Deep Learning for Data Science
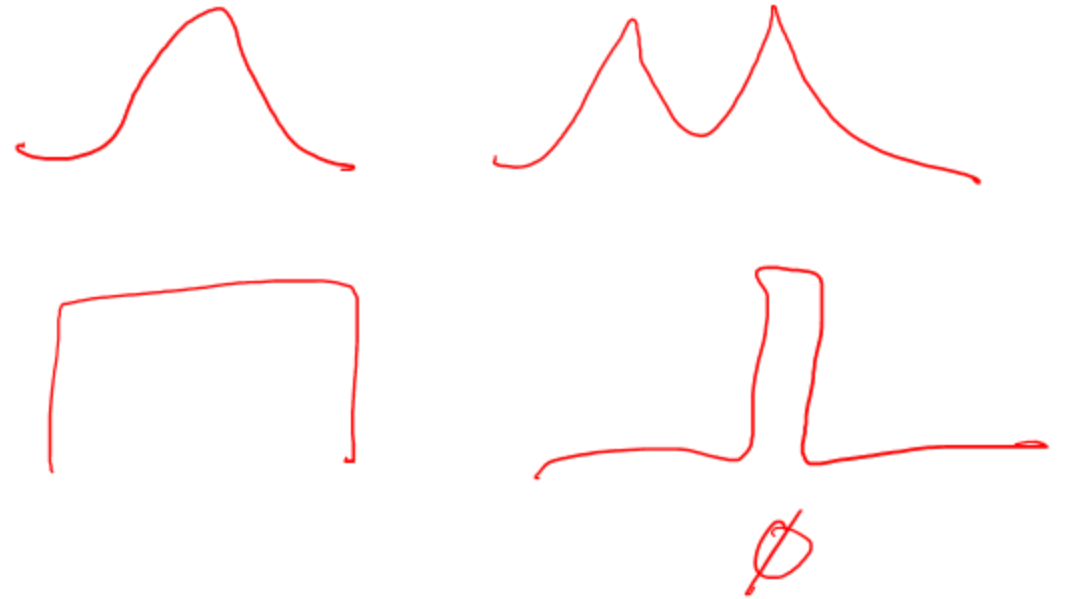# DS 542

https://dl4ds.github.io/fa2025/

Backpropagation

Plan for Today

- Motivation for backpropagation
- Intuition for backpropagation
- Toy model
- Matrix calculus
- Neural network forward pass
- Neural network backward pass

How do we efficiently compute the gradient over deep networks?

Loss function

- Training dataset of I pairs of input/output examples:

$$\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{I}$$

- Loss function or cost function measures how bad model is:

$$L\big[\boldsymbol{\phi}, \mathrm{f}\big[\mathbf{x}_i, \boldsymbol{\phi}\big], \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{I}\big]$$

or for short:

$$L\big[\boldsymbol{\phi}\big]$$

Returns a scalar that is smaller when model maps inputs to outputs better

Gradient descent algorithm

Step 1. Compute the derivatives of the loss with respect to the parameters:

$$\frac{\partial L}{\partial \phi} = \begin{bmatrix} \frac{\partial L}{\partial \phi_0} \\ \frac{\partial L}{\partial \phi_1} \\ \vdots \\ \frac{\partial L}{\partial \phi_N} \end{bmatrix}.$$

Also notated as $\nabla_w L$

Step 2. Update the parameters according to the rule:

$$\phi \longleftarrow \phi - \alpha \frac{\partial L}{\partial \phi},$$

where the positive scalar α determines the magnitude of the change.

But so far, we looked at simple models that were easy to calculate gradients

For example, linear, 1-layer models.

$$L[\phi] \;=\; \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} (\mathrm{f}[x_i, \phi] - y_i)^2$$

$$= \sum_{i=1}^{I} (\phi_0 + \phi_1 x_i - y_i)^2$$

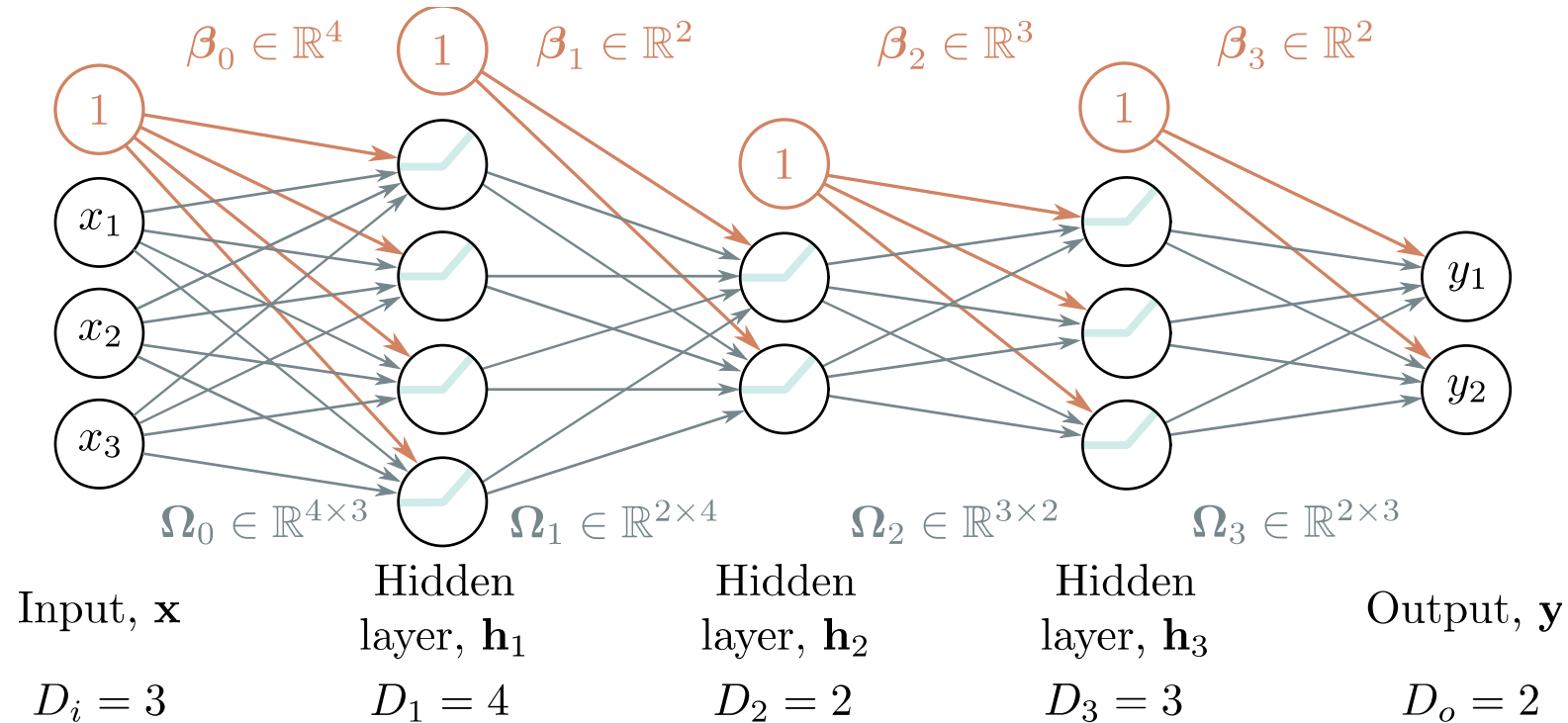Least squares loss for linear regression

$$f[x_i, \phi] = \phi_0 + \phi_1 x_i$$

$$\frac{\partial L}{\partial \phi} = \frac{\partial}{\partial \phi} \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} \frac{\partial \ell_i}{\partial \phi}$$

$$\frac{\partial \ell_i}{\partial \phi} = \begin{bmatrix} \frac{\partial \ell_i}{\partial \phi_0} \\ \frac{\partial \ell_i}{\partial \phi_1} \end{bmatrix} = \begin{bmatrix} 2(\phi_0 + \phi_1 x_i - y_i) \\ 2x_i(\phi_0 + \phi_1 x_i - y_i) \end{bmatrix}$$

Partial derivative w.r.t. each parameter

What about deep learning models?

$$\mathbf{h}_1 = \mathbf{a}[\boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0\mathbf{x}]$$

$$\mathbf{h}_2 = \mathbf{a}[\boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1\mathbf{h}_1]$$

$$\mathbf{h}_3 = \mathbf{a}[\boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2\mathbf{h}_2]$$

$$\mathbf{f}[\mathbf{x}, \boldsymbol{\phi}] = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3\mathbf{h}_3$$

We need to compute partial derivatives w.r.t. every parameter!

Loss: sum of individual terms:

$$L[\boldsymbol{\phi}] = \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} l[f[\mathbf{x}_i, \boldsymbol{\phi}], y_i]$$

SGD Algorithm:

$$\boldsymbol{\phi}_{t+1} \longleftarrow \boldsymbol{\phi}_t - \alpha \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$$

Millions and even *billions* of parameters:

$$\boldsymbol{\phi} = \{\boldsymbol{\beta}_0, \boldsymbol{\Omega}_0, \boldsymbol{\beta}_1, \boldsymbol{\Omega}_1, \boldsymbol{\beta}_2, \boldsymbol{\Omega}_2, \ldots\}$$
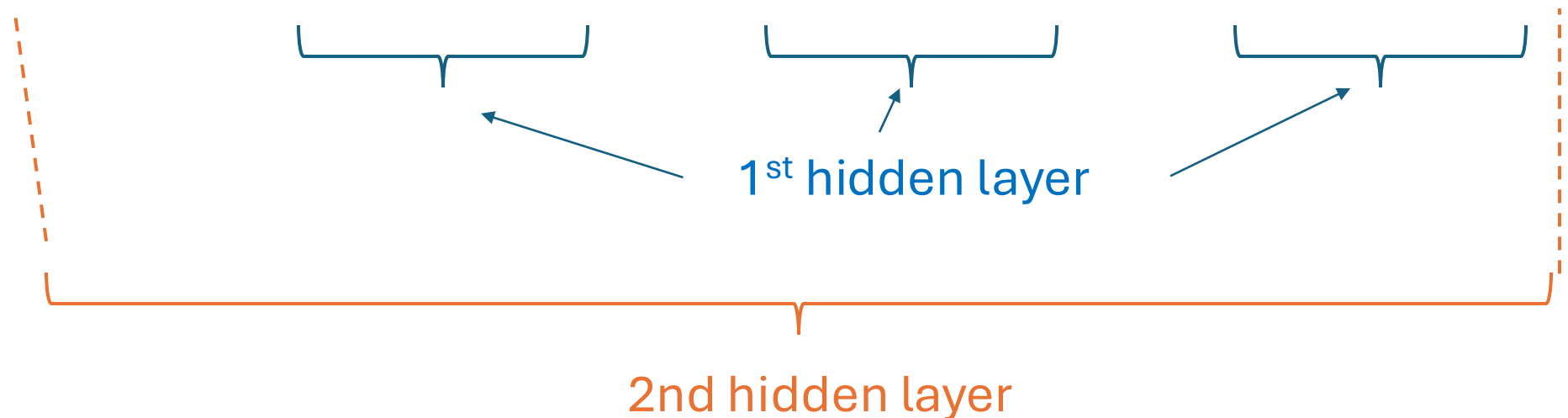
We need the partial derivative with respect to every weight and bias we want to update for every sample in the batch.

$$\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_k} \qquad \text{and} \qquad \frac{\partial \ell_i}{\partial \boldsymbol{\Omega}_k}$$

Network equation gets unwieldy even for small models

- Model equation for 2 hidden layers of 3 units each:

$$y' = \phi'_0 + \phi'_1 \mathrm{a}\left[\psi_{10} + \psi_{11}\mathrm{a}[\theta_{10} + \theta_{11}x] + \psi_{12}\mathrm{a}[\theta_{20} + \theta_{21}x] + \psi_{13}\mathrm{a}[\theta_{30} + \theta_{31}x]\right]$$
$$+ \phi'_2 \mathrm{a}[\psi_{20} + \psi_{21}\mathrm{a}[\theta_{10} + \theta_{11}x] + \psi_{22}\mathrm{a}[\theta_{20} + \theta_{21}x] + \psi_{23}\mathrm{a}[\theta_{30} + \theta_{31}x]]$$
$$+ \phi'_3 \mathrm{a}[\psi_{30} + \psi_{31}\mathrm{a}[\theta_{10} + \theta_{11}x] + \psi_{32}\mathrm{a}[\theta_{20} + \theta_{21}x] + \psi_{33}\mathrm{a}[\theta_{30} + \theta_{31}x]]$$

1st hidden layer

2nd hidden layer

Don't We Have Auto Grad?

