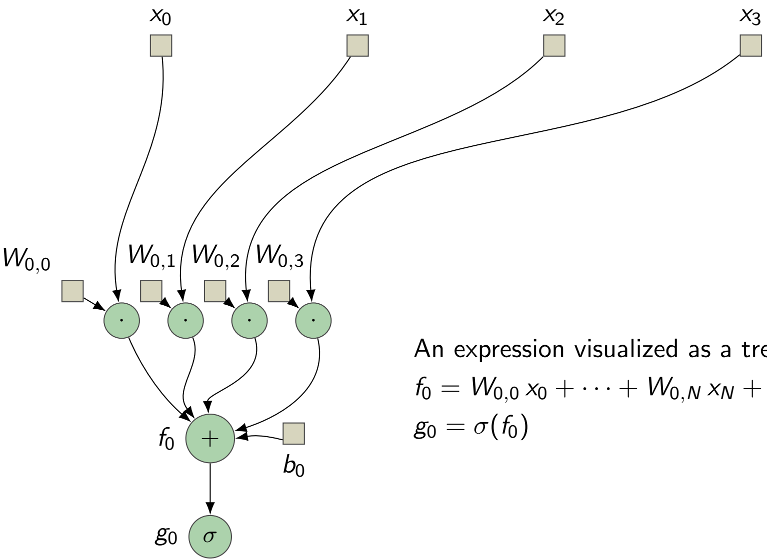
A neuron

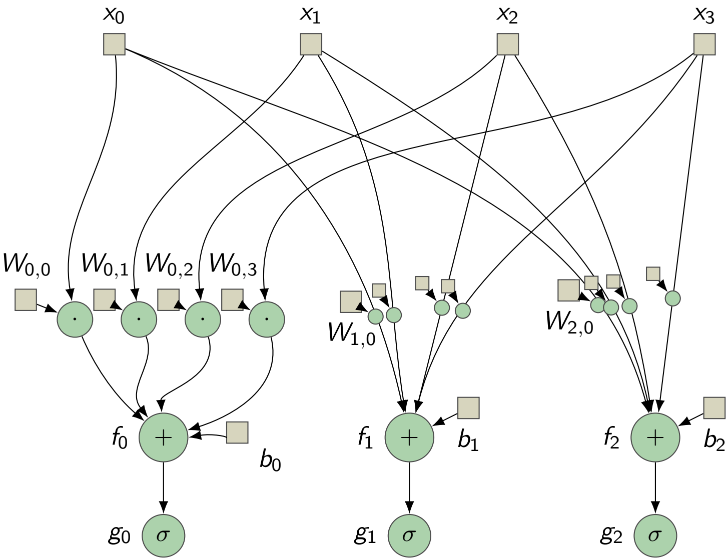


An expression visualized as a tree

$f_0 = W_{0,0}\, x_0 + \cdots + W_{0,N}\, x_N + b_0$

$g_0 = \sigma(f_0)$

A layer of neurons

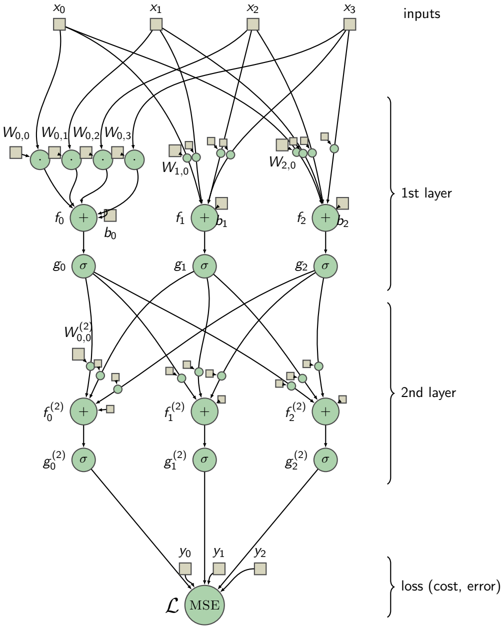


An expression with re-used values gives a DAG (directed acyclic graph).

