

Tópicos en Inteligencia Artificial Deep Learning

Backpropagation

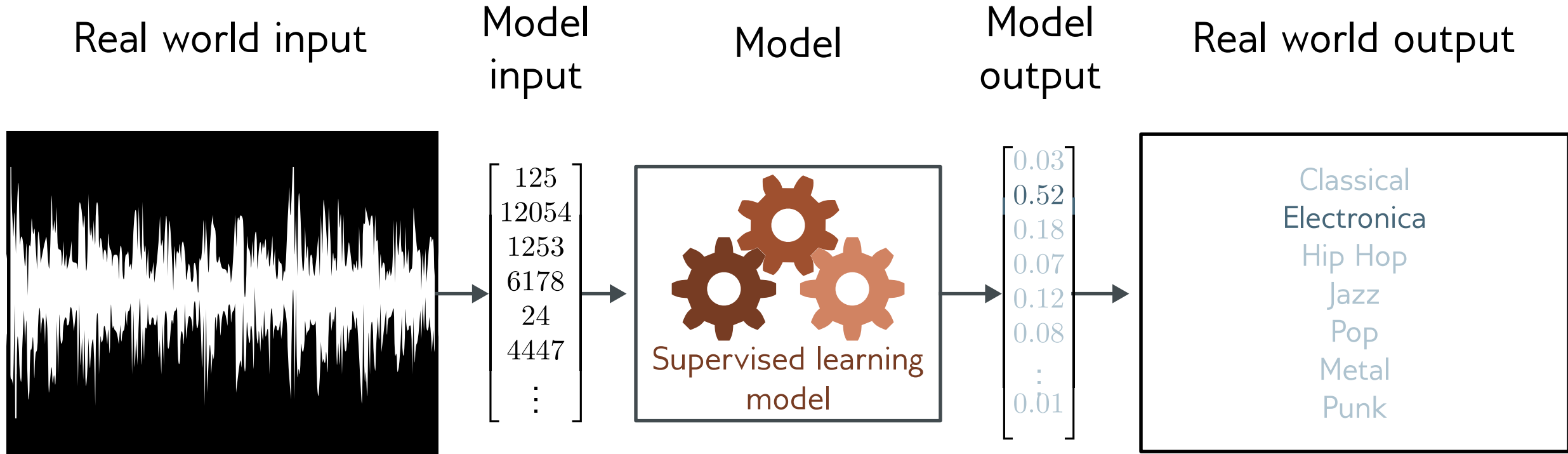
Basado en el ppt 07 – Gradients de

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:

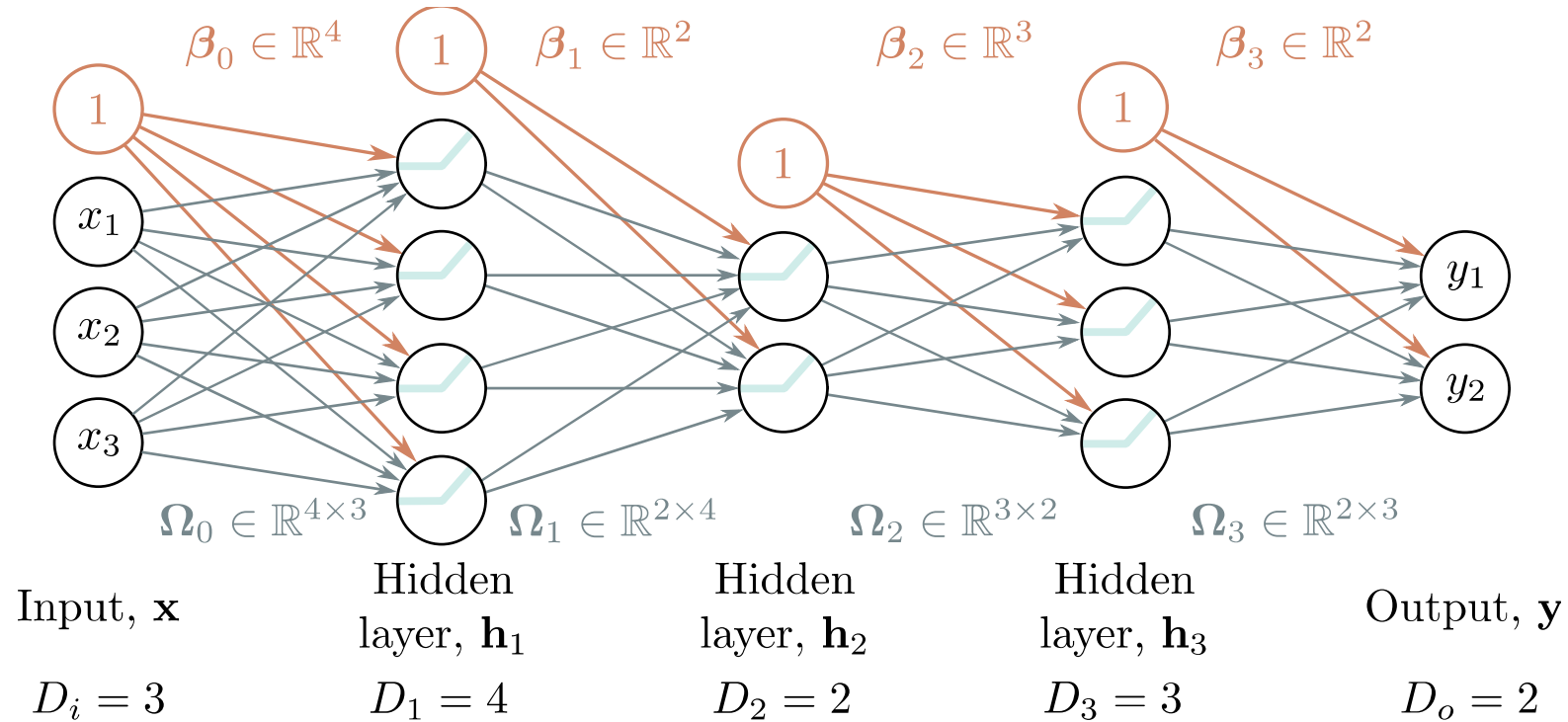
$$L[\phi, f[\mathbf{x}_i, \phi], \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^I]$$

o para abreviar:

$$L[\phi]$$

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

Ejemplo



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

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

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

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

Problema 1: Cómo calcular los gradientes?

