



Lecture 08

Gradients

DL4DS – Spring 2026

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:

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

or for short:

$$L[\phi]$$

← 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:

$$\frac{\partial L}{\partial \phi} = \begin{bmatrix} \frac{\partial L}{\partial \phi_0} \\ \frac{\partial L}{\partial \phi_1} \\ \vdots \\ \frac{\partial L}{\partial \phi_N} \end{bmatrix}. \quad \text{Also notated as } \nabla_w L$$

Step 2. Update the parameters according to the rule:

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

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.

$$L[\phi] = \sum_{i=1}^I \ell_i = \sum_{i=1}^I (f[x_i, \phi] - y_i)^2 \\ = \sum_{i=1}^I (\phi_0 + \phi_1 x_i - y_i)^2$$

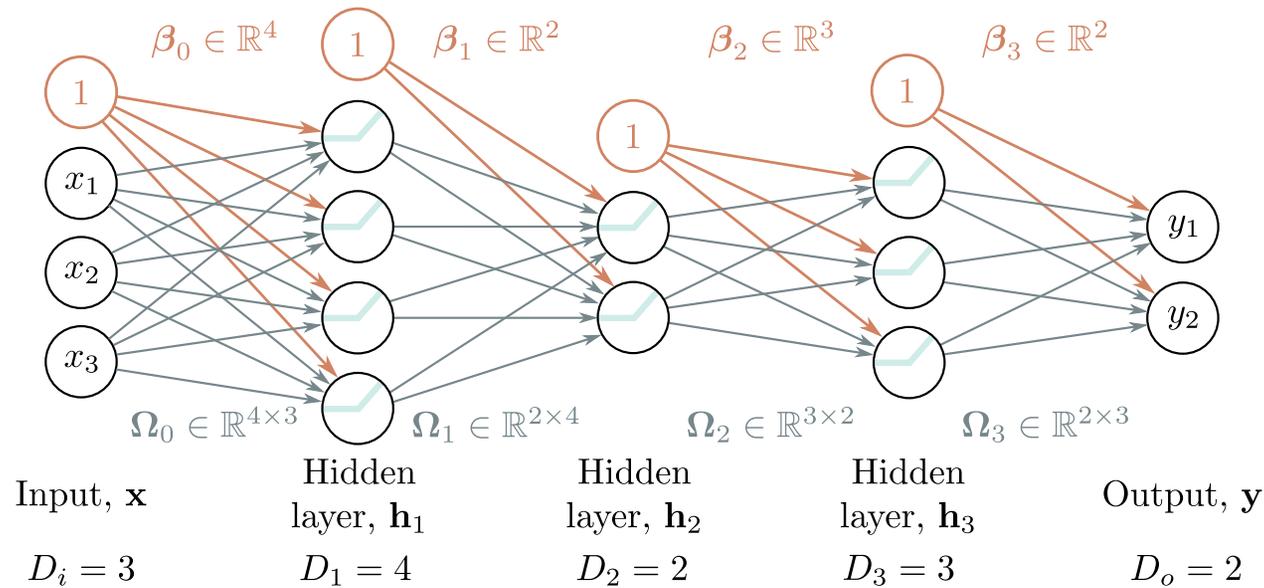
Least squares loss for linear regression

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

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

Partial derivative w.r.t. each parameter

What about deep learning models?



$$\mathbf{h}_1 = \mathbf{a}[\beta_0 + \Omega_0 \mathbf{x}]$$

$$\mathbf{h}_2 = \mathbf{a}[\beta_1 + \Omega_1 \mathbf{h}_1]$$

$$\mathbf{h}_3 = \mathbf{a}[\beta_2 + \Omega_2 \mathbf{h}_2]$$

$$\mathbf{f}[\mathbf{x}, \phi] = \beta_3 + \Omega_3 \mathbf{h}_3$$

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

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:

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

Millions and even *billions* of parameters:

$$\phi = \{\beta_0, \Omega_0, \beta_1, \Omega_1, \beta_2, \Omega_2, \dots\}$$

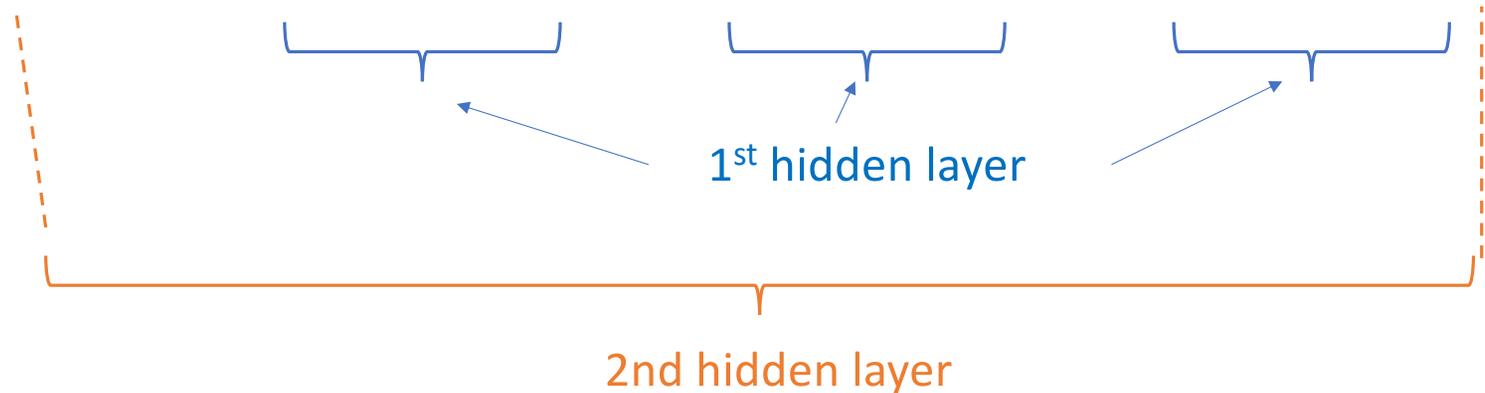
We need the partial derivative with respect to every weight and bias we want to update for every sample in the batch.

$$\frac{\partial \ell_i}{\partial \beta_k} \quad \text{and} \quad \frac{\partial \ell_i}{\partial \Omega_k}$$

Network equation gets unwieldy even for small models

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

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



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:

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

Parameters:

$$\phi = \{\beta_0, \Omega_0, \beta_1, \Omega_1, \beta_2, \Omega_2, \beta_3, \Omega_3\}$$

Need to compute gradients

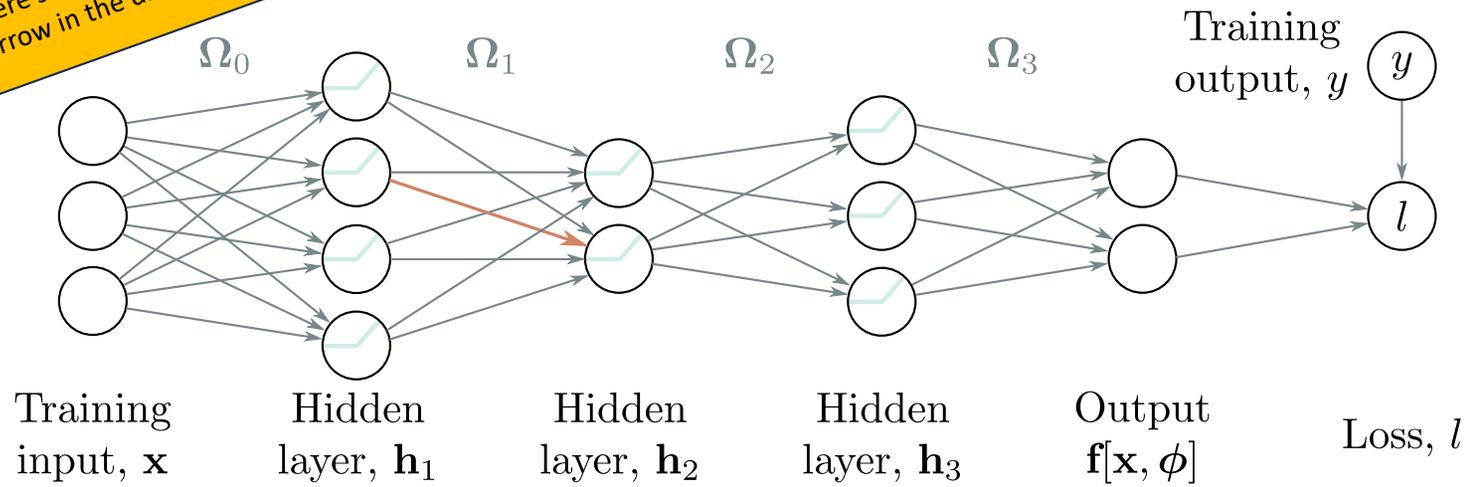
$$\frac{\partial \ell_i}{\partial \beta_k} \quad \text{and} \quad \frac{\partial \ell_i}{\partial \Omega_k}$$

