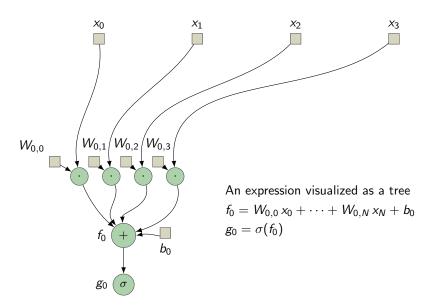
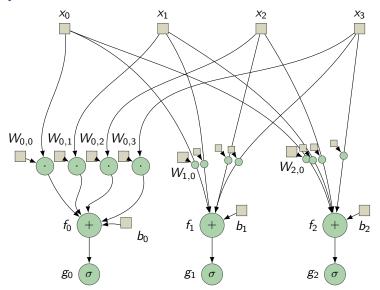
A neuron



A layer of neurons

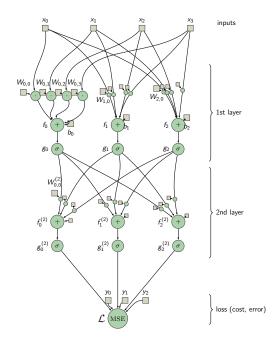


An expression with re-used values gives a DAG (directed acyclic graph).

A neural network

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

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

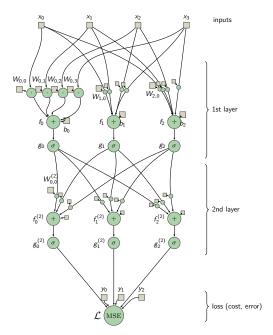


A neural network

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

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

$$\mathcal{L} = rac{1}{N^{(2)}} \sum_{i} (g_{i}^{(2)} - y_{i})^{2}$$
 where $g_{i}^{(2)} = \sigma(f_{i}^{(2)})$, ...



Gradients

Gradient = vector 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

Gradient = vector 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 ε increases f by $\approx a\varepsilon$

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

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 ε 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 ε increases f by $a_f \varepsilon =: \varepsilon'$,

which increases g by $a_g \varepsilon'$. 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 ε 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 ε increases f by $a_f \varepsilon =: \varepsilon'$,

which increases g by $a_g \varepsilon'$. 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 ε 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 ε increases f by $a_f \varepsilon =: \varepsilon'$,

which increases g by $a_g \varepsilon'$. 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 ε 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 ε increases f by $a_f \varepsilon =: \varepsilon'$,

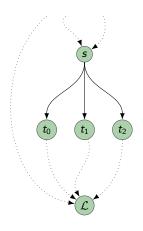
which increases g by $a_g \varepsilon'$. So $\frac{\partial g}{\partial x} = a_g a_f$. This proves $\frac{\partial g}{\partial x} = \frac{\partial g}{\partial f} \frac{\partial f}{\partial x}$.

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

Derivates – in a graph

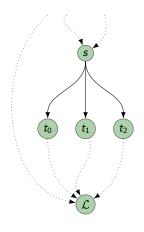


$$\frac{\partial \mathcal{L}}{\partial s} = \sum_{t \in \text{out}(s)} \frac{\partial \mathcal{L}}{\partial t} \cdot \frac{\partial t}{\partial s}$$

Because increasing s by ε 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

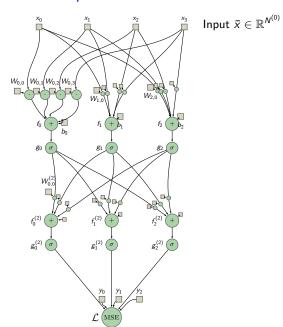


$$\frac{\partial \mathcal{L}}{\partial s} = \sum_{t \in \text{out}(s)} \frac{\partial \mathcal{L}}{\partial t} \cdot \frac{\partial t}{\partial s}$$

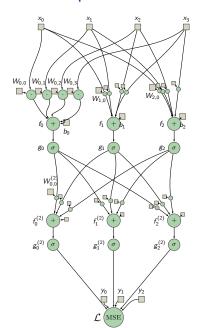
Because increasing s by ε increases each t_i by $\frac{\partial t_i}{\partial s}\varepsilon$.

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 (that is, $\frac{\partial t_i}{\partial s}$).