Pérdida:
suma de los términos individuales

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

Algoritmo SGD:

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

Parámetros:

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

Necesidad de calcular gradientes

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

Qué son los Batches?



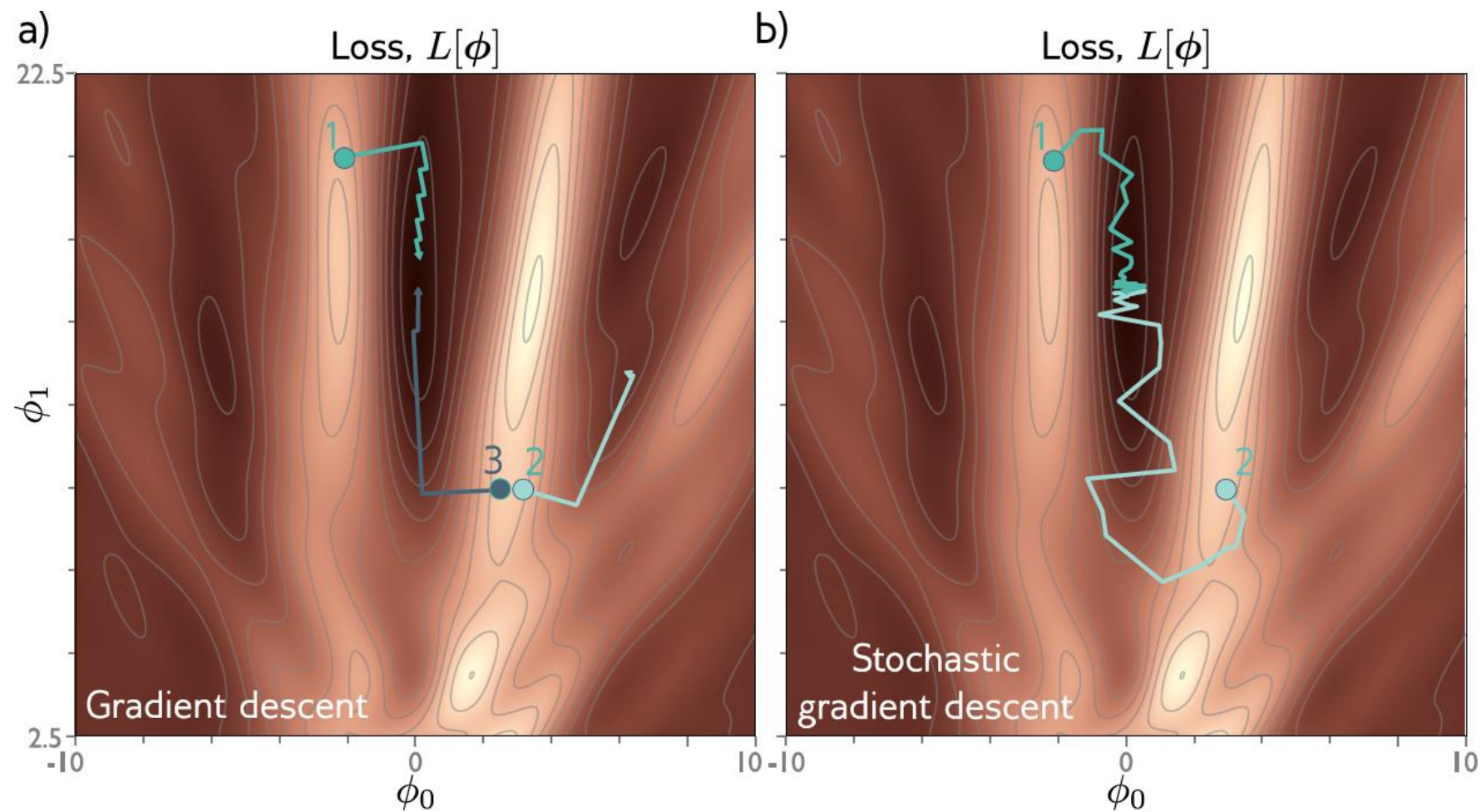
¿Por qué es tan importante?

- Una red neuronal no es más que una ecuación:

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

- 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 de *backpropagation*
- Paso hacia atrás por *backpropagation*
- Diferenciación algorítmica
- Código

Problema 1: Cómo calcular los gradientes?

Pérdida:
suma de los términos individuales

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

Algoritmo SGD:

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

Parámetros:

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

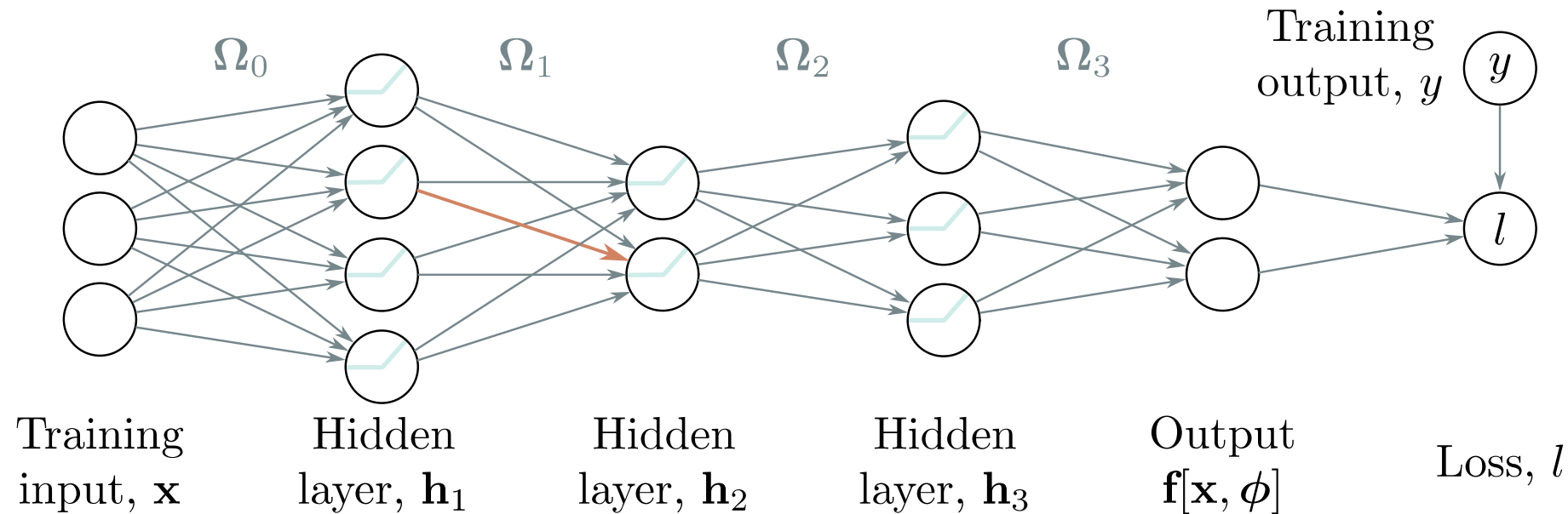
Necesidad de calcular gradientes

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

Algoritmo para calcular el gradiente de forma eficiente

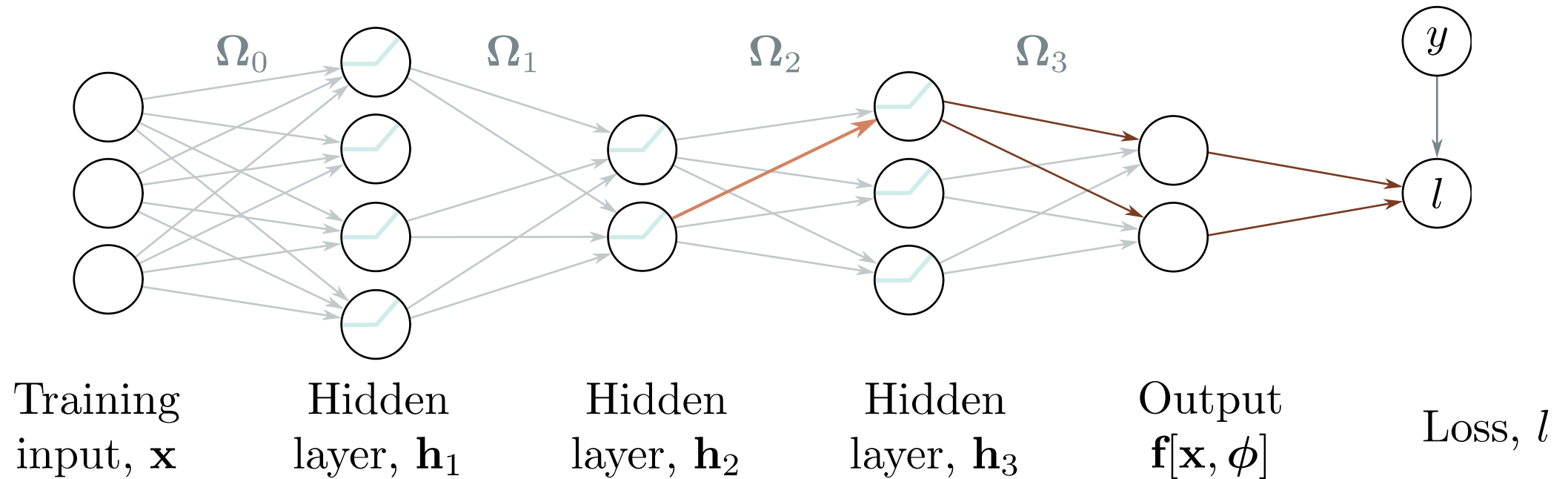
- “Backpropagation algorithm” Rumelhart, Hinton y Williams (1986)

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.

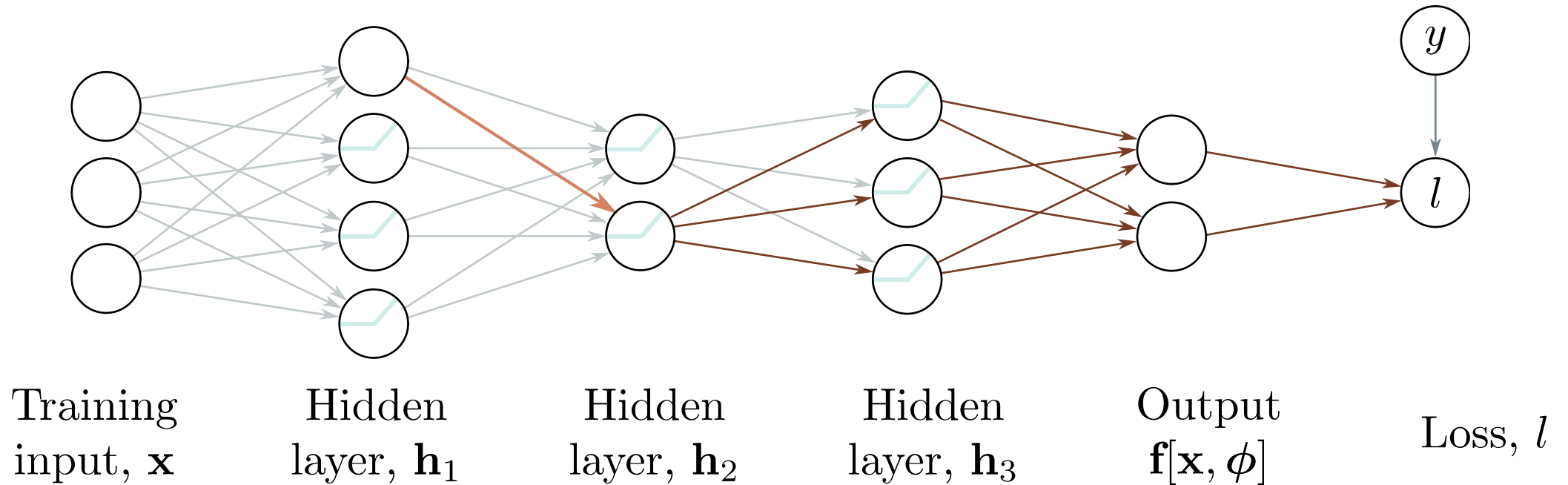
Intuición BackProp nº 2: el pase hacia atrás