Algorithm to compute gradient efficiently

- “Backpropagation algorithm”
- Rumelhart, Hinton, and Williams (1986)

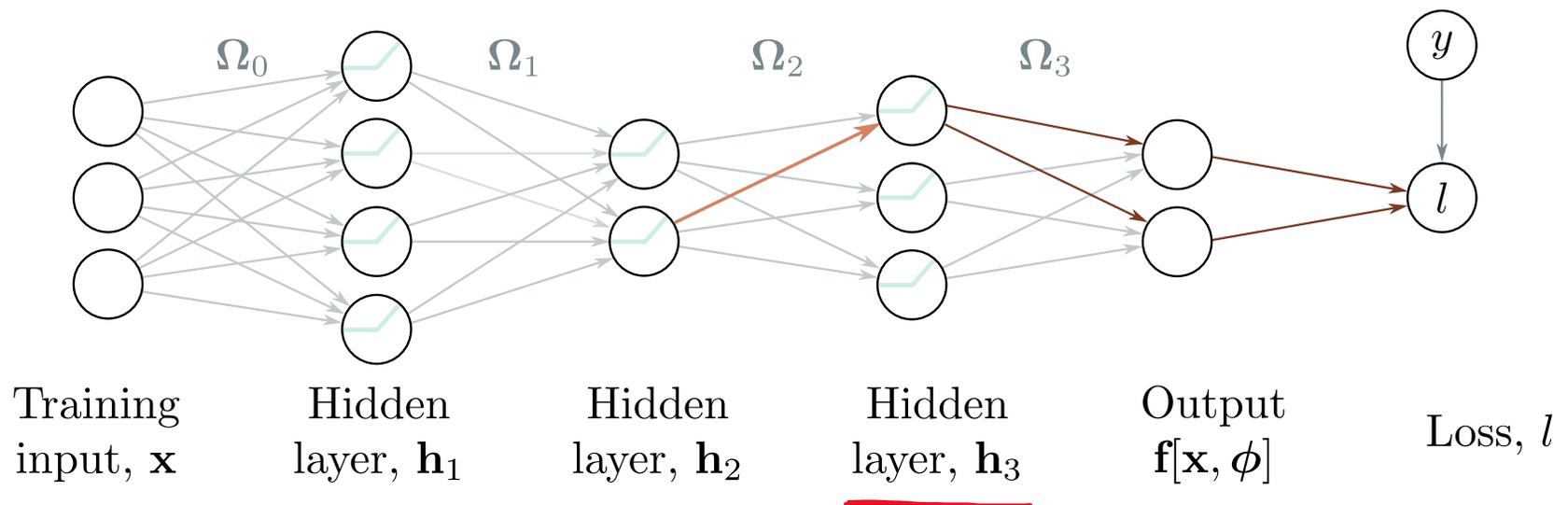
BackProp intuition #1: the forward pass

Remember! There's an implied weight on every arrow in the diagram



- The weight on the orange arrow multiplies activation (ReLU output) of previous layer
- We want to know how change (*partial derivative*) w.r.t. 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

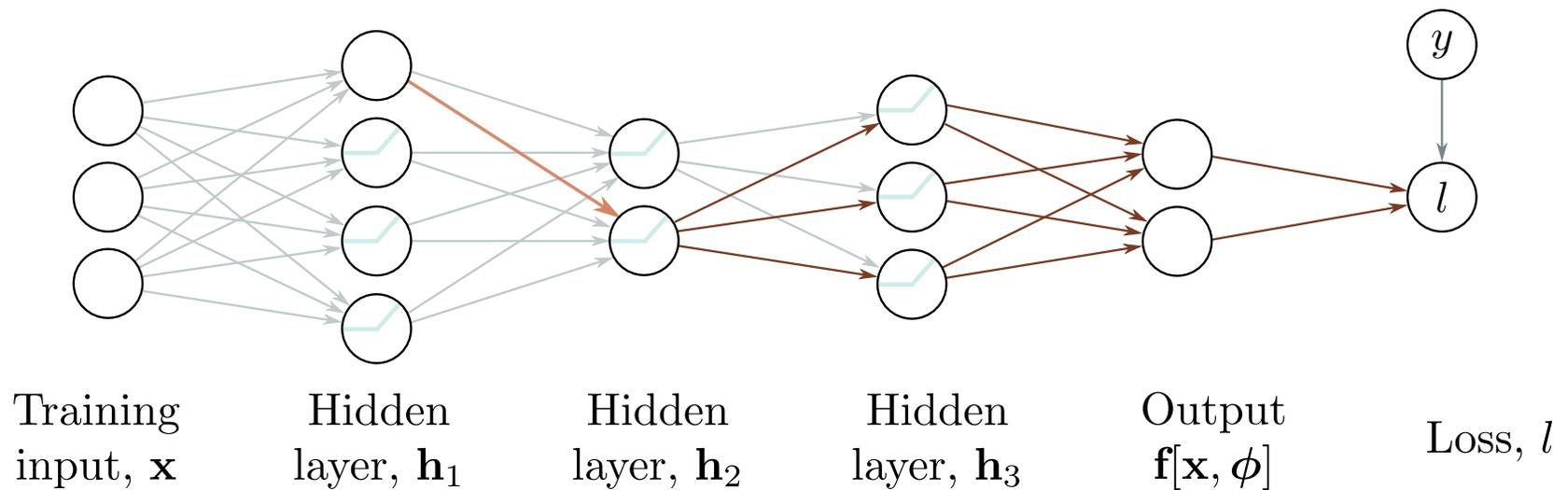
BackProp intuition #2: the backward pass



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



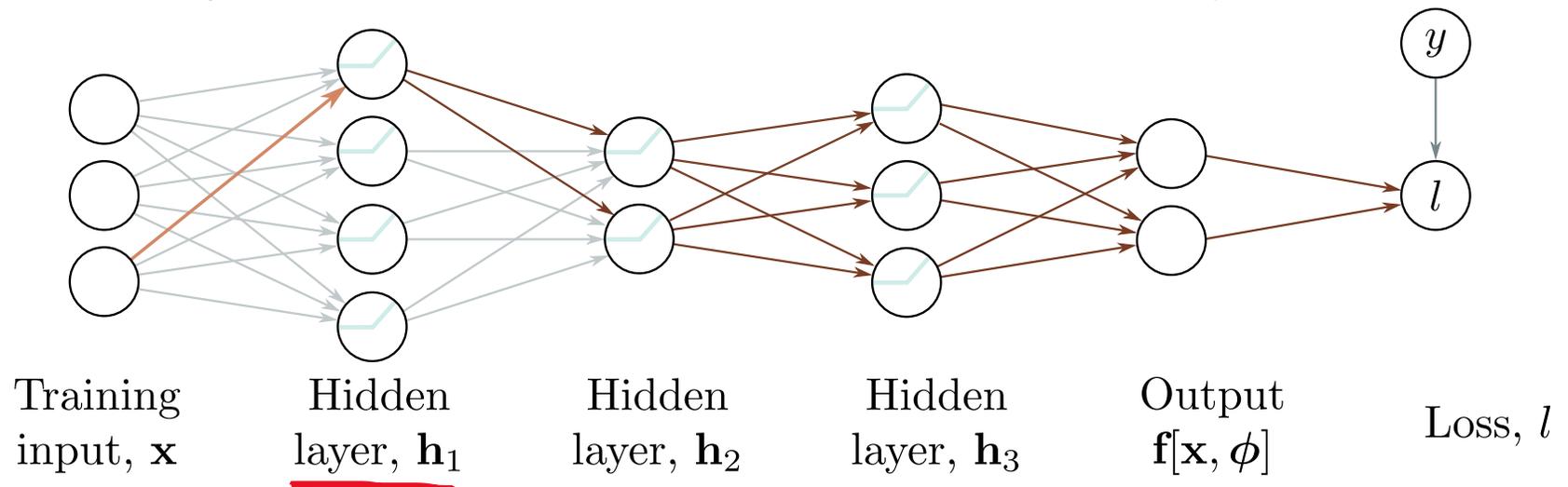
To calculate how a small change in a weight or bias feeding into hidden layer \mathbf{h}_2 modifies the loss, we need to know:

- 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

BackProp intuition #2: the backward pass



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

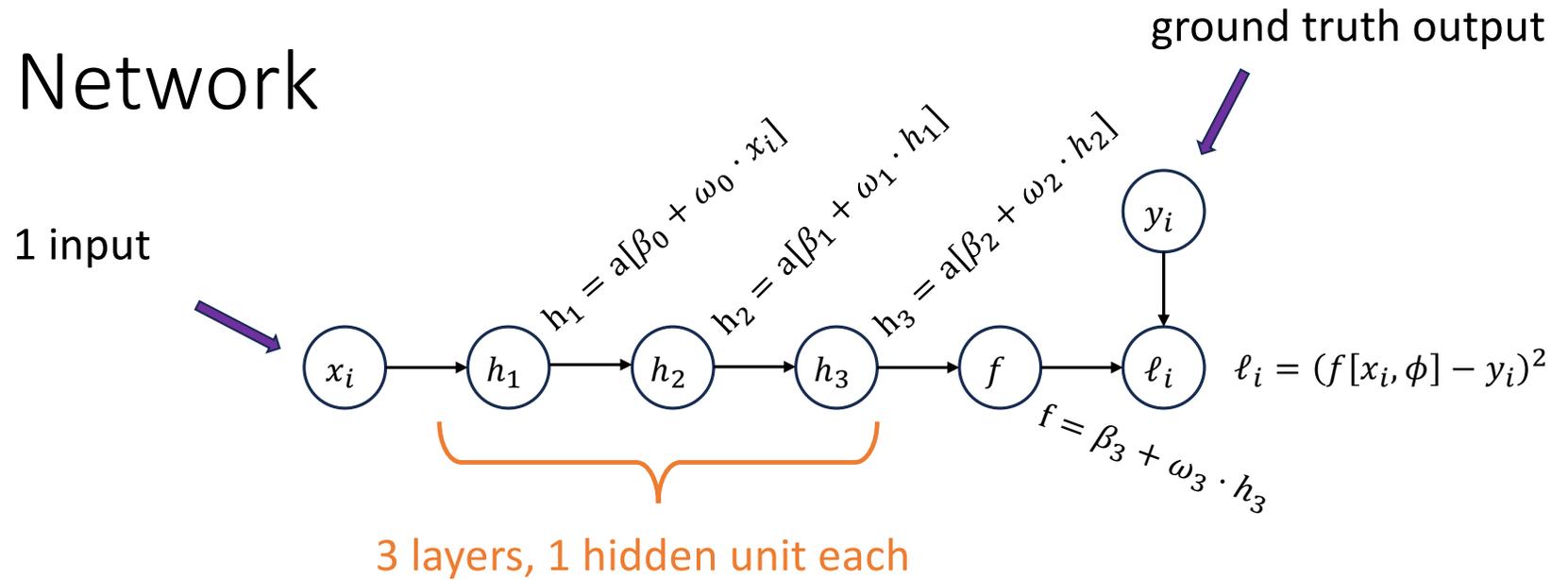


We know these from the previous steps

Gradients

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

Toy Network



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

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

Gradients of toy function

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

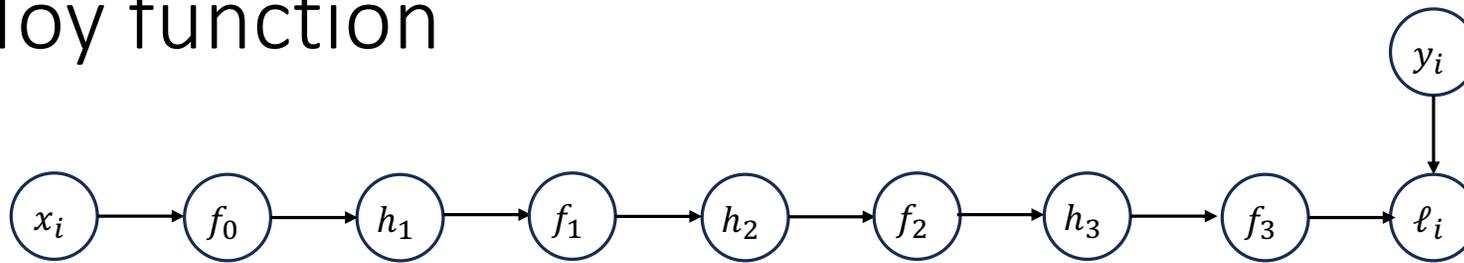
We want to calculate each partial:

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

Tells us how a small change in β_i or ω_i change the loss ℓ_i for the i^{th} example



Toy function



Pre-Activations

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

$$h_1 = a[f_0]$$

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

$$h_2 = a[f_1]$$

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

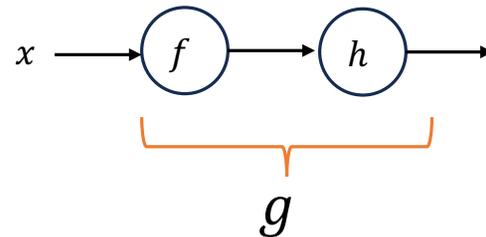
$$h_3 = a[f_2]$$

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

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

Intermediate values

Refresher: The Chain Rule



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

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

Or can be written equivalently as

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

Leibniz's Notation

Lagrange's Notation

Forward pass

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

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

1. Write this as a series of intermediate calculations

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

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

2. Compute these intermediate quantities

$$h_1 = a[f_0]$$

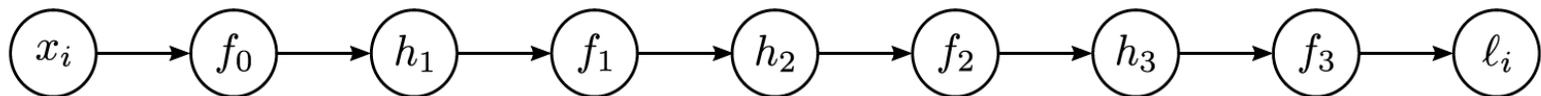
$$h_3 = a[f_2]$$

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

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

$$h_2 = a[f_1]$$

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



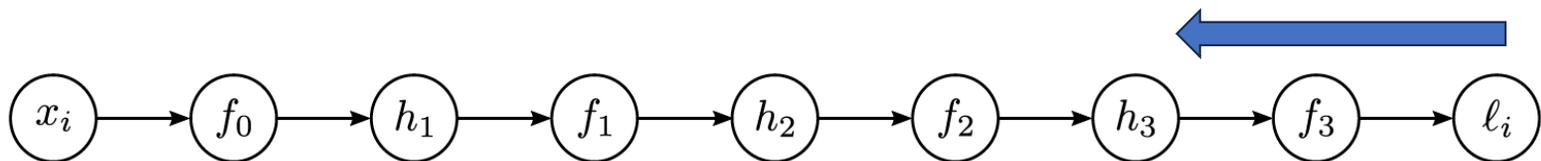
Backward pass

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

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

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

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

Backward pass

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

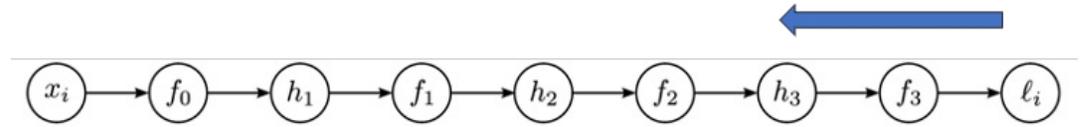
$$l_i = (f[x_i, \phi] - y_i)^2$$

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

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



Backward pass



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

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

$$h_1 = a[f_0]$$

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

$$h_2 = a[f_1]$$

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

$$h_3 = a[f_2]$$

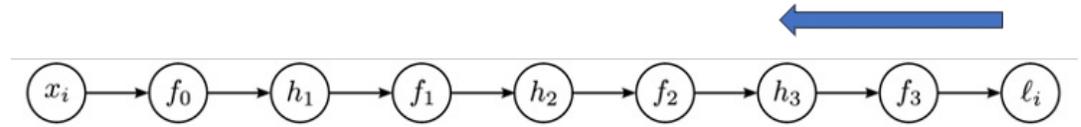
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$