- The backpropagation formulas for gradients are going to guide us to better initializations next lecture.

- Many problems with neural network training are due to poor gradient management.

Gradients

- Backpropagation intuition
- Toy model
- Matrix calculus
- Backpropagation matrix forward pass
- Backpropagation matrix backward pass

Problem 1: Computing gradients

Loss: sum of individual terms:

$$L[\phi] = \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} l[f[\mathbf{x}_i, \phi], y_i]$$

SGD Algorithm:

$$\phi_{t+1} \longleftarrow \phi_t - \alpha \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i[\phi_t]}{\partial \phi}$$

Parameters:

$$\phi = \{\boldsymbol{\beta}_0, \boldsymbol{\Omega}_0, \boldsymbol{\beta}_1, \boldsymbol{\Omega}_1, \boldsymbol{\beta}_2, \boldsymbol{\Omega}_2, \boldsymbol{\beta}_3, \boldsymbol{\Omega}_3\}$$

Need to compute gradients

$$\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_k} \qquad \text{and} \qquad \frac{\partial \ell_i}{\partial \boldsymbol{\Omega}_k}$$

Algorithm to compute gradient efficiently

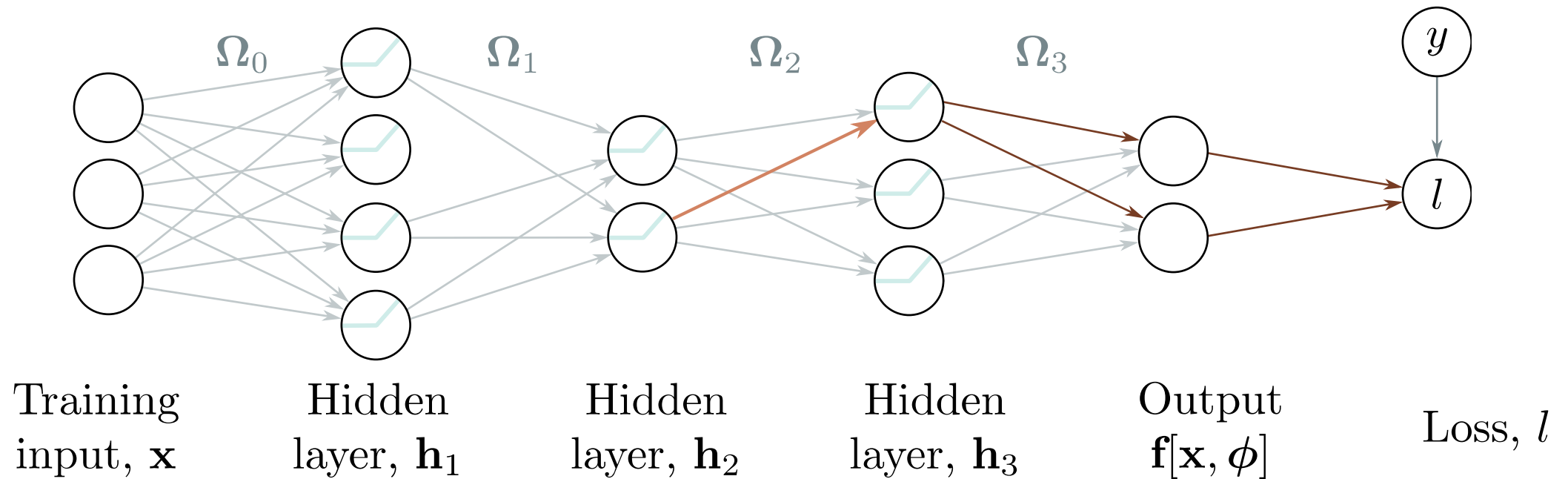- "Backpropagation algorithm"
- Rumelhart, Hinton, and Williams (1986)

BackProp intuition #1: the forward pass

weight of 2 means 2× influence on right side vs left

Training output, y

Ω_0 Ω_1 Ω_2 Ω_3

Training input, $\mathbf{x}$ Hidden layer, $\mathbf{h}_1$ Hidden layer, $\mathbf{h}_2$ Hidden layer, $\mathbf{h}_3$ Output $\mathbf{f}[\mathbf{x}, \phi]$ Loss, l

- The weight on the orange arrow multiplies activation (ReLU output) of previous layer
- We want to know how change *(partial derivative)* in orange weight affects loss
- If we double activation in previous layer, weight will have twice the effect
- Conclusion: we need to know the activations at each layer.
- Put another way: we need to evaluate each partial derivatives for each input

14

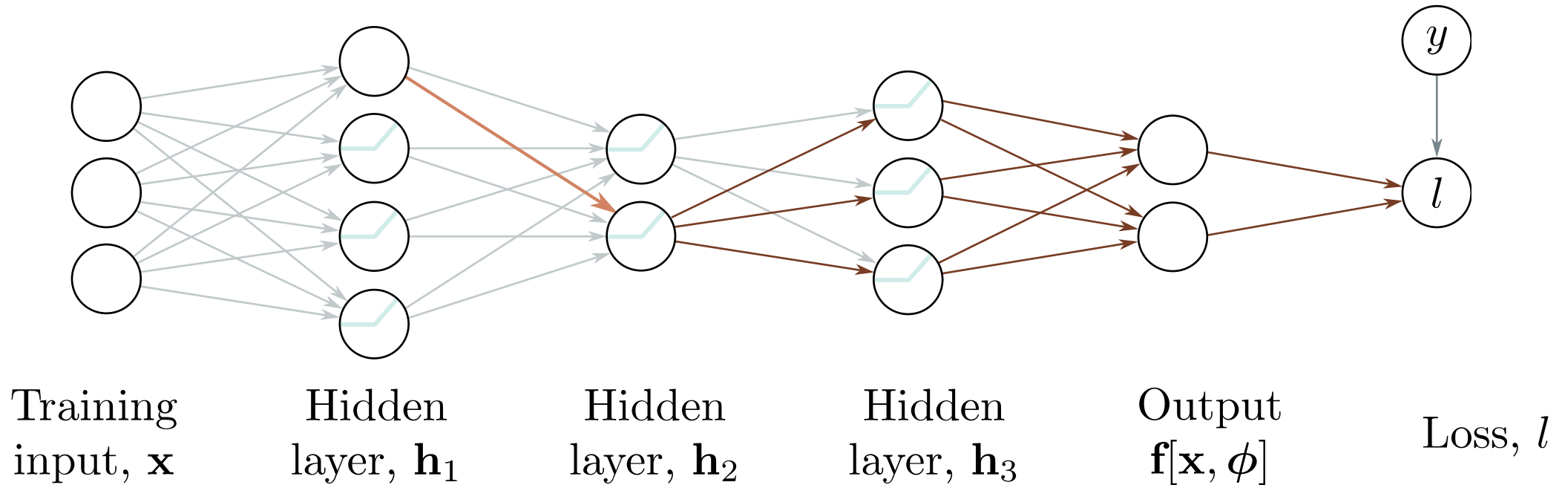# BackProp intuition #2: the backward pass

To calculate how a small change in a weight or bias feeding into hidden layer $\mathbf{h}_3$ modifies the loss, we need to know:
- how a change in layer $\mathbf{h}_3$ changes the model output $\mathbf{f}$
- how a change in the model output changes the loss l

BackProp intuition #2: the backward pass

$$\text{Chain rule: } \frac{d\,f(g(x))}{dx} = \frac{df}{dx}f'(g(x)) \cdot g'(x)$$

Training input, $\mathbf{x}$ Hidden layer, $\mathbf{h}_1$ Hidden layer, $\mathbf{h}_2$ Hidden layer, $\mathbf{h}_3$ Output $\mathbf{f}[\mathbf{x}, \phi]$ Loss, l
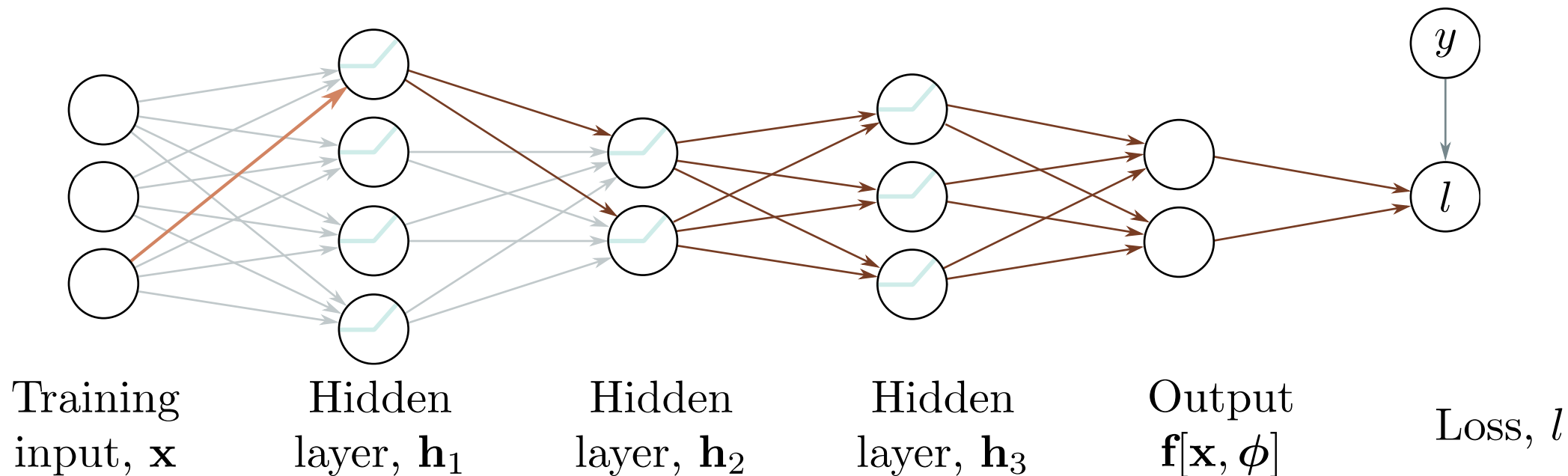
To calculate how a small change in a weight or bias feeding into hidden layer $\mathbf{h}_2$ modifies the loss, we need to know:
- how a change in layer $\mathbf{h}_2$ affects $\mathbf{h}_3$
- how $\mathbf{h}_3$ changes the model output $\mathbf{f}$
- how a change in the model output $\mathbf{f}$ changes the loss l

