\ell)} &= \sum_j W_{i,j}^{(\ell)} g_j^{(\ell-1)} + b_i^{(\ell)} \\ \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} &= \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} \sigma'(f_i^{(\ell)}) = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} g_i^{(\ell)} (1 - g_i^{(\ell)}) \\ g_i^{(\ell)} &= \sigma(f_i^{(\ell)}) \end{split}$$

 $\frac{\partial \mathcal{L}}{\partial \sigma_{i}(L)} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right)$ 

$$\begin{split} \frac{\partial \mathcal{L}}{\partial W_{i,j}^{(\ell)}} &= \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \frac{\partial f_i^{(\ell)}}{\partial W_{i,j}^{(\ell)}} \\ \frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} &= \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \\ \frac{\partial \mathcal{L}}{\partial g_j^{(\ell-1)}} &= \sum_i \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} W_{i,j}^{(\ell)} \\ f_i^{(\ell)} &= \sum_j W_{i,j}^{(\ell)} g_j^{(\ell-1)} + b_i^{(\ell)} \end{split}$$

$$rac{\partial \mathcal{L}}{\partial oldsymbol{\sigma}_{i}^{(L)}} = rac{2}{N^{(L)}} \left( oldsymbol{g}_{i}^{(L)} - y_{i} 
ight)$$

 $rac{\partial \mathcal{L}}{\partial f^{(\ell)}} = rac{\partial \mathcal{L}}{\partial \sigma^{(\ell)}} \sigma'(f^{(\ell)}_i) = rac{\partial \mathcal{L}}{\partial \sigma^{(\ell)}} g_i^{(\ell)} (1 - g_i^{(\ell)})$ 

$$\mathcal{L} = \frac{1}{M^{(L)}} \sum_{i} (g_i^{(L)} - y_i)^2$$

 $g_i^{(\ell)} = \sigma(f_i^{(\ell)})$ 

$$\frac{\partial \mathcal{L}}{\partial W_{i,j}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} g_j^{(\ell-1)}$$

$$\frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}}$$

$$\frac{\partial \mathcal{L}}{\partial g_j^{(\ell-1)}} = \sum_i \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} W_{i,j}^{(\ell)}$$

$$f_i^{(\ell)} = \sum_j W_{i,j}^{(\ell)} g_j^{(\ell-1)} + b_i^{(\ell)}$$

$$egin{aligned} rac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} &= rac{\partial \mathcal{L}}{\partial oldsymbol{g}_i^{(\ell)}} \sigma'(f_i^{(\ell)}) = rac{\partial \mathcal{L}}{\partial oldsymbol{g}_i^{(\ell)}} oldsymbol{g}_i^{(\ell)} (1 - oldsymbol{g}_i^{(\ell)}) \ &= rac{\partial \mathcal{L}}{\partial oldsymbol{g}_i^{(\ell)}} &= rac{2}{2\Omega} \left( oldsymbol{g}_i^{(L)} - oldsymbol{v}_i 
ight) \end{aligned}$$

$$rac{\partial \mathcal{L}}{\partial g_i^{(L)}} = rac{2}{\mathsf{N}^{(L)}} \left( g_i^{(L)} - y_i 
ight)$$
 $\mathcal{L} = rac{1}{\mathsf{N}^{(L)}} \sum_i (g_i^{(L)} - y_i)^2$ 

Weight and biases

$$\begin{split} \frac{\partial \mathcal{L}}{\partial W_{i,j}(\ell)} &= \frac{\partial \mathcal{L}}{\partial f_i(\ell)} \, g_j^{(\ell-1)} \\ \frac{\partial \mathcal{L}}{\partial b_i(\ell)} &= \frac{\partial \mathcal{L}}{\partial f_i(\ell)} \end{split}$$

Activations

$$\frac{\partial \mathcal{L}}{\partial g_{j}^{(\ell-1)}} = \sum_{i} \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} W_{i,j}^{(\ell)}$$

Pre-activations

$$rac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} = rac{\partial \mathcal{L}}{\partial g_{i}^{(\ell)}} \, g_{i}^{(\ell)} oldsymbol{(1-g_{i}^{(\ell)})}$$

$$\frac{\partial \mathcal{L}}{\partial g_{i}^{(L)}} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right)$$