$$l_i = (f_3 - y_i)^2$$

- The first of these derivatives is trivial

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

Backward pass



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

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

$$h_1 = a[f_0]$$

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

$$h_2 = a[f_1]$$

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

$$h_3 = a[f_2]$$

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

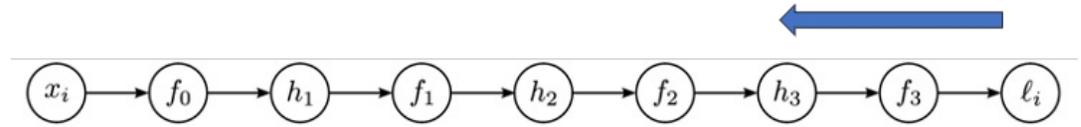
$$\ell_i = (y_i - f_3)^2$$

- 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

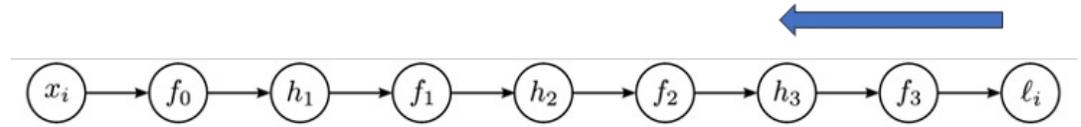
$$\frac{\partial \ell_i}{\partial h_3} = \frac{\partial f_3}{\partial h_3} \frac{\partial \ell_i}{\partial f_3}$$

How does a small change in h_3 change ℓ_i ?

How does a small change in h_3 change f_3 ?

How does a small change in f_3 change ℓ_i ?

Backward pass



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

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

$$h_1 = a[f_0]$$

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

$$h_2 = a[f_1]$$

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

$$h_3 = a[f_2]$$

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

$$l_i = (y_i - f_3)^2$$

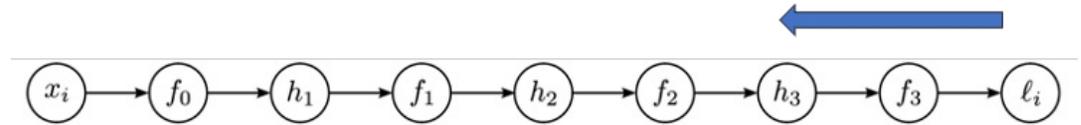
- The second of these derivatives is computed via the chain rule

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



Already computed!

Backward pass



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

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

$$h_1 = a[f_0]$$

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

$$h_2 = a[f_1]$$

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

$$h_3 = a[f_2]$$

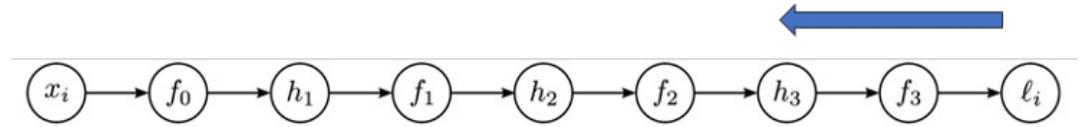
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$

$$l_i = (y_i - f_3)^2$$

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

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

Backward pass



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

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

$$h_1 = a[f_0]$$

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

$$h_2 = a[f_1]$$

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

$$h_3 = a[f_2]$$

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

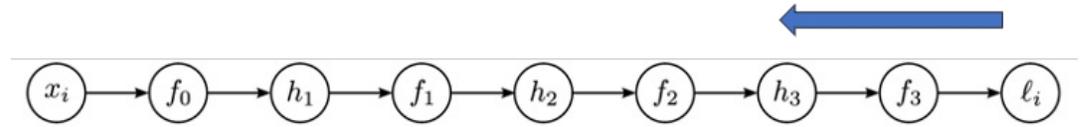
$$l_i = (y_i - f_3)^2$$

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

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

Already computed!

Backward pass



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

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

$$h_1 = a[f_0]$$

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

$$h_2 = a[f_1]$$

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

$$h_3 = a[f_2]$$

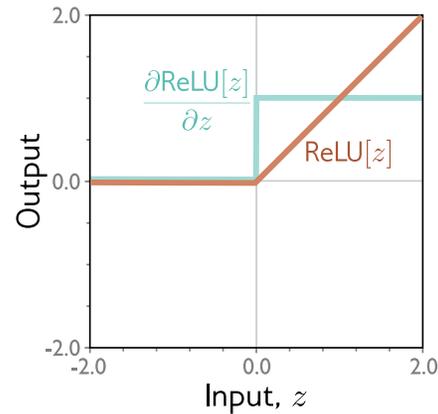
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$

$$l_i = (y_i - f_3)^2$$

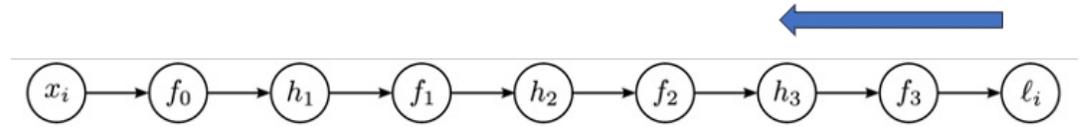
- The remaining derivatives also calculated by further use of chain rule

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

Indicator function: $\mathbf{1}_{\{z>0\}}$



Backward pass



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

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

$$h_1 = a[f_0]$$

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

$$h_2 = a[f_1]$$

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

$$h_3 = a[f_2]$$

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

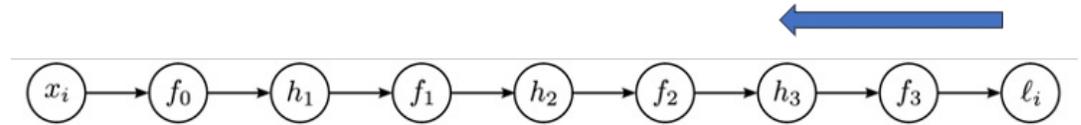
$$l_i = (y_i - f_3)^2$$

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

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

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

Backward pass



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

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

$$h_1 = a[f_0]$$

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

$$h_2 = a[f_1]$$

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

$$h_3 = a[f_2]$$

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

$$l_i = (y_i - f_3)^2$$

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

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

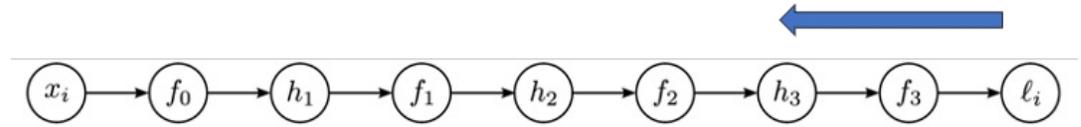
$$\frac{\partial l_i}{\partial h_2} = \frac{\partial f_2}{\partial h_2} \left(\frac{\partial h_3}{\partial f_2} \frac{\partial f_3}{\partial h_3} \frac{\partial l_i}{\partial f_3} \right)$$