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

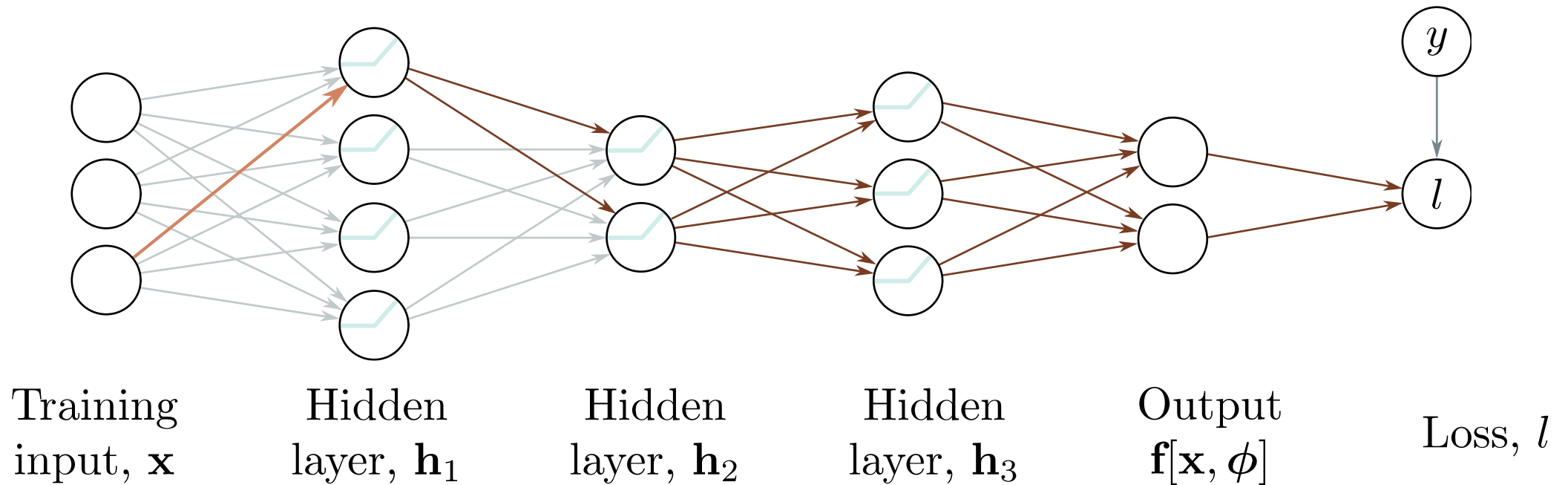
Intuición BackProp nº 2: el pase hacia atrás



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 retropropagación
- Modelo de juguete
- Matemáticas de base
- Paso adelante de retropropagación
- Paso hacia atrás por retropropagación
- 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

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

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 $u = g(x)$ $\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$

- Sea $y = f(u)$ y $u = g(x)$ $\frac{d}{dx} f(g(x)) = \frac{df}{dg} \cdot \frac{dg}{dx}$

- Sea $y = f(u)$ y $u = g(x)$ $f(g(x))' = f'(g(x)) \cdot g'(x)$

Composición de funciones: Ejemplo

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

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

$$\begin{aligned} \frac{df}{dy} &= 2y = 2(2x + 1) & \frac{dh}{dx} &= 2(2x + 1) \cdot 2 = 4(2x + 1) \\ \frac{dy}{dx} &= \frac{dg}{dx} = 2 & \frac{d}{dx}(f(g(x))) &= \frac{d}{dx}[(2x + 1)^2] = 4(2x + 1) \end{aligned}$$

Función de juguete

$$f[x, \phi] = \beta_3 + \omega_3 \cdot \cos \left[\beta_2 + \omega_2 \cdot \exp \left[\beta_1 + \omega_1 \cdot \sin [\beta_0 + \omega_0 \cdot x] \right] \right]$$

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

- 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

“Red” Neuronal

$$f[x, \phi] = \beta_3 + \omega_3 \cdot \cos \left[\beta_2 + \omega_2 \cdot \exp \left[\beta_1 + \omega_1 \cdot \sin [\beta_0 + \omega_0 \cdot x] \right] \right]$$



Función de juguete

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

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

Derivadas

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

Gradientes de la función juguete

$$f[x, \phi] = \beta_3 + \omega_3 \cdot \cos \left[\beta_2 + \omega_2 \cdot \exp \left[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x] \right] \right]$$

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

Queremos calcular:

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

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


Gradientes de funciones compuestas

$$f[x, \phi] = \beta_3 + \omega_3 \cdot \cos \left[\beta_2 + \omega_2 \cdot \exp \left[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x] \right] \right]$$

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

Calcular expresiones a mano:

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

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

Gradientes de funciones compuestas

$$f[x, \phi] = \beta_3 + \omega_3 \cdot \cos \left[\beta_2 + \omega_2 \cdot \exp \left[\beta_1 + \omega_1 \cdot \sin[\beta_0 + \omega_0 \cdot x] \right] \right]$$

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

Calcular expresiones a mano:

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

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

Pase hacia delante (*forward pass*)

$$f[x, \phi] = \beta_3 + \omega_3 \cdot \cos \left[\beta_2 + \omega_2 \cdot \exp \left[\beta_1 + \omega_1 \cdot \sin [\beta_0 + \omega_0 \cdot x] \right] \right]$$

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

1. Escribe esto como una serie de cálculos intermedios
2. Calcula estas cantidades intermedias

Pase hacia delante

$$f[x, \phi] = \beta_3 + \omega_3 \cdot \cos \left[\beta_2 + \omega_2 \cdot \exp \left[\beta_1 + \omega_1 \cdot \sin [\beta_0 + \omega_0 \cdot x] \right] \right]$$

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

1. Escribe esto como una serie de cálculos intermedios

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

$$h_1 = \sin[f_0]$$

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

$$h_2 = \exp[f_1]$$

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

$$h_3 = \cos[f_2]$$

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

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

2. Calcula estas cantidades intermedias

Pase hacia delante

$$f[x, \phi] = \beta_3 + \omega_3 \cdot \cos \left[\beta_2 + \omega_2 \cdot \exp \left[\beta_1 + \omega_1 \cdot \sin [\beta_0 + \omega_0 \cdot x] \right] \right]$$

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

1. Escribe esto como una serie de cálculos intermedios

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

$$h_1 = \sin[f_0]$$

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

$$h_2 = \exp[f_1]$$

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

$$h_3 = \cos[f_2]$$

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

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

2. Calcula estas cantidades intermedias



Paso atrás (*backward pass*)

$$f[x, \phi] = \beta_3 + \omega_3 \cdot \cos \left[\beta_2 + \omega_2 \cdot \exp \left[\beta_1 + \omega_1 \cdot \sin [\beta_0 + \omega_0 \cdot x] \right] \right]$$

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

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*)

$$f[x, \phi] = \beta_3 + \omega_3 \cdot \cos \left[\beta_2 + \omega_2 \cdot \exp \left[\beta_1 + \omega_1 \cdot \sin [\beta_0 + \omega_0 \cdot x] \right] \right]$$

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

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.