We know this from the previous step

16

BackProp intuition #2: the backward pass

Training input, $\mathbf{x}$ Hidden layer, $\mathbf{h}_1$ Hidden layer, $\mathbf{h}_2$ Hidden layer, $\mathbf{h}_3$ Output $\mathbf{f}[\mathbf{x}, \phi]$ Loss, l

To calculate how a small change in a weight or bias feeding into hidden layer $\mathbf{h}_1$ modifies the loss, we need to know:
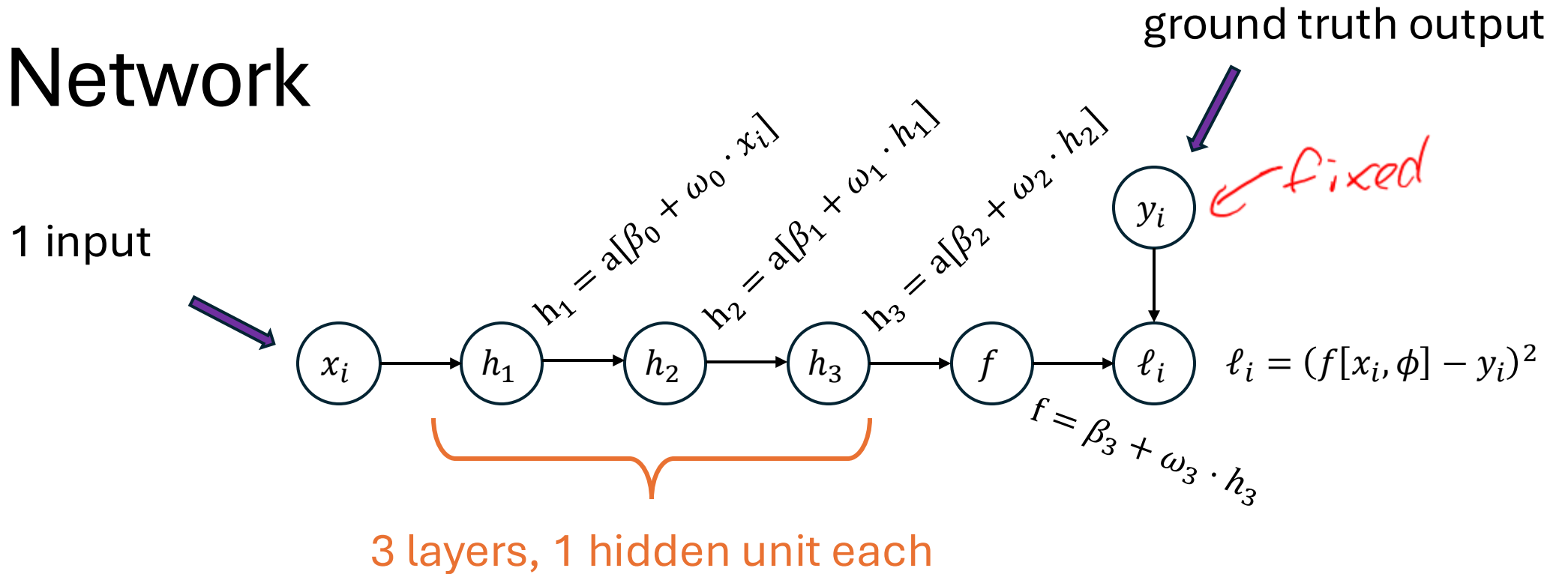- how a change in layer $\mathbf{h}_1$ affects $\mathbf{h}_2$
- how a change in layer $\mathbf{h}_2$ affects $\mathbf{h}_3$
- how $\mathbf{h}_3$ changes the model output $\mathbf{f}$
- how a change in the model output $\mathbf{f}$ changes the loss l

We know these from the previous steps

Gradients

- Backpropagation intuition
- Toy model
- Matrix calculus
- Backpropagation matrix forward pass
- Backpropagation matrix backward pass

Toy Network

ground truth output

fixed

1 input

$h_1 = a[\beta_0 + \omega_0 \cdot x_i]$

$h_2 = a[\beta_1 + \omega_1 \cdot h_1]$

$h_3 = a[\beta_2 + \omega_2 \cdot h_2]$

$x_i \rightarrow h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow f \rightarrow \ell_i$

y_i

$\ell_i = (f[x_i, \phi] - y_i)^2$

$f = \beta_3 + \omega_3 \cdot h_3$

3 layers, 1 hidden unit each

$$f[x_i, \phi] = \beta_3 + \omega_3 \cdot a\Big[\beta_2 + \omega_2 \cdot a[\beta_1 + \omega_1 \cdot a[\beta_0 + \omega_0 \cdot x_i]]\Big]$$
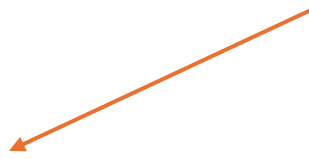
$$\ell_i = (f[x_i, \phi] - y_i)^2$$

Gradients of toy function

$$f[x_i, \phi] = \beta_3 + \omega_3 \cdot a\Big[\beta_2 + \omega_2 \cdot a[\beta_1 + \omega_1 \cdot a[\beta_0 + \omega_0 \cdot x_i]]\Big]$$

$$\ell_i = (f[x_i, \phi] - y_i)^2$$

Tells us how a small change in β_j or ω_j change the loss ℓ_i for the ith example
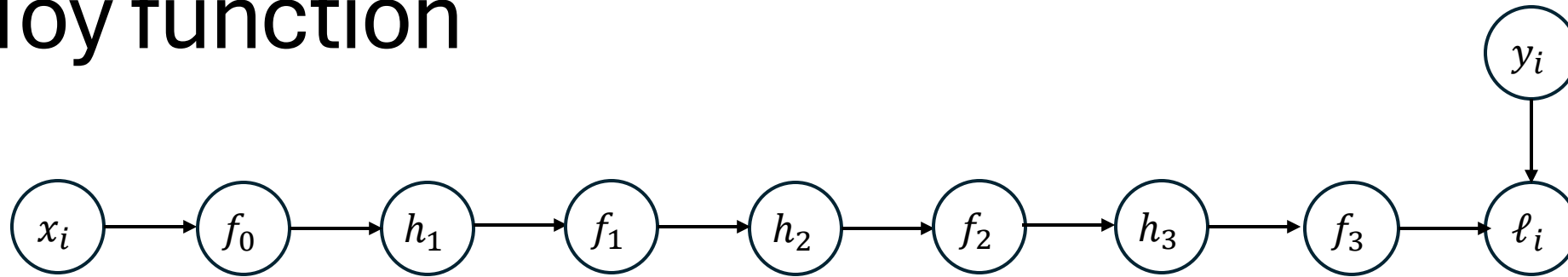
We want to calculate each partial:

$$\frac{\partial \ell_i}{\partial \beta_0}, \quad \frac{\partial \ell_i}{\partial \omega_0}, \quad \frac{\partial \ell_i}{\partial \beta_1}, \quad \frac{\partial \ell_i}{\partial \omega_1}, \quad \frac{\partial \ell_i}{\partial \beta_2}, \quad \frac{\partial \ell_i}{\partial \omega_2}, \quad \frac{\partial \ell_i}{\partial \beta_3}, \quad \text{and} \quad \frac{\partial \ell_i}{\partial \omega_3}$$

Toy function

Pre-Activations

Intermediate values

$$f_0 = \beta_0 + \omega_0 \cdot x$$
$$h_1 = a[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = a[f_1]$$

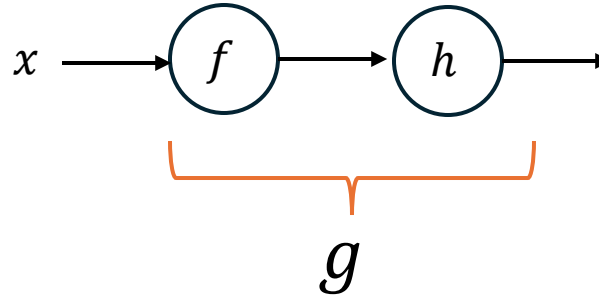$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = a[f_2]$$
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (y_i - f_3)^2$$

Refresher: The Chain Rule

$$x \longrightarrow \boxed{f} \longrightarrow \boxed{h} \longrightarrow$$

$$\underbrace{\qquad}_{g}$$

For $g(x) = h\big(f(x)\big)$

then $g'(x) = h'\big(f(x)\big) f'(x)$, where $g'(x)$ is the derivative of $g(x)$.

Lagrange's Notation

Or can be written equivalently as

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

Leibniz's Notation

Forward pass

$$f[x_i, \phi] = \beta_3 + \omega_3 \cdot a\left[\beta_2 + \omega_2 \cdot a[\beta_1 + \omega_1 \cdot a[\beta_0 + \omega_0 \cdot x_i]]\right]$$

$$\ell_i = (f[x_i, \phi] - y_i)^2$$

1. Write this as a series of intermediate calculations

2. Compute these intermediate quantities

$$f_0 = \beta_0 + \omega_0 \cdot x_i$$
$$h_1 = a[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = a[f_1]$$

$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = a[f_2]$$
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (y_i - f_3)^2$$

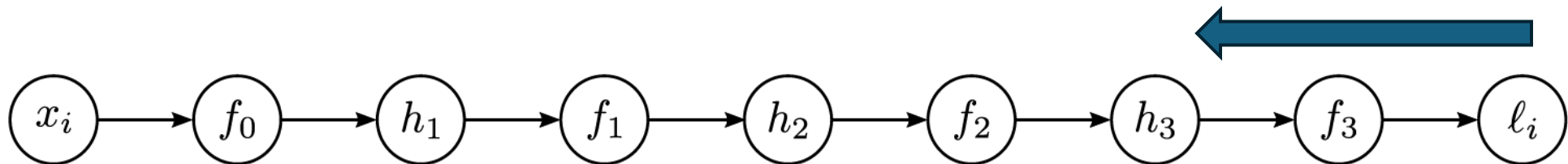$$\frac{\partial \ell_i}{\partial f_3} = 2(y_i - f_3)$$
$$\circ (-1)$$

Backward pass

$$f[x_i, \phi] = \beta_3 + \omega_3 \cdot a\left[\beta_2 + \omega_2 \cdot a[\beta_1 + \omega_1 \cdot a[\beta_0 + \omega_0 \cdot x_i]]\right]$$

$$\ell_i = (f[x_i, \phi] - y_i)^2$$

1. Compute the derivatives of the *loss* with respect to these intermediate quantities, but in reverse order.

$$\frac{\partial \ell_i}{\partial f_3}, \quad \frac{\partial \ell_i}{\partial h_3}, \quad \frac{\partial \ell_i}{\partial f_2}, \quad \frac{\partial \ell_i}{\partial h_2}, \quad \frac{\partial \ell_i}{\partial f_1}, \quad \frac{\partial \ell_i}{\partial h_1}, \quad \text{and} \quad \frac{\partial \ell_i}{\partial f_0}$$

$$x_i \rightarrow f_0 \rightarrow h_1 \rightarrow f_1 \rightarrow h_2 \rightarrow f_2 \rightarrow h_3 \rightarrow f_3 \rightarrow \ell_i$$