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

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

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

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

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

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

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

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

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

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

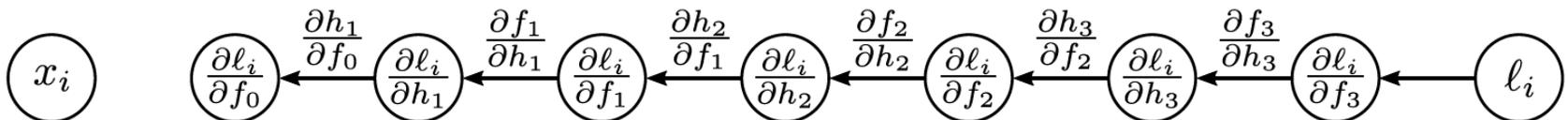
$$\frac{\partial l_i}{\partial f_0} = \frac{\partial h_1}{\partial f_0} \left(\frac{\partial f_1}{\partial h_1} \frac{\partial h_2}{\partial f_1} \frac{\partial f_2}{\partial h_2} \frac{\partial h_3}{\partial f_2} \frac{\partial f_3}{\partial h_3} \frac{\partial l_i}{\partial f_3} \right)$$

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

$$\begin{aligned} \frac{\partial l_i}{\partial f_3} &= 2(f_3 - y_i) \\ \frac{\partial l_i}{\partial h_3} &= \frac{\partial f_3}{\partial h_3} \frac{\partial l_i}{\partial f_3} \\ \frac{\partial l_i}{\partial f_2} &= \frac{\partial h_3}{\partial f_2} \left(\frac{\partial f_3}{\partial h_3} \frac{\partial l_i}{\partial f_3} \right) \\ \frac{\partial l_i}{\partial h_2} &= \frac{\partial f_2}{\partial h_2} \left(\frac{\partial h_3}{\partial f_2} \frac{\partial f_3}{\partial h_3} \frac{\partial l_i}{\partial f_3} \right) \\ \frac{\partial l_i}{\partial f_1} &= \frac{\partial h_2}{\partial f_1} \left(\frac{\partial f_2}{\partial h_2} \frac{\partial h_3}{\partial f_2} \frac{\partial f_3}{\partial h_3} \frac{\partial l_i}{\partial f_3} \right) \\ \frac{\partial l_i}{\partial h_1} &= \frac{\partial f_1}{\partial h_1} \left(\frac{\partial h_2}{\partial f_1} \frac{\partial f_2}{\partial h_2} \frac{\partial h_3}{\partial f_2} \frac{\partial f_3}{\partial h_3} \frac{\partial l_i}{\partial f_3} \right) \\ \frac{\partial l_i}{\partial f_0} &= \frac{\partial h_1}{\partial f_0} \left(\frac{\partial f_1}{\partial h_1} \frac{\partial h_2}{\partial f_1} \frac{\partial f_2}{\partial h_2} \frac{\partial h_3}{\partial f_2} \frac{\partial f_3}{\partial h_3} \frac{\partial l_i}{\partial f_3} \right) \end{aligned}$$



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

Backward pass

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

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

$$h_1 = a[f_0]$$

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

$$h_2 = a[f_1]$$

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

$$h_3 = a[f_2]$$

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

$$l_i = (y_i - f_3)^2$$

- Another application of the chain rule

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

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 .

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

$$h_1 = a[f_0]$$

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

$$h_2 = a[f_1]$$

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

$$h_3 = a[f_2]$$

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

$$l_i = (y_i - f_3)^2$$

- Another application of the chain rule

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

How does a small change in ω_k change l_i ?

Already calculated in part 1.

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

We have from the forward pass.

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 .

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

$$h_1 = a[f_0]$$

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

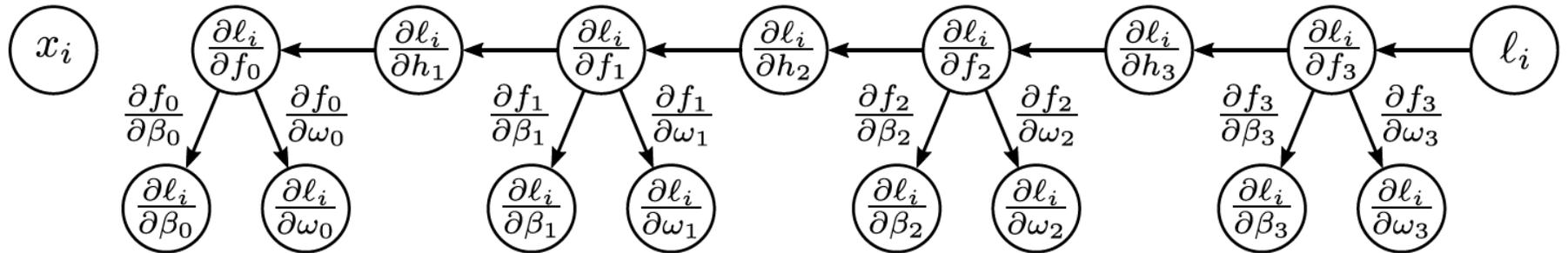
$$h_2 = a[f_1]$$

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

$$h_3 = a[f_2]$$

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

$$l_i = (y_i - f_3)^2$$



slido

Please download and install the Slido app on all computers you use



Which statement best describes the purpose of backpropagation in training a neural network?

① Start presenting to display the poll results on this slide.

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}

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

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

The derivative is a vector of shape \mathbf{a}

Matrix calculus

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

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

The partial derivative is a matrix of shape same as \mathbf{A}

Matrix calculus

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

$$\mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} \quad \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}$$

Vector of scalar
valued functions

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

Columns are each
element function

Rows are each
variable element

Comparing vector and matrix

Scalar derivatives:

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

Comparing vector and matrix

Scalar derivatives:

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

Matrix derivatives:

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

Comparing vector and matrix

Scalar derivatives:

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

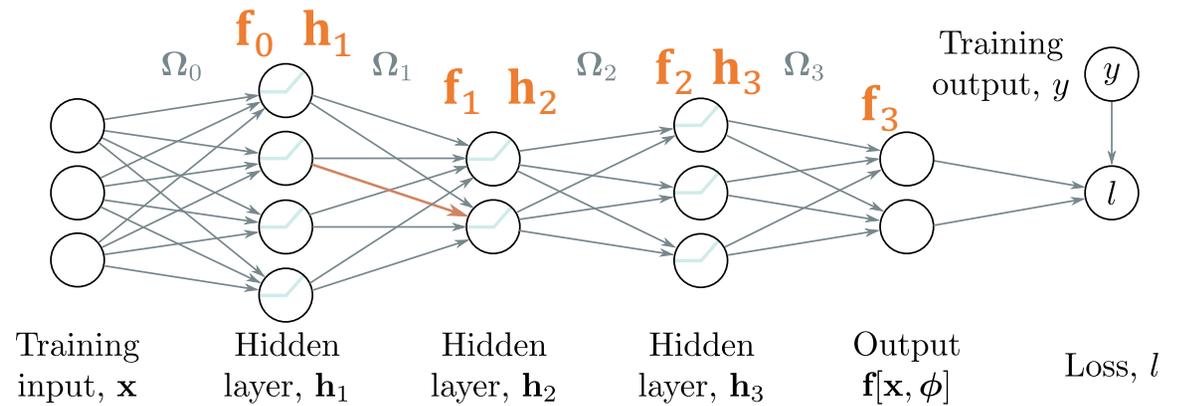
Matrix derivatives:

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

Gradients

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

The forward pass



1. Write this as a series of intermediate calculations

$$\mathbf{f}_0 = \beta_0 + \Omega_0 \mathbf{x}_i$$

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

$$\mathbf{f}_1 = \beta_1 + \Omega_1 \mathbf{h}_1$$

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

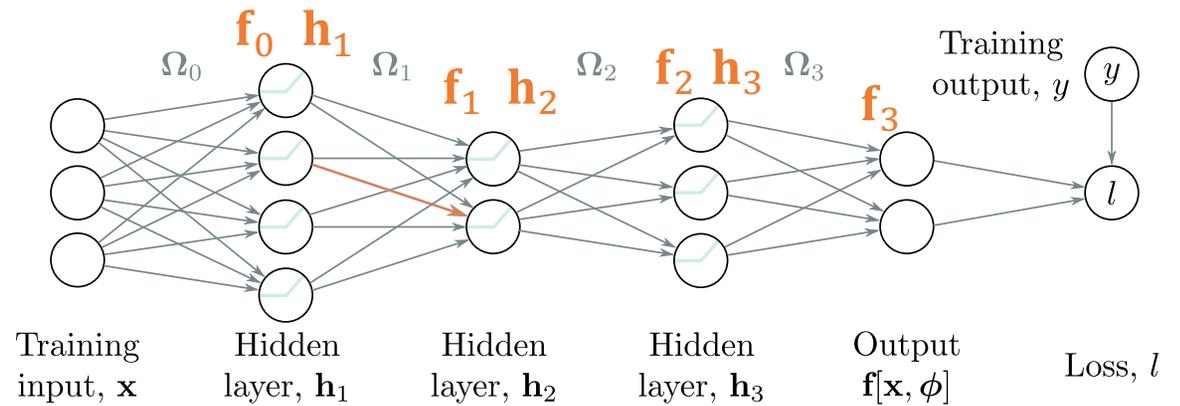
$$\mathbf{f}_2 = \beta_2 + \Omega_2 \mathbf{h}_2$$