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

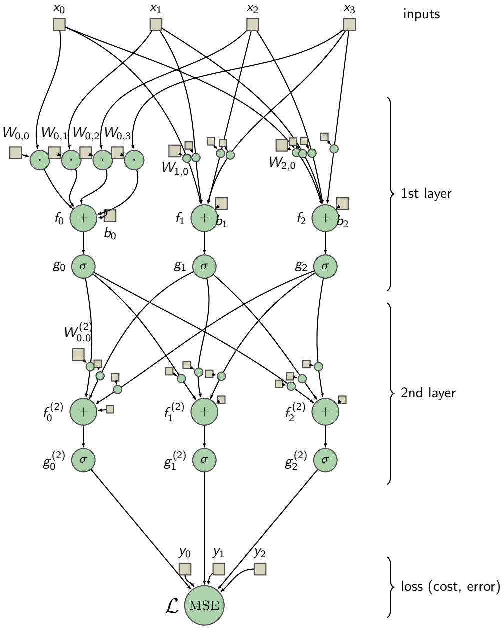
For computing gradients,
a neural network =
one big expression for the loss.

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

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

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

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

Gradients

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

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

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

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

Derivates – chain rule

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

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

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

Derivates – chain rule

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

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

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

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

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

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

Derivates – chain rule

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

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

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

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

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 $\frac{\partial g}{\partial x} = \frac{\partial g}{\partial f}\frac{\partial f}{\partial x}$.

Derivates – chain rule

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

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

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

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

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.

Derivates – chain rule

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

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

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

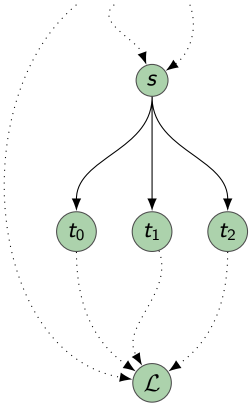
Let $\quad g = g(f(x, y, z)), \quad \frac{\partial f}{\partial x} = a_f, \quad \frac{\partial g}{\partial f} = a_g.$

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

Derivates – in a graph



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

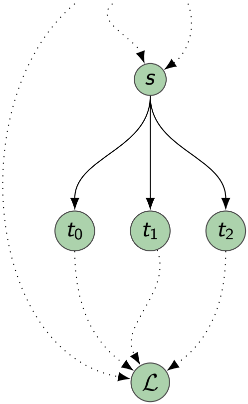
Because increasing s by ε

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



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

Because increasing s by ε

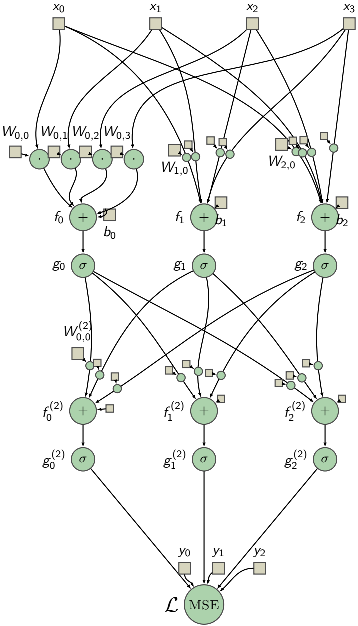
increases each t_i by $\frac{\partial t_i}{\partial s}\varepsilon$.

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, $\frac{\partial t_i}{\partial s}$).