Backward pass

$$f[x_i, \phi] = \beta_3 + \omega_3 \cdot a\Big[\beta_2 + \omega_2 \cdot a\big[\beta_1 + \omega_1 \cdot a[\beta_0 + \omega_0 \cdot x_i]\big]\Big]$$
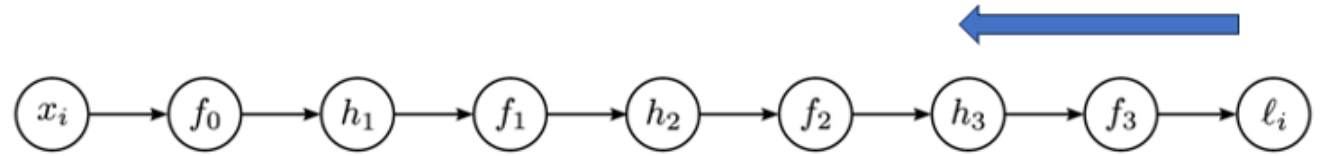
$$\ell_i = (f[x_i, \phi] - y_i)^2$$

1. Compute the derivatives of the loss with respect to these intermediate quantities, but in reverse order.

$$\frac{\partial \ell_i}{\partial f_3}, \quad \frac{\partial \ell_i}{\partial h_3}, \quad \frac{\partial \ell_i}{\partial f_2}, \quad \frac{\partial \ell_i}{\partial h_2}, \quad \frac{\partial \ell_i}{\partial f_1}, \quad \frac{\partial \ell_i}{\partial h_1}, \quad \text{and} \quad \frac{\partial \ell_i}{\partial f_0}$$

$$x_i \qquad \boxed{\frac{\partial \ell_i}{\partial f_0}} \leftarrow \boxed{\frac{\partial \ell_i}{\partial h_1}} \leftarrow \boxed{\frac{\partial \ell_i}{\partial f_1}} \leftarrow \boxed{\frac{\partial \ell_i}{\partial h_2}} \leftarrow \boxed{\frac{\partial \ell_i}{\partial f_2}} \leftarrow \boxed{\frac{\partial \ell_i}{\partial h_3}} \leftarrow \boxed{\frac{\partial \ell_i}{\partial f_3}} \leftarrow \ell_i$$

Backward pass

1. Compute the derivatives of the loss with respect to these intermediate quantities, but in reverse order.

$$f_0 = \beta_0 + \omega_0 \cdot x$$
$$h_1 = a[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = a[f_1]$$

$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
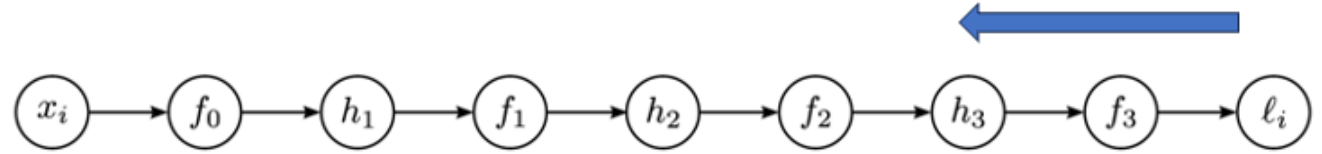$$h_3 = a[f_2]$$
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (f_3 - y_i)^2$$

- The first of these derivatives is trivial

$$\frac{\partial \ell_i}{\partial f_3} = 2(f_3 - y_i)$$

Backward pass

1. Compute the derivatives of the loss with respect to these intermediate quantities, but in reverse order.

$$f_0 = \beta_0 + \omega_0 \cdot x$$
$$h_1 = a[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = a[f_1]$$

$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = a[f_2]$$
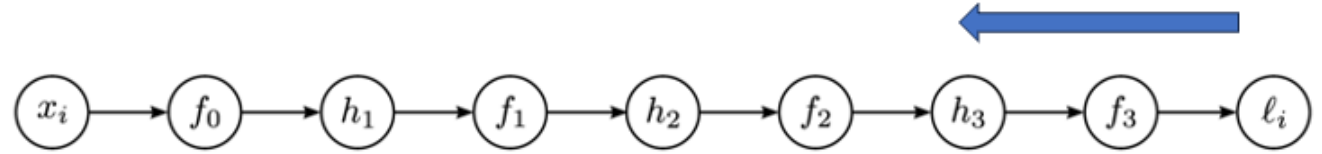$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (y_i - f_3)^2$$

- The second of these derivatives is computed via the chain rule

$$\frac{\partial \ell_i}{\partial h_3} = \frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3}$$

How does a small change in h_3 change ℓ_i?

Backward pass

1. Compute the derivatives of the loss with respect to these intermediate quantities, but in reverse order.

$$f_0 = \beta_0 + \omega_0 \cdot x$$
$$h_1 = a[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = a[f_1]$$

$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = a[f_2]$$
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (y_i - f_3)^2$$

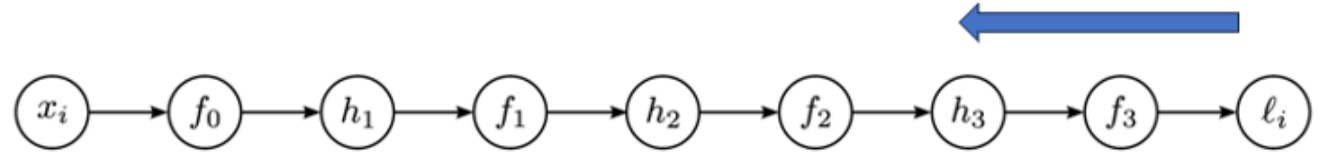- The second derivative is computed via the chain rule

$$\frac{\partial \ell_i}{\partial h_3} = \frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3}$$

How does a small change in f_3 change ℓ_i?

How does a small change in h_3 change ℓ_i?

How does a small change in h_3 change f_3?

Backward pass

1. Compute the derivatives of the loss with respect to these intermediate quantities, but in reverse order.

$$f_0 = \beta_0 + \omega_0 \cdot x$$

$$h_1 = \mathrm{a}[f_0]$$

$$f_1 = \beta_1 + \omega_1 \cdot h_1$$

$$h_2 = \mathrm{a}[f_1]$$

$$f_2 = \beta_2 + \omega_2 \cdot h_2$$

$$h_3 = \mathrm{a}[f_2]$$

$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
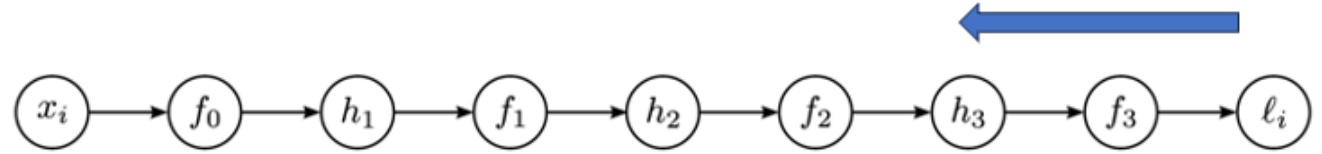
$$\ell_i = (y_i - f_3)^2$$

- The second of these derivatives is computed via the chain rule

$$\frac{\partial \ell_i}{\partial h_3} = \frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3}$$

$$\frac{\delta f_3}{\partial h_3} = \frac{\partial (\beta_3 + \omega_3 \cdot h_3)}{\partial h_3}$$

$$= \omega_3$$

Already computed!

Backward pass

1. Compute the derivatives of the loss with respect to these intermediate quantities, but in reverse order.

$$f_0 = \beta_0 + \omega_0 \cdot x$$

$$h_1 = a[f_0]$$

$$f_1 = \beta_1 + \omega_1 \cdot h_1$$

$$h_2 = a[f_1]$$

$$f_2 = \beta_2 + \omega_2 \cdot h_2$$

$$\checkmark h_3 = a[f_2]$$

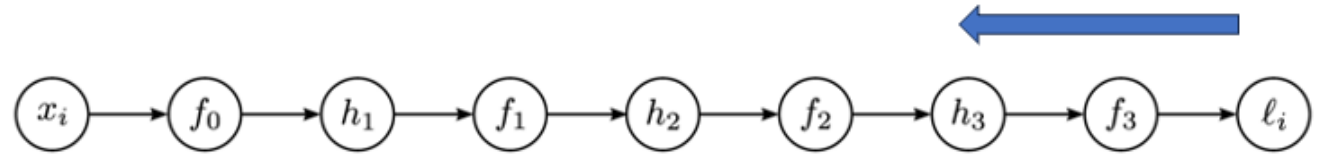$$\checkmark f_3 = \beta_3 + \omega_3 \cdot h_3$$

$$\ell_i = (y_i - f_3)^2$$

- The remaining derivatives also calculated by further use of chain rule

$$\frac{\partial \ell_i}{\partial f_2} = \frac{\partial h_3}{\partial f_2} \left(\frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3} \right)$$

$$\frac{\partial \ell_i}{\partial f_2} = \frac{\partial h_3}{\partial f_2} \cdot \frac{\partial \ell_i}{\partial h_3}$$

30

Backward pass

1. Compute the derivatives of the loss with respect to these intermediate quantities, but in reverse order.

$$f_0 = \beta_0 + \omega_0 \cdot x$$
$$h_1 = a[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = a[f_1]$$

$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = a[f_2]$$
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (y_i - f_3)^2$$

- The remaining derivatives also calculated by further use of chain rule

$$\frac{\partial \ell_i}{\partial f_2} = \frac{\partial h_3}{\partial f_2} \left(\frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3} \right)$$

Already computed!

Backward pass

1. Compute the derivatives of the loss with respect to these intermediate quantities, but in reverse order.

$$f_0 = \beta_0 + \omega_0 \cdot x$$

$$h_1 = a[f_0]$$

$$f_1 = \beta_1 + \omega_1 \cdot h_1$$

$$h_2 = a[f_1]$$

$$f_2 = \beta_2 + \omega_2 \cdot h_2$$

$$h_3 = a[f_2]$$
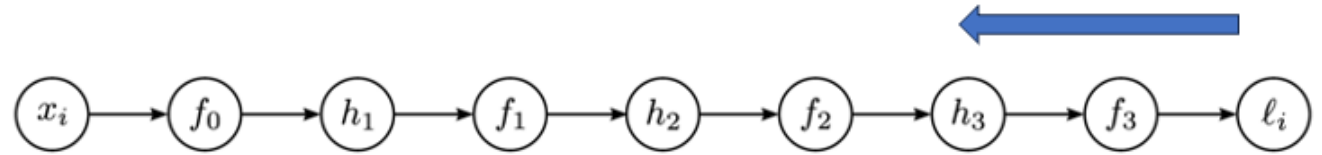
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$

$$\ell_i = (y_i - f_3)^2$$

- The remaining derivatives also calculated by further use of chain rule

$$\frac{\partial \ell_i}{\partial f_2} = \frac{\partial h_3}{\partial f_2} \left(\frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3} \right)$$

$$\frac{\partial \ell_i}{\partial h_2} = \frac{\partial f_2}{\partial h_2} \left(\frac{\partial h_3}{\partial f_2} \frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3} \right)$$

Backward pass

1. Compute the derivatives of the loss with respect to these intermediate quantities, but in reverse order.

$$f_0 = \beta_0 + \omega_0 \cdot x$$
$$h_1 = a[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = a[f_1]$$