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



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

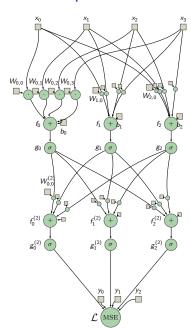
 $x_0,\ldots,x_{N^{(0)}-1}$

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

Pre-activations

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

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



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

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

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

Pre-activations

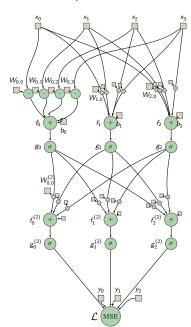
$$ar{f}^{(1)} = W^{(1)} ar{x} + ar{b}^{(1)} \in \mathbb{R}^{N^{(1)}}$$

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

Activations

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

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



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

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

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

Pre-activations

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

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

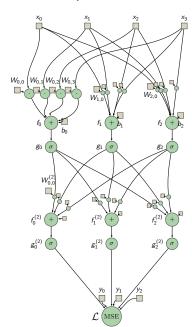
Activations

$$oldsymbol{g}^{(1)} \equiv \sigma(oldsymbol{f}^{(1)})$$

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

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

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



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

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

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

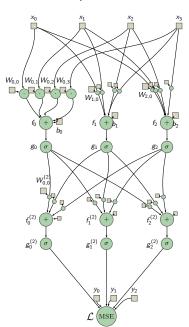
$$g^{(1)} = \sigma(\bar{f}^{(1)})$$

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

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

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

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



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

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

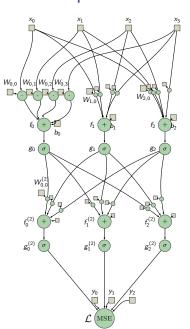
Activations

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

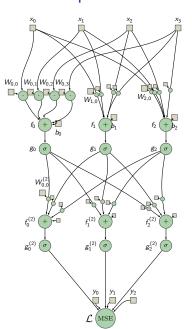
$$\bar{f}^{(2)} = W^{(2)}\bar{g}^{(1)} + \bar{b}^{(2)} \in \mathbb{R}^{N^{(2)}}$$

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

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



import numpy as np

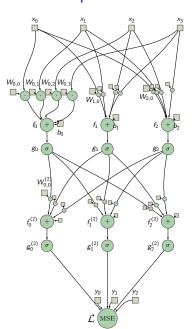


import numpy as np

When batched:

x has shape $(B, N^{(0)})$

g has shape before iteration i

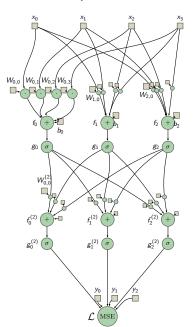


import numpy as np

When batched:

x has shape $(B, N^{(0)})$

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

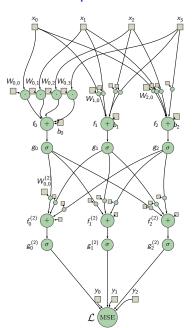


```
import numpy as np
```

```
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 $(N^{(i+1)}, N^{(i)})$
- w @ g.T has shape



import numpy as np

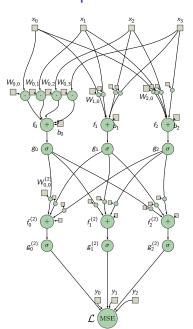
When batched:

x has shape $(B, N^{(0)})$

g has shape $(B, N^{(i)})$ before iteration i

w has shape $(N^{(i+1)}, N^{(i)})$

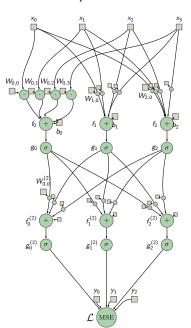
 $w \otimes g.T$ has shape $(N^{(i+1)}, B)$



import numpy as np

When batched:

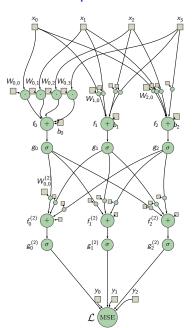
x has shape $(B, N^{(0)})$ g has shape $(B, N^{(i)})$ before iteration i w has shape $(N^{(i+1)}, N^{(i)})$ w @ g.T has shape $(N^{(i+1)}, B)$ (w @ g.T).T has shape $(B, N^{(i+1)})$



import numpy as np

When batched:

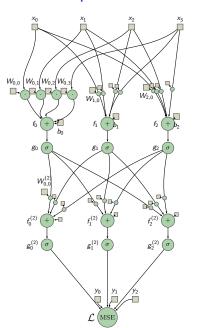
x has shape $(B, N^{(0)})$ g has shape $(B, N^{(i)})$ before iteration i w has shape $(N^{(i+1)}, N^{(i)})$ $w \otimes g.T$ has shape $(N^{(i+1)}, B)$ (w @ g.T).T has shape $(B, N^{(i+1)})$ ъ has shape



import numpy as np

When batched:

when batched: \mathbf{x} has shape $(B, N^{(0)})$ \mathbf{g} has shape $(B, N^{(i)})$ before iteration \mathbf{i} \mathbf{w} has shape $(N^{(i+1)}, N^{(i)})$ \mathbf{w} @ \mathbf{g} .T has shape $(N^{(i+1)}, B)$ $(\mathbf{w}$ @ \mathbf{g} .T).T has shape $(B, N^{(i+1)})$ \mathbf{b} has shape $(N^{(i+1)})$

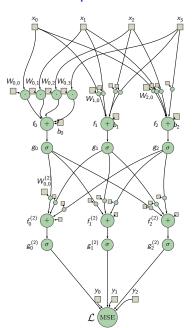


import numpy as np

When batched:

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

When batched:

x has shape $(B, N^{(0)})$

g has shape $(B, N^{(i)})$ before iteration i

w has shape $(N^{(i+1)}, N^{(i)})$

 $w \otimes g.T$ has shape $(N^{(i+1)}, B)$

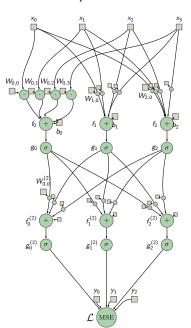
(w @ g.T).T has shape $(B, N^{(i+1)})$

b has shape $(N^{(i+1)})$

f = (w @ g.T).T + b

(with b broadcasted along the batch)

@ is the same as np.matmul



import numpy as np

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 $(N^{(i+1)})$

@ is the same as np.matmul loss = ((g - y) ** 2).mean() over outputs and batch, $\frac{1}{B.M^{(2)}}$

(with b broadcasted along the batch)

f = (w @ g.T).T + b

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

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

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

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

$$egin{align} egin{align} egin{align} egin{align} egin{align} egin{align} egin{align} egin{align} egin{align} rac{\partial \mathcal{L}}{\partial oldsymbol{g}_i^{(L)}} = egin{align} egin{align} \mathcal{L} & = rac{1}{M^{(L)}} \sum_i (oldsymbol{g}_i^{(L)} - y_i)^2 \ \end{pmatrix} \end{aligned}$$

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

$$\begin{split} \boldsymbol{g_i}^{(\ell)} &= \sigma(f_i^{(\ell)}) \\ &\frac{\partial \mathcal{L}}{\partial \boldsymbol{g_i}^{(L)}} = \frac{2}{N^{(L)}} \left(\boldsymbol{g_i}^{(L)} - \boldsymbol{y_i} \right) \frac{\partial \left(\boldsymbol{g_i}^{(L)} - \boldsymbol{y_i} \right)}{\partial \boldsymbol{g_i}^{(L)}} \\ \mathcal{L} &= \frac{1}{N^{(L)}} \sum_i (\boldsymbol{g_i}^{(L)} - \boldsymbol{y_i})^2 \end{split}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

$$rac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} = rac{\partial \mathcal{L}}{\partial oldsymbol{g}_i^{(\ell)}} oldsymbol{g}_i^{(\ell)} oldsymbol{(1-oldsymbol{g}_i^{(\ell)})}$$

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

Weight and biases

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

$$\frac{\partial \mathcal{L}}{\partial b_{i}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_{i}^{(\ell)}}$$

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

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

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

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

Weight and biases

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

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

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

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

Weight and biases

$$\frac{\partial \mathcal{L}}{\partial \textit{W}_{\textit{i},\textit{j}}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial \textit{f}_{\textit{i}}^{(\ell)}} \, \textit{g}_{\textit{j}}^{(\ell-1)}$$

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

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

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

