UNIVERSITY OF BATH

Tópicos en Inteligencia Artificial
Deep Learning

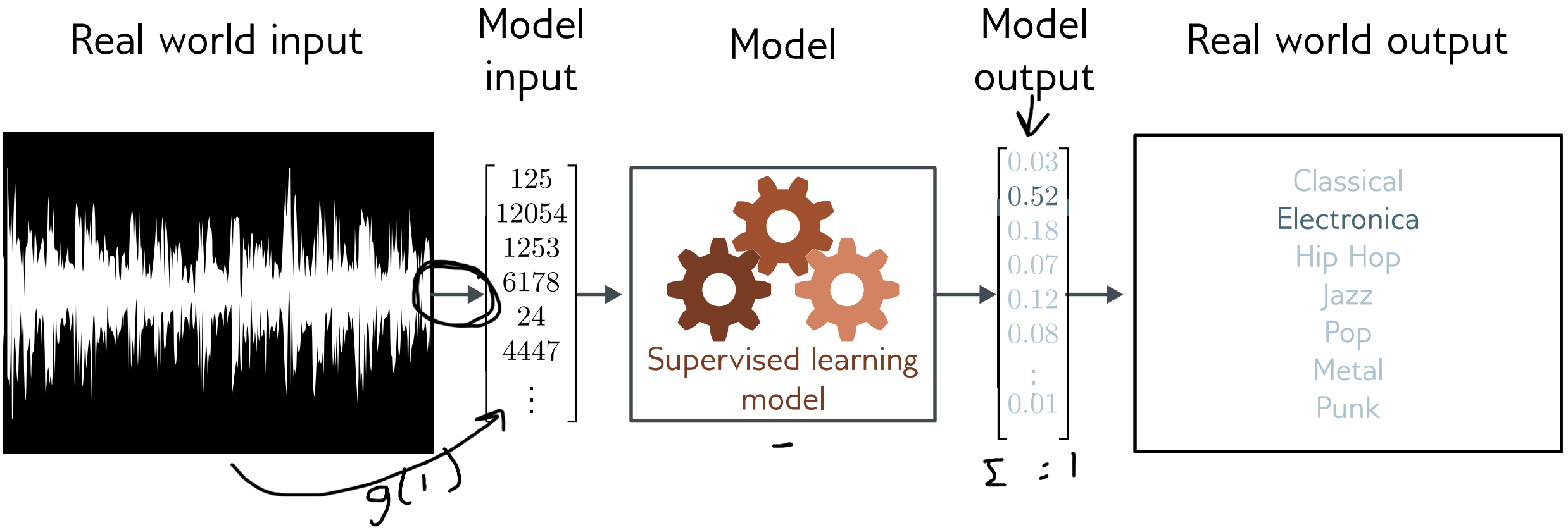
Backpropagation

Basado en el **ppt 07 – Gradients** de

Prof. Simon Prince

Adaptado por Prof. Fernando Crema García

Clasificación de los géneros musicales



- Problema de clasificación multiclase (clases discretas, >2 valores posibles)
- Red convolucional

Función de pérdida

- Conjunto de datos de entrenamiento de I pares de ejemplos de entrada/salida:

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

- La función de pérdida o función de coste mide lo malo que es el modelo:

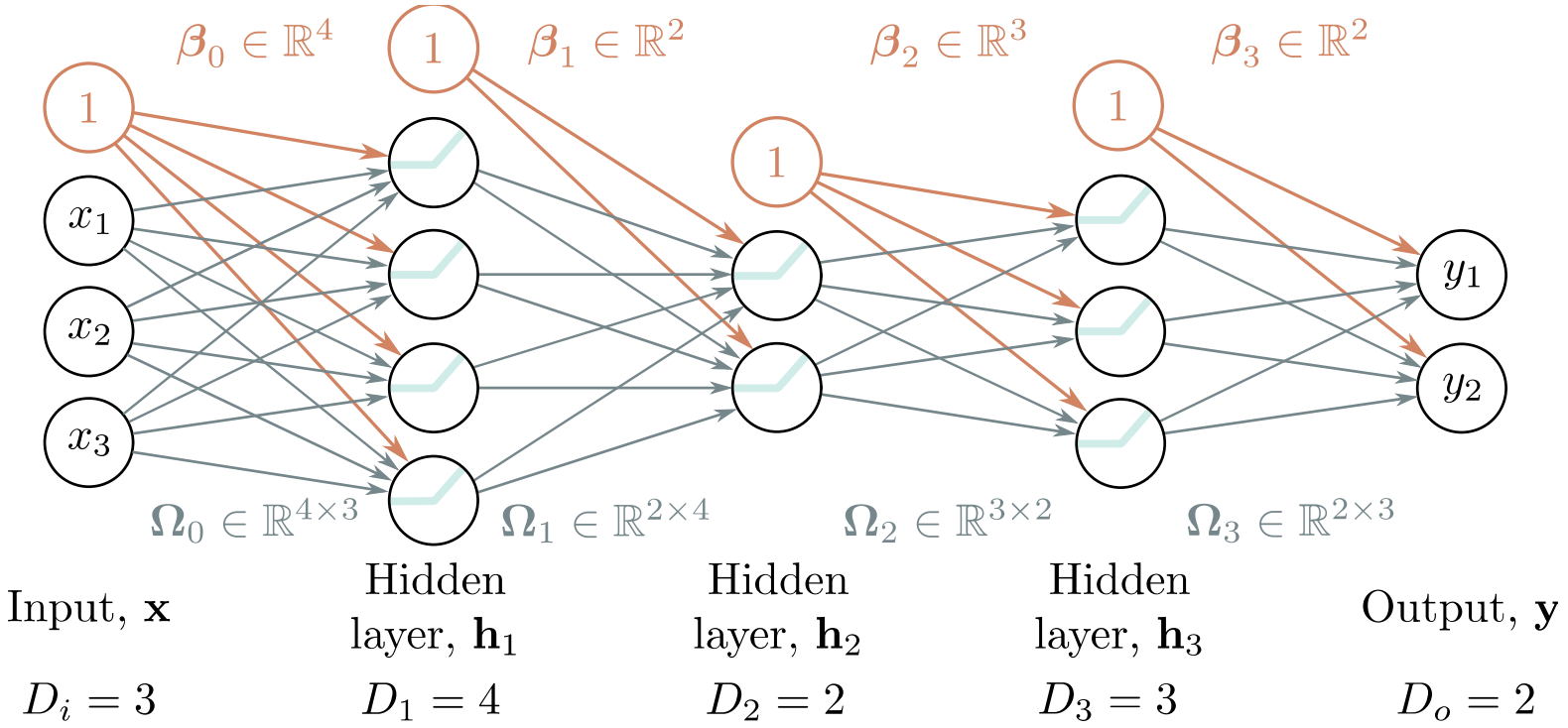
$$L[\boldsymbol{\phi}, \mathrm{f}\left[\mathbf{x}_i, \boldsymbol{\phi}\right], \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{I}]$$

o para abreviar:

$$L\left[\boldsymbol{\phi}\right]$$

Devuelve un escalar que es más pequeño cuando el modelo asigna mejor las entradas a las salidas

Ejemplo



$$\begin{aligned}
\mathbf{h}_1 &= \mathbf{a}[\boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0\mathbf{x}] \\
\mathbf{h}_2 &= \mathbf{a}[\boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1\mathbf{h}_1] \\
\mathbf{h}_3 &= \mathbf{a}[\boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2\mathbf{h}_2] \\
\mathbf{f}[\mathbf{x}, \boldsymbol{\phi}] &= \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3\mathbf{h}_3
\end{aligned}$$

Problema 1: Cómo calcular los gradientes?

Pérdida:
suma de los términos individuales

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

Algoritmo SGD:

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

Parámetros:

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

Necesidad de calcular gradientes

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

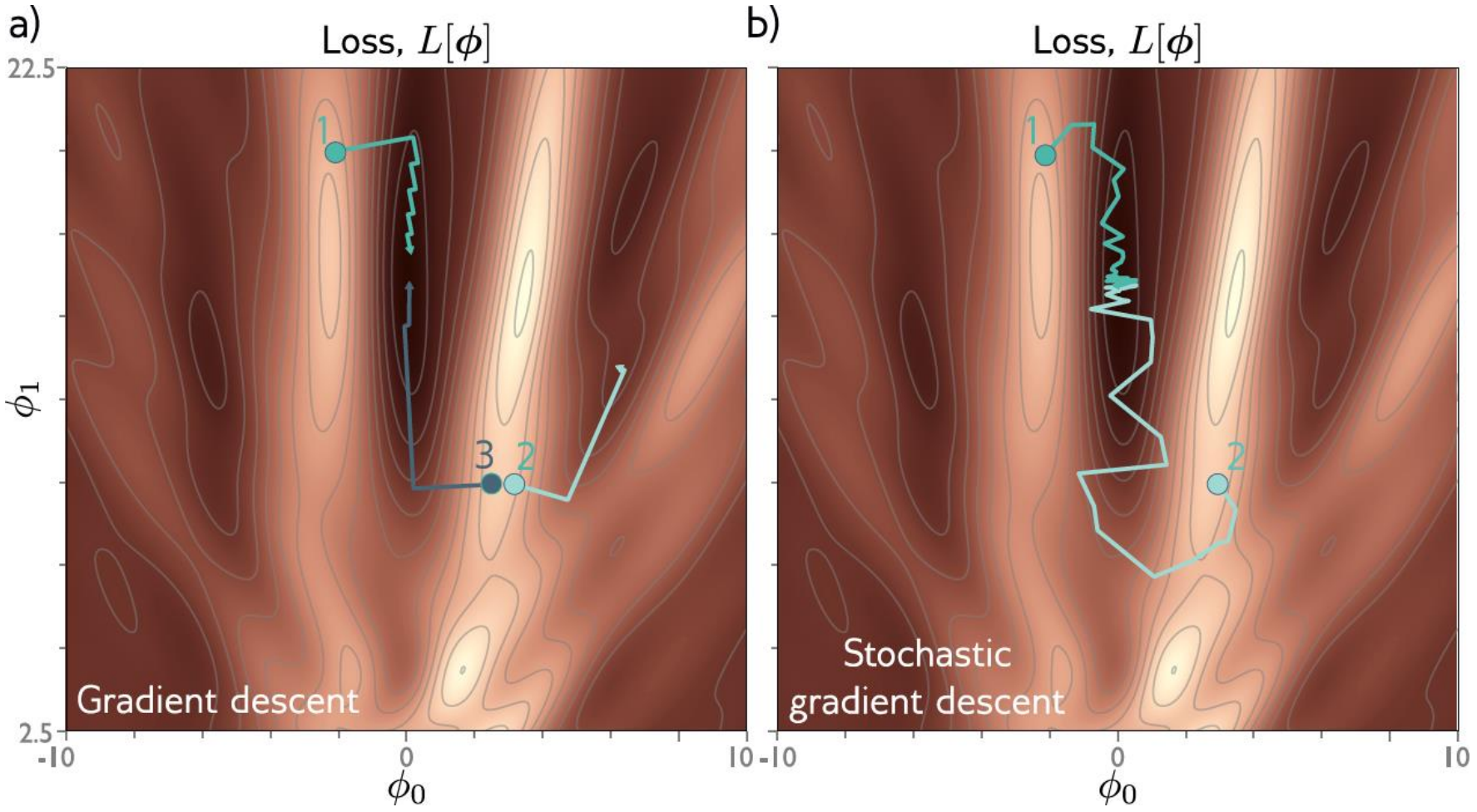
¿Por qué es tan importante?

- Una red neuronal no es más que una ecuación:

$$
\begin{aligned}
y' = \phi_0' + \phi_1' \mathrm{a}\left[\psi_{10} + \psi_{11}\mathrm{a}[\theta_{10} + \theta_{11}x] + \psi_{12}\mathrm{a}[\theta_{20} + \theta_{21}x] + \psi_{13}\mathrm{a}[\theta_{30} + \theta_{31}x]\right] \\
+ \phi_2'\mathrm{a}[\psi_{20} + \psi_{21}\mathrm{a}[\theta_{10} + \theta_{11}x] + \psi_{22}\mathrm{a}[\theta_{20} + \theta_{21}x] + \psi_{23}\mathrm{a}[\theta_{30} + \theta_{31}x]] \\
+ \phi_3'\mathrm{a}[\psi_{30} + \psi_{31}\mathrm{a}[\theta_{10} + \theta_{11}x] + \psi_{32}\mathrm{a}[\theta_{20} + \theta_{21}x] + \psi_{33}\mathrm{a}[\theta_{30} + \theta_{31}x]]
\end{aligned}
$$

- Pero es una ecuación enorme, y necesitamos calcular la derivada
  - para cada parámetro
  - para cada punto del *batch*
  - para cada iteración de *SGD*

Problema 2: inicialización



¿Por dónde deberíamos empezar los parámetros antes de iniciar el SGD?

Gradientes

- La intuición de *backpropagation (propagación hacia atrás de los errores)*
- Modelo de juguete
- Matemáticas necesarias
- Paso adelante (forward pass)de *backpropagation*
- Paso hacia atrás (backward pass)por *backpropagation*
- Diferenciación algorítmica
- Código

Problema 1: Cómo calcular los gradientes?