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

$$h_1 = \sin[f_0]$$

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

$$h_2 = \exp[f_1]$$

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

$$h_3 = \cos[f_2]$$

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

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

- 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.

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

$$h_1 = \sin[f_0]$$

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

$$h_2 = \exp[f_1]$$

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

$$h_3 = \cos[f_2]$$

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

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

- 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 ?

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

Expresiones equivalentes

Backward pass

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

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

$$h_1 = \sin[f_0]$$

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

$$h_2 = \exp[f_1]$$

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

$$h_3 = \cos[f_2]$$

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

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

- 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.

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

$$h_1 = \sin[f_0]$$

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

$$h_2 = \exp[f_1]$$

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

$$h_3 = \cos[f_2]$$

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

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

- 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 en el primer paso del **backward pass!**

Backward pass

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

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

$$h_1 = \sin[f_0]$$

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

$$h_2 = \exp[f_1]$$

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

$$h_3 = \cos[f_2]$$

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

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

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

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

$$\frac{\partial f_3}{\partial h_3} = \omega_3$$

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

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:

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

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

$$\frac{\partial f_3}{\partial h_3} = \omega_3$$

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.

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

$$h_1 = \sin[f_0]$$

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

$$h_2 = \exp[f_1]$$

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

$$h_3 = \cos[f_2]$$

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

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

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

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

$$-\sin(f_2)$$

Ya está computado.

Backward pass

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

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

$$h_1 = \sin[f_0]$$

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

$$h_2 = \exp[f_1]$$

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

$$h_3 = \cos[f_2]$$

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

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

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

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

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

Backward pass

1. 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

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

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

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

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

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

Backward pass

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

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

$$h_1 = \sin[f_0]$$

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

$$h_2 = \exp[f_1]$$

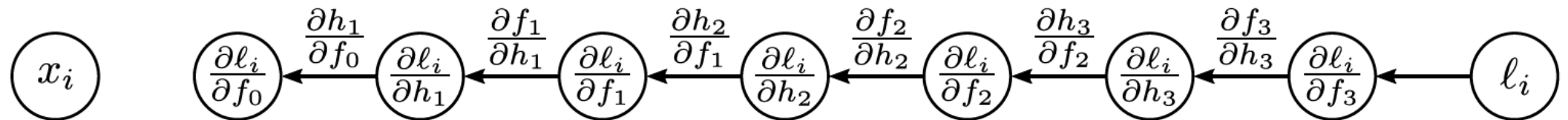
$$f_2 = \beta_2 + \omega_2 \cdot h_2$$

$$h_3 = \cos[f_2]$$

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

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

- 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 .

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

$$h_1 = \sin[f_0]$$

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

$$h_2 = \exp[f_1]$$

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

$$h_3 = \cos[f_2]$$

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

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

- 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 .

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

$$h_1 = \sin[f_0]$$

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

$$h_2 = \exp[f_1]$$

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

$$h_3 = \cos[f_2]$$

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

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

- 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 .

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

$$h_1 = \sin[f_0]$$

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

$$h_2 = \exp[f_1]$$

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

$$h_3 = \cos[f_2]$$

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

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

- 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 .

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

$$h_1 = \sin[f_0]$$

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

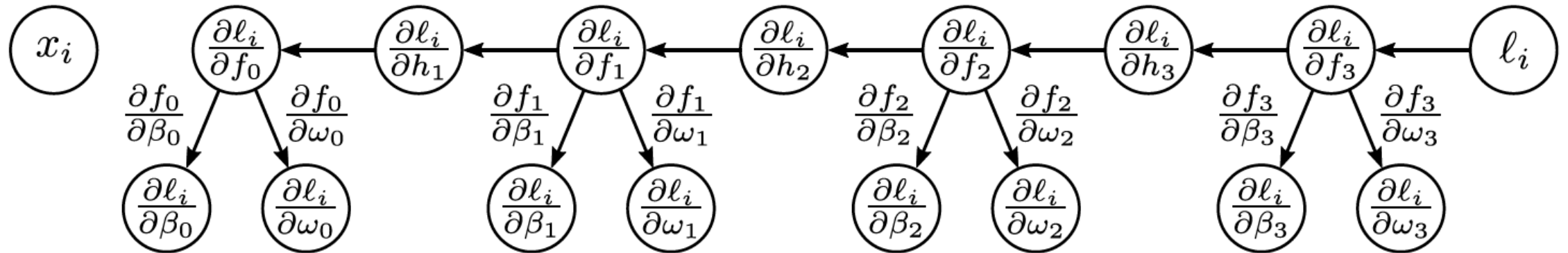
$$h_2 = \exp[f_1]$$

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

$$h_3 = \cos[f_2]$$

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

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



Gradientes

- La intuición de la retropropagación
- Modelo de juguete
- Matemáticas de base
- Paso adelante de retropropagación
- Paso hacia atrás por retropropagación
- 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?

Cálculo matricial

Adaptado de 23.5 Matrix Calculus

- For a function $y = f[\mathbf{x}]$ where $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^D$, the derivative $\partial y / \partial \mathbf{x}$ is also a D -dimensional vector, where the i^{th} element is computed as $\partial y / \partial x_i$.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{x}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{x} \in \mathbb{R}^{D_x}$, the derivative $\partial \mathbf{y} / \partial \mathbf{x}$ is a $D_x \times D_y$ matrix where element (i, j) contains the derivative $\partial y_j / \partial x_i$. This is known as a *Jacobian* and is sometimes written as $\nabla_{\mathbf{x}} \mathbf{y}$ in other documents.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y} / \partial \mathbf{X}$ is a 3D tensor containing the derivatives $\partial y_i / \partial x_{jk}$.