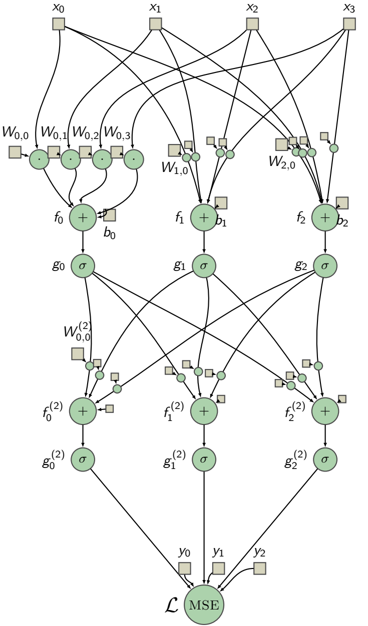
Forward pass



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

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

Forward pass



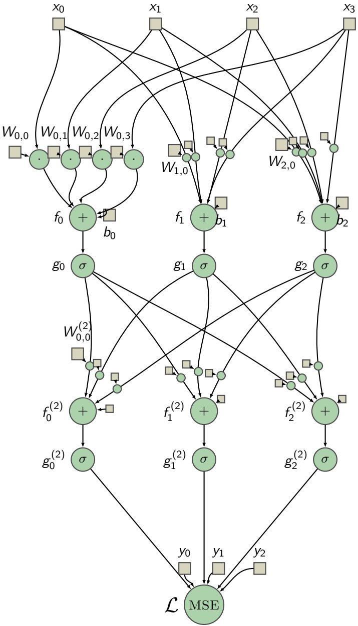
Input $\bar{x} \in \mathbb{R}^{N^{(0)}}$ $\quad x_0, \ldots, x_{N^{(0)}-1}$