Pérdida:
suma de los términos individuales

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

Algoritmo SGD:

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

Parámetros:

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

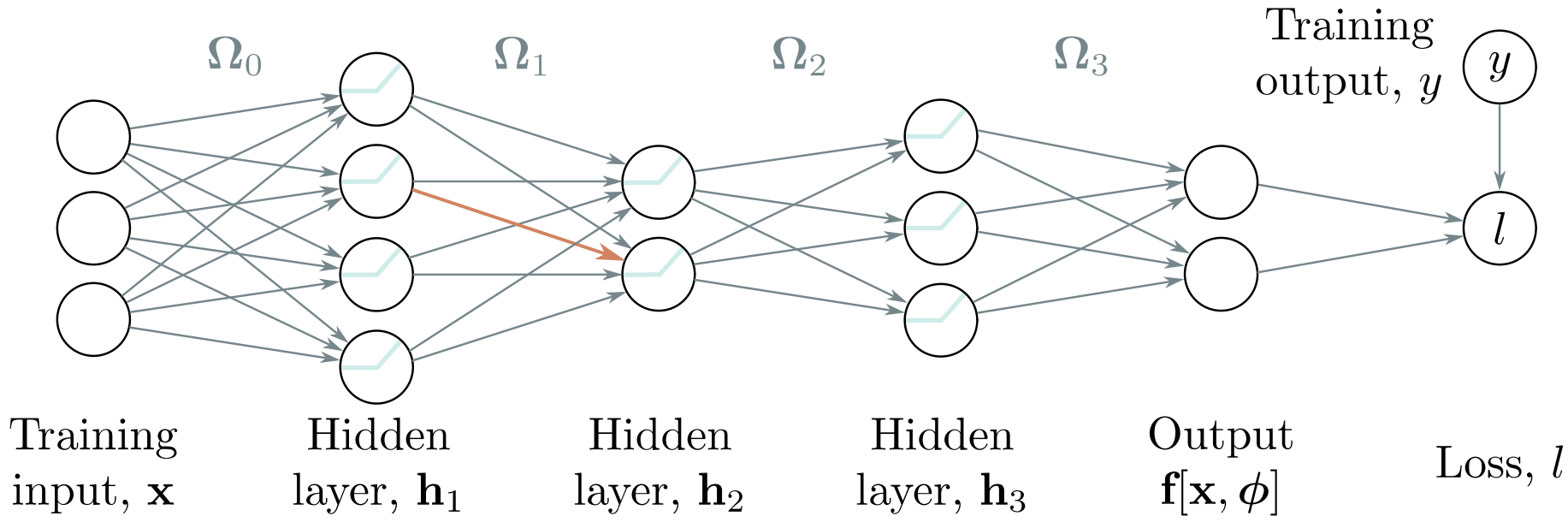
Necesidad de calcular gradientes

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

Algoritmo para calcular el gradiente de forma eficiente

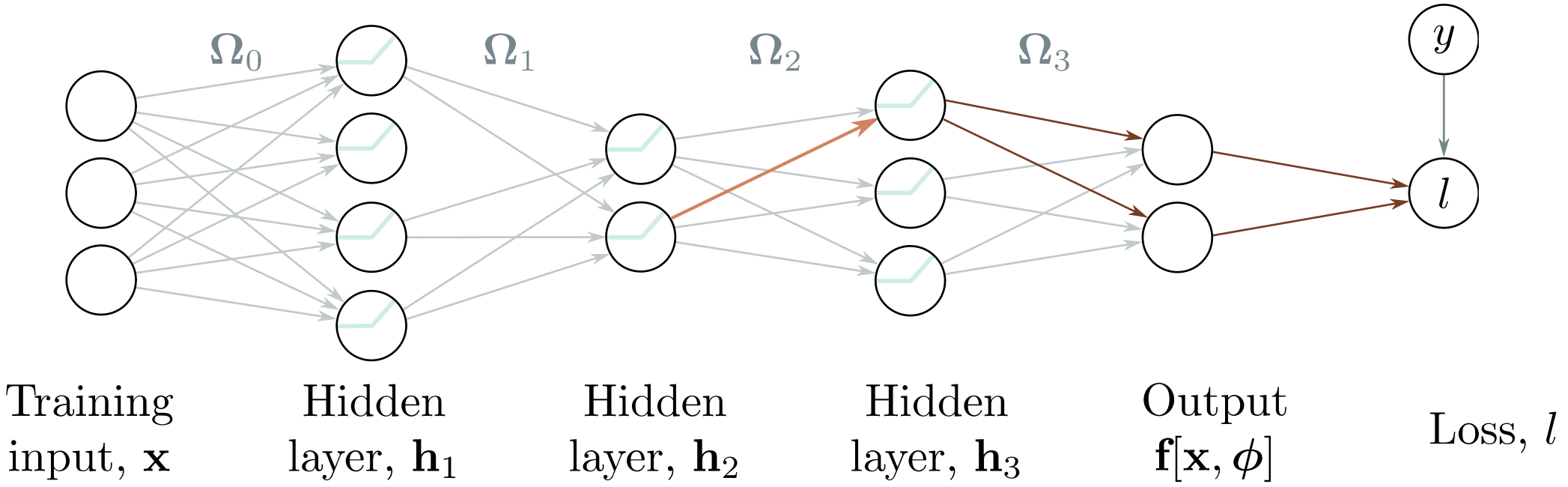
- “Backpropagation algorithm“ Rumelhart, Hinton y Williams (1986)

Intuición BackProp nº 1: el pase hacia delante



- El peso naranja multiplica la activación (salida ReLU) en la capa anterior
- Queremos saber cómo afecta el cambio de peso de la naranja a la pérdida
- Si duplicamos la activación en la capa anterior, el peso tendrá el doble de efecto
- Conclusión: necesitamos conocer las activaciones en cada capa.

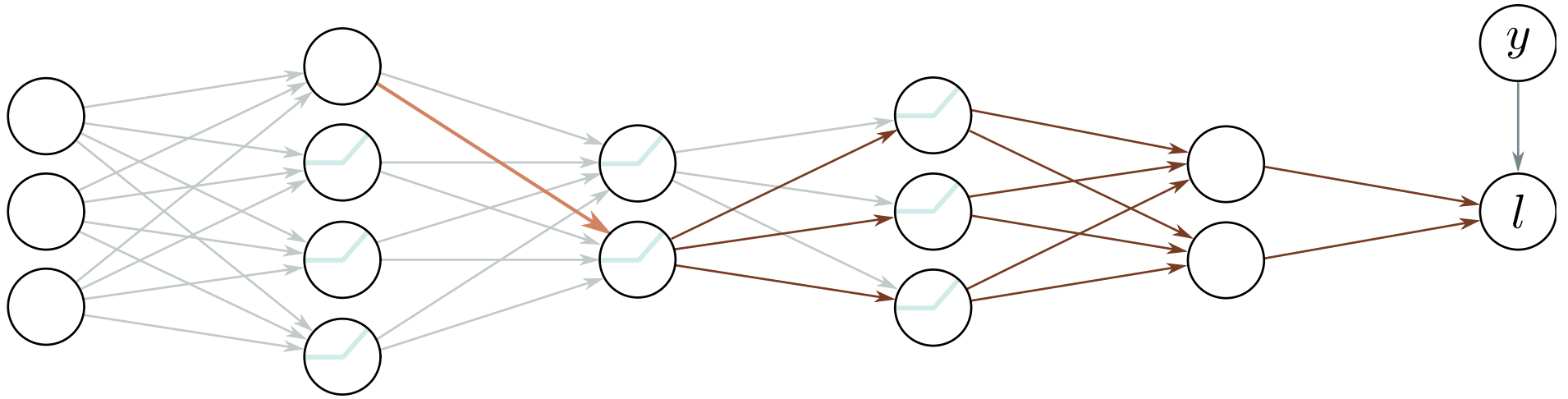
Intuición BackProp nº 2: el pase hacia atrás



Para calcular cómo un pequeño cambio en un peso o sesgo que alimenta la capa oculta $\mathbf{h}_3$ modifica la función de pérdida, necesitamos saber:

- cómo un cambio en la capa $\mathbf{h}_3$ cambia la salida del modelo $\mathbf{f}$
- cómo un cambio en el resultado del modelo modifica la pérdida l

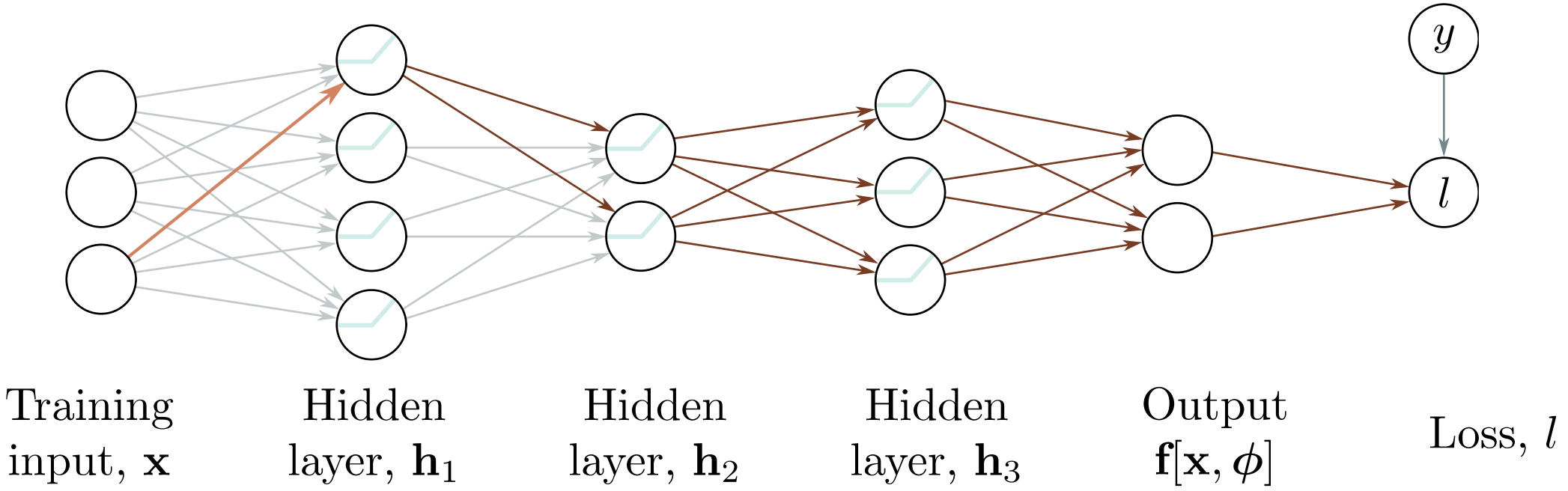
Intuición BackProp nº 2: el pase hacia atrás



Para calcular cómo un pequeño cambio en un peso o sesgo que alimenta la capa oculta $\mathbf{h}_2$ modifica la pérdida, necesitamos saber:

- cómo afecta a $\mathbf{h}_2$ un cambio en la capa $\mathbf{h}_3$
- cómo $\mathbf{h}_3$ cambia el resultado del modelo
- cómo esta salida cambia la pérdida

Intuición BackProp nº 2: el pase hacia atrás



Para calcular cómo un pequeño cambio en un peso o sesgo que alimenta la capa oculta $\mathbf{h}_1$ modifica la pérdida, necesitamos saber:

- cómo un cambio en la capa $\mathbf{h}_1$ afecta la capa $\mathbf{h}_2$
- cómo un cambio en la capa $\mathbf{h}_2$ afecta la capa $\mathbf{h}_3$
- cómo la capa $\mathbf{h}_3$ modifica los resultados del modelo
- cómo el resultado del modelo modifica la pérdida

Gradientes

- La intuición de la backpropagation
- Modelo de juguete
- Matemáticas de base
- Paso adelante (forward pass)de backpropagation
- Paso hacia atrás (backward pass)por backpropagation
- Diferenciación algorítmica
- Código

Composición de funciones

Operación matemática que involucra dos funciones f, g y produce una nueva función que aplica primero g y luego f. Se denota como:

$$(f \circ g)(x) = f(g(x))$$

Para que la composición $f \circ g$ esté bien definida, el codominio de g debe coincidir con el dominio de f

$$g(x) \in B \subseteq \mathrm{dom}(f)$$

Composición de funciones

- Ejemplo $g(x) = 2x + 1$

  $f(y) = y^2$

- Entonces

$$(f \circ g)(x) = f(g(x)) = f(2x + 1) = (2x + 1)^2$$

Derivada de $f \circ g$: regla de la cadena

- Sea la composición de funciones

$$h(x) = (f \circ g)(x) = f(g(x))$$

- Entonces

$$f(g(x))' = f'(g(x)) \cdot g'(x)$$

Derivada de $f \circ g$: Notaciones

- Todas las siguientes expresiones son equivalentes
- Sea $y = f(u)$ y $u = g(x)$ $\quad \dfrac{\mathrm{d}y}{\mathrm{d}x} = \dfrac{\mathrm{d}y}{\mathrm{d}u} \cdot \dfrac{\mathrm{d}u}{\mathrm{d}x}$
- Sea $y = f(u)$ y $u = g(x)$ $\quad \dfrac{\mathrm{d}}{\mathrm{d}x} f(g(x)) = \dfrac{\mathrm{d}f}{\mathrm{d}g} \cdot \dfrac{\mathrm{d}g}{\mathrm{d}x}$
- Sea $y = f(u)$ y $u = g(x)$ $\quad f(g(x))' = f'(g(x)) \cdot g'(x)$

Composición de funciones: Ejemplo

- Ejemplo $g(x) = 2x + 1$ $\quad h(x) = f(g(x)) = (2x+1)^2$

  $f(y) = y^2$

- Entonces $f(g(x))' = f'(g(x)) \cdot g'(x)$

$\frac{\mathrm{d}f}{\mathrm{d}y} = 2y = 2(2x+1)$

$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}g}{\mathrm{d}x} = 2$

$$\frac{\mathrm{d}h}{\mathrm{d}x} = 2(2x+1) \cdot 2 = 4(2x+1)$$

$$\frac{\mathrm{d}}{\mathrm{d}x}(f(g(x))) = \frac{\mathrm{d}}{\mathrm{d}x}\left[(2x+1)^2\right] = 4(2x+1)$$

Función de juguete

$$\mathrm{f}[x, \boldsymbol{\phi}] = \beta_3 + \omega_3 \cdot \cos\Big[\beta_2 + \omega_2 \cdot \exp\big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x]\big]\Big]$$

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

- Consiste en una serie de funciones que se componen entre sí.
- Al contrario que en las redes neuronales sólo utiliza escalares (no vectores)
- "Funciones de activación" sen, exp, cos

Función de juguete

$$\mathrm{f}[x, \boldsymbol{\phi}] = \beta_3 + \omega_3 \cdot \cos\Big[\beta_2 + \omega_2 \cdot \exp\Big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x]\Big]\Big]$$

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

Derivadas

$$\frac{\partial \cos[z]}{\partial z} = -\sin[z] \qquad \frac{\partial \exp[z]}{\partial z} = \exp[z] \qquad \frac{\partial \sin[z]}{\partial z} = \cos[z]$$

Gradientes de la función juguete

$$\mathrm{f}[x, \boldsymbol{\phi}] = \beta_3 + \omega_3 \cdot \cos\Big[\beta_2 + \omega_2 \cdot \exp\big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x]\big]\Big]$$

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

¿Cómo puede un pequeño cambio en β_3 cambiar la pérdida l_i para el i'ésimo ejemplo?