$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = a[f_2]$$
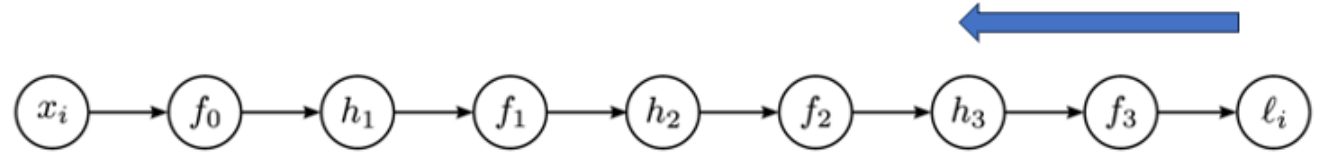$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (y_i - f_3)^2$$

- The remaining derivatives also calculated by further use of chain rule

$$\frac{\partial \ell_i}{\partial f_2} = \frac{\partial h_3}{\partial f_2}\left(\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)$$
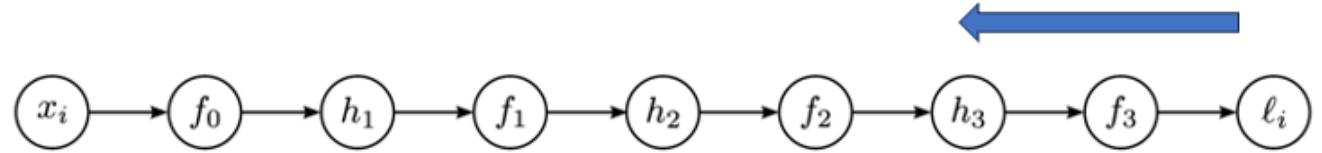
$$\frac{\partial \ell_i}{\partial h_2} = \frac{\partial f_2}{\partial h_2}\left(\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)$$

$$\frac{\partial \ell_i}{\partial f_1} = \frac{\partial h_2}{\partial f_1}\left(\frac{\partial f_2}{\partial h_2}\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)$$

$$\frac{\partial \ell_i}{\partial h_1} = \frac{\partial f_1}{\partial h_1}\left(\frac{\partial h_2}{\partial f_1}\frac{\partial f_2}{\partial h_2}\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)$$

$$\frac{\partial \ell_i}{\partial f_0} = \frac{\partial h_1}{\partial f_0}\left(\frac{\partial f_1}{\partial h_1}\frac{\partial h_2}{\partial f_1}\frac{\partial f_2}{\partial h_2}\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)$$

Backward pass

1. Compute the derivatives of the loss with respect to these intermediate quantities, but in reverse order.

- The remaining derivatives also calculated by further use of chain rule

$$\frac{\partial \ell_i}{\partial f_3} = 2(f_3 - y_i)$$

$$\frac{\partial \ell_i}{\partial h_3} = \frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3}$$

$$\frac{\partial \ell_i}{\partial f_2} = \frac{\partial h_3}{\partial f_2} \left(\frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3} \right)$$

$$\frac{\partial \ell_i}{\partial h_2} = \frac{\partial f_2}{\partial h_2} \left(\frac{\partial h_3}{\partial f_2} \frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3} \right)$$

$$\frac{\partial \ell_i}{\partial f_1} = \frac{\partial h_2}{\partial f_1} \left(\frac{\partial f_2}{\partial h_2} \frac{\partial h_3}{\partial f_2} \frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3} \right)$$

$$\frac{\partial \ell_i}{\partial h_1} = \frac{\partial f_1}{\partial h_1} \left(\frac{\partial h_2}{\partial f_1} \frac{\partial f_2}{\partial h_2} \frac{\partial h_3}{\partial f_2} \frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3} \right)$$

$$\frac{\partial \ell_i}{\partial f_0} = \frac{\partial h_1}{\partial f_0} \left(\frac{\partial f_1}{\partial h_1} \frac{\partial h_2}{\partial f_1} \frac{\partial f_2}{\partial h_2} \frac{\partial h_3}{\partial f_2} \frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3} \right)$$

Backward pass

1. Compute the derivatives of the loss with respect to these intermediate quantities, but in reverse order.

- The remaining derivatives also calculated by further use of chain rule

$$\frac{\partial \ell_i}{\partial f_3} = 2(f_3 - y_i)$$

$$\frac{\partial \ell_i}{\partial h_3} = \frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}$$

$$\frac{\partial \ell_i}{\partial f_2} = \frac{\partial h_3}{\partial f_2}\left(\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)$$

$$\frac{\partial \ell_i}{\partial h_2} = \frac{\partial f_2}{\partial h_2}\left(\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)$$

$$\frac{\partial \ell_i}{\partial f_1} = \frac{\partial h_2}{\partial f_1}\left(\frac{\partial f_2}{\partial h_2}\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)$$

$$\frac{\partial \ell_i}{\partial h_1} = \frac{\partial f_1}{\partial h_1}\left(\frac{\partial h_2}{\partial f_1}\frac{\partial f_2}{\partial h_2}\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)$$
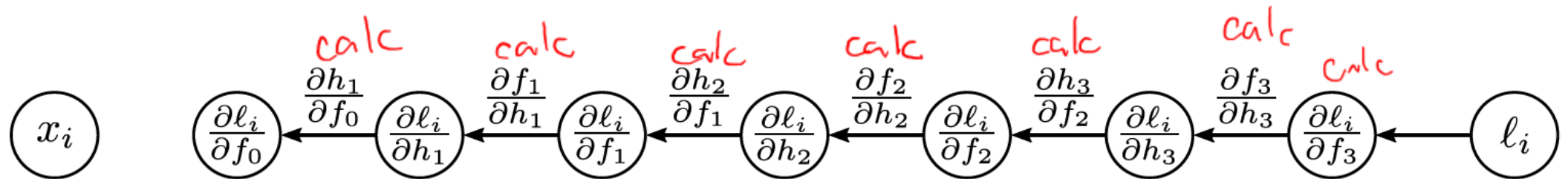
$$\frac{\partial \ell_i}{\partial f_0} = \frac{\partial h_1}{\partial f_0}\left(\frac{\partial f_1}{\partial h_1}\frac{\partial h_2}{\partial f_1}\frac{\partial f_2}{\partial h_2}\frac{\partial h_3}{\partial f_2}\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right)$$

We extend this to get to the parameters ω's and β's

Previously calculated partial derivatives w/ respect to pre + post activations

Backward pass

2. Find how the loss changes as a function of the parameters β and ω.

$$f_0 = \beta_0 + \omega_0 \cdot x$$
$$h_1 = a[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = a[f_1]$$

$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = a[f_2]$$
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (y_i - f_3)^2$$

- Another application of the chain rule

$$\frac{\partial \ell_i}{\partial \omega_k} = \frac{\partial f_k}{\partial \omega_k} \frac{\partial \ell_i}{\partial f_k}$$

How does a small change in ω_k change l_i?

How does a small change in ω_k change f_k?

How does a small change in f_k change l_i?

Backward pass

2. Find how the loss changes as a function of the parameters β and ω.

$f_0 = \beta_0 + \omega_0 \cdot x$

$h_1 = a[f_0]$

$f_1 = \beta_1 + \omega_1 \cdot h_1$

$h_2 = a[f_1]$

$f_2 = \beta_2 + \omega_2 \cdot h_2$

$h_3 = a[f_2]$

$f_3 = \beta_3 + \omega_3 \cdot h_3$

$\ell_i = (y_i - f_3)^2$

- Another application of the chain rule

$$\frac{\partial \ell_i}{\partial \omega_k} = \frac{\partial f_k}{\partial \omega_k} \frac{\partial \ell_i}{\partial f_k}$$

$\frac{\partial f_k}{\partial \beta_k} = 1$

$f_k = \beta_k + \omega_k \cdot h_k$

$\frac{\partial f_k}{\partial \omega_k} = \emptyset + h_k$

Already calculated in part 1.

How does a small change in ω_k change l_i?

$$\frac{\partial f_k}{\partial \omega_k} = h_k$$

38

Backward pass

2. Find how the loss changes as a function of the parameters β and ω.

$$f_0 = \beta_0 + \omega_0 \cdot x$$
$$h_1 = a[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = a[f_1]$$

$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = a[f_2]$$
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (y_i - f_3)^2$$

- Another application of the chain rule
- Similarly for β parameters

$$\frac{\partial \ell_i}{\partial \omega_k} = \frac{\partial f_k}{\partial \omega_k} \frac{\partial \ell_i}{\partial f_k}$$

$$\frac{\partial \ell_i}{\partial \beta_k} = \frac{\partial f_k}{\partial \beta_k} \frac{\partial \ell_i}{\partial f_k}$$

1

Backward pass

2. Find how the loss changes as a function of the parameters β and ω.

$$f_0 = \beta_0 + \omega_0 \cdot x$$
$$h_1 = a[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = a[f_1]$$

$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = a[f_2]$$
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (y_i - f_3)^2$$

Gradients

- Backpropagation intuition
- Toy model
- Matrix calculus
- Backpropagation matrix forward pass
- Backpropagation matrix backward pass

Matrix calculus

Scalar function $f[\cdot]$ of a *vector* $\mathbf{a}$

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}$$

$$\frac{\partial f}{\partial \mathbf{a}} = \begin{bmatrix} \dfrac{\partial f}{\partial a_1} \\[1em] \dfrac{\partial f}{\partial a_2} \\[1em] \dfrac{\partial f}{\partial a_3} \\[1em] \dfrac{\partial f}{\partial a_4} \end{bmatrix}$$

The derivative with respect to vector $\mathbf{a}$ is a vector of the same shape as $\mathbf{a}$.

Matrix calculus

Scalar function $f[\cdot]$ of a *matrix* $\mathbf{A}$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{bmatrix}$$

$$\frac{\partial f}{\partial \mathbf{A}} = \begin{bmatrix} \frac{\partial f}{\partial a_{11}} & \frac{\partial f}{\partial a_{12}} & \frac{\partial f}{\partial a_{13}} \\ \frac{\partial f}{\partial a_{21}} & \frac{\partial f}{\partial a_{22}} & \frac{\partial f}{\partial a_{23}} \\ \frac{\partial f}{\partial a_{31}} & \frac{\partial f}{\partial a_{32}} & \frac{\partial f}{\partial a_{33}} \\ \frac{\partial f}{\partial a_{41}} & \frac{\partial f}{\partial a_{42}} & \frac{\partial f}{\partial a_{43}} \end{bmatrix}$$

The derivative with respect to matrix $\mathbf{A}$ is a matrix of the same shape as $\mathbf{A}.$

Matrix calculus

Columns are each
element function

Rows are each
variable element

Vector function $\mathbf{f}[\cdot]$ of a *vector* $\mathbf{a}$

$$\mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} \quad \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} \quad \frac{\partial \mathbf{f}}{\partial \mathbf{a}} = \begin{bmatrix} \dfrac{\partial f_1}{\partial a_1} & \dfrac{\partial f_2}{\partial a_1} & \dfrac{\partial f_3}{\partial a_1} \\ \dfrac{\partial f_1}{\partial a_2} & \dfrac{\partial f_2}{\partial a_2} & \dfrac{\partial f_3}{\partial a_2} \\ \dfrac{\partial f_1}{\partial a_3} & \dfrac{\partial f_2}{\partial a_3} & \dfrac{\partial f_3}{\partial a_3} \\ \dfrac{\partial f_1}{\partial a_4} & \dfrac{\partial f_2}{\partial a_4} & \dfrac{\partial f_4}{\partial a_4} \end{bmatrix}$$