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

Pre-activations

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

Forward pass



Input $\bar{x} \in \mathbb{R}^{N^{(0)}}$ $\qquad x_0, \ldots, x_{N^{(0)}-1}$

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

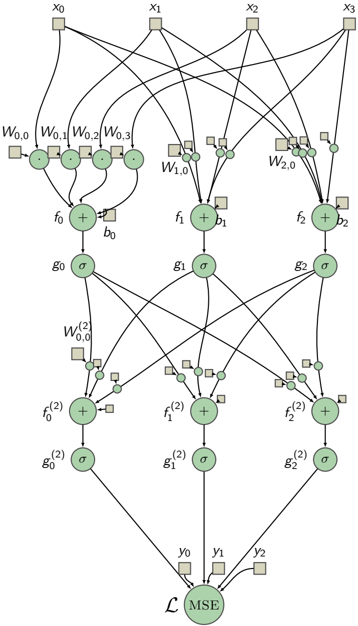
Pre-activations

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

Activations

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

Forward pass



Input $\bar{x} \in \mathbb{R}^{N^{(0)}}$ $\qquad x_0, \ldots, x_{N^{(0)}-1}$

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

Pre-activations

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

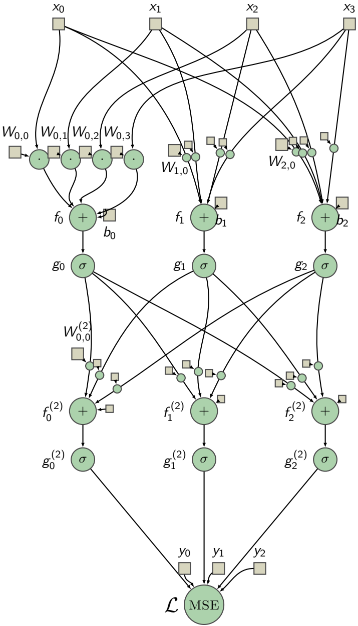
Activations

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

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

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

Forward pass



Input $\bar{x} \in \mathbb{R}^{N^{(0)}}$ $\qquad x_0, \ldots, x_{N^{(0)}-1}$

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

Pre-activations

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

Activations

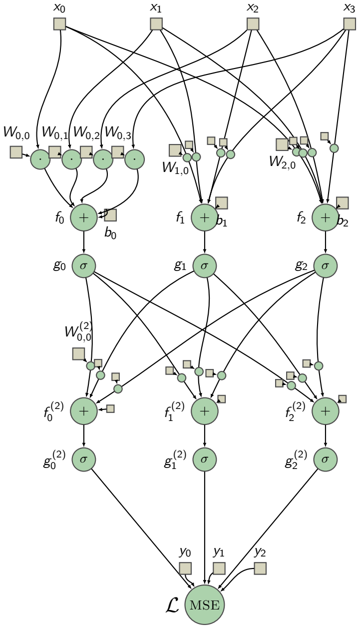
$$\bar{g}^{(1)} = \sigma(\bar{f}^{(1)}) \qquad g_i^{(1)} = \sigma(f_i^{(1)})$$

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

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

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

Forward pass



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

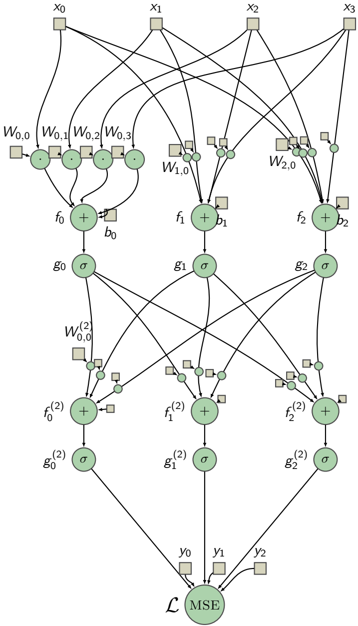
$\bar{g}^{(1)} = \sigma(\bar{f}^{(1)})$

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

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

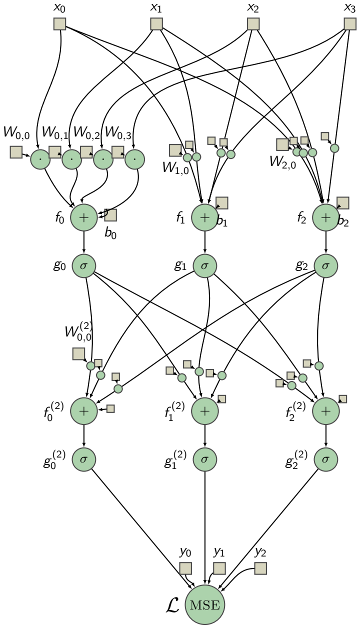
$\mathcal{L} = \mathrm{mean}\left[(\bar{g}^{(2)} - \bar{y})^2\right]$

Forward pass – code




```
import numpy as np

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

Forward pass – code



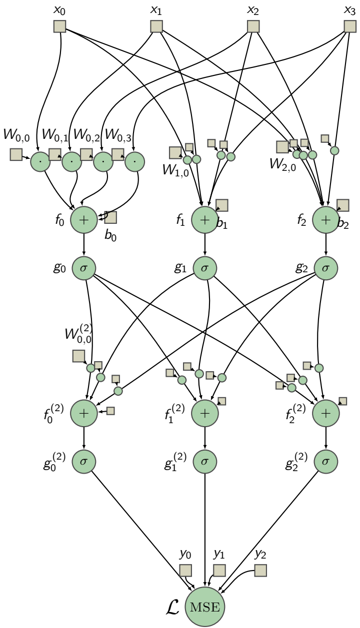

```
import numpy as np

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 before iteration i

Forward pass – code




```
import numpy as np

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

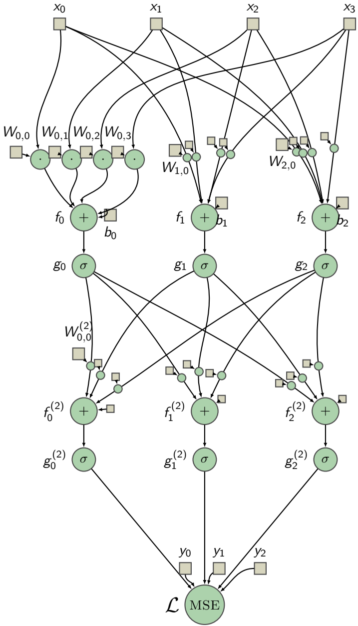
When batched:

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

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

`w` has shape

Forward pass – code