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

$$\mathbf{f}_3 = \beta_3 + \Omega_3 \mathbf{h}_3$$

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

The forward pass



1. Write this as a series of intermediate calculations

$$\mathbf{f}_0 = \beta_0 + \Omega_0 \mathbf{x}_i$$

2. Compute these

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

intermediate quantities

$$\mathbf{f}_1 = \beta_1 + \Omega_1 \mathbf{h}_1$$

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

$$\mathbf{f}_2 = \beta_2 + \Omega_2 \mathbf{h}_2$$

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

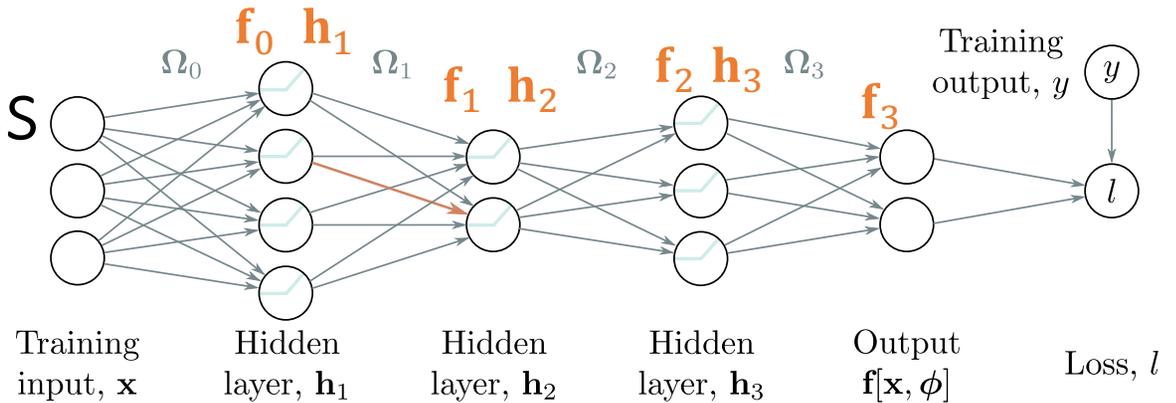
$$\mathbf{f}_3 = \beta_3 + \Omega_3 \mathbf{h}_3$$

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

Gradients

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

The backward pass



1. Write this as a series of intermediate calculations

$$\mathbf{f}_0 = \beta_0 + \mathbf{\Omega}_0 \mathbf{x}_i$$

2. Compute these intermediate quantities

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

$$\mathbf{f}_1 = \beta_1 + \mathbf{\Omega}_1 \mathbf{h}_1$$

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

$$\mathbf{f}_2 = \beta_2 + \mathbf{\Omega}_2 \mathbf{h}_2$$

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

$$\mathbf{f}_3 = \beta_3 + \mathbf{\Omega}_3 \mathbf{h}_3$$

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

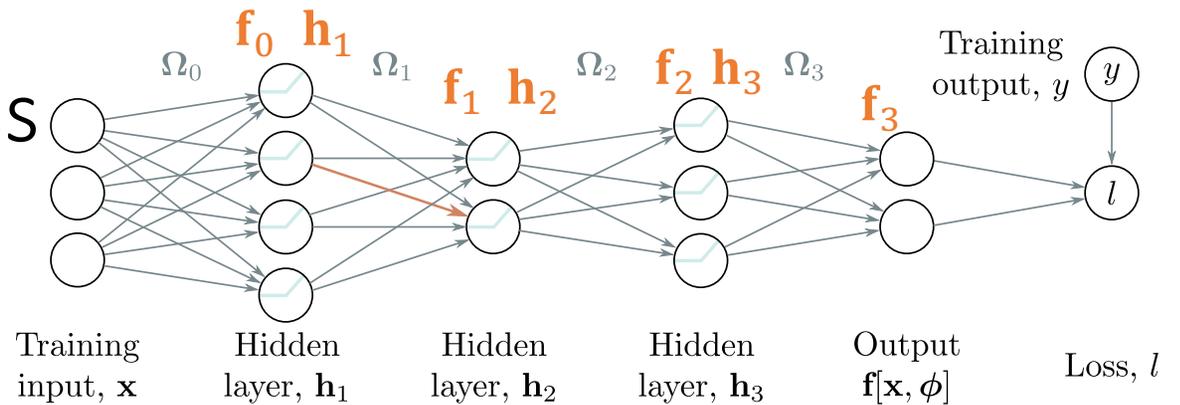
$$\frac{\partial l_i}{\partial \mathbf{f}_3}$$

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

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

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

The backward pass



1. Write this as a series of intermediate calculations

$$\mathbf{f}_0 = \beta_0 + \mathbf{\Omega}_0 \mathbf{x}_i$$

2. Compute these intermediate quantities

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

$$\mathbf{f}_1 = \beta_1 + \mathbf{\Omega}_1 \mathbf{h}_1$$

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

$$\mathbf{f}_2 = \beta_2 + \mathbf{\Omega}_2 \mathbf{h}_2$$

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

$$\mathbf{f}_3 = \beta_3 + \mathbf{\Omega}_3 \mathbf{h}_3$$

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

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

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

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

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

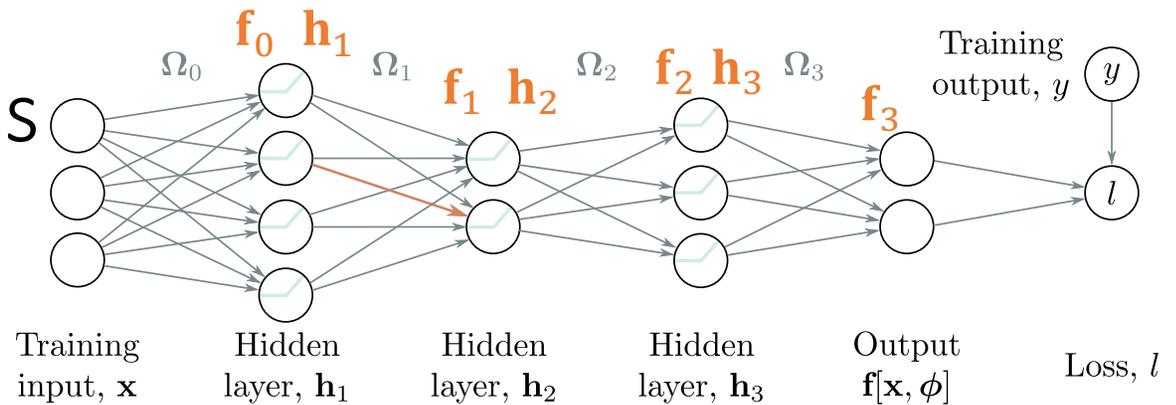
- But:

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

- Quite similar to:

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

The backward pass



1. Write this as a series of intermediate calculations

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

2. Compute these intermediate quantities

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

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

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

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

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

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

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

3. Take derivatives of output with respect to intermediate quantities

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

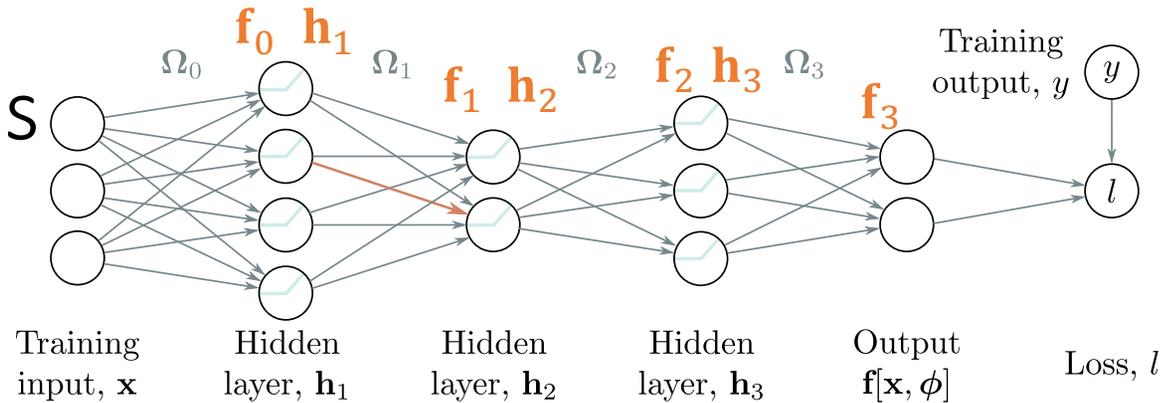
$$\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} = \frac{\partial}{\partial \mathbf{h}_3} (\boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3) = \boldsymbol{\Omega}_3^T$$

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

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

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

The backward pass



1. Write this as a series of intermediate calculations

$$\mathbf{f}_0 = \beta_0 + \Omega_0 \mathbf{x}_i$$

2. Compute these intermediate quantities

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

$$\mathbf{f}_1 = \beta_1 + \Omega_1 \mathbf{h}_1$$

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

$$\mathbf{f}_2 = \beta_2 + \Omega_2 \mathbf{h}_2$$

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

$$\mathbf{f}_3 = \beta_3 + \Omega_3 \mathbf{h}_3$$

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

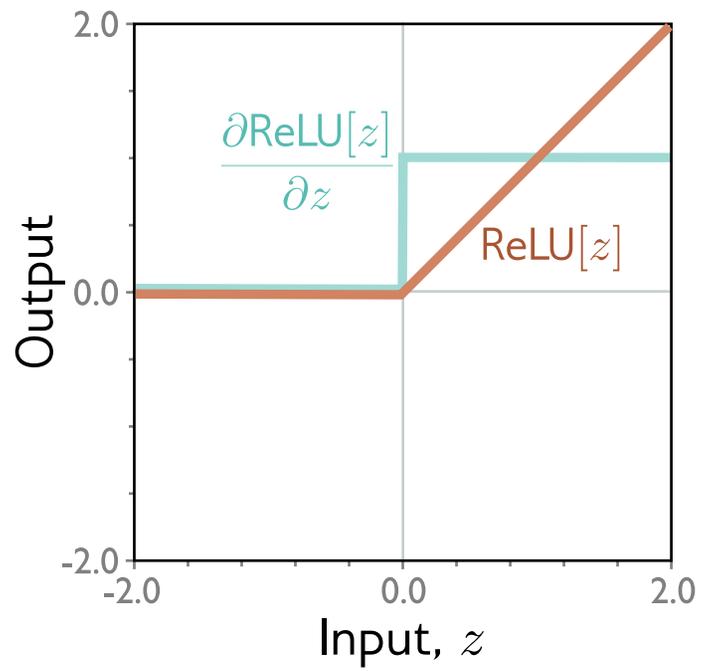
$$\frac{\partial l_i}{\partial \mathbf{f}_3}$$

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

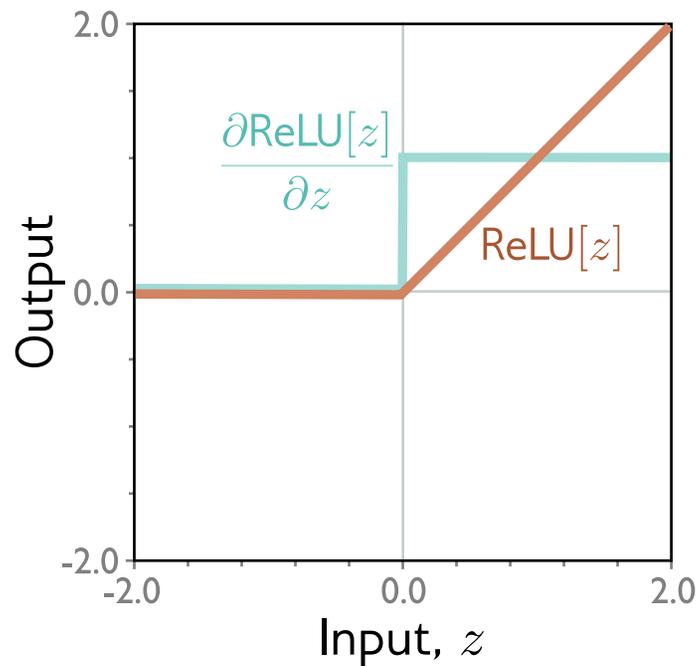
$$\frac{\partial l_i}{\partial \mathbf{f}_1} = \frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \left(\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial l_i}{\partial \mathbf{f}_3}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

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

Derivative of ReLU



Derivative of ReLU



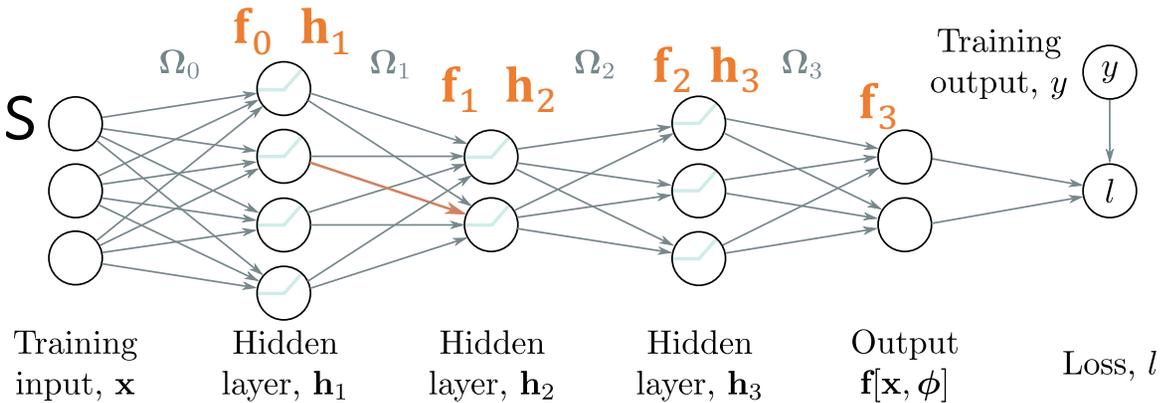
$$\text{ReLU}[z] = \max(0, z)$$

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

Vector Indicator function:

Vector containing ones where $z > 0$ and zeros elsewhere.