Cálculo matricial

Función escalar - vector

- Sea $\mathbf{f}$ una función escalar $y = \mathbf{f}[\mathbf{x}]$ con $\mathbf{x} \in \mathbb{R}^D$, entonces $\partial y / \partial \mathbf{x}$ respeta la dimensión del denominador $\mathbf{x}$. Esto es:

$$\frac{\partial y}{\partial \mathbf{x}} \in \mathbb{R}^D$$

Ejemplo:

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

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

Cálculo matricial

Función escalar - vector

- For a function $y = f[\mathbf{x}]$ where $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^D$, the derivative $\partial y / \partial \mathbf{x}$ is also a D -dimensional vector, where the i^{th} element is computed as $\partial y / \partial x_i$.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{x}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{x} \in \mathbb{R}^{D_x}$, the derivative $\partial \mathbf{y} / \partial \mathbf{x}$ is a $D_x \times D_y$ matrix where element (i, j) contains the derivative $\partial y_j / \partial x_i$. This is known as a *Jacobian* and is sometimes written as $\nabla_{\mathbf{x}} \mathbf{y}$ in other documents.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y} / \partial \mathbf{X}$ is a 3D tensor containing the derivatives $\partial y_i / \partial x_{jk}$.

Cálculo matricial

Función vectorial - vector

- Sea $\mathbf{f}$ una función vectorial con $\mathbf{y} = \mathbf{f}[\mathbf{x}]$ con $\mathbf{y} \in \mathbb{R}^{D_y}$ y $\mathbf{x} \in \mathbb{R}^{D_x}$, entonces $\partial \mathbf{y} / \partial \mathbf{x}$ es una matriz de dimensión $D_x \times D_y$ de donde cada entrada (i, j) contiene $\partial y_j / \partial x_i$ denominador. Esta matriz $\mathbf{J}_f$ es conocida como el **Jacobiano** escrita como $\nabla_{\mathbf{x}} \mathbf{y}$ en algunas fuentes.

Ejemplo:

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

Cálculo matricial

Función vectorial - vector

- Fíjense que $\mathbf{J}_f \in \mathbb{R}^{D_x \times D_y}$ con $\mathbf{y} \in \mathbb{R}^{D_y}$ y $\mathbf{x} \in \mathbb{R}^{D_x}$ se pudiese reescribir como

$$\mathbf{J}_f = \begin{bmatrix} \frac{\partial f^\top}{\partial x_1} \\ \vdots \\ \frac{\partial f^\top}{\partial x_n} \end{bmatrix} = [\nabla f_1 \quad \cdots \quad \nabla f_m] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

- De dónde es fácil ver que aplicamos m veces la propiedad 1.

Cálculo matricial

Notaciones

- Si quisiéramos estudiar, por ejemplo, usando Wikipedia encontramos:

Definition [\[edit \]](#)

Let $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a function such that each of its first-order partial derivatives exists on $\mathbb{R}^n$. This function takes a vector $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ as input and produces the vector $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x})) \in \mathbb{R}^m$ as output. Then the Jacobian matrix of $\mathbf{f}$, denoted $\mathbf{J}_{\mathbf{f}}$, is the $m \times n$ matrix whose (i, j) entry is $\frac{\partial f_i}{\partial x_j}$; explicitly

$$\mathbf{J}_{\mathbf{f}} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^T f_1 \\ \vdots \\ \nabla^T f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

where $\nabla^T f_i$ is the transpose (row vector) of the [gradient](#) of the i -th component.

Cálculo matricial

Notaciones numerador y denominador

- **(Numerador)** Wikipedia nos dice:

- $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$

- $\mathbf{J}_f$ es la matriz $m \times n$ cuya entrada (i, j) es $\partial f_i / \partial x_j$.

- **(Denominador)** Prince nos dice:

- $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$

- $\mathbf{J}_f$ es la matriz $m \times n$ cuya entrada (j, i) es $\partial f_i / \partial x_j$.

$$\underbrace{\frac{\partial \mathbf{y}}{\partial \mathbf{x}}}_{\text{Prince, denominador}} = \underbrace{\mathbf{J}_f^T}_{\text{Wikipedia, numerador transpuesta}}$$

Cálculo matricial

Notaciones numerador y denominador

Vista	Significado
$\begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}$	Stack de vectores columna $\partial f / \partial x_j$ lado a lado
$\begin{bmatrix} \nabla^\top f_1 \\ \vdots \\ \nabla^\top f_m \end{bmatrix}$	Stack de gradientes fila $\nabla^\top f_i$ uno encima de otro
Elemento $(i, j) = \partial f_i / \partial x_j$	Fila = índice de salida i , columna = índice de entrada j

Vista	Significado
$\begin{bmatrix} \frac{\partial f^\top}{\partial x_1} \\ \vdots \\ \frac{\partial f^\top}{\partial x_n} \end{bmatrix}$	Stack de vectores fila $\partial f^\top / \partial x_j$ uno encima de otro
$\begin{bmatrix} \nabla f_1 & \cdots & \nabla f_m \end{bmatrix}$	Stack de gradientes columna ∇f_i lado a lado
Elemento $(j, i) = \partial f_i / \partial x_j$	Fila = índice de entrada j , columna = índice de salida i

Cálculo matricial

Notación: Función escalar – vector en notación numerador

- Sea $\mathbf{f}$ una función escalar $y = \mathbf{f}[\mathbf{x}]$ con $\mathbf{x} \in \mathbb{R}^D$, entonces $\partial y / \partial \mathbf{x}$ respeta la dimensión del numerador y . Esto es:

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

- Salvo excepciones, en el curso usaremos notación **denominador**

Cálculo matricial

Función vectorial - Matriz

- For a function $y = f[\mathbf{x}]$ where $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^D$, the derivative $\partial y / \partial \mathbf{x}$ is also a D -dimensional vector, where the i^{th} element is computed as $\partial y / \partial x_i$.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{x}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{x} \in \mathbb{R}^{D_x}$, the derivative $\partial \mathbf{y} / \partial \mathbf{x}$ is a $D_x \times D_y$ matrix where element (i, j) contains the derivative $\partial y_j / \partial x_i$. This is known as a *Jacobian* and is sometimes written as $\nabla_{\mathbf{x}} \mathbf{y}$ in other documents.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y} / \partial \mathbf{X}$ is a 3D tensor containing the derivatives $\partial y_i / \partial x_{jk}$.