Queremos calcular:

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

Gradientes de funciones compuestas

$$\mathrm{f}[x, \boldsymbol{\phi}] = \beta_3 + \omega_3 \cdot \cos\Big[\beta_2 + \omega_2 \cdot \exp\big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x]\big]\Big]$$

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

Calcular expresiones a mano:

- algunas expresiones muy complicadas.
- redundancia obvia (mira los términos sin en la ecuación inferior)

$$\begin{aligned}\frac{\partial \ell_i}{\partial \omega_0} = \quad & -2\left(\beta_3 + \omega_3 \cdot \cos\Big[\beta_2 + \omega_2 \cdot \exp\big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x_i]\big]\Big] - y_i\right) \\ & \cdot \omega_1\omega_2\omega_3 \cdot x_i \cdot \cos[\beta_0 + \omega_0 \cdot x_i] \cdot \exp\Big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x_i]\Big] \\ & \cdot \sin\Big[\beta_2 + \omega_2 \cdot \exp\Big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x_i]\Big]\Big]\end{aligned}$$

Gradientes de funciones compuestas

$$\mathrm{f}[x, \boldsymbol{\phi}] = \beta_3 + \omega_3 \cdot \cos\Big[\beta_2 + \omega_2 \cdot \exp\Big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x]\Big]\Big]$$

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

Calcular expresiones a mano:

- algunas expresiones muy complicadas.
- redundancia obvia (mira los términos sin en la ecuación inferior)

$$\begin{aligned}\frac{\partial \ell_i}{\partial \omega_0} = \quad & -2\left(\beta_3 + \omega_3 \cdot \cos\Big[\beta_2 + \omega_2 \cdot \exp\Big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x_i]\Big]\Big] - y_i\right) \\ & \cdot \omega_1\omega_2\omega_3 \cdot x_i \cdot \cos[\beta_0 + \omega_0 \cdot x_i] \cdot \exp\Big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x_i]\Big] \\ & \cdot \sin\Big[\beta_2 + \omega_2 \cdot \exp\Big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x_i]\Big]\Big]\end{aligned}$$

Pase hacia delante (*forward pass*)

$$\mathrm{f}[x, \boldsymbol{\phi}] = \beta_3 + \omega_3 \cdot \cos\Big[\beta_2 + \omega_2 \cdot \exp\Big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x]\Big]\Big]$$

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

1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

Pase hacia delante

$$\mathrm{f}[x, \boldsymbol{\phi}] = \beta_3 + \omega_3 \cdot \cos\Big[\beta_2 + \omega_2 \cdot \exp\Big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x]\Big]\Big]$$

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

1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

$$f_0 = \beta_0 + \omega_0 \cdot x_i$$
$$h_1 = \sin[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = \exp[f_1]$$
$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = \cos[f_2]$$
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (f_3 - y_i)^2.$$

Pase hacia delante

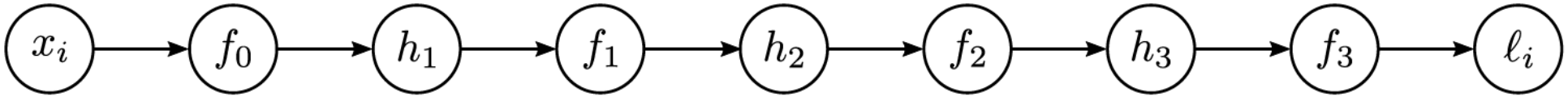
$$\mathrm{f}[x, \boldsymbol{\phi}] = \beta_3 + \omega_3 \cdot \cos\Big[\beta_2 + \omega_2 \cdot \exp\Big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x]\Big]\Big]$$

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

1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

$$\begin{aligned} f_0 &= \beta_0 + \omega_0 \cdot x_i \\ h_1 &= \sin[f_0] \\ f_1 &= \beta_1 + \omega_1 \cdot h_1 \\ h_2 &= \exp[f_1] \end{aligned} \qquad \begin{aligned} f_2 &= \beta_2 + \omega_2 \cdot h_2 \\ h_3 &= \cos[f_2] \\ f_3 &= \beta_3 + \omega_3 \cdot h_3 \\ \ell_i &= (f_3 - y_i)^2. \end{aligned}$$



Paso atrás *(backward pass)*

$$\mathrm{f}[x, \boldsymbol{\phi}] = \beta_3 + \omega_3 \cdot \cos\Big[\beta_2 + \omega_2 \cdot \exp\Big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x]\Big]\Big]$$

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

1. Calcula las derivadas de la pérdida con respecto a estas cantidades intermedias, pero en orden inverso.

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

Paso atrás *(backward pass)*

$$\mathrm{f}[x, \boldsymbol{\phi}] = \beta_3 + \omega_3 \cdot \cos\Big[\beta_2 + \omega_2 \cdot \exp\Big[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x]\Big]\Big]$$

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

1. Calcula las derivadas de la pérdida con respecto a estas cantidades intermedias, pero en orden inverso.

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

Backward pass

1. Calcula las derivadas de la pérdida con respecto a estas cantidades intermedias, pero en orden inverso.

$$
\begin{aligned}
f_0 &= \beta_0 + \omega_0 \cdot x_i \\
h_1 &= \sin[f_0] \\
f_1 &= \beta_1 + \omega_1 \cdot h_1 \\
h_2 &= \exp[f_1]
\end{aligned}
\qquad
\begin{aligned}
f_2 &= \beta_2 + \omega_2 \cdot h_2 \\
h_3 &= \cos[f_2] \\
f_3 &= \beta_3 + \omega_3 \cdot h_3 \\
\ell_i &= (f_3 - y_i)^2.
\end{aligned}
$$

- La primera de estas derivadas es trivial

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

Backward pass

1. Calcula las derivadas de la pérdida con respecto a estas cantidades intermedias, pero en orden inverso.

$$
\begin{aligned}
f_0 &= \beta_0 + \omega_0 \cdot x_i \\
h_1 &= \sin[f_0] \\
f_1 &= \beta_1 + \omega_1 \cdot h_1 \\
h_2 &= \exp[f_1]
\end{aligned}
\qquad
\begin{aligned}
f_2 &= \beta_2 + \omega_2 \cdot h_2 \\
h_3 &= \cos[f_2] \\
f_3 &= \beta_3 + \omega_3 \cdot h_3 \\
\ell_i &= (f_3 - y_i)^2.
\end{aligned}
$$

- La segunda de estas derivadas se calcula mediante la regla de la cadena

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

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

¿Cómo puede un pequeño cambio en h_3 cambiar l_i?

Expresiones equivalentes

Backward pass

1. Calcula las derivadas de la pérdida con respecto a estas cantidades intermedias, pero en orden inverso.

$$
\begin{aligned}
f_0 &= \beta_0 + \omega_0 \cdot x_i \\
h_1 &= \sin[f_0] \\
f_1 &= \beta_1 + \omega_1 \cdot h_1 \\
h_2 &= \exp[f_1]
\end{aligned}
\qquad
\begin{aligned}
f_2 &= \beta_2 + \omega_2 \cdot h_2 \\
h_3 &= \cos[f_2] \\
f_3 &= \beta_3 + \omega_3 \cdot h_3 \\
\ell_i &= (f_3 - y_i)^2.
\end{aligned}
$$

- La segunda derivada se calcula mediante la regla de la cadena

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

¿Cómo puede un pequeño cambio en h_3 cambiar l_i?

¿Cómo puede un pequeño cambio h_3 cambiar f_3?

¿Cómo puede un pequeño cambio f_3 cambiar l_i?

Backward pass

1. Calcula las derivadas de la pérdida con respecto a estas cantidades intermedias, pero en orden inverso.

$$
\begin{aligned}
f_0 &= \beta_0 + \omega_0 \cdot x_i \\
h_1 &= \sin[f_0] \\
f_1 &= \beta_1 + \omega_1 \cdot h_1 \\
h_2 &= \exp[f_1]
\end{aligned}
\qquad
\begin{aligned}
f_2 &= \beta_2 + \omega_2 \cdot h_2 \\
h_3 &= \cos[f_2] \\
f_3 &= \beta_3 + \omega_3 \cdot h_3 \\
\ell_i &= (f_3 - y_i)^2.
\end{aligned}
$$

- La segunda de estas derivadas se calcula mediante la regla de la cadena

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

¿Cómo puede un pequeño cambio en h_3 cambiar l_i?

ω_3

Disponible en el **forward pass!**

Backward pass

1. Calcula las derivadas de la pérdida con respecto a estas cantidades intermedias, pero en orden inverso.

$$
\begin{aligned}
f_0 &= \beta_0 + \omega_0 \cdot x_i \\
h_1 &= \sin[f_0] \\
f_1 &= \beta_1 + \omega_1 \cdot h_1 \\
h_2 &= \exp[f_1]
\end{aligned}
\qquad
\begin{aligned}
f_2 &= \beta_2 + \omega_2 \cdot h_2 \\
h_3 &= \cos[f_2] \\
f_3 &= \beta_3 + \omega_3 \cdot h_3 \\
\ell_i &= (f_3 - y_i)^2.
\end{aligned}
$$

- La segunda de estas derivadas se calcula mediante la regla de la cadena

$$
\frac{\partial \ell_i}{\partial h_3} = \boxed{\frac{\partial f_3}{\partial h_3}}\boxed{\frac{\partial \ell_i}{\partial f_3}} = \omega_3 \cdot 2(f_3 - y_i)
$$

$$
\frac{\partial f_3}{\partial h_3} = \omega_3 \qquad \frac{\partial \ell_i}{\partial f_3} = 2(f_3 - y_i)
$$

Backward pass

1. Calcula las derivadas de la pérdida con respecto a estas cantidades intermedias, pero en orden inverso.

$$f_0 = \beta_0 + \omega_0 \cdot x_i$$
$$h_1 = \sin[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = \exp[f_1]$$
$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = \cos[f_2]$$
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (f_3 - y_i)^2.$$

- Que usando la notación convencional sería:

$$\frac{\partial \ell_i}{\partial h_3} = \boxed{\frac{\partial \ell_i}{\partial f_3}} \boxed{\frac{\partial f_3}{\partial h_3}} = 2(f_3 - y_i) \ \omega_3$$

$$\frac{\partial \ell_i}{\partial f_3} = 2(f_3 - y_i) \qquad \frac{\partial f_3}{\partial h_3} = \omega_3$$

Backward pass

1. Calcula las derivadas de la pérdida con respecto a estas cantidades intermedias, pero en orden inverso.

$$f_0 = \beta_0 + \omega_0 \cdot x_i$$
$$h_1 = \sin[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = \exp[f_1]$$
$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = \cos[f_2]$$
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (f_3 - y_i)^2.$$

- Las derivadas restantes también se calculan utilizando la regla de la cadena

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

Backward pass

1. Calcula las derivadas de la pérdida con respecto a estas cantidades intermedias, pero en orden inverso.

$$
\begin{aligned}
f_0 &= \beta_0 + \omega_0 \cdot x_i \\
h_1 &= \sin[f_0] \\
f_1 &= \beta_1 + \omega_1 \cdot h_1 \\
h_2 &= \exp[f_1]
\end{aligned}
\qquad
\begin{aligned}
f_2 &= \beta_2 + \omega_2 \cdot h_2 \\
h_3 &= \cos[f_2] \\
f_3 &= \beta_3 + \omega_3 \cdot h_3 \\
\ell_i &= (f_3 - y_i)^2.
\end{aligned}
$$

- Las derivadas restantes también se calculan utilizando la regla de la cadena

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

$-\sin(f_2)$

Ya está computado.