The backward pass



1. Write this as a series of intermediate calculations

$$\mathbf{f}_0 = \beta_0 + \mathbf{\Omega}_0 \mathbf{x}_i$$

2. Compute these intermediate quantities

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

$$\mathbf{f}_1 = \beta_1 + \mathbf{\Omega}_1 \mathbf{h}_1$$

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

$$\mathbf{f}_2 = \beta_2 + \mathbf{\Omega}_2 \mathbf{h}_2$$

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

$$\mathbf{f}_3 = \beta_3 + \mathbf{\Omega}_3 \mathbf{h}_3$$

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

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

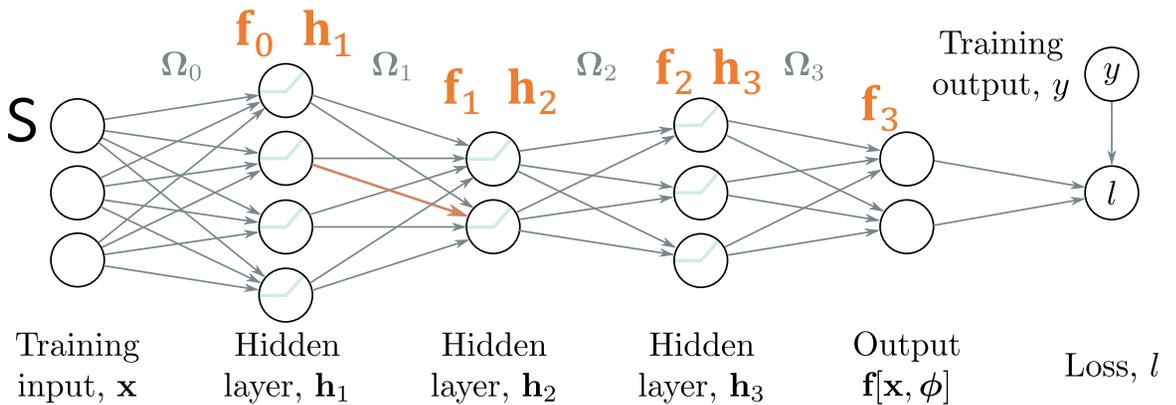
$$\mathbb{I}[\mathbf{f}_2 > 0]$$

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

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

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

The backward pass



1. Write this as a series of intermediate calculations

$$\mathbf{f}_0 = \beta_0 + \mathbf{\Omega}_0 \mathbf{x}_i$$

2. Compute these intermediate quantities

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

$$\mathbf{f}_1 = \beta_1 + \mathbf{\Omega}_1 \mathbf{h}_1$$

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

$$\mathbf{f}_2 = \beta_2 + \mathbf{\Omega}_2 \mathbf{h}_2$$

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

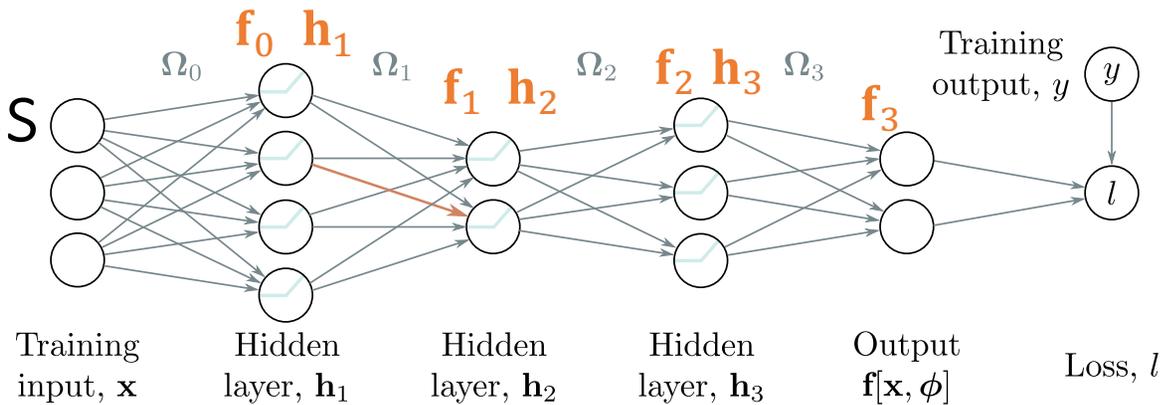
$$\mathbf{f}_3 = \beta_3 + \mathbf{\Omega}_3 \mathbf{h}_3$$

4. Take derivatives w.r.t. parameters

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

$$\begin{aligned} \frac{\partial l_i}{\partial \beta_k} &= \frac{\partial \mathbf{f}_k}{\partial \beta_k} \frac{\partial l_i}{\partial \mathbf{f}_k} \\ &= \frac{\partial}{\partial \beta_k} (\beta_k + \mathbf{\Omega}_k \mathbf{h}_k) \frac{\partial l_i}{\partial \mathbf{f}_k} \\ &= \frac{\partial l_i}{\partial \mathbf{f}_k}, \end{aligned}$$

The backward pass



1. Write this as a series of intermediate calculations

$$\mathbf{f}_0 = \beta_0 + \mathbf{\Omega}_0 \mathbf{x}_i$$

2. Compute these intermediate quantities

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

$$\mathbf{f}_1 = \beta_1 + \mathbf{\Omega}_1 \mathbf{h}_1$$

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

$$\mathbf{f}_2 = \beta_2 + \mathbf{\Omega}_2 \mathbf{h}_2$$

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

$$\mathbf{f}_3 = \beta_3 + \mathbf{\Omega}_3 \mathbf{h}_3$$

4. Take derivatives w.r.t. parameters

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

$$\begin{aligned} \frac{\partial \ell_i}{\partial \mathbf{\Omega}_k} &= \frac{\partial \mathbf{f}_k}{\partial \mathbf{\Omega}_k} \frac{\partial \ell_i}{\partial \mathbf{f}_k} \\ &= \frac{\partial}{\partial \mathbf{\Omega}_k} (\beta_k + \mathbf{\Omega}_k \mathbf{h}_k) \frac{\partial \ell_i}{\partial \mathbf{f}_k} \\ &= \frac{\partial \ell_i}{\partial \mathbf{f}_k} \mathbf{h}_k^T \end{aligned}$$

Do not edit
How to change the design



In the matrix form of backpropagation, what is the primary role of the Jacobian (or derivative matrices) during the backward pass?

① The Slido app must be installed on every computer you're presenting from

slido

Do not edit
How to change the design



What is the primary purpose of the loss function during neural network training?

① The Slido app must be installed on every computer you're presenting from

slido

Do not edit
How to change the design



During the forward pass, why must we store activations at every layer?

① The Slido app must be installed on every computer you're presenting from

slido

Do not edit
How to change the design



Why is the chain rule essential for backpropagation?

① The Slido app must be installed on every computer you're presenting from

slido

Do not edit
How to change the design



**What is the derivative of the ReLU
activation function?**

① The Slido app must be installed on every computer you're presenting from

slido

Gradients

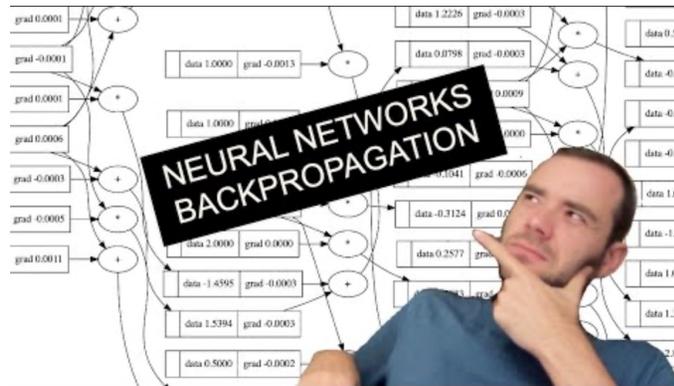
- Backpropagation intuition
- Toy model
- Matrix calculus
- Backpropagation matrix forward pass
- Backpropagation matrix backward pass
- Matrix backprop summary

Pros and cons

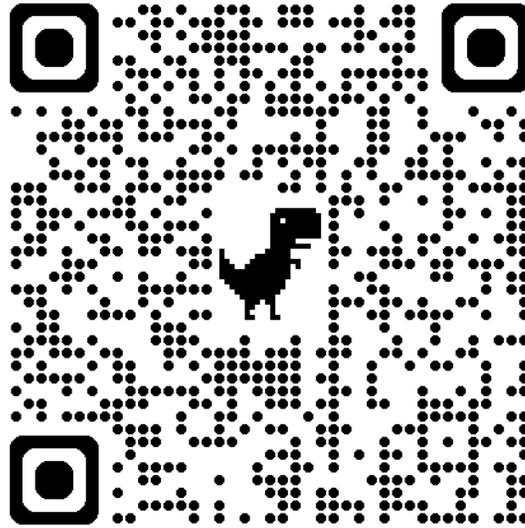
- Extremely efficient
 - Only need matrix multiplication and thresholding for ReLU functions
- Memory hungry – must store all the intermediate quantities
- Sequential
 - can process multiple batches in parallel
 - but things get harder if the whole model doesn't fit on one machine.

More practice

- You'll get to practice the concepts here in the homework notebooks.
- For another tutorial that mimics the PyTorch autograd notation is [Andrej Karpathy's Micrograd Tutorial](#)



Feedback?



<https://forms.gle/pXHM5nx1Ti9aFmpw6>