Cálculo matricial

Función vectorial - Matriz

- Sea $\mathbf{f}$ una función vectorial $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ con $\mathbf{y} \in \mathbb{R}^{D_y}$ y $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, entonces $\partial \mathbf{y} / \partial \mathbf{X}$ es un tensor 3D que contiene las derivadas $\partial y_i / \partial x_{jk}$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{X}} \in \mathbb{R}^{D_1 \times D_2 \times D_y}$$

Cada elemento (j, k, i)

$$\left[\frac{\partial \mathbf{y}}{\partial \mathbf{X}} \right]_{j,k,i} = \frac{\partial y_i}{\partial x_{jk}}$$

Cálculo matricial

Función vectorial - Matriz

- En este caso $\frac{\partial \mathbf{y}}{\partial \mathbf{X}} \in \mathbb{R}^{D_1 \times D_2 \times D_y}$ podemos interpretar la derivada de una función escalar por una matriz para cada dimensión i

$$\frac{\partial \mathbf{y}}{\partial \mathbf{X}} = \left\{ \underbrace{\begin{bmatrix} \frac{\partial y_i}{\partial x_{11}} & \frac{\partial y_i}{\partial x_{12}} & \cdots & \frac{\partial y_i}{\partial x_{1D_2}} \\ \frac{\partial y_i}{\partial x_{21}} & \frac{\partial y_i}{\partial x_{22}} & \cdots & \frac{\partial y_i}{\partial x_{2D_2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_i}{\partial x_{D_1 1}} & \frac{\partial y_i}{\partial x_{D_1 2}} & \cdots & \frac{\partial y_i}{\partial x_{D_1 D_2}} \end{bmatrix}}_{\text{slice } i} \right\}_{i=1}^{D_y}$$

Cálculo matricial

Función vectorial - Matriz

- En este caso $\frac{\partial \mathbf{y}}{\partial \mathbf{X}} \in \mathbb{R}^{D_1 \times D_2 \times D_y}$ podemos interpretar la derivada de una función escalar por una matriz para cada dimensión i

$$\frac{\partial \mathbf{y}}{\partial \mathbf{X}} = \left[\begin{bmatrix} \frac{\partial y_1}{\partial x_{11}} & \cdots & \frac{\partial y_1}{\partial x_{1D_2}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_{D_1}} & \cdots & \frac{\partial y_1}{\partial x_{D_1 D_2}} \end{bmatrix}, \cdots, \begin{bmatrix} \frac{\partial y_{D_y}}{\partial x_{11}} & \cdots & \frac{\partial y_{D_y}}{\partial x_{1D_2}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_{D_y}}{\partial x_{D_1}} & \cdots & \frac{\partial y_{D_y}}{\partial x_{D_1 D_2}} \end{bmatrix} \right]$$

Cálculo matricial

Función escalar – Matriz: **La oculta**

- For a function $y = f[\mathbf{x}]$ where $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^D$, the derivative $\partial y / \partial \mathbf{x}$ is also a D -dimensional vector, where the i^{th} element is computed as $\partial y / \partial x_i$.
- For a function $\mathbf{y} = f[\mathbf{x}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{x} \in \mathbb{R}^{D_x}$, the derivative $\partial \mathbf{y} / \partial \mathbf{x}$ is a $D_x \times D_y$ matrix where element (i, j) contains the derivative $\partial y_j / \partial x_i$. This is known as a *Jacobian* and is sometimes written as $\nabla_{\mathbf{x}} \mathbf{y}$ in other documents.
- For a function $\mathbf{y} = f[\mathbf{X}]$ where $y \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y} / \partial \mathbf{X}$ is a matrix where element (i, j) contains the derivative $\partial y / \partial x_{ij}$.
- For a function $\mathbf{y} = f[\mathbf{X}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y} / \partial \mathbf{X}$ is a 3D tensor containing the derivatives $\partial y_i / \partial x_{jk}$.

Cálculo matricial

La oculta: Función escalar - Matriz

- Sea $\mathbf{f}$ una función escalar $y = \mathbf{f}[\mathbf{x}]$ con $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, entonces $\partial \mathbf{y} / \partial \mathbf{X}$ respeta la dimensión del denominador $\mathbf{X}$. Esto es:

$$\frac{\partial y}{\partial \mathbf{X}} \in \mathbb{R}^{D_1 \times D_2}$$

Cada elemento (i, j) de $\partial y / \partial x_{ij}$

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1D_2}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{2D_2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{D_1 1}} & \frac{\partial y}{\partial x_{D_1 2}} & \cdots & \frac{\partial y}{\partial x_{D_1 D_2}} \end{bmatrix}$$

Cálculo matricial

Función escalar - Matriz

- Sea $\mathbf{f}$ una función escalar $y = \mathbf{f}[\mathbf{x}]$ con $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, entonces $\partial \mathbf{y} / \partial \mathbf{X}$ respeta la dimensión del denominador $\mathbf{X}$. Esto es:

$$\frac{\partial y}{\partial \mathbf{X}} \in \mathbb{R}^{D_1 \times D_2}$$

Ejemplo

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

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

De cálculo escalar a vectorial

Derivadas escalares:

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

Derivadas vectoriales:

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

Cálculo matricial

En qué caso estamos?

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \quad \frac{\partial \mathbf{f}_3}{\partial \boldsymbol{\beta}_3} ?$$

- For a function $y = f[\mathbf{x}]$ where $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^D$, the derivative $\partial y / \partial \mathbf{x}$ is also a D -dimensional vector, where the i^{th} element is computed as $\partial y / \partial x_i$.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{x}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{x} \in \mathbb{R}^{D_x}$, the derivative $\partial \mathbf{y} / \partial \mathbf{x}$ is a $D_x \times D_y$ matrix where element (i, j) contains the derivative $\partial y_j / \partial x_i$. This is known as a *Jacobian* and is sometimes written as $\nabla_{\mathbf{x}} \mathbf{y}$ in other documents.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y} / \partial \mathbf{X}$ is a matrix where element (i, j) contains the derivative $\partial y / \partial x_{ij}$.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y} / \partial \mathbf{X}$ is a 3D tensor containing the derivatives $\partial y_i / \partial x_{jk}$.