Backward pass

1. Calcula las derivadas de la pérdida con respecto a estas cantidades intermedias, pero en orden inverso.

$$
\begin{aligned}
f_0 &= \beta_0 + \omega_0 \cdot x_i \\
h_1 &= \sin[f_0] \\
f_1 &= \beta_1 + \omega_1 \cdot h_1 \\
h_2 &= \exp[f_1]
\end{aligned}
\qquad
\begin{aligned}
f_2 &= \beta_2 + \omega_2 \cdot h_2 \\
h_3 &= \cos[f_2] \\
f_3 &= \beta_3 + \omega_3 \cdot h_3 \\
\ell_i &= (f_3 - y_i)^2.
\end{aligned}
$$

- Las derivadas restantes también se calculan utilizando la regla de la cadena

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

Backward pass

1. Calcula las derivadas de la pérdida con respecto a estas cantidades intermedias, pero en orden inverso.

$$
\begin{aligned}
f_0 &= \beta_0 + \omega_0 \cdot x_i \\
h_1 &= \sin[f_0] \\
f_1 &= \beta_1 + \omega_1 \cdot h_1 \\
h_2 &= \exp[f_1]
\end{aligned}
\qquad
\begin{aligned}
f_2 &= \beta_2 + \omega_2 \cdot h_2 \\
h_3 &= \cos[f_2] \\
f_3 &= \beta_3 + \omega_3 \cdot h_3 \\
\ell_i &= (f_3 - y_i)^2.
\end{aligned}
$$

- Las derivadas restantes también se calculan utilizando la regla de la cadena

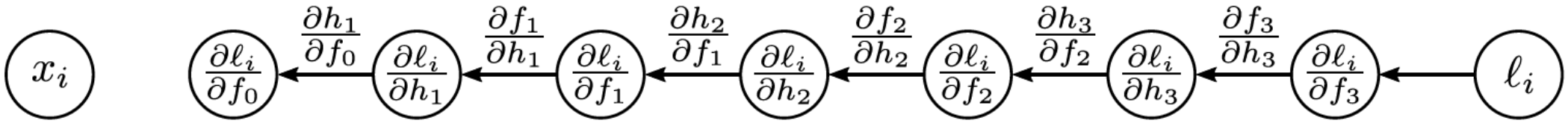
$$
\begin{aligned}
\frac{\partial \ell_i}{\partial f_2} &= \frac{\partial h_3}{\partial f_2}\left(\frac{\partial f_3}{\partial h_3}\frac{\partial \ell_i}{\partial f_3}\right) \\
\frac{\partial \ell_i}{\partial h_2} &= \frac{\partial f_2}{\partial h_2}\left(\frac{\partial h_3}{\partial f_2}\frac{\partial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

Backward pass

1. Calcula las derivadas de la pérdida con respecto a estas cantidades intermedias, pero en orden inverso.

$$f_0 = \beta_0 + \omega_0 \cdot x_i$$
$$h_1 = \sin[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = \exp[f_1]$$
$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = \cos[f_2]$$
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (f_3 - y_i)^2.$$

- Las derivadas restantes también se calculan utilizando la regla de la cadena



Backward pass

2. Hallar cómo cambia la pérdida en función de los parámetros β y ω.

$$
\begin{aligned}
f_0 &= \beta_0 + \omega_0 \cdot x_i \\
h_1 &= \sin[f_0] \\
f_1 &= \beta_1 + \omega_1 \cdot h_1 \\
h_2 &= \exp[f_1]
\end{aligned}
\qquad
\begin{aligned}
f_2 &= \beta_2 + \omega_2 \cdot h_2 \\
h_3 &= \cos[f_2] \\
f_3 &= \beta_3 + \omega_3 \cdot h_3 \\
\ell_i &= (f_3 - y_i)^2.
\end{aligned}
$$

- Otra aplicación de la regla de la cadena

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

¿Cómo puede un pequeño cambio en ω_k cambiar l_i?

¿Cómo puede un pequeño cambio en ω_k cambiar f_k?

¿Cómo puede un pequeño cambio en f_k cambiar l_i?

Backward pass

2. Hallar cómo cambia la pérdida en función de los parámetros β y ω.

$$
\begin{aligned}
f_0 &= \beta_0 + \omega_0 \cdot x_i \\
h_1 &= \sin[f_0] \\
f_1 &= \beta_1 + \omega_1 \cdot h_1 \\
h_2 &= \exp[f_1]
\end{aligned}
\qquad
\begin{aligned}
f_2 &= \beta_2 + \omega_2 \cdot h_2 \\
h_3 &= \cos[f_2] \\
f_3 &= \beta_3 + \omega_3 \cdot h_3 \\
\ell_i &= (f_3 - y_i)^2.
\end{aligned}
$$

- Otra aplicación de la regla de la cadena

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

¿Cómo puede un pequeño cambio en ω_k cambiar l_i?

h_k

Ya calculado en la parte 1.

Backward pass

2. Hallar cómo cambia la pérdida en función de los parámetros β y ω.

$$
\begin{aligned}
f_0 &= \beta_0 + \omega_0 \cdot x_i \\
h_1 &= \sin[f_0] \\
f_1 &= \beta_1 + \omega_1 \cdot h_1 \\
h_2 &= \exp[f_1]
\end{aligned}
\qquad
\begin{aligned}
f_2 &= \beta_2 + \omega_2 \cdot h_2 \\
h_3 &= \cos[f_2] \\
f_3 &= \beta_3 + \omega_3 \cdot h_3 \\
\ell_i &= (f_3 - y_i)^2.
\end{aligned}
$$

- Otra aplicación de la regla de la cadena
- Del mismo modo, para los parámetros β

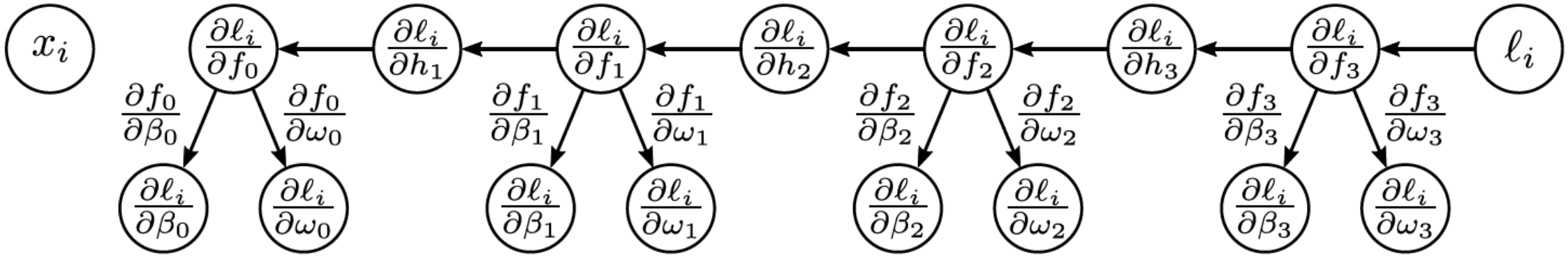
$$\frac{\partial \ell_i}{\partial \omega_k} = \frac{\partial f_k}{\partial \omega_k} \frac{\partial \ell_i}{\partial f_k}$$

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

Paso atrás

2. Hallar cómo cambia la pérdida en función de los parámetros β y ω.

$$f_0 = \beta_0 + \omega_0 \cdot x_i$$
$$h_1 = \sin[f_0]$$
$$f_1 = \beta_1 + \omega_1 \cdot h_1$$
$$h_2 = \exp[f_1]$$
$$f_2 = \beta_2 + \omega_2 \cdot h_2$$
$$h_3 = \cos[f_2]$$
$$f_3 = \beta_3 + \omega_3 \cdot h_3$$
$$\ell_i = (f_3 - y_i)^2.$$



Gradientes

- La intuición de la backpropagation
- Modelo de juguete
- Matemáticas de base
- Paso adelante (forward pass)de backpropagation
- Paso hacia atrás (backward pass)por backpropagation
- Diferenciación algorítmica
- Código

Próxima clase

- Cómo podemos generalizar esta idea para redes neuronales?
- Cómo escribir todo en notación matricial?
- Breve introducción a regla de la cadena para cálculo matricial
- Derivando el algoritmo para la función ReLU? (Tarea 1)

Cálculo matricial

Función escalar f de un vector $\boldsymbol{a}$ $\quad \boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}$

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

Cálculo matricial: Gradiente

Sean f una **función escalar** $f : \mathbb{R}^n \to \mathbb{R}$ y un vector **x** de entrada $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

La derivada de f respecto al vector x es el **gradiente** que es un **vector fila** $\nabla_{\mathbf{x}} f$ (o vector columna, según convención) con las derivadas parciales de f respecto a cada componente de $\boldsymbol{x}$:

$$\nabla_{\mathbf{x}} f = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right] \quad \in \mathbb{R}^{1 \times n}$$

$$\nabla_{\mathbf{x}} f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \quad \in \mathbb{R}^{n \times 1}$$

Cálculo matricial

Función escalar f[] de una matriz **A**

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

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

Cálculo matricial:
Gradiente matricial o derivada matricial

Función escalar f[] de una matriz **A**

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

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

Cálculo matricial

Función vectorial **f[]** del vector **a**

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

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

Cálculo con vectores y matrices

Derivadas escalares:

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

Cálculo con vectores y matrices

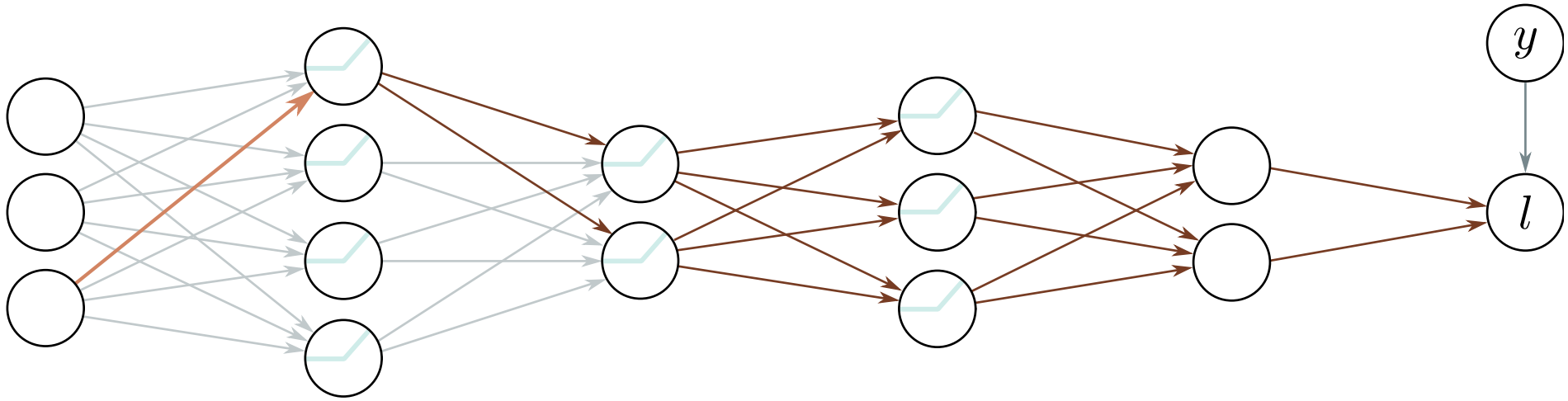
Derivadas escalares:

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

Derivados de la matriz:

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

Cálculo con vectores y matrices: Derivando $\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}$



Training input, $\mathbf{x}$

Hidden layer, $\mathbf{h}_1$

Hidden layer, $\mathbf{h}_2$

Hidden layer, $\mathbf{h}_3$

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

Loss, l

Cálculo con vectores y matrices: Derivando $\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}$

Derivadas de la matriz:

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

Recordemos 00_Álgebra_Lineal 8.1

$$y = Ax = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} * x = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{bmatrix}$$

Cálculo con vectores y matrices: Derivando $\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}$

$\Omega_3 \in \mathbb{R}^{2\times 3}$

Derivadas de la matriz:

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

Recordemos 00_Álgebra_Lineal 8.1

$$f_3 = \beta_3 + \Omega_3 h_3 = \begin{bmatrix}\beta_{13}\\ \beta_{23}\end{bmatrix} + \begin{bmatrix}\Omega_{13}^T\\ \Omega_{23}^T\end{bmatrix} * \begin{bmatrix}h_{13}\\ h_{23}\\ h_{33}\end{bmatrix} = \begin{bmatrix}\beta_{13} + \Omega_{13}^T h_3\\ \beta_{23} + \Omega_{23}^T h_3\end{bmatrix}$$

Cálculo con vectores y matrices: Derivando $\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}$

Derivadas de la matriz:

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

Recordemos 00_Álgebra_Lineal 8.1

$$\mathbf{f_3} = \begin{bmatrix} \beta_{13} + \Omega_{13}^T h_3 \\ \beta_{23} + \Omega_{23}^T h_3 \end{bmatrix} \mathbf{h_3} = \begin{bmatrix} h_{13} \\ h_{23} \\ h_{33} \end{bmatrix} \qquad \frac{\partial \mathbf{f_3}}{\partial \mathbf{h_3}} = \begin{bmatrix} \frac{\partial f_{13}}{\partial h_{13}} & \frac{\partial f_{23}}{\partial h_{13}} \\ \frac{\partial f_{13}}{\partial h_{23}} & \frac{\partial f_{23}}{\partial h_{33}} \\ \frac{\partial f_{13}}{\partial h_{33}} & \frac{\partial f_{23}}{\partial h_{33}} \end{bmatrix} = \begin{bmatrix} \frac{\partial [\beta_{13} + \Omega_{13}^T h_3]}{\partial h_{13}} & \frac{\partial [\beta_{23} + \Omega_{23}^T]}{\partial h_{13}} \\ \frac{\partial [\beta_{13} + \Omega_{13}^T h_3]}{\partial h_{23}} & \frac{\partial [\beta_{23} + \Omega_{23}^T]}{\partial h_{33}} \\ \frac{\partial [\beta_{13} + \Omega_{13}^T h_3]}{\partial h_{33}} & \frac{\partial [\beta_{23} + \Omega_{23}^T]}{\partial h_{33}} \end{bmatrix}$$

Cálculo con vectores y matrices: Derivando $\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}$

Derivadas de la matriz:

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

Recordemos 00_Álgebra_Lineal 8.1

$$\begin{bmatrix} \frac{\partial[\beta_{13}+\Omega_{13}^T h_3]}{\partial h_{13}} & \frac{\partial[\beta_{13}+\Omega_{23}^T]}{\partial h_{13}} \\ \frac{\partial[\beta_{13}+\Omega_{13}^T h_3]}{\partial h_{23}} & \frac{\partial[\beta_{23}+\Omega_{23}^T]}{\partial h_{33}} \\ \frac{\partial[\beta_{13}+\Omega_{13}^T h_3]}{\partial h_{33}} & \frac{\partial[\beta_{23}+\Omega_{23}^T]}{\partial h_{33}} \end{bmatrix} = \begin{bmatrix} \frac{\partial[\beta_{13}+\Omega_{13}^T h_3]}{\partial h_{13}} & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} = \begin{bmatrix} \frac{\partial[\beta_{13}+\omega_{113}h_{13}+\omega_{213}h_{23}+\omega_{313}h_{33}]}{\partial h_{13}} = \omega_{113} & \cdot \\ \cdot & \cdot \end{bmatrix}$$