```
import numpy as np

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

When batched:

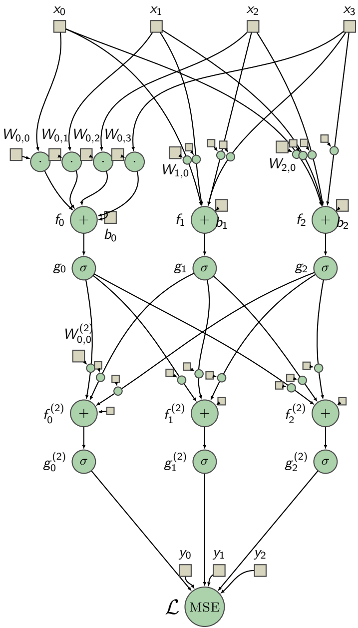
`x` has shape $(B, N^{(0)})$

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

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

`w @ g.T` has shape

Forward pass – code




```
import numpy as np

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

When batched:

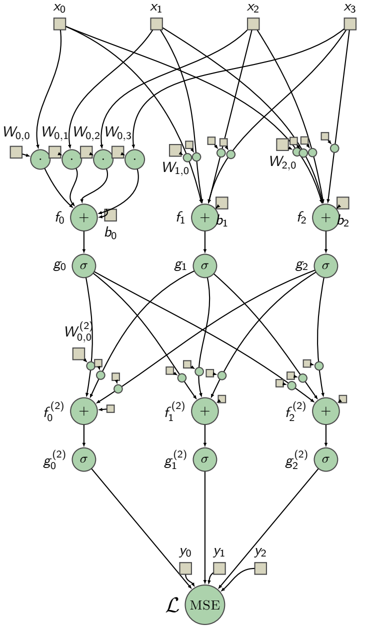
`x` has shape $(B, N^{(0)})$

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

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

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

Forward pass – code




```
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 $(B, N^{(0)})$

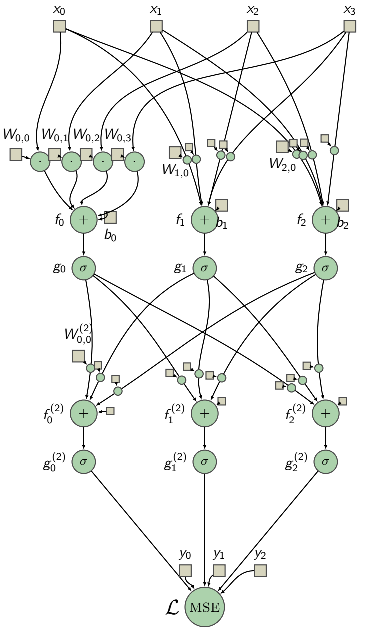
`g` has shape $(B, N^{(i)})$ before iteration `i`

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

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

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

Forward pass – code




```
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`

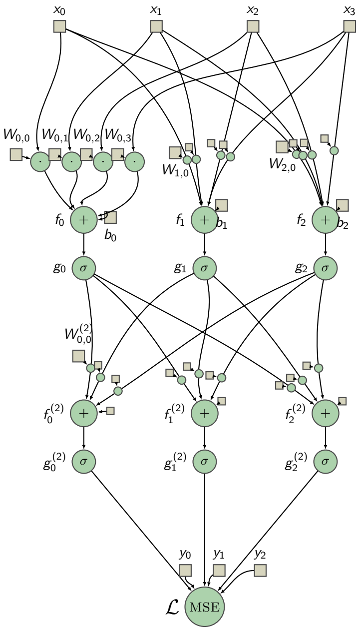
`w` has shape $(N^{(i+1)}, N^{(i)})$

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

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

`b` has shape

Forward pass – code