Weight and biases

$$\begin{split} \frac{\partial \mathcal{L}}{\partial W_{i,j}(\ell)} &= \frac{\partial \mathcal{L}}{\partial f_i(\ell)} \, g_j^{(\ell-1)} \\ \frac{\partial \mathcal{L}}{\partial b_i(\ell)} &= \frac{\partial \mathcal{L}}{\partial f_i(\ell)} \end{split}$$

$$\nabla_{\vec{\textit{b}}^{(\ell)}} \mathcal{L} = \nabla_{\vec{\textit{f}}^{(\ell)}} \mathcal{L} \ \in \mathbb{R}^{\textit{N}^{(\ell)}}$$

Activations

$$\frac{\partial \mathcal{L}}{\partial g_{j}^{(\ell-1)}} = \sum_{i} \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} W_{i,j}^{(\ell)}$$

Pre-activations

$$rac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} = rac{\partial \mathcal{L}}{\partial g_{i}^{(\ell)}} \, g_{i}^{(\ell)} oldsymbol{(1-g_{i}^{(\ell)})}$$

$$\frac{\partial \mathcal{L}}{\partial g_{i}^{(L)}} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right)$$

Weight and biases

$$\begin{split} \frac{\partial \mathcal{L}}{\partial W_{i,j}(\ell)} &= \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \, g_j^{(\ell-1)} & \nabla_{W^{(\ell)}} \mathcal{L} = \\ \frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} &= \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} & \nabla_{\overline{b}^{(\ell)}} \mathcal{L} = \nabla_{\overline{f}^{(\ell)}} \mathcal{L} \, \in \mathbb{R}^{N^{(\ell)}} \end{split}$$

Activations

$$\frac{\partial \mathcal{L}}{\partial g_{j}^{(\ell-1)}} = \sum_{i} \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} W_{i,j}^{(\ell)}$$

Pre-activations

$$rac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} = rac{\partial \mathcal{L}}{\partial oldsymbol{g}_{i}^{(\ell)}} \, oldsymbol{g}_{i}^{(\ell)} oldsymbol{(1-g_{i}^{(\ell)})}$$

$$\frac{\partial \mathcal{L}}{\partial g_{i}^{(L)}} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right) \frac{\partial \left( g_{i}^{(L)} - y_{i} \right)}{\partial g_{i}^{(L)}}$$

Weight and biases

$$\begin{split} &\frac{\partial \mathcal{L}}{\partial W_{i,j}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \, g_j^{(\ell-1)} \\ &\frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \end{split}$$

$$abla_{ec{b}^{(\ell)}}\mathcal{L} = 
abla_{ec{ar{e}}^{(\ell)}}\mathcal{L} \ \in \mathbb{R}^{oldsymbol{N}^{(\ell)}}$$

 $\nabla_{\mathcal{W}^{(\ell)}}\mathcal{L} =$ 

$$\in \mathbb{R}^{\textit{N}^{(\ell)} \times \textit{N}^{(\ell-1)}}$$

Activations

$$\frac{\partial \mathcal{L}}{\partial g_{j}^{(\ell-1)}} = \sum_{i} \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} W_{i,j}^{(\ell)}$$

Pre-activations

$$rac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} = rac{\partial \mathcal{L}}{\partial g_{i}^{(\ell)}} \, g_{i}^{(\ell)} oldsymbol{\left(1 - g_{i}^{(\ell)}
ight)}$$

$$\frac{\partial \mathcal{L}}{\partial g_i^{(L)}} = \frac{2}{N^{(L)}} \left( g_i^{(L)} - y_i \right)$$

Weight and biases

$$rac{\partial \mathcal{L}}{\partial W_{i,j}^{(\ell)}} = rac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} g_j^{(\ell-1)}$$

$$\frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}}$$

Activations

$$\frac{\partial \mathcal{L}}{\partial g_{j}^{(\ell-1)}} = \sum_{i} \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} W_{i,j}^{(\ell)}$$