Cálculo con vectores y matrices: Derivando $\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3}$

Derivadas de la matriz:

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

Recordemos 00_Álgebra_Lineal 8.1

$$\begin{bmatrix} \frac{\partial \Omega_{13}^T h_3}{\partial h_{13}} & \frac{\partial \Omega_{23}^T}{\partial h_{13}} \\ \frac{\partial \Omega_{13}^T h_3}{\partial h_{23}} & \frac{\partial \Omega_{23}^T}{\partial h_{33}} \\ \frac{\partial \Omega_{13}^T h_3}{\partial h_{33}} & \frac{\partial \Omega_{23}^T}{\partial h_{33}} \end{bmatrix} = \begin{bmatrix} \omega_{113} & \omega_{123} \\ \omega_{213} & \omega_{223} \\ \omega_{313} & \omega_{323} \end{bmatrix} = \begin{bmatrix} \Omega_{13} & \Omega_{23} \end{bmatrix} = \Omega_3^T$$

Cálculo con vectores y matrices: Derivando $\frac{\partial \mathbf{f}_3}{\partial \boldsymbol{\beta}_3}$

Derivadas escalares:

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

Derivados de la matriz:

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

Cálculo con vectores y matrices: Derivando $\frac{\partial \mathbf{f}_3}{\partial \boldsymbol{\beta}_3}$

Derivados de la matriz:

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

$$\begin{bmatrix} \frac{\partial[\beta_{13}+\Omega_{13}^T h_3]}{\partial \beta_{13}} & \frac{\partial[\beta_{13}+\Omega_{23}^T]}{\partial \beta_{13}} \\ \frac{\partial[\beta_{13}+\Omega_{13}^T h_3]}{\partial \beta_{23}} & \frac{\partial[\beta_{23}+\Omega_{23}^T]}{\partial \beta_{23}} \end{bmatrix} = \begin{bmatrix} \frac{\partial[\beta_{13}+\Omega_{13}^T h_3]}{\partial \beta_{13}} & . \\ \frac{\partial[\beta_{13}+\Omega_{13}^T h_3]}{\partial \beta_{23}} & . \end{bmatrix} = \begin{bmatrix} \frac{\partial[\beta_{13}+\omega_{113} h_{13}+\omega_{213} h_{23}+\omega_{313} h_{33}]}{\partial \beta_{13}} = 1 & . \\ \frac{\partial[\beta_{13}+\omega_{113} h_{13}+\omega_{213} h_{23}+\omega_{313} h_{33}]}{\partial \beta_{23}} = 0 & . \end{bmatrix}$$

Tareas para casa:

- Considera la función: $\mathbf{f} = \mathbf{B}\mathbf{a}$

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

- Se puede escribir como $f_i = \sum_j B_{ij} a_j = b_i^T a_j$
- Ahora calcula:

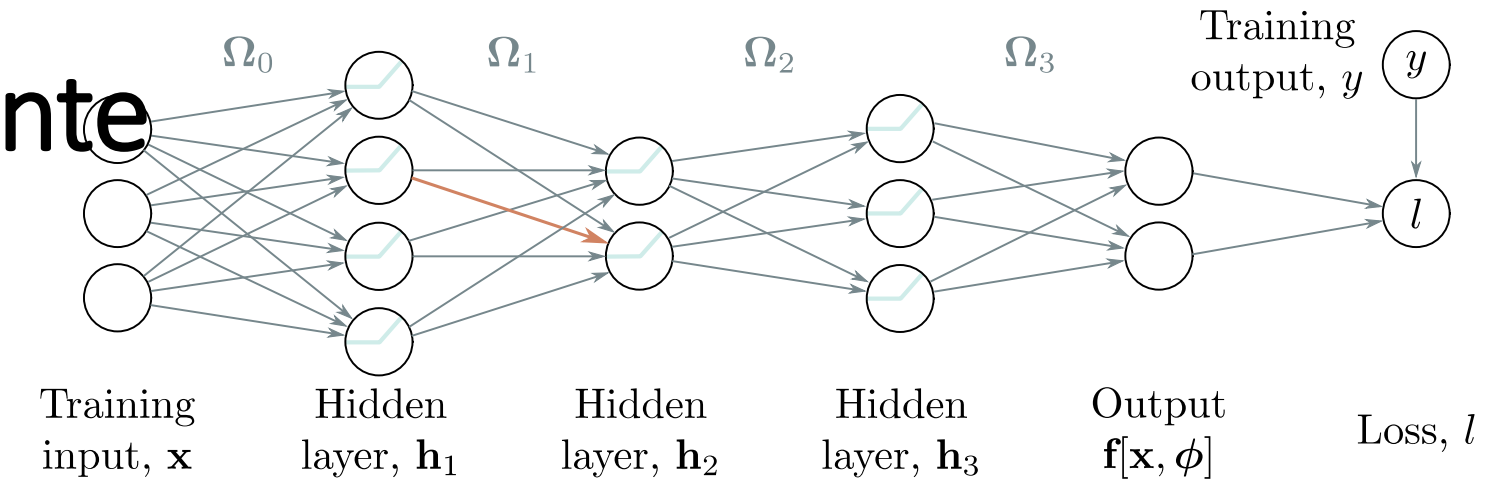
$$\frac{\partial \mathbf{f}}{\partial \mathbf{a}} = \begin{bmatrix} \frac{\partial f_1}{\partial a_1} & \frac{\partial f_2}{\partial a_1} & \frac{\partial f_3}{\partial a_1} \\ \frac{\partial f_1}{\partial a_2} & \frac{\partial f_2}{\partial a_2} & \frac{\partial f_3}{\partial a_2} \\ \frac{\partial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

- Escribe la expresión final como una matriz

Gradientes

- La intuición de la backpropagation
- Modelo de juguete
- Matemáticas de base
- Paso adelante (forward pass)de backpropagation
- Paso hacia atrás (backward pass)por backpropagation
- Diferenciación algorítmica
- Código

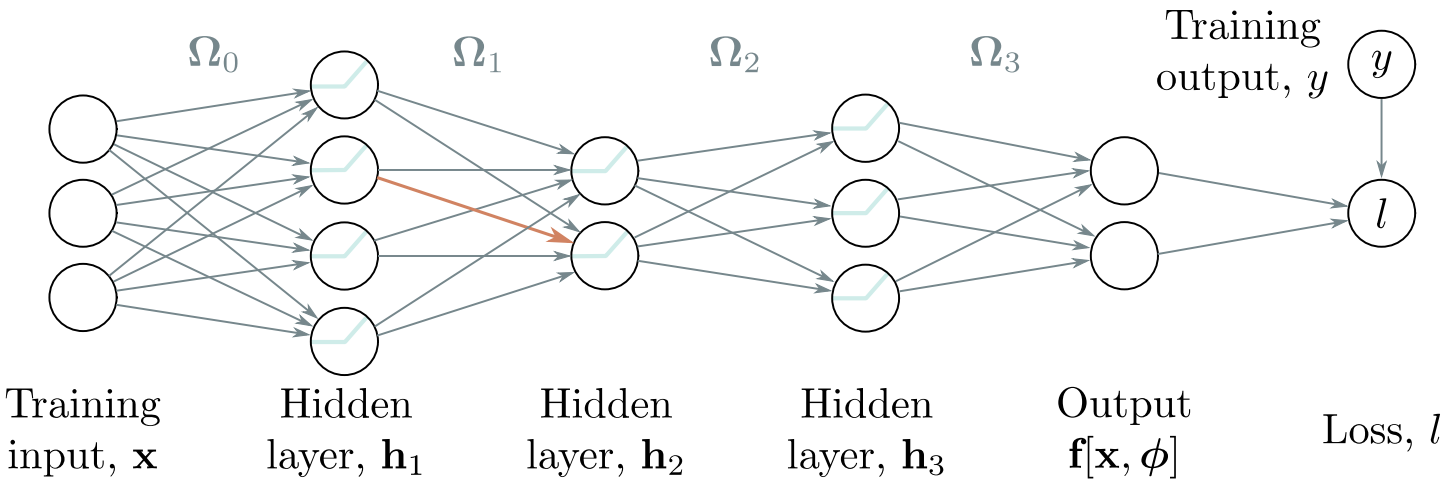
El pase hacia delante



1. Escribe esto como una serie de cálculos intermedios

$$
\begin{aligned}
\mathbf{f}_0 &= \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i \\
\mathbf{h}_1 &= \mathbf{a}[\mathbf{f}_0] \\
\mathbf{f}_1 &= \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1 \\
\mathbf{h}_2 &= \mathbf{a}[\mathbf{f}_1] \\
\mathbf{f}_2 &= \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2 \\
\mathbf{h}_3 &= \mathbf{a}[\mathbf{f}_2] \\
\mathbf{f}_3 &= \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \\
\ell_i &= \mathrm{l}[\mathbf{f}_3, y_i]
\end{aligned}
$$

El pase hacia delante

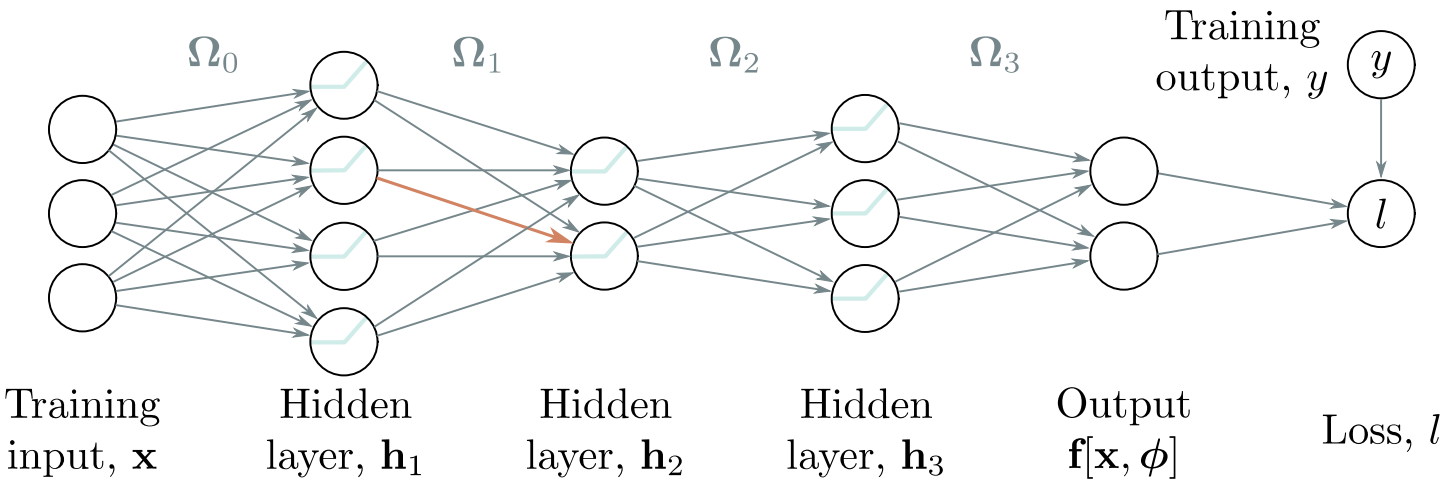


1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

$$\begin{aligned}
\mathbf{f}_0 &= \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0\mathbf{x}_i \\
\mathbf{h}_1 &= \mathbf{a}[\mathbf{f}_0] \\
\mathbf{f}_1 &= \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1\mathbf{h}_1 \\
\mathbf{h}_2 &= \mathbf{a}[\mathbf{f}_1] \\
\mathbf{f}_2 &= \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2\mathbf{h}_2 \\
\mathbf{h}_3 &= \mathbf{a}[\mathbf{f}_2] \\
\mathbf{f}_3 &= \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3\mathbf{h}_3 \\
\ell_i &= \mathrm{l}[\mathbf{f}_3, y_i]
\end{aligned}$$

El pase hacia atrás



1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

3. Tomar derivadas de la producción con respecto a las cantidades intermedias

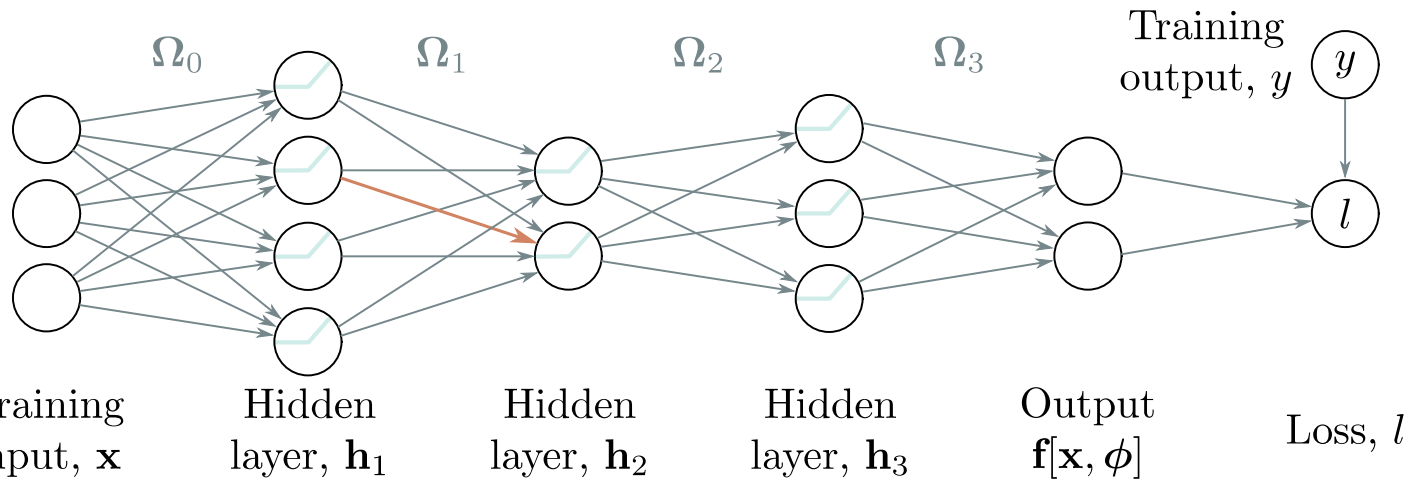
$$
\begin{aligned}
\mathbf{f}_0 &= \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i \\
\mathbf{h}_1 &= \mathbf{a}[\mathbf{f}_0] \\
\mathbf{f}_1 &= \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1 \\
\mathbf{h}_2 &= \mathbf{a}[\mathbf{f}_1] \\
\mathbf{f}_2 &= \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2 \\
\mathbf{h}_3 &= \mathbf{a}[\mathbf{f}_2] \\
\mathbf{f}_3 &= \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \\
\ell_i &= \mathrm{l}[\mathbf{f}_3, y_i]
\end{aligned}
$$