```
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`

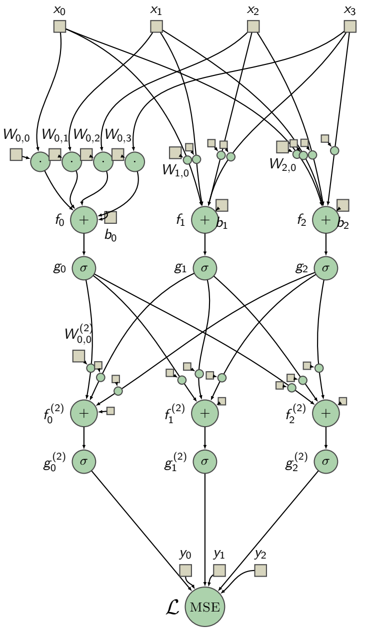
`w` has shape $(N^{(i+1)}, N^{(i)})$

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

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

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

Forward pass – code




```
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 $(N^{(i+1)}, B)$

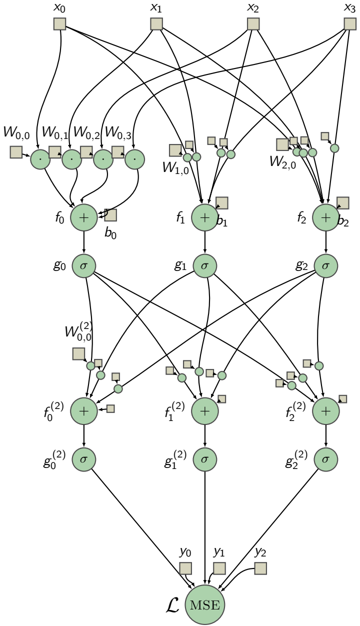
`(w @ g.T).T` has shape $(B, N^{(i+1)})$

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

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

(with `b` broadcasted along the batch)

Forward pass – code




```
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 $(N^{(i+1)}, B)$

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

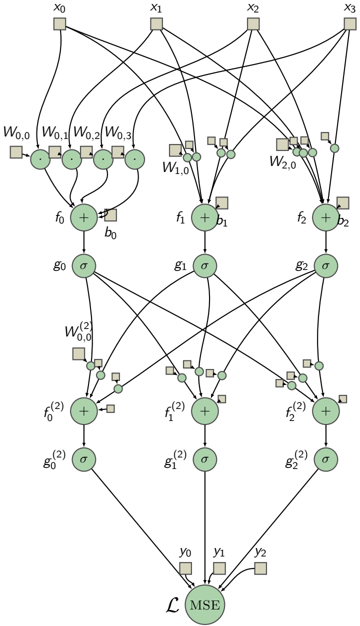
`b` has shape $(N^{(i+1)})$

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

(with `b` broadcasted along the batch)