Vector of scalar valued functions

$$\frac{\partial f_3}{\partial \beta_3} = \frac{\partial}{\partial \beta_3}\left(\beta_3 + \Omega_3 h_3\right) = \frac{\partial}{\partial \beta_3}\left(\beta_3\right)$$

$$\beta_3 = \begin{pmatrix} \beta_{3,1} \\ \beta_{3,2} \\ \beta_{3,3} \end{pmatrix}$$

$$g(\beta) = \beta$$

44

Comparing vector and matrix

Scalar derivatives:

$$f_3 = \beta_3 + \omega_3 h_3$$

$$\frac{\partial f_3}{\partial h_3} = \frac{\partial}{\partial h_3}(\beta_3 + \omega_3 h_3) = \omega_3$$

Comparing vector and matrix

Ω h

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} = \begin{bmatrix} w_{11}h_1 + w_{12}h_2 + w_{13}h_3 \\ w_{21}h_1 + w_{22}h_2 + w_{23}h_3 \end{bmatrix}$$

Scalar derivatives:

$$f_3 = \beta_3 + \omega_3 h_3$$

$\omega_{i,j}$

$$\frac{\partial f_3}{\partial h_3} = \frac{\partial}{\partial h_3}(\beta_3 + \omega_3 h_3) = \omega_3$$

$$\frac{\partial f_3}{\partial \beta_3} = \frac{\partial}{\partial \beta_3}(\beta_3 + \Omega_3 h_3) = I$$

Matrix derivatives:

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$$

$$\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} = \frac{\partial}{\partial \mathbf{h}_3}(\boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3) = \boldsymbol{\Omega}_3^T$$

$$\frac{\partial f}{\partial h_1} = \begin{bmatrix} w_{11} & w_{12} \end{bmatrix} \quad \frac{\partial f}{\partial h} = \begin{bmatrix} w_{11} & w_{12} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{bmatrix}$$

Comparing vector and matrix

Scalar derivatives:

$$f_3 = \beta_3 + \omega_3 h_3$$

$$\frac{\partial f_3}{\partial \beta_3} = \frac{\partial}{\partial \omega_3}\beta_3 + \omega_3 h_3 = 1$$

Matrix derivatives:

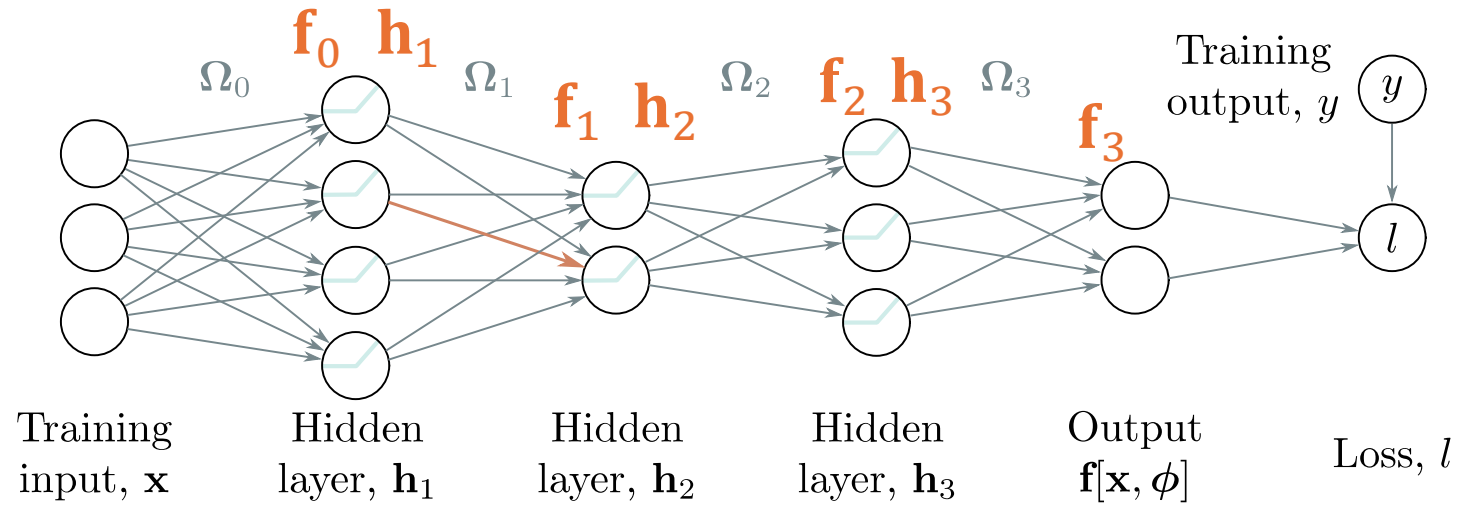$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$$

$$\frac{\partial \mathbf{f}_3}{\partial \boldsymbol{\beta}_3} = \frac{\partial}{\partial \beta_3}(\boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3) = \mathbf{I}$$

Gradients

- Backpropagation intuition
- Toy model
- Matrix calculus
- Backpropagation matrix forward pass
- Backpropagation matrix backward pass

The forward pass

Training input, $\mathbf{x}$ — Hidden layer, $\mathbf{h}_1$ — Hidden layer, $\mathbf{h}_2$ — Hidden layer, $\mathbf{h}_3$ — Output $\mathbf{f}[\mathbf{x}, \phi]$

Training output, y

Loss, l

1. Write this as a series of intermediate calculations

$$\mathbf{f}_0 = \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i$$

$$\mathbf{h}_1 = \mathbf{a}[\mathbf{f}_0]$$

$$\mathbf{f}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1$$

$$\mathbf{h}_2 = \mathbf{a}[\mathbf{f}_1]$$

$$\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2$$

$$\mathbf{h}_3 = \mathbf{a}[\mathbf{f}_2]$$

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$$

$$\ell_i = \mathrm{l}[\mathbf{f}_3, y_i]$$

49

The forward pass

1. Write this as a series of intermediate calculations

$$\mathbf{f}_0 = \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i$$

$$\mathbf{h}_1 = \mathbf{a}[\mathbf{f}_0]$$

$$\mathbf{f}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1$$

2. Compute these intermediate quantities

$$\mathbf{h}_2 = \mathbf{a}[\mathbf{f}_1]$$

$$\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2$$

$$\mathbf{h}_3 = \mathbf{a}[\mathbf{f}_2]$$

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$$

$$\ell_i = \mathrm{l}[\mathbf{f}_3, y_i]$$

The backward pass

1. Write this as a series of intermediate calculations

2. Compute these intermediate quantities

3. Take derivatives of output with respect to intermediate quantities

$$\mathbf{f}_0 = \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i$$

$$\mathbf{h}_1 = \mathbf{a}[\mathbf{f}_0]$$

$$\mathbf{f}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1$$

$$\mathbf{h}_2 = \mathbf{a}[\mathbf{f}_1]$$

$$\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2$$

$$\mathbf{h}_3 = \mathbf{a}[\mathbf{f}_2]$$

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$$

$$\ell_i = l[\mathbf{f}_3, y_i]$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_3}$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_2} = \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3}$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_1} = \frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \left(\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_0} = \frac{\partial \mathbf{h}_1}{\partial \mathbf{f}_0} \frac{\partial \mathbf{f}_1}{\partial \mathbf{h}_1} \left(\frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$
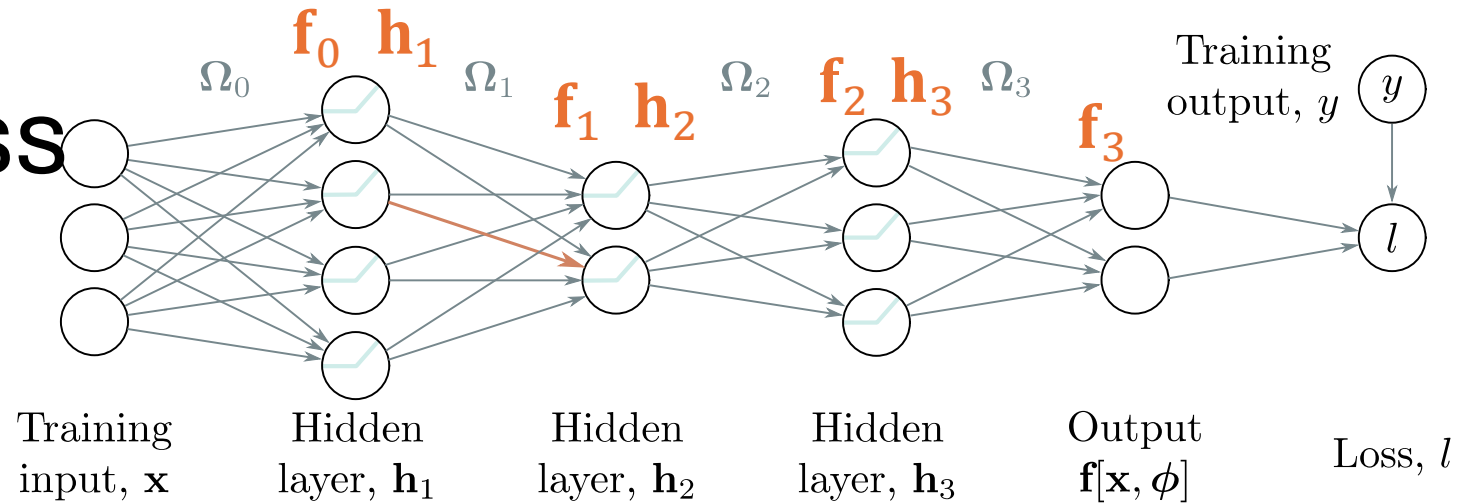
Gradients

- Backpropagation intuition
- Toy model
- Matrix calculus
- Backpropagation matrix forward pass
- Backpropagation matrix backward pass