Pre-activations

$$rac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} = rac{\partial \mathcal{L}}{\partial g_{i}^{(\ell)}} \, g_{i}^{(\ell)} oldsymbol{(1-g_{i}^{(\ell)})}$$

$$\frac{\partial \mathcal{L}}{\partial g_{i}^{(L)}} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right)$$

$$\nabla_{\vec{\textit{b}}^{(\ell)}}\mathcal{L} = \nabla_{\vec{\textit{f}}^{(\ell)}}\mathcal{L} \ \in \mathbb{R}^{\textit{N}^{(\ell)}}$$

Weight and biases

$$\frac{\partial \mathcal{L}}{\partial W_{i,j}(\ell)} = \frac{\partial \mathcal{L}}{\partial f_i(\ell)} \, \mathbf{g}_j^{(\ell-1)}$$

$$rac{\partial W_{i,j}(\ell)}{\partial b_i(\ell)} = rac{\partial \mathcal{L}}{\partial f_i(\ell)}$$

Activations

$$\frac{\partial \mathcal{L}}{\partial g_{j}^{(\ell-1)}} = \sum_{i} \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} W_{i,j}^{(\ell)}$$

$$rac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} = rac{\partial \mathcal{L}}{\partial oldsymbol{g}_{i}^{(\ell)}} \, oldsymbol{g}_{i}^{(\ell)} oldsymbol{(1-g_{i}^{(\ell)})}$$

Output activations

$$\frac{\partial \mathcal{L}}{\partial g_{i}^{(L)}} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right)$$

 $\nabla_{\textbf{\textit{W}}^{(\ell)}} \mathcal{L} = (\nabla_{\boldsymbol{\bar{\textit{f}}}(\ell)} \mathcal{L})^{\top} \; \boldsymbol{\bar{\textit{g}}}^{(\ell-1)} \; \in \mathbb{R}^{\textbf{\textit{N}}^{(\ell)} \times \textbf{\textit{N}}^{(\ell-1)}}$ 

$$abla_{ar{b}^{(\ell)}}\mathcal{L} = 
abla_{ar{f}^{(\ell)}}\mathcal{L} \ \in \mathbb{R}^{N^{(\ell)}}$$

$$abla_{ec{p}(\ell-1)}\mathcal{L} = \mathcal{W}^{(\ell) op}\left(
abla_{ec{f}(\ell)}\mathcal{L}
ight)$$

Weight and biases

$$\begin{split} \frac{\partial \mathcal{L}}{\partial W_{i,j}(\ell)} &= \frac{\partial \mathcal{L}}{\partial f_i(\ell)} \, g_j^{(\ell-1)} \\ \frac{\partial \mathcal{L}}{\partial b_i(\ell)} &= \frac{\partial \mathcal{L}}{\partial f_i(\ell)} \end{split}$$

$$abla_{ar{\mathcal{B}}^{(\ell)}}\mathcal{L} = 
abla_{ar{\mathcal{B}}^{(\ell)}}\mathcal{L} \ \in \mathbb{R}^{oldsymbol{N}^{(\ell)}}$$

 $\nabla_{\textbf{\textit{W}}^{(\ell)}} \mathcal{L} = (\nabla_{\boldsymbol{\bar{\textit{f}}}(\ell)} \mathcal{L})^{\top} \; \boldsymbol{\bar{\textit{g}}}^{(\ell-1)} \; \in \mathbb{R}^{\textbf{\textit{N}}^{(\ell)} \times \textbf{\textit{N}}^{(\ell-1)}}$ 

Activations

$$\frac{\partial \mathcal{L}}{\partial g_{j}^{(\ell-1)}} = \sum_{i} \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} W_{i,j}^{(\ell)}$$

$$abla_{ar{g}^{(\ell-1)}}\mathcal{L} = \mathit{W}^{(\ell) op}\left(
abla_{ar{f}^{(\ell)}}\mathcal{L}
ight)$$

Pre-activations

$$rac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} = rac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} \, g_i^{(\ell)} ig(1 - g_i^{(\ell)}ig)$$

$$\nabla_{\vec{\textit{f}}^{(\ell)}}\!\mathcal{L} =$$

$$\frac{\partial \mathcal{L}}{\partial g_i^{(L)}} = \frac{2}{N^{(L)}} \left( g_i^{(L)} - y_i \right)$$