`@` is the same as `np.matmul`

Forward pass – code




```python
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 $(N^{(i+1)}, B)$

`(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`

`loss = ((g - y) ** 2).mean()`

over outputs and batch, $\frac{1}{B \cdot N^{(2)}}$

Backward pass

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

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

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

Backward pass

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

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

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

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

Backward pass

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

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

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

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

Backward pass

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

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

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

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

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

Backward pass

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

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

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

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

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

Backward pass

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

$$\frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} \frac{\partial g_i^{(\ell)}}{\partial f_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} \sigma'(f_i^{(\ell)}) =$$

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

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

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

Backward pass

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

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

$$\mathcal{L} = \tfrac{1}{N^{(L)}} \sum_i (g_i^{(L)} - y_i)^2$$

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

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

Backward pass

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

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

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

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

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

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

Backward pass

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

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

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

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

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

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

Backward pass

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

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

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

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

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

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

Backward pass

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

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

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

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

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

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

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

Backward pass

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

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

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

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

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

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

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

Backward pass

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

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

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

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

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

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

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

Backward pass

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

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

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

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

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

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

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

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

Backward pass

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

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

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

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

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

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

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

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

Backward pass

Weight and biases

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

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

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

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

Output activations

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

Backward pass

Weight and biases

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

$$\frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \qquad \nabla_{\bar{b}^{(\ell)}} \mathcal{L} = \nabla_{\bar{f}^{(\ell)}} \mathcal{L} \ \in \mathbb{R}^{N^{(\ell)}}$$

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

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