The backward pass

Training output, y

Training input, $\mathbf{x}$

Hidden layer, $\mathbf{h}_1$

Hidden layer, $\mathbf{h}_2$

Hidden layer, $\mathbf{h}_3$

Output $\mathbf{f}[\mathbf{x}, \phi]$

Loss, l

1. Write this as a series of intermediate calculations

2. Compute these intermediate quantities

3. Take derivatives of output with respect to intermediate quantities

$$\mathbf{f}_0 = \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i$$

$$\mathbf{h}_1 = \mathbf{a}[\mathbf{f}_0]$$

$$\mathbf{f}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1$$

$$\mathbf{h}_2 = \mathbf{a}[\mathbf{f}_1]$$

$$\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2$$

$$\mathbf{h}_3 = \mathbf{a}[\mathbf{f}_2]$$

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$$

$$\ell_i = \mathbf{l}[\mathbf{f}_3, y_i]$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_3}$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_2} = \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \boxed{\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}} \frac{\partial \ell_i}{\partial \mathbf{f}_3}$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_1} = \frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \left(\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_0} = \frac{\partial \mathbf{h}_1}{\partial \mathbf{f}_0} \frac{\partial \mathbf{f}_1}{\partial \mathbf{h}_1} \left(\frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$
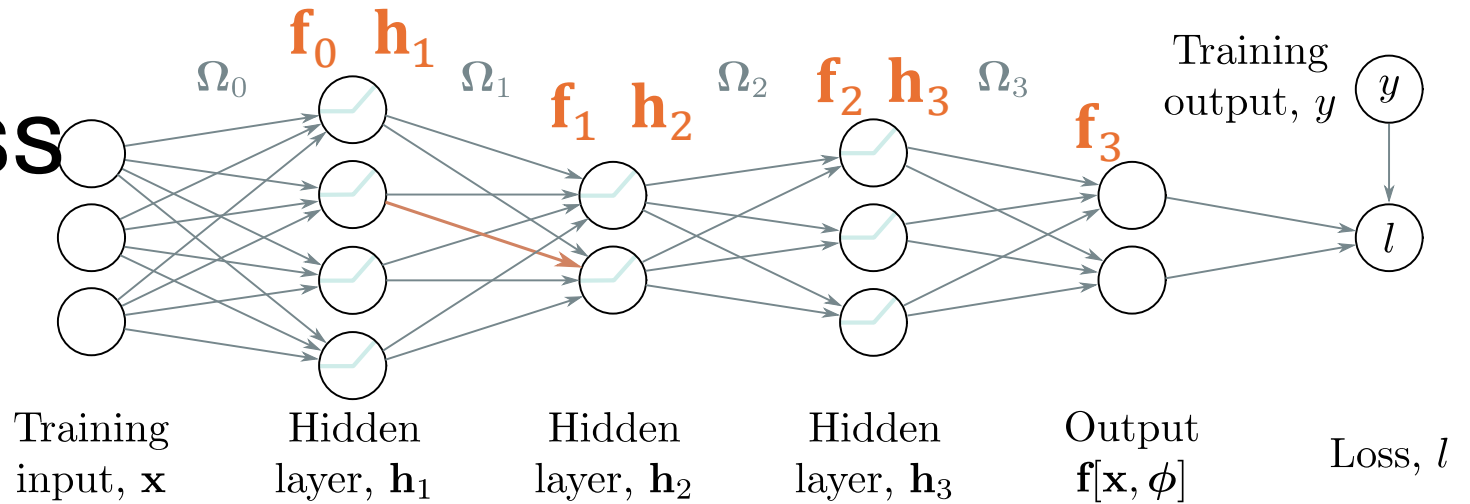
Yikes!

- But:

$$\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} = \frac{\partial}{\partial \mathbf{h}_3} \left(\boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \right) = \boldsymbol{\Omega}_3^T$$

- Quite similar to:

$$\frac{\partial f_3}{\partial h_3} = \frac{\partial}{\partial h_3} \left(\beta_3 + \omega_3 h_3 \right) = \omega_3$$

The backward pass

Training input, **x** | Hidden layer, $\mathbf{h}_1$ | Hidden layer, $\mathbf{h}_2$ | Hidden layer, $\mathbf{h}_3$ | Output $\mathbf{f}[\mathbf{x}, \phi]$ | Loss, l

1. Write this as a series of intermediate calculations

2. Compute these intermediate quantities

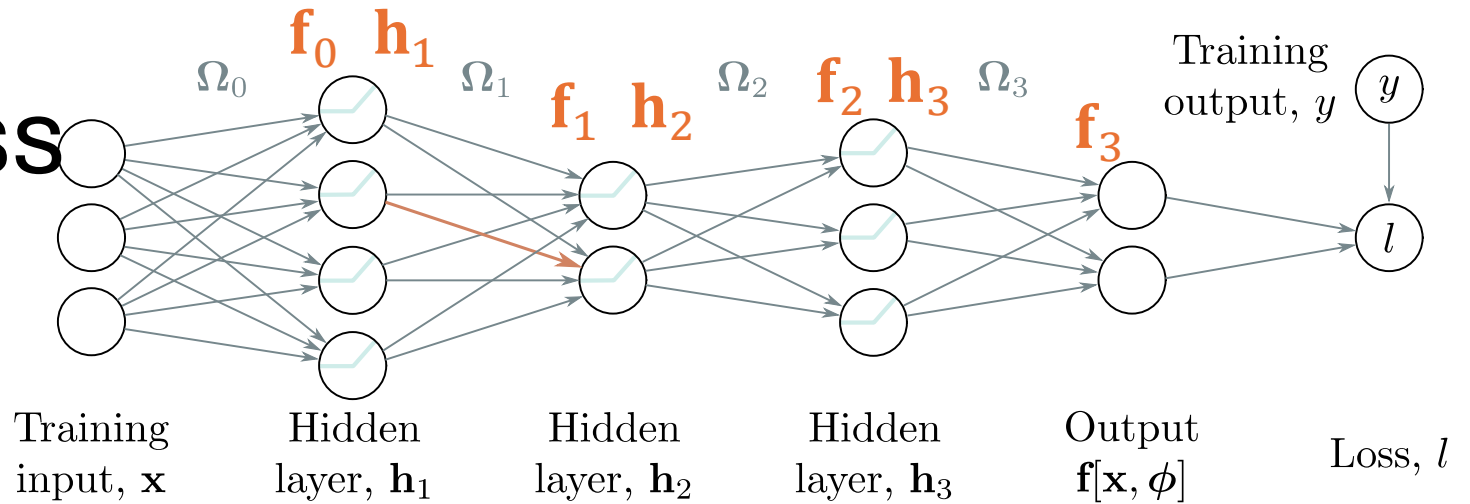3. Take derivatives of output with respect to intermediate quantities

$$\mathbf{f}_0 = \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i$$

$$\mathbf{h}_1 = \mathbf{a}[\mathbf{f}_0]$$

$$\mathbf{f}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1$$

$$\mathbf{h}_2 = \mathbf{a}[\mathbf{f}_1]$$

$$\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2$$

$$\mathbf{h}_3 = \mathbf{a}[\mathbf{f}_2]$$

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$$

$$\ell_i = \mathbf{l}[\mathbf{f}_3, y_i]$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_3}$$

$$\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} = \frac{\partial}{\partial \mathbf{h}_3} \left(\boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \right) = \boldsymbol{\Omega}_3^T$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_2} = \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \boxed{\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}} \frac{\partial \ell_i}{\partial \mathbf{f}_3}$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_1} = \frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \left(\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_0} = \frac{\partial \mathbf{h}_1}{\partial \mathbf{f}_0} \frac{\partial \mathbf{f}_1}{\partial \mathbf{h}_1} \left(\frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$

The backward pass

1. Write this as a series of intermediate calculations

2. Compute these intermediate quantities

3. Take derivatives of output with respect to intermediate quantities

$$\mathbf{f}_0 = \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i$$

$$\mathbf{h}_1 = \mathbf{a}[\mathbf{f}_0]$$

$$\mathbf{f}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1$$

$$\mathbf{h}_2 = \mathbf{a}[\mathbf{f}_1]$$

$$\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2$$

$$\mathbf{h}_3 = \mathbf{a}[\mathbf{f}_2]$$

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$$

$$\ell_i = \mathrm{l}[\mathbf{f}_3, y_i]$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_3}$$

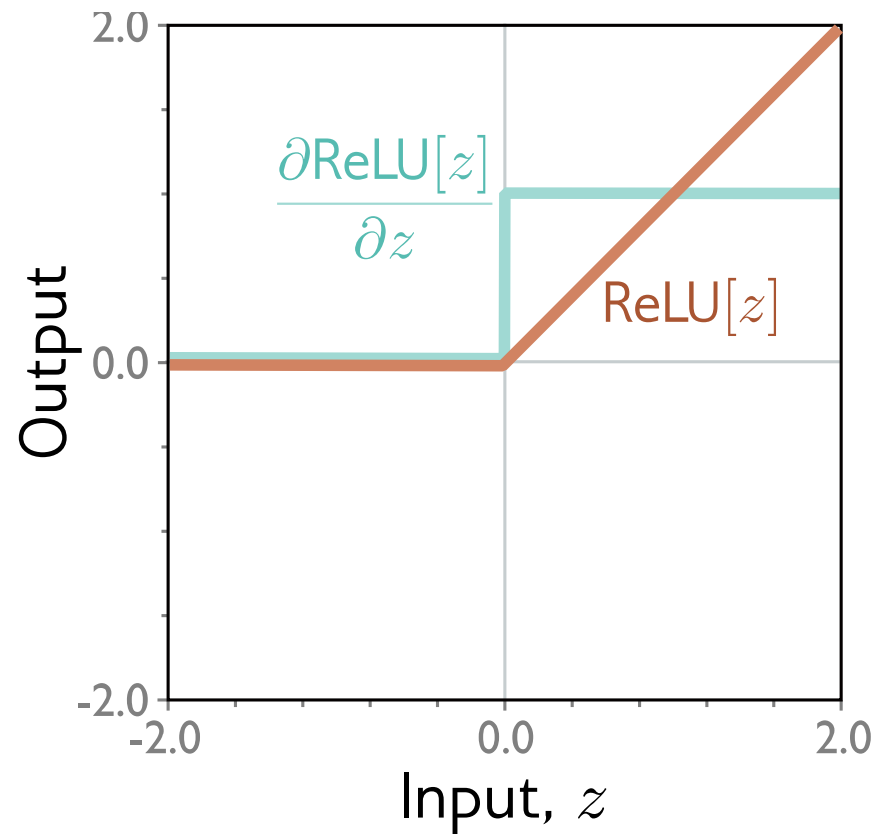$$\frac{\partial \ell_i}{\partial \mathbf{f}_2} = \boxed{\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2}} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3}$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_1} = \frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \left(\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$
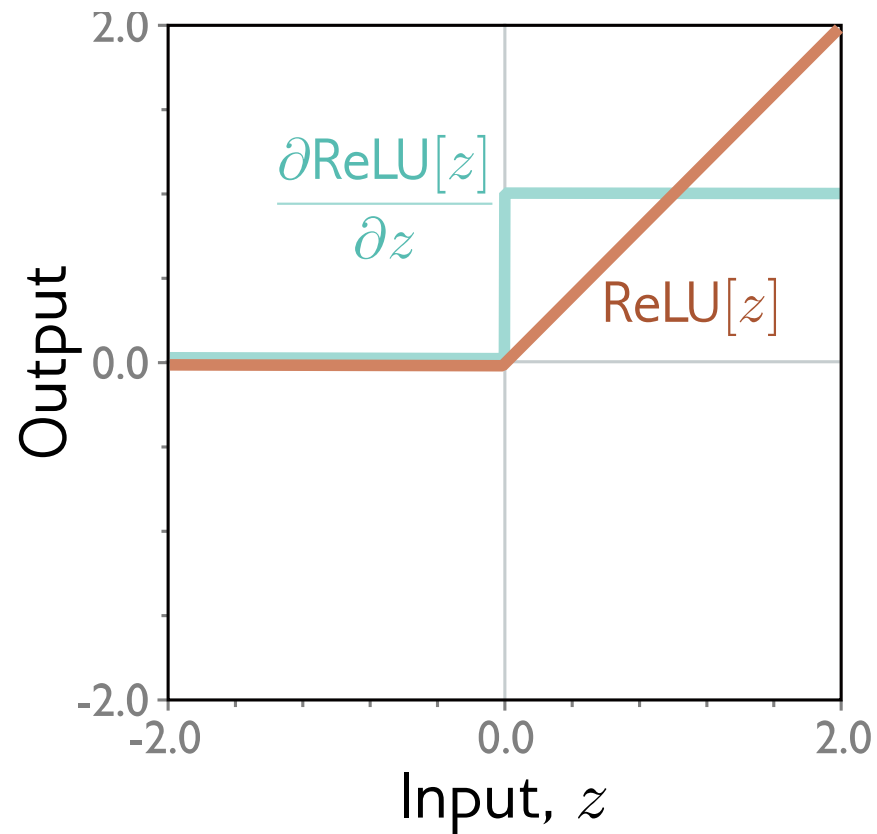
$$\frac{\partial \ell_i}{\partial \mathbf{f}_0} = \frac{\partial \mathbf{h}_1}{\partial \mathbf{f}_0} \frac{\partial \mathbf{f}_1}{\partial \mathbf{h}_1} \left(\frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$