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

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

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

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

Weight and biases

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

$$\frac{\partial \mathcal{L}}{\partial \textit{b}_{\textit{i}}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial \textit{f}_{\textit{i}}^{(\ell)}}$$

Activations

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

Pre-activations

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

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

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

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

Weight and biases

$$\frac{\partial \mathcal{L}}{\partial \textit{W}_{\textit{i},\textit{j}}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial \textit{f}_{\textit{i}}^{(\ell)}} \, \textit{g}_{\textit{j}}^{(\ell-1)}$$

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

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

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

Activations

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

$$abla_{ar{g}^{(\ell-1)}}\mathcal{L} = W^{(\ell)\top} \left(
abla_{ar{f}^{(\ell)}}\mathcal{L} \right)$$

Pre-activations

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

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

Weight and biases

 $\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{f}^{(\ell)}}$

$$\frac{\partial \mathcal{L}}{\partial \textit{W}_{\textit{i},\textit{j}}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial \textit{f}_{\textit{i}}^{(\ell)}} \, \textit{g}_{\textit{j}}^{(\ell-1)}$$

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

 $\boldsymbol{\nabla}_{\textit{W}^{(\ell)}} \mathcal{L} = (\boldsymbol{\nabla}_{\vec{\textbf{f}}^{\,\ell}}) \mathcal{L})^\top \ \vec{\textbf{g}}^{(\ell-1)} \ \in \mathbb{R}^{\textit{N}^{(\ell)} \times \textit{N}^{(\ell-1)}}$

Activations

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

$$abla_{ar{g}^{(\ell-1)}}\mathcal{L} = W^{(\ell)\top} \left(
abla_{ar{f}^{(\ell)}} \mathcal{L} \right)$$

Pre-activations

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

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

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

Weight and biases

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

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}_{i}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{f}_{i}^{(\ell)}} \qquad \qquad \nabla_{\vec{\mathbf{b}}^{(\ell)}} \mathcal{L} = \nabla_{\vec{\mathbf{f}}^{(\ell)}} \mathcal{L} \ \in \mathbb{R}^{N^{(\ell)}}$$

Activations

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

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

Pre-activations

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

$$abla_{ec{f}^{(\ell)}}\mathcal{L} = \left(
abla_{ec{g}^{(\ell)}}\mathcal{L}
ight)\odotec{g}^{(\ell)}\odot\left(1-ec{g}^{(\ell)}
ight)$$

 $\boldsymbol{\nabla}_{\textit{W}^{(\ell)}} \mathcal{L} = (\boldsymbol{\nabla}_{\vec{\textbf{f}}^{\,\ell}}) \mathcal{L})^\top \ \vec{\textbf{g}}^{(\ell-1)} \ \in \mathbb{R}^{\textit{N}^{(\ell)} \times \textit{N}^{(\ell-1)}}$

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

Weight and biases

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

$$\boldsymbol{\nabla}_{\boldsymbol{W}^{(\ell)}} \mathcal{L} = (\boldsymbol{\nabla}_{\vec{\boldsymbol{f}}^{(\ell)}} \mathcal{L})^\top \ \boldsymbol{\bar{\boldsymbol{g}}}^{(\ell-1)} \ \in \mathbb{R}^{\boldsymbol{N}^{(\ell)} \times \boldsymbol{N}^{(\ell-1)}}$$

$$\frac{\partial \mathcal{L}}{\partial \textit{b}_{\textit{i}}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial \textit{f}_{\textit{i}}^{(\ell)}}$$

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

Activations

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

$$abla_{ar{g}^{(\ell-1)}}\mathcal{L} = W^{(\ell)\top} \left(
abla_{ar{f}^{(\ell)}} \mathcal{L} \right)$$

Pre-activations

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

$$\nabla_{\vec{\mathbf{f}}^{(\ell)}}\mathcal{L} = \, \left(\nabla_{\overline{\mathbf{g}}^{(\ell)}}\mathcal{L}\right)\odot\overline{\mathbf{g}}^{(\ell)}\odot(\mathbf{1}-\overline{\mathbf{g}}^{(\ell)})$$

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

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