Output activations

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

Backward pass

Weight and biases

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

$$\frac{\partial \mathcal{L}}{\partial b_i{}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i{}^{(\ell)}} \qquad\qquad \nabla_{\bar{b}^{(\ell)}}\mathcal{L} = \nabla_{\bar{f}^{(\ell)}}\mathcal{L} \;\in \mathbb{R}^{N^{(\ell)}}$$

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

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

Output activations

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

Backward pass

Weight and biases

$$\frac{\partial \mathcal{L}}{\partial W_{i,j}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} g_j^{(\ell-1)} \qquad \nabla_{W^{(\ell)}} \mathcal{L} = \qquad \in \mathbb{R}^{N^{(\ell)} \times N^{(\ell-1)}}$$

$$\frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \qquad \nabla_{\bar{b}^{(\ell)}} \mathcal{L} = \nabla_{\bar{f}^{(\ell)}} \mathcal{L} \ \in \mathbb{R}^{N^{(\ell)}}$$

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

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

Output activations

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

Backward pass

Weight and biases

$$\frac{\partial \mathcal{L}}{\partial W_{i,j}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} g_j^{(\ell-1)} \qquad \nabla_{W^{(\ell)}} \mathcal{L} = \left(\nabla_{\bar{f}^{(\ell)}} \mathcal{L}\right)^\top \bar{g}^{(\ell-1)} \in \mathbb{R}^{N^{(\ell)} \times N^{(\ell-1)}}$$

$$\frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \qquad \nabla_{\bar{b}^{(\ell)}} \mathcal{L} = \nabla_{\bar{f}^{(\ell)}} \mathcal{L} \in \mathbb{R}^{N^{(\ell)}}$$

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

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

Output activations

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

Backward pass

Weight and biases

$$\frac{\partial \mathcal{L}}{\partial W_{i,j}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} g_j^{(\ell-1)} \qquad \nabla_{W^{(\ell)}} \mathcal{L} = (\nabla_{\bar{f}^{(\ell)}} \mathcal{L})^\top \ \bar{g}^{(\ell-1)} \ \in \mathbb{R}^{N^{(\ell)} \times N^{(\ell-1)}}$$

$$\frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \qquad \nabla_{\bar{b}^{(\ell)}} \mathcal{L} = \nabla_{\bar{f}^{(\ell)}} \mathcal{L} \ \in \mathbb{R}^{N^{(\ell)}}$$

Activations

$$\frac{\partial \mathcal{L}}{\partial g_j^{(\ell-1)}} = \sum_i \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} W_{i,j}^{(\ell)} \qquad \nabla_{\bar{g}^{(\ell-1)}} \mathcal{L} = W^{(\ell)\top} (\nabla_{\bar{f}^{(\ell)}} \mathcal{L})$$

Pre-activations

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

Output activations

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

Backward pass

Weight and biases

$$\frac{\partial \mathcal{L}}{\partial W_{i,j}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} g_j^{(\ell-1)} \qquad \nabla_{W^{(\ell)}} \mathcal{L} = (\nabla_{\bar{f}^{(\ell)}} \mathcal{L})^\top \ \bar{g}^{(\ell-1)} \ \in \mathbb{R}^{N^{(\ell)} \times N^{(\ell-1)}}$$

$$\frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \qquad \nabla_{\bar{b}^{(\ell)}} \mathcal{L} = \nabla_{\bar{f}^{(\ell)}} \mathcal{L} \ \in \mathbb{R}^{N^{(\ell)}}$$

Activations

$$\frac{\partial \mathcal{L}}{\partial g_j^{(\ell-1)}} = \sum_i \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} W_{i,j}^{(\ell)} \qquad \nabla_{\bar{g}^{(\ell-1)}} \mathcal{L} = W^{(\ell)\top} (\nabla_{\bar{f}^{(\ell)}} \mathcal{L})$$

Pre-activations

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

Output activations

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

Backward pass

Weight and biases

$$\frac{\partial \mathcal{L}}{\partial W_{i,j}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} g_j^{(\ell-1)} \qquad \nabla_{W^{(\ell)}} \mathcal{L} = (\nabla_{\bar{f}^{(\ell)}} \mathcal{L})^\top \; \bar{g}^{(\ell-1)} \;\; \in \mathbb{R}^{N^{(\ell)} \times N^{(\ell-1)}}$$

$$\frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \qquad \nabla_{\bar{b}^{(\ell)}} \mathcal{L} = \nabla_{\bar{f}^{(\ell)}} \mathcal{L} \;\; \in \mathbb{R}^{N^{(\ell)}}$$

Activations

$$\frac{\partial \mathcal{L}}{\partial g_j^{(\ell-1)}} = \sum_i \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} W_{i,j}^{(\ell)} \qquad \nabla_{\bar{g}^{(\ell-1)}} \mathcal{L} = W^{(\ell)\top} (\nabla_{\bar{f}^{(\ell)}} \mathcal{L})$$

Pre-activations

$$\frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}} g_i^{(\ell)} (1 - g_i^{(\ell)}) \qquad \nabla_{\bar{f}^{(\ell)}} \mathcal{L} = \left(\nabla_{\bar{g}^{(\ell)}} \mathcal{L}\right) \odot \bar{g}^{(\ell)} \odot (1 - \bar{g}^{(\ell)})$$

Output activations

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

Backward pass

Weight and biases

$$\frac{\partial \mathcal{L}}{\partial W_{i,j}^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}}\, g_j^{(\ell-1)} \qquad \nabla_{W^{(\ell)}} \mathcal{L} = (\nabla_{\bar{f}^{(\ell)}} \mathcal{L})^\top \ \bar{g}^{(\ell-1)} \ \in \mathbb{R}^{N^{(\ell)} \times N^{(\ell-1)}}$$

$$\frac{\partial \mathcal{L}}{\partial b_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} \qquad \nabla_{\bar{b}^{(\ell)}} \mathcal{L} = \nabla_{\bar{f}^{(\ell)}} \mathcal{L} \ \in \mathbb{R}^{N^{(\ell)}}$$

Activations

$$\frac{\partial \mathcal{L}}{\partial g_j^{(\ell-1)}} = \sum_i \frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} W_{i,j}^{(\ell)} \qquad \nabla_{\bar{g}^{(\ell-1)}} \mathcal{L} = W^{(\ell)\top} (\nabla_{\bar{f}^{(\ell)}} \mathcal{L})$$

Pre-activations

$$\frac{\partial \mathcal{L}}{\partial f_i^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial g_i^{(\ell)}}\, g_i^{(\ell)} (1 - g_i^{(\ell)}) \qquad \nabla_{\bar{f}^{(\ell)}} \mathcal{L} = \left(\nabla_{\bar{g}^{(\ell)}} \mathcal{L}\right) \odot \bar{g}^{(\ell)} \odot (1 - \bar{g}^{(\ell)})$$

Output activations

$$\frac{\partial \mathcal{L}}{\partial g_i^{(L)}} = \tfrac{2}{N^{(L)}} \left(g_i^{(L)} - y_i\right) \qquad \nabla_{\bar{g}^{(L)}} \mathcal{L} = \tfrac{2}{N^{(L)}} \left(\bar{g}^{(L)} - \bar{y}\right)$$