Derivative of ReLU

Derivative of ReLU

$$\text{ReLU}[z] = \max(0, z)$$

$$\frac{\partial \text{ReLU}[z]}{\partial x} = \mathbb{I}[z > 0]$$

"Indicator function"

Derivative of ReLU

1. Consider:

$$\mathbf{a} = \mathbf{ReLU}[\mathbf{b}]$$

where:

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$(b > 0) = \begin{bmatrix} b_1 > 0 \\ b_2 > 0 \\ b_3 > 0 \end{bmatrix}$$

2. We could equivalently write:

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \mathrm{ReLU}[b_1] \\ \mathrm{ReLU}[b_2] \\ \mathrm{ReLU}[b_3] \end{bmatrix}$$

3. Taking the derivative
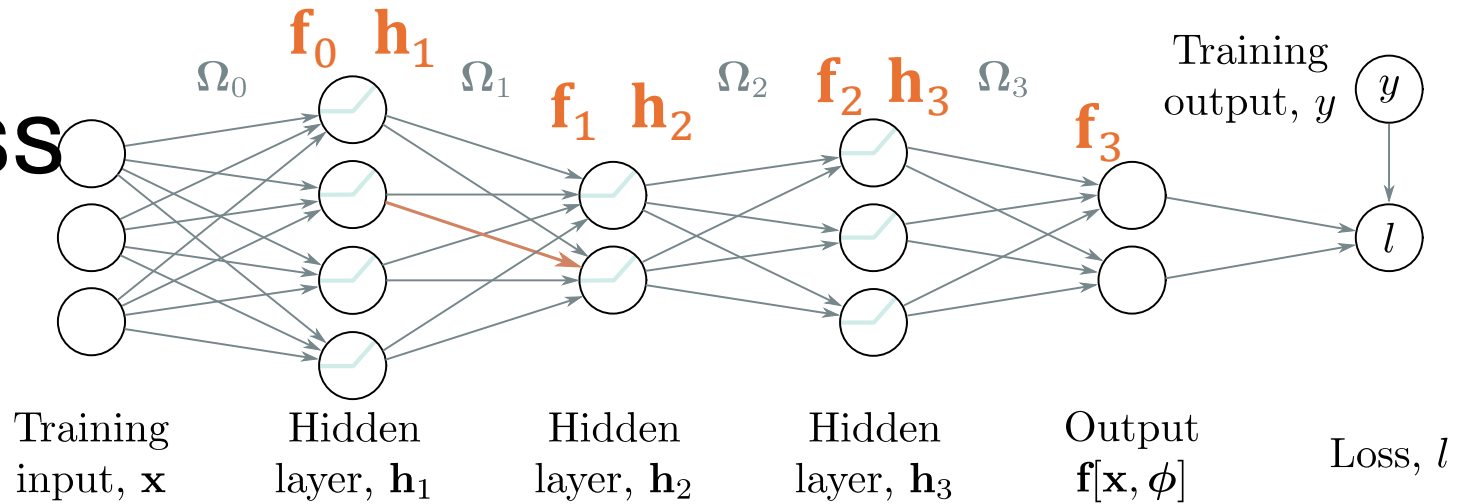
$$\mathbb{I}[\mathbf{b} > 0]$$

$$\frac{\partial \mathbf{a}}{\partial \mathbf{b}} = \begin{bmatrix} \frac{\partial a_1}{\partial b_1} & \frac{\partial a_2}{\partial b_1} & \frac{\partial a_3}{\partial b_1} \\ \frac{\partial a_1}{\partial b_2} & \frac{\partial a_2}{\partial b_2} & \frac{\partial a_3}{\partial b_2} \\ \frac{\partial a_1}{\partial b_3} & \frac{\partial a_2}{\partial b_3} & \frac{\partial a_3}{\partial b_3} \end{bmatrix} = \begin{bmatrix} \mathbb{I}[b_1 > 0] & 0 & 0 \\ 0 & \mathbb{I}[[b_2 > 0] & 0 \\ 0 & 0 & \mathbb{I}[b_3 > 0] \end{bmatrix}$$

4. We can equivalently pointwise multiply by diagonal

$$\mathbb{I}[\mathbf{b} > 0] \odot \quad \longleftarrow \text{ turn vector into diagonal of matrix}$$

vector w/ 1 if true

59

The backward pass

1. Write this as a series of intermediate calculations

2. Compute these intermediate quantities

3. Take derivatives of output with respect to intermediate quantities

$$\mathbf{f}_0 = \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i$$

$$\mathbf{h}_1 = \mathbf{a}[\mathbf{f}_0]$$

$$\mathbf{f}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1$$

$$\mathbf{h}_2 = \mathbf{a}[\mathbf{f}_1]$$

$$\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2$$

$$\mathbf{h}_3 = \mathbf{a}[\mathbf{f}_2]$$

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$$

$$\ell_i = \mathrm{l}[\mathbf{f}_3, y_i]$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_3}$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_2} = \boxed{\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2}} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3}$$
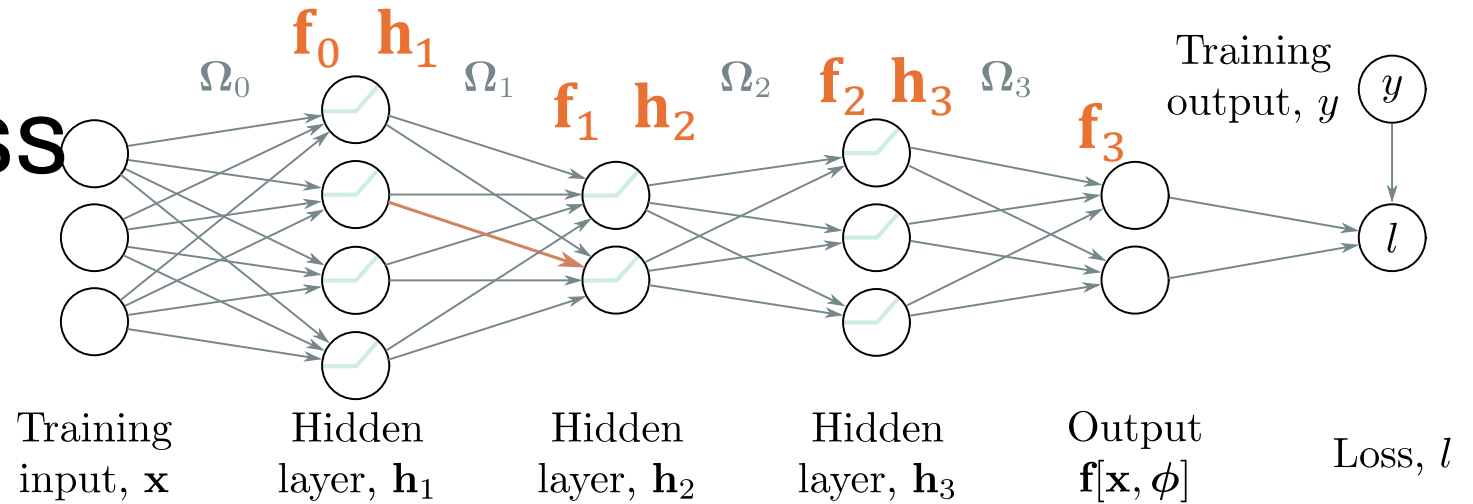
$$\frac{\partial \ell_i}{\partial \mathbf{f}_1} = \frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \left(\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_0} = \frac{\partial \mathbf{h}_1}{\partial \mathbf{f}_0} \frac{\partial \mathbf{f}_1}{\partial \mathbf{h}_1} \left(\frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$

$$\boxed{\mathbb{I}[\mathbf{f}_2 > 0]}$$

The backward pass

1. Write this as a series of intermediate calculations

2. Compute these intermediate quantities

3. Take derivatives of output with respect to intermediate quantities

4. Take derivatives w.r.t. parameters

$$\mathbf{f}_0 = \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i$$

$$\mathbf{h}_1 = \mathbf{a}[\mathbf{f}_0]$$

$$\mathbf{f}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1$$

$$\mathbf{h}_2 = \mathbf{a}[\mathbf{f}_1]$$

$$\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2$$

$$\mathbf{h}_3 = \mathbf{a}[\mathbf{f}_2]$$

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$$

$$\ell_i = \mathbf{l}[\mathbf{f}_3, y_i]$$

$$\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_k} = \frac{\partial \mathbf{f}_k}{\partial \boldsymbol{\beta}_k} \frac{\partial \ell_i}{\partial \mathbf{f}_k}$$

$$= \frac{\partial}{\partial \boldsymbol{\beta}_k} (\boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \mathbf{h}_k) \frac{\partial \ell_i}{\partial \mathbf{f}_k}$$

$$= \frac{\partial \ell_i}{\partial \mathbf{f}_k},$$

The backward pass

1. Write this as a series of intermediate calculations

2. Compute these intermediate quantities

3. Take derivatives of output with respect to intermediate quantities

4. Take derivatives w.r.t. parameters

$$\mathbf{f}_0 = \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i$$

$$\mathbf{h}_1 = \mathbf{a}[\mathbf{f}_0]$$

$$\mathbf{f}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1$$

$$\mathbf{h}_2 = \mathbf{a}[\mathbf{f}_1]$$

$$\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2$$

$$\mathbf{h}_3 = \mathbf{a}[\mathbf{f}_2]$$

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3$$

$$\ell_i = \mathrm{l}[\mathbf{f}_3, y_i]$$

repeat w Ω_k

$$\frac{\partial \ell_i}{\partial \boldsymbol{\Omega}_k} = \frac{\partial \mathbf{f}_k}{\partial \boldsymbol{\Omega}_k} \frac{\partial \ell_i}{\partial \mathbf{f}_k}$$

$$= \frac{\partial}{\partial \boldsymbol{\Omega}_k} \left(\boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \mathbf{h}_k \right) \frac{\partial \ell_i}{\partial \mathbf{f}_k}$$

$$= \frac{\partial \ell_i}{\partial \mathbf{f}_k} \mathbf{h}_k^T$$

62

Pros and cons

- Extremely efficient
  - Only need matrix multiplication and thresholding for ReLU functions
- Memory hungry – must store all the intermediate quantities
- Sequential
  - can process multiple batches in parallel
  - but things get harder if the whole model doesn't fit on one machine.

Looking Ahead to Initialization

The chain rule tells us to multiply all these "local" partial derivatives together...

$$\frac{\partial \ell_i}{\partial \mathbf{f}_1} = \frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \left(\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_0} = \frac{\partial \mathbf{h}_1}{\partial \mathbf{f}_0} \frac{\partial \mathbf{f}_1}{\partial \mathbf{h}_1} \left(\frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$

- What happens when most of those values are >2.0?
- What happens when most of those values are <0.5?

Our initialization will be setting the initial local partial derivatives.