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

Gradientes

- La intuición de la backpropagation
- Modelo de juguete
- Matemáticas de fondo
- Paso adelante (forward pass)de backpropagation
- Paso hacia atrás (backward pass)por backpropagation
- Diferenciación algorítmica
- Código

El pase hacia atrás



1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

3. Tomar derivadas de la producción con respecto a las cantidades intermedias

$$
\begin{aligned}
\mathbf{f}_0 &= \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i \\
\mathbf{h}_1 &= \mathbf{a}[\mathbf{f}_0] \\
\mathbf{f}_1 &= \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1 \\
\mathbf{h}_2 &= \mathbf{a}[\mathbf{f}_1] \\
\mathbf{f}_2 &= \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2 \\
\mathbf{h}_3 &= \mathbf{a}[\mathbf{f}_2] \\
\mathbf{f}_3 &= \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \\
\ell_i &= \mathrm{l}[\mathbf{f}_3, y_i]
\end{aligned}
$$

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

No es tan difícil!

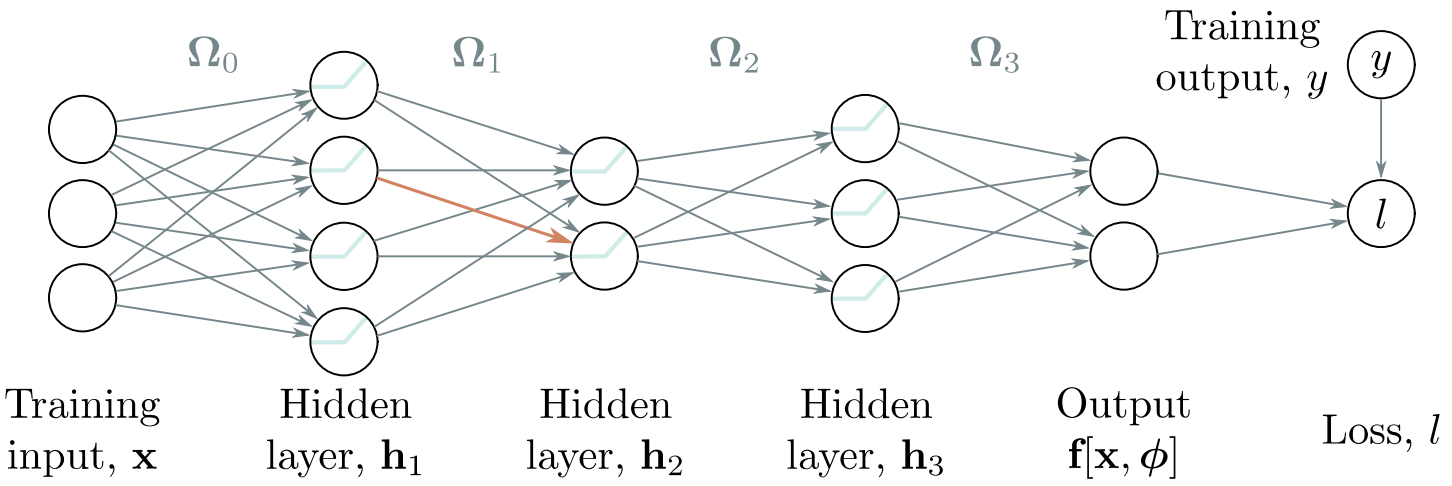
- Pero..:

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

- Muy similar a:

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

El pase hacia atrás



1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

3. Tomar derivadas de la producción con respecto a las cantidades intermedias

$$
\begin{aligned}
\mathbf{f}_0 &= \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i \\
\mathbf{h}_1 &= \mathbf{a}[\mathbf{f}_0] \\
\mathbf{f}_1 &= \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1 \\
\mathbf{h}_2 &= \mathbf{a}[\mathbf{f}_1] \\
\mathbf{f}_2 &= \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2 \\
\mathbf{h}_3 &= \mathbf{a}[\mathbf{f}_2] \\
\mathbf{f}_3 &= \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \\
\ell_i &= \mathrm{l}[\mathbf{f}_3, y_i]
\end{aligned}
$$

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

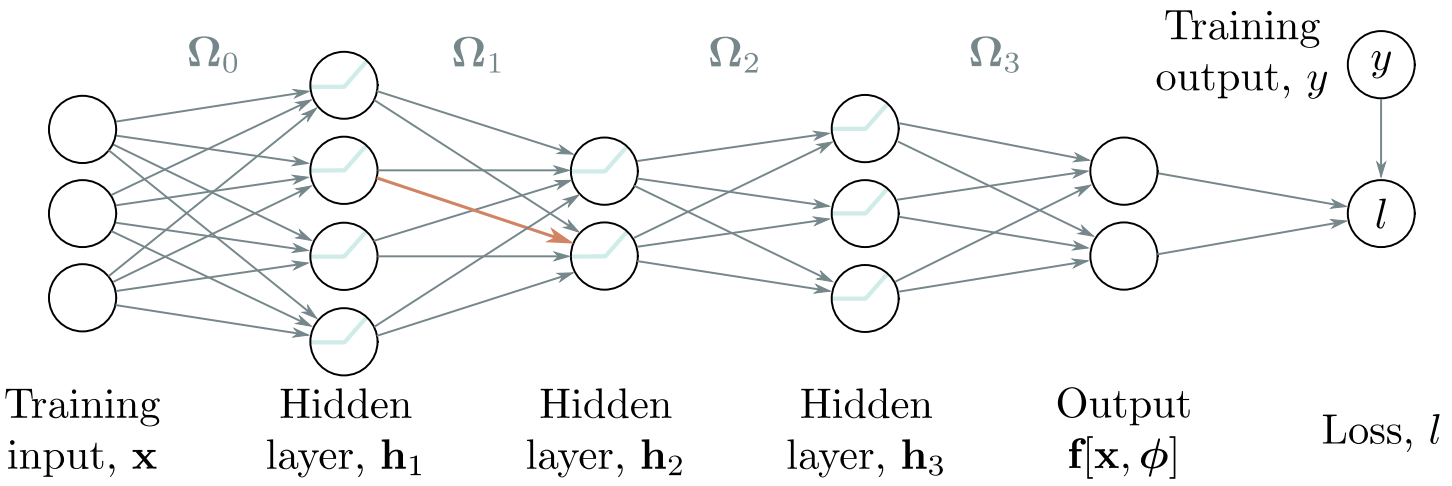
$$
\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} = \frac{\partial}{\partial \mathbf{h}_3}\left(\boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3\right) = \boldsymbol{\Omega}_3^T
$$

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

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

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

El pase hacia atrás



1. Escribe esto como una serie de cálculos intermedios

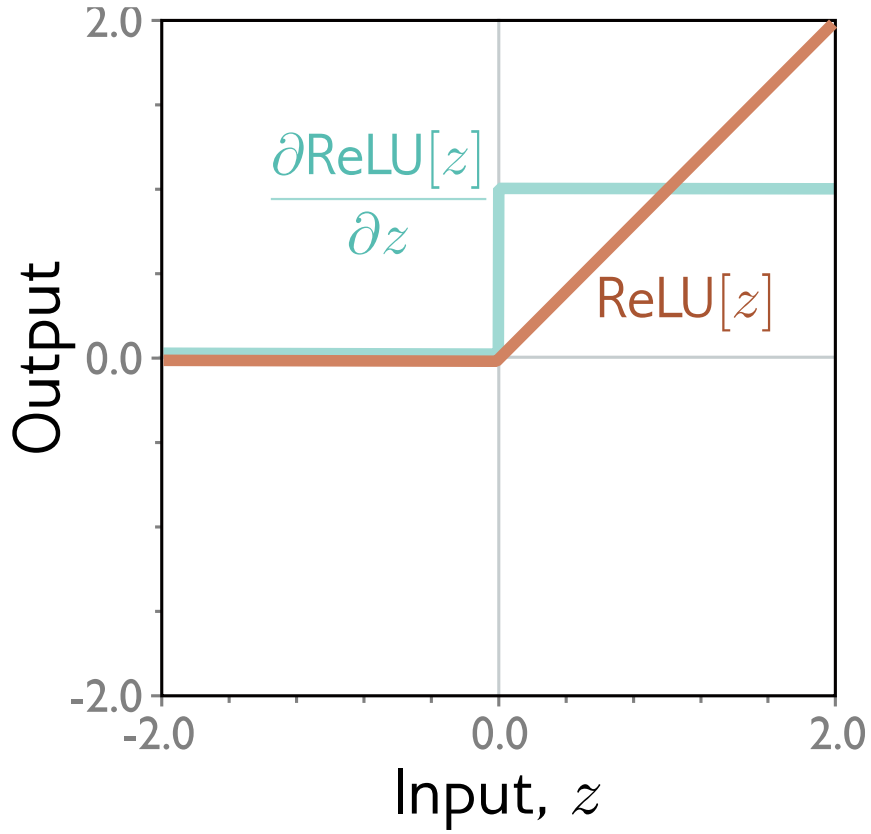
2. Calcula estas cantidades intermedias

3. Tomar derivadas de la producción con respecto a las cantidades intermedias

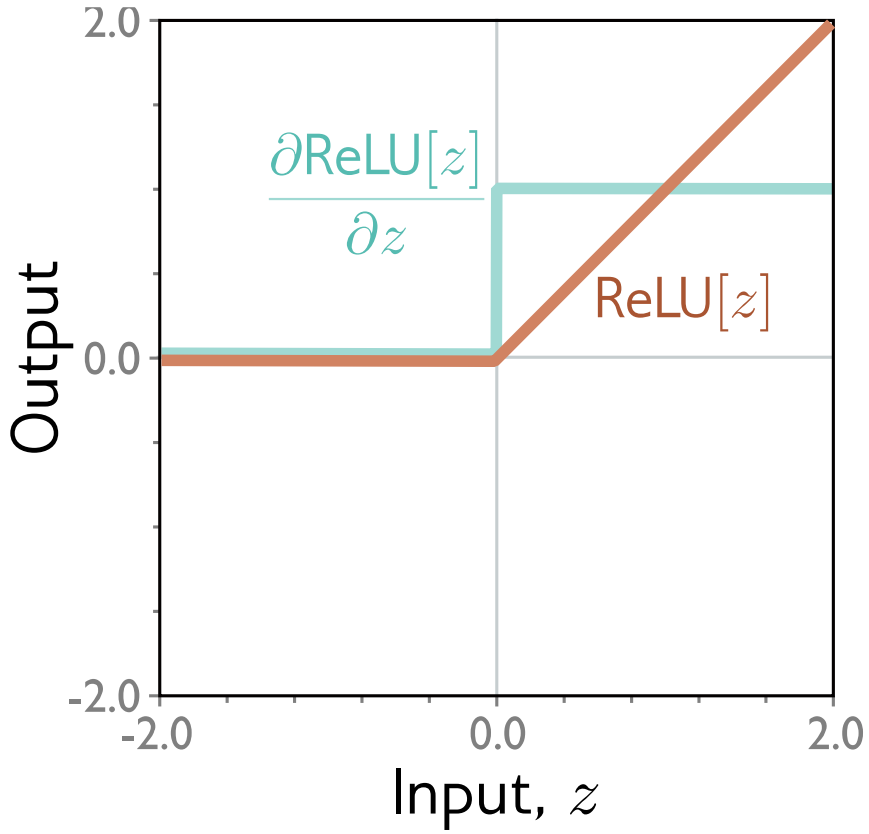
$$
\begin{aligned}
\mathbf{f}_0 &= \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i \\
\mathbf{h}_1 &= \mathbf{a}[\mathbf{f}_0] \\
\mathbf{f}_1 &= \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1 \\
\mathbf{h}_2 &= \mathbf{a}[\mathbf{f}_1] \\
\mathbf{f}_2 &= \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2 \\
\mathbf{h}_3 &= \mathbf{a}[\mathbf{f}_2] \\
\mathbf{f}_3 &= \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \\
\ell_i &= \mathrm{l}[\mathbf{f}_3, y_i]
\end{aligned}
$$

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

Derivada de ReLU



Derivada de ReLU



$$\mathbb{I}[z > 0]$$

"Función de indicador"

Derivada de RELU

1. Considera:

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

donde:

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

2. Podríamos escribir equivalentemente:

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

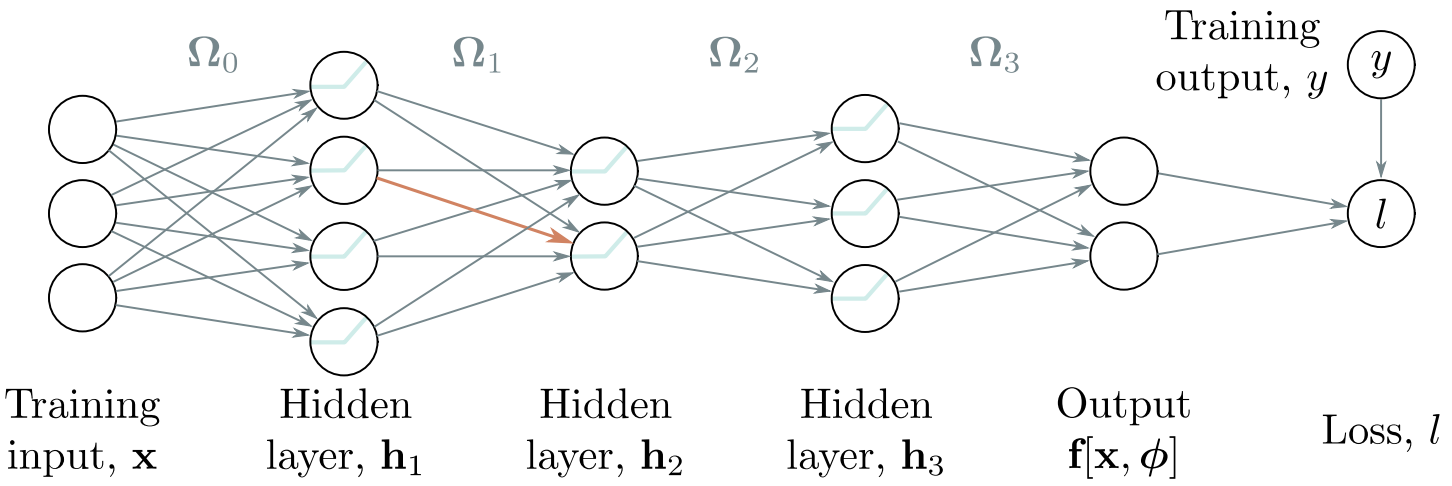
3. Tomando la derivada

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

4. De forma equivalente, podemos multiplicar puntualmente por la diagonal

$$\mathbb{I}[\mathbf{b} > 0] \odot$$

El pase hacia atrás



1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

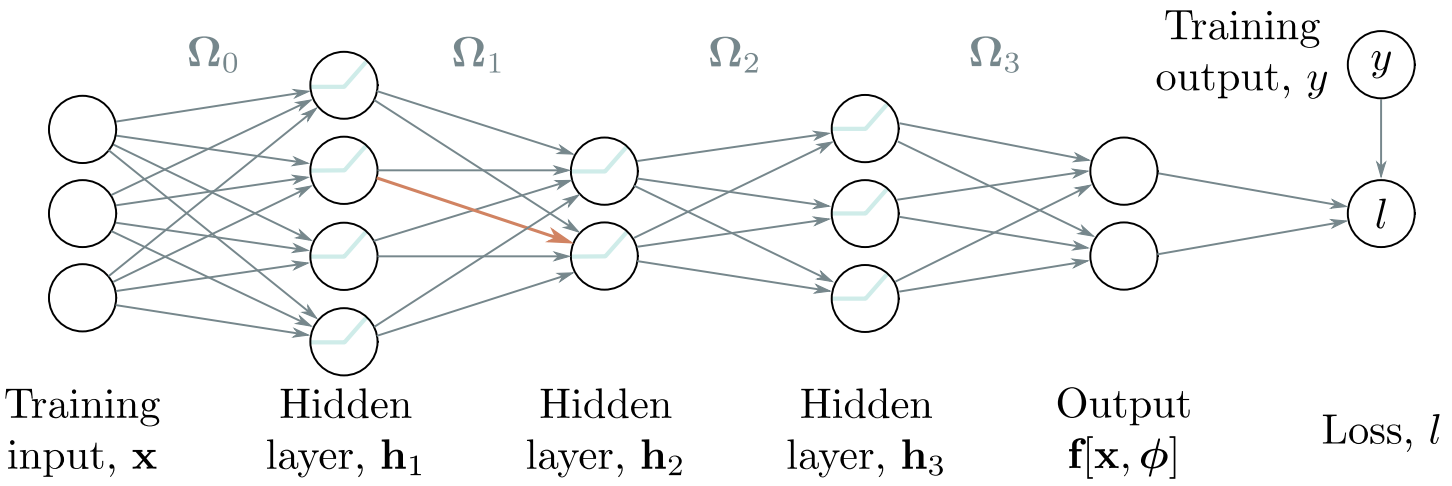
3. Tomar derivadas de la producción con respecto a las cantidades intermedias

$$
\begin{aligned}
\mathbf{f}_0 &= \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i \\
\mathbf{h}_1 &= \mathbf{a}[\mathbf{f}_0] \\
\mathbf{f}_1 &= \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1 \\
\mathbf{h}_2 &= \mathbf{a}[\mathbf{f}_1] \\
\mathbf{f}_2 &= \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2 \\
\mathbf{h}_3 &= \mathbf{a}[\mathbf{f}_2] \\
\mathbf{f}_3 &= \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \\
\ell_i &= \mathrm{l}[\mathbf{f}_3, y_i]
\end{aligned}
$$

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

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

El pase hacia atrás



1. Escribe esto como una serie de cálculos intermedios

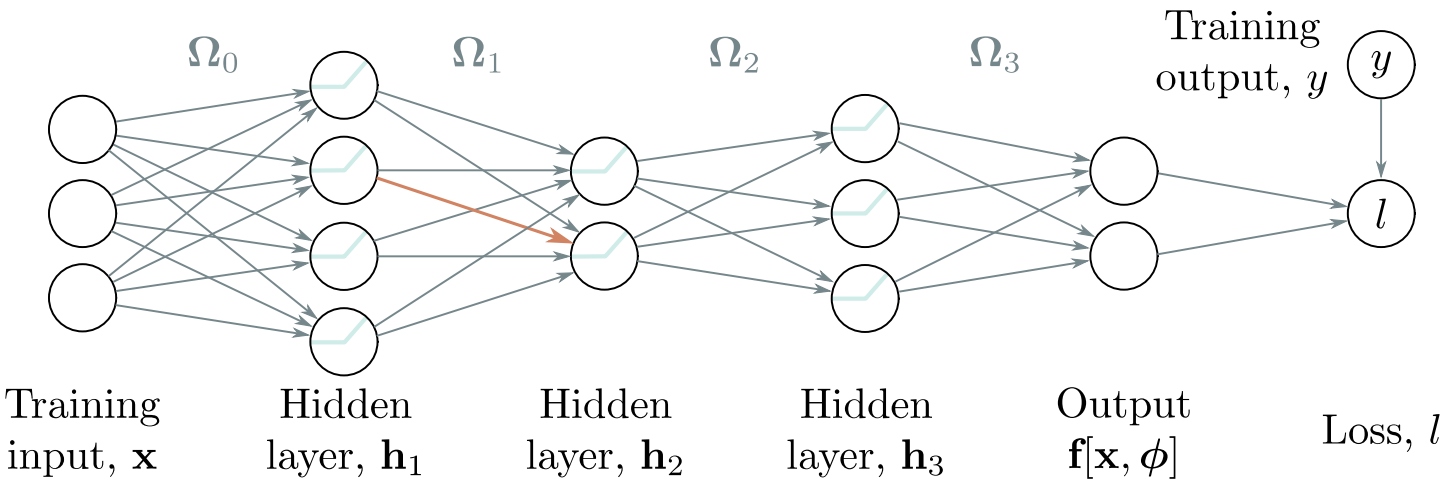
2. Calcula estas cantidades intermedias

3. Tomar derivadas de la producción con respecto a las cantidades intermedias

$$
\begin{aligned}
\mathbf{f}_0 &= \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i \\
\mathbf{h}_1 &= \mathbf{a}[\mathbf{f}_0] \\
\mathbf{f}_1 &= \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1 \\
\mathbf{h}_2 &= \mathbf{a}[\mathbf{f}_1] \\
\mathbf{f}_2 &= \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2 \\
\mathbf{h}_3 &= \mathbf{a}[\mathbf{f}_2] \\
\mathbf{f}_3 &= \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \\
\ell_i &= \mathrm{l}[\mathbf{f}_3, y_i]
\end{aligned}
$$

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

El pase hacia atrás



1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

3. Tomar derivadas de la producción con respecto a las cantidades intermedias

$$
\begin{aligned}
\mathbf{f}_0 &= \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i \\
\mathbf{h}_1 &= \mathbf{a}[\mathbf{f}_0] \\
\mathbf{f}_1 &= \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1 \\
\mathbf{h}_2 &= \mathbf{a}[\mathbf{f}_1] \\
\mathbf{f}_2 &= \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2 \\
\mathbf{h}_3 &= \mathbf{a}[\mathbf{f}_2] \\
\mathbf{f}_3 &= \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \\
\ell_i &= \mathrm{l}[\mathbf{f}_3, y_i]
\end{aligned}
$$

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

Resumen del backprop

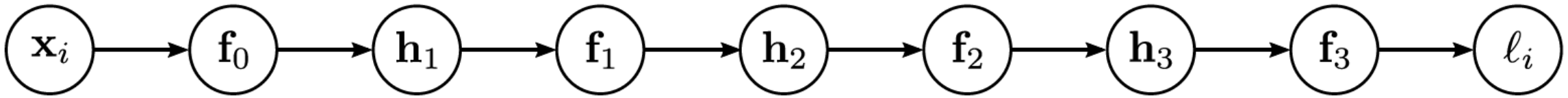
Forward pass: We compute and store the following quantities:

$$
\begin{aligned}
\mathbf{f}_0 &= \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i \\
\mathbf{h}_k &= \mathbf{a}[\mathbf{f}_{k-1}] && k \in \{1, 2, \ldots K\} \\
\mathbf{f}_k &= \boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \mathbf{h}_k. && k \in \{1, 2, \ldots K\}
\end{aligned}
$$

Resumen del backprop

Forward pass: We compute and store the following quantities:

$$
\begin{aligned}
\mathbf{f}_0 &= \boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0 \mathbf{x}_i \\
\mathbf{h}_k &= \mathbf{a}[\mathbf{f}_{k-1}] && k \in \{1, 2, \ldots K\} \\
\mathbf{f}_k &= \boldsymbol{\beta}_k + \boldsymbol{\Omega}_k \mathbf{h}_k. && k \in \{1, 2, \ldots K\}
\end{aligned}
$$



Resumen del backprop

Backward pass: We start with the derivative $\partial \ell_i / \partial \mathbf{f}_K$ of the loss function ℓ_i with respect to the network output $\mathbf{f}_K$ and work backward through the network:

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_k} &= \frac{\partial \ell_i}{\partial \mathbf{f}_k} && k \in \{K, K-1, \ldots 1\} \\
\frac{\partial \ell_i}{\partial \boldsymbol{\Omega}_k} &= \frac{\partial \ell_i}{\partial \mathbf{f}_k}\mathbf{h}_k^T && k \in \{K, K-1, \ldots 1\} \\
\frac{\partial \ell_i}{\partial \mathbf{f}_{k-1}} &= \mathbb{I}[\mathbf{f}_{k-1} > 0] \odot \left(\boldsymbol{\Omega}_k^T \frac{\partial \ell_i}{\partial \mathbf{f}_k}\right), && k \in \{K, K-1, \ldots 1\} \qquad (7.13)
\end{aligned}$$

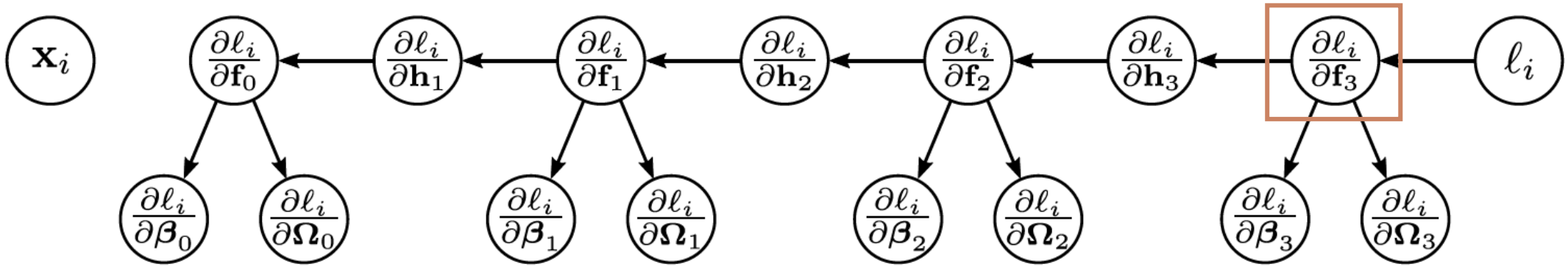
where $\odot$ denotes pointwise multiplication and $\mathbb{I}[\mathbf{f}_{k-1} > 0]$ is a vector containing ones where $\mathbf{f}_{k-1}$ is greater than zero and zeros elsewhere.

Resumen del backprop

Backward pass: We start with the derivative $\partial \ell_i / \partial \mathbf{f}_K$ of the loss function ℓ_i with respect to the network output $\mathbf{f}_K$ and work backward through the network:

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_k} &= \frac{\partial \ell_i}{\partial \mathbf{f}_k} && k \in \{K, K-1, \ldots 1\} \\
\frac{\partial \ell_i}{\partial \boldsymbol{\Omega}_k} &= \frac{\partial \ell_i}{\partial \mathbf{f}_k}\mathbf{h}_k^T && k \in \{K, K-1, \ldots 1\} \\
\frac{\partial \ell_i}{\partial \mathbf{f}_{k-1}} &= \mathbb{I}[\mathbf{f}_{k-1} > 0] \odot \left(\boldsymbol{\Omega}_k^T \frac{\partial \ell_i}{\partial \mathbf{f}_k}\right), && k \in \{K, K-1, \ldots 1\} \qquad (7.13)
\end{aligned}$$

where $\odot$ denotes pointwise multiplication and $\mathbb{I}[\mathbf{f}_{k-1} > 0]$ is a vector containing ones where $\mathbf{f}_{k-1}$ is greater than zero and zeros elsewhere.