Weight and biases

$$\begin{split} &\frac{\partial \mathcal{L}}{\partial W_{i,j}(\ell)} = \frac{\partial \mathcal{L}}{\partial f_i(\ell)} \, \mathbf{g}_j^{(\ell-1)} \\ &\frac{\partial \mathcal{L}}{\partial \mathbf{b}_i(\ell)} = \frac{\partial \mathcal{L}}{\partial f_i(\ell)} \end{split}$$

$$abla_{ar{m{eta}}^{(\ell)}} \mathcal{L} = 
abla_{ar{m{ar{f}}}^{(\ell)}} \mathcal{L} \ \in \mathbb{R}^{m{\mathsf{N}}^{(\ell)}}$$

Activations

$$\frac{\partial \mathcal{L}}{\partial g_{j}^{(\ell-1)}} = \sum_{i} \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} W_{i,j}^{(\ell)}$$

$$abla_{ar{g}^{(\ell-1)}}\mathcal{L} = \mathit{W}^{(\ell) op}\left(
abla_{ar{f}^{(\ell)}}\mathcal{L}
ight)$$

Pre-activations

$$rac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} = rac{\partial \mathcal{L}}{\partial g_{i}^{(\ell)}} g_{i}^{(\ell)} (1 - g_{i}^{(\ell)})$$

$$abla_{ar{\mathbf{f}}^{(\ell)}}\mathcal{L} = \left(
abla_{ar{\mathbf{g}}^{(\ell)}}\mathcal{L}
ight)\odotar{\mathbf{g}}^{(\ell)}\odot\left(1-ar{\mathbf{g}}^{(\ell)}
ight)$$

 $\nabla_{\textbf{W}^{(\ell)}} \mathcal{L} = (\nabla_{\boldsymbol{\bar{r}}^{(\ell)}} \mathcal{L})^{\top} \ \boldsymbol{\bar{g}}^{(\ell-1)} \ \boldsymbol{\in} \ \mathbb{R}^{\textbf{N}^{(\ell)} \times \textbf{N}^{(\ell-1)}}$ 

$$\frac{\partial \mathcal{L}}{\partial g_{i}^{(L)}} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right)$$

Weight and biases

$$\frac{\partial \mathcal{L}}{\partial W_{i,j}(\ell)} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} g_j^{(\ell-1)}$$
$$\frac{\partial \mathcal{L}}{\partial b^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f^{(\ell)}}$$

$$abla_{ar{m{eta}}^{(\ell)}} \mathcal{L} = 
abla_{ar{m{ar{m{\ell}}}}^{(\ell)}} \mathcal{L} \ \in \mathbb{R}^{m{N}^{(\ell)}}$$

Activations

$$\frac{\partial \mathcal{L}}{\partial g_{i}^{(\ell-1)}} = \sum_{i} \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}} W_{i,j}^{(\ell)}$$

$$abla_{ar{oldsymbol{g}}^{(\ell-1)}}\mathcal{L} = oldsymbol{W}^{(\ell) op}\left(
abla_{ar{oldsymbol{f}}^{(\ell)}}\mathcal{L}
ight)$$

Pre-activations

$$\frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} g_i^{(\ell)} (1 - g_i^{(\ell)})$$

$$abla_{ar{m{g}}^{(\ell)}}\mathcal{L} = \left(
abla_{ar{m{g}}^{(\ell)}}\mathcal{L}
ight)\odotar{m{g}}^{(\ell)}\odot\left(1-ar{m{g}}^{(\ell)}
ight)$$

 $\nabla_{\boldsymbol{W}^{(\ell)}} \mathcal{L} = (\nabla_{\boldsymbol{\bar{r}}^{(\ell)}} \mathcal{L})^{\top} \; \boldsymbol{\bar{g}}^{(\ell-1)} \; \in \mathbb{R}^{\boldsymbol{N}^{(\ell)} \times \boldsymbol{N}^{(\ell-1)}}$ 

$$\frac{\partial \mathcal{L}}{\partial \sigma_{i}(L)} = \frac{2}{N^{(L)}} \left( g_{i}^{(L)} - y_{i} \right)$$

$$abla_{ar{oldsymbol{g}}^{(L)}} \mathcal{L} = rac{2}{oldsymbol{N}^{(L)}} \left(ar{oldsymbol{g}}^{(L)} - ar{oldsymbol{y}}
ight)$$