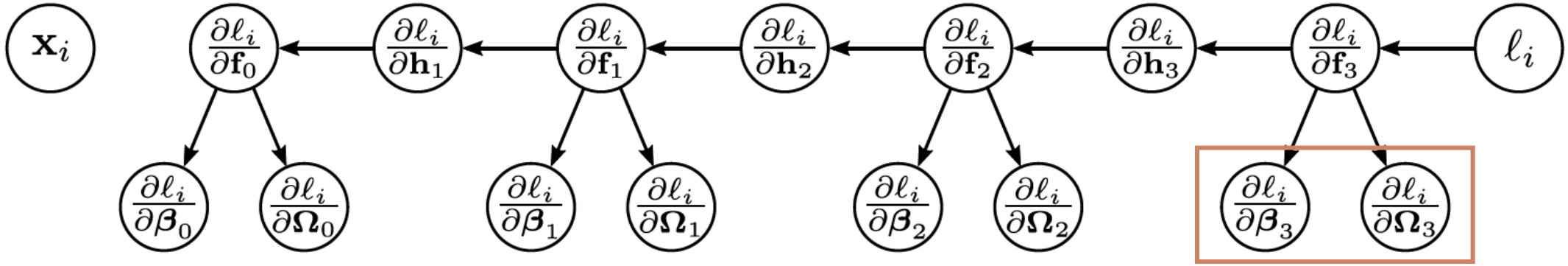


Resumen del backprop

Backward pass: We start with the derivative $\partial \ell_i / \partial \mathbf{f}_K$ of the loss function ℓ_i with respect to the network output $\mathbf{f}_K$ and work backward through the network:

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_k} &= \frac{\partial \ell_i}{\partial \mathbf{f}_k} && k \in \{K, K-1, \ldots 1\} \\
\frac{\partial \ell_i}{\partial \boldsymbol{\Omega}_k} &= \frac{\partial \ell_i}{\partial \mathbf{f}_k}\mathbf{h}_k^T && k \in \{K, K-1, \ldots 1\} \\
\frac{\partial \ell_i}{\partial \mathbf{f}_{k-1}} &= \mathbb{I}[\mathbf{f}_{k-1} > 0] \odot \left(\boldsymbol{\Omega}_k^T \frac{\partial \ell_i}{\partial \mathbf{f}_k}\right), && k \in \{K, K-1, \ldots 1\} \qquad (7.13)
\end{aligned}$$

where $\odot$ denotes pointwise multiplication and $\mathbb{I}[\mathbf{f}_{k-1} > 0]$ is a vector containing ones where $\mathbf{f}_{k-1}$ is greater than zero and zeros elsewhere.

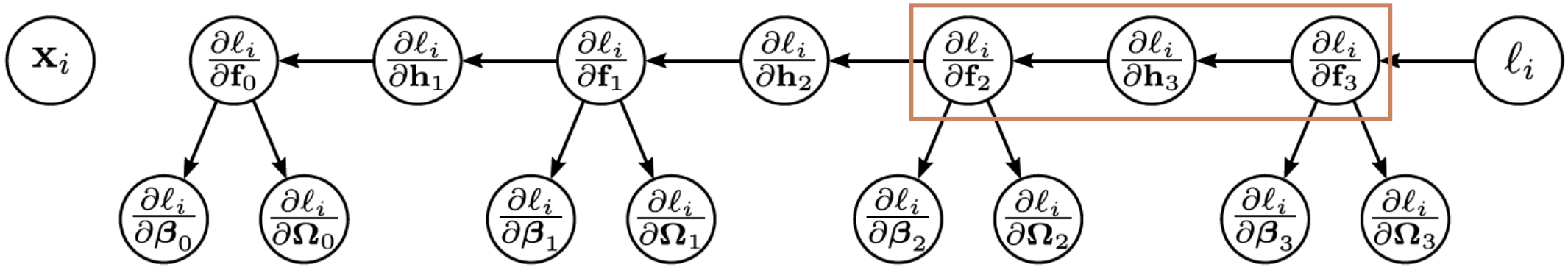


Resumen del backprop

Backward pass: We start with the derivative $\partial \ell_i / \partial \mathbf{f}_K$ of the loss function ℓ_i with respect to the network output $\mathbf{f}_K$ and work backward through the network:

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_k} &= \frac{\partial \ell_i}{\partial \mathbf{f}_k} && k \in \{K, K-1, \ldots 1\} \\
\frac{\partial \ell_i}{\partial \boldsymbol{\Omega}_k} &= \frac{\partial \ell_i}{\partial \mathbf{f}_k}\mathbf{h}_k^T && k \in \{K, K-1, \ldots 1\} \\
\frac{\partial \ell_i}{\partial \mathbf{f}_{k-1}} &= \mathbb{I}[\mathbf{f}_{k-1} > 0] \odot \left(\boldsymbol{\Omega}_k^T \frac{\partial \ell_i}{\partial \mathbf{f}_k}\right), && k \in \{K, K-1, \ldots 1\} \qquad (7.13)
\end{aligned}$$

where $\odot$ denotes pointwise multiplication and $\mathbb{I}[\mathbf{f}_{k-1} > 0]$ is a vector containing ones where $\mathbf{f}_{k-1}$ is greater than zero and zeros elsewhere.



Resumen del backprop

Backward pass: We start with the derivative $\partial \ell_i / \partial \mathbf{f}_K$ of the loss function ℓ_i with respect to the network output $\mathbf{f}_K$ and work backward through the network:

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_k} &= \frac{\partial \ell_i}{\partial \mathbf{f}_k} && k \in \{K, K-1, \ldots 1\} \\
\frac{\partial \ell_i}{\partial \boldsymbol{\Omega}_k} &= \frac{\partial \ell_i}{\partial \mathbf{f}_k}\mathbf{h}_k^T && k \in \{K, K-1, \ldots 1\} \\
\frac{\partial \ell_i}{\partial \mathbf{f}_{k-1}} &= \mathbb{I}[\mathbf{f}_{k-1} > 0] \odot \left(\boldsymbol{\Omega}_k^T \frac{\partial \ell_i}{\partial \mathbf{f}_k}\right), && k \in \{K, K-1, \ldots 1\} \qquad (7.13)
\end{aligned}$$

where $\odot$ denotes pointwise multiplication and $\mathbb{I}[\mathbf{f}_{k-1} > 0]$ is a vector containing ones where $\mathbf{f}_{k-1}$ is greater than zero and zeros elsewhere. Finally, we compute the derivatives with respect to the first set of biases and weights:

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_0} &= \frac{\partial \ell_i}{\partial \mathbf{f}_0} \\
\frac{\partial \ell_i}{\partial \boldsymbol{\Omega}_0} &= \frac{\partial \ell_i}{\partial \mathbf{f}_0}\mathbf{x}_i^T
\end{aligned}$$

Ventajas e inconvenientes

- Extremadamente eficiente
  - Las funciones ReLU sólo necesitan multiplicación de matrices y aplicar umbrales
- Método hambriento de memoria: debe almacenar todas las cantidades intermedias
- Secuencial
  - puede procesar varios lotes en paralelo
  - pero las cosas se complican si todo el modelo no cabe en una sola máquina.

Gradientes

- La intuición de la backpropagation
- Modelo de juguete
- Matemáticas de fondo
- Paso adelante (forward pass)de backpropagation
- Paso hacia atrás (backward pass)por backpropagation
- Diferenciación algorítmica
- Código

Diferenciación algorítmica

- Los marcos modernos de aprendizaje profundo calculan las derivadas automáticamente
- Sólo tienes que especificar el modelo y la pérdida
- ¿Cómo? Diferenciación algorítmica
  - Cada componente sabe calcular su propia derivada
    - ReLU sabe cómo calcular la derivada de la salida con respecto a la entrada
    - La función lineal sabe calcular la derivada de la salida respecto a la entrada
    - La función lineal sabe calcular la derivada de la salida con respecto al parámetro
  - Se especifica el orden de los componentes
  - Puede calcular la cadena de derivadas
- Funciona con ramas siempre que siga siendo un grafo acíclico

Gradientes

- La intuición de la *backpropagation*
- Modelo de juguete
- Matemáticas de fondo
- Paso adelante (forward pass)de *backpropagation*
- Paso hacia atrás (backward pass)por backpropagation
- Diferenciación algorítmica
- Código

Código PyTorch

- Definir una red neuronal
- Inicializar params con inicialización He
- Definir la función de pérdida
- Elegir algoritmo de optimización
- Elija el ritmo de aprendizaje inicial
- Elija el calendario de aprendizaje
- Hacer algunos datos aleatorios
- Formación para 100 lotes

```
import torch, torch.nn as nn
from torch.utils.data import TensorDataset, DataLoader
from torch.optim.lr_scheduler import StepLR

# define input size, hidden layer size, output size
D_i, D_k, D_o = 10, 40, 5
# create model with two hidden layers
model = nn.Sequential(
   nn.Linear(D_i, D_k),
   nn.ReLU(),
   nn.Linear(D_k, D_k),
   nn.ReLU(),
   nn.Linear(D_k, D_o))

# He initialization of weights
def weights_init(layer_in):
   if isinstance(layer_in, nn.Linear):
      nn.init.kaiming_uniform(layer_in.weight)
      layer_in.bias.data.fill_(0.0)
model.apply(weights_init)

# choose least squares loss function
criterion = nn.MSELoss()
# construct SGD optimizer and initialize learning rate and momentum
optimizer = torch.optim.SGD(model.parameters(), lr = 0.01, momentum=0.9)
# object that decreases learning rate by half every 10 epochs
scheduler = StepLR(optimizer, step_size=10, gamma=0.5)

# create 100 dummy data points and store in data loader class
x = torch.randn(100, D_i)
y = torch.randn(100, D_o)
data_loader = DataLoader(TensorDataset(x,y), batch_size=10, shuffle=True)

# loop over the dataset 100 times
for epoch in range(100):
   epoch_loss = 0.0
   # loop over batches
   for i, data in enumerate(data_loader):
      # retrieve inputs and labels for this batch
      x_batch, y_batch = data
      # zero the parameter gradients
      optimizer.zero_grad()
      # forward pass
      pred = model(x_batch)
      loss = criterion(pred, y_batch)
      # backward pass
      loss.backward()
      # SGD update
      optimizer.step()
      # update statistics
      epoch_loss += loss.item()
   # print error
   print(f'Epoch {epoch:5d}, loss {epoch_loss:.3f}')
   # tell scheduler to consider updating learning rate
   scheduler.step()
```

Código PyTorch

- Definir una red neuronal
- Inicializar params con inicialización He
- Definir la función de pérdida
- Elegir algoritmo de optimización
- Elija el ritmo de aprendizaje inicial
- Elija el calendario de aprendizaje
- Hacer algunos datos aleatorios
- Formación para 100 lotes

```
import torch, torch.nn as nn
from torch.utils.data import TensorDataset, DataLoader
from torch.optim.lr_scheduler import StepLR

# define input size, hidden layer size, output size
D_i, D_k, D_o = 10, 40, 5
# create model with two hidden layers
model = nn.Sequential(
   nn.Linear(D_i, D_k),
   nn.ReLU(),
   nn.Linear(D_k, D_k),
   nn.ReLU(),
   nn.Linear(D_k, D_o))

# He initialization of weights
def weights_init(layer_in):
   if isinstance(layer_in, nn.Linear):
      nn.init.kaiming_uniform(layer_in.weight)
      layer_in.bias.data.fill_(0.0)
model.apply(weights_init)

# choose least squares loss function
criterion = nn.MSELoss()
# construct SGD optimizer and initialize learning rate and momentum
optimizer = torch.optim.SGD(model.parameters(), lr = 0.01, momentum=0.9)
# object that decreases learning rate by half every 10 epochs
scheduler = StepLR(optimizer, step_size=10, gamma=0.5)

# create 100 dummy data points and store in data loader class
x = torch.randn(100, D_i)
y = torch.randn(100, D_o)
data_loader = DataLoader(TensorDataset(x,y), batch_size=10, shuffle=True)

# loop over the dataset 100 times
for epoch in range(100):
   epoch_loss = 0.0
   # loop over batches
```

Código PyTorch

- Definir una red neuronal
- Inicializar params con inicialización He
- Definir la función de pérdida
- Elegir algoritmo de optimización
- Elija el ritmo de aprendizaje inicial
- Elija el calendario de aprendizaje
- Hacer algunos datos aleatorios
- Formación para 100 lotes

```
model.apply(weights_init)

# choose least squares loss function
criterion = nn.MSELoss()
# construct SGD optimizer and initialize learning rate and momentum
optimizer = torch.optim.SGD(model.parameters(), lr = 0.01, momentum=0.9)
# object that decreases learning rate by half every 10 epochs
scheduler = StepLR(optimizer, step_size=10, gamma=0.5)

# create 100 dummy data points and store in data loader class
x = torch.randn(100, D_i)
y = torch.randn(100, D_o)
data_loader = DataLoader(TensorDataset(x,y), batch_size=10, shuffle=True)

# loop over the dataset 100 times
for epoch in range(100):
    epoch_loss = 0.0
    # loop over batches
    for i, data in enumerate(data_loader):
        # retrieve inputs and labels for this batch
        x_batch, y_batch = data
        # zero the parameter gradients
        optimizer.zero_grad()
        # forward pass
        pred = model(x_batch)
        loss = criterion(pred, y_batch)
        # backward pass
        loss.backward()
        # SGD update
        optimizer.step()
        # update statistics
        epoch_loss += loss.item()
    # print error
    print(f'Epoch {epoch:5d}, loss {epoch_loss:.3f}')
    # tell scheduler to consider updating learning rate
    scheduler.step()
```