Cálculo matricial

En qué caso estamos?

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \quad \frac{\partial \mathbf{f}_3}{\partial \boldsymbol{\beta}_3} ?$$

- For a function $y = f[\mathbf{x}]$ where $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^D$, the derivative $\partial y / \partial \mathbf{x}$ is also a D -dimensional vector, where the i^{th} element is computed as $\partial y / \partial x_i$.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{x}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{x} \in \mathbb{R}^{D_x}$, the derivative $\partial \mathbf{y} / \partial \mathbf{x}$ is a $D_x \times D_y$ matrix where element (i, j) contains the derivative $\partial y_j / \partial x_i$. This is known as a *Jacobian* and is sometimes written as $\nabla_{\mathbf{x}} \mathbf{y}$ in other documents.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y} / \partial \mathbf{X}$ is a matrix where element (i, j) contains the derivative $\partial y / \partial x_{ij}$.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y} / \partial \mathbf{X}$ is a 3D tensor containing the derivatives $\partial y_i / \partial x_{jk}$.

Calculo matricial en capas lineales

Caso para β

- Supongamos que la capa 3 recibe 2 entradas y produce 3. Dimensiones:

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

$\mathbf{h}_3 \in \mathbb{R}^{2 \times 1}$ - vector de entrada con componentes h_1, h_2

$\Omega_3 \in \mathbb{R}^{3 \times 2}$ - matriz de pesos con entradas ω_{ij}

$\beta_3 \in \mathbb{R}^{3 \times 1}$ - vector de sesgos con componentes $\beta_1, \beta_2, \beta_3$

$\mathbf{f}_3 \in \mathbb{R}^{3 \times 1}$ - vector de salida con componentes f_1, f_2, f_3

Calculo matricial en capas lineales

Caso para β

- Supongamos que la capa 3 recibe 2 entradas y produce 3. Dimensiones:

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

$$\begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \\ \omega_{31} & \omega_{32} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} \beta_1 + \omega_{11}h_1 + \omega_{12}h_2 \\ \beta_2 + \omega_{21}h_1 + \omega_{22}h_2 \\ \beta_3 + \omega_{31}h_1 + \omega_{32}h_2 \end{bmatrix}$$

$$f_i = \beta_i + \omega_{i1}h_1 + \omega_{i2}h_2 = \beta_i + \sum_j \omega_{ij}h_j \quad f_i = \beta_i + \underbrace{\sum_j \omega_{ij}h_j}_{\text{no depende de } \beta_3}$$

Calculo matricial en capas lineales

Caso para β

- Supongamos que la capa 3 recibe 2 entradas y produce 3. Dimensiones:

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

$$\frac{\partial \mathbf{f}_3}{\partial \beta_3} = \begin{bmatrix} \frac{\partial f_1}{\partial \beta_1} & \frac{\partial f_2}{\partial \beta_1} & \frac{\partial f_3}{\partial \beta_1} \\ \frac{\partial f_1}{\partial \beta_2} & \frac{\partial f_2}{\partial \beta_2} & \frac{\partial f_3}{\partial \beta_2} \\ \frac{\partial f_1}{\partial \beta_3} & \frac{\partial f_2}{\partial \beta_3} & \frac{\partial f_3}{\partial \beta_3} \end{bmatrix} \quad \begin{array}{l} \frac{\partial f_1}{\partial \beta_1} = 1, \quad \frac{\partial f_1}{\partial \beta_2} = 0, \quad \frac{\partial f_1}{\partial \beta_3} = 0 \\ \frac{\partial f_2}{\partial \beta_1} = 0, \quad \frac{\partial f_2}{\partial \beta_2} = 1, \quad \frac{\partial f_2}{\partial \beta_3} = 0 \\ \frac{\partial f_3}{\partial \beta_1} = 0, \quad \frac{\partial f_3}{\partial \beta_2} = 0, \quad \frac{\partial f_3}{\partial \beta_3} = 1 \end{array}$$

Calculo matricial en capas lineales

Caso para β

- Supongamos que la capa 3 recibe 2 entradas y produce 3. Dimensiones:

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

$$\frac{\partial f_3}{\partial \beta_3} = \begin{bmatrix} \frac{\partial f_1}{\partial \beta_1} & \frac{\partial f_2}{\partial \beta_1} & \frac{\partial f_3}{\partial \beta_1} \\ \frac{\partial f_1}{\partial \beta_2} & \frac{\partial f_2}{\partial \beta_2} & \frac{\partial f_3}{\partial \beta_2} \\ \frac{\partial f_1}{\partial \beta_3} & \frac{\partial f_2}{\partial \beta_3} & \frac{\partial f_3}{\partial \beta_3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{I}$$

Calculo matricial en capas lineales

Caso para β

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:

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

De cálculo escalar a vectorial

Derivadas escalares:

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

Derivadas vectoriales:

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

Cálculo matricial

En qué caso estamos?

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \quad \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} ?$$

- For a function $y = f[\mathbf{x}]$ where $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^D$, the derivative $\partial y / \partial \mathbf{x}$ is also a D -dimensional vector, where the i^{th} element is computed as $\partial y / \partial x_i$.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{x}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{x} \in \mathbb{R}^{D_x}$, the derivative $\partial \mathbf{y} / \partial \mathbf{x}$ is a $D_x \times D_y$ matrix where element (i, j) contains the derivative $\partial y_j / \partial x_i$. This is known as a *Jacobian* and is sometimes written as $\nabla_{\mathbf{x}} \mathbf{y}$ in other documents.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y} / \partial \mathbf{X}$ is a matrix where element (i, j) contains the derivative $\partial y / \partial x_{ij}$.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y} / \partial \mathbf{X}$ is a 3D tensor containing the derivatives $\partial y_i / \partial x_{jk}$.

Cálculo matricial

En qué caso estamos?

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \quad \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} ?$$

- For a function $y = f[\mathbf{x}]$ where $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^D$, the derivative $\partial y / \partial \mathbf{x}$ is also a D -dimensional vector, where the i^{th} element is computed as $\partial y / \partial x_i$.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{x}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{x} \in \mathbb{R}^{D_x}$, the derivative $\partial \mathbf{y} / \partial \mathbf{x}$ is a $D_x \times D_y$ matrix where element (i, j) contains the derivative $\partial y_j / \partial x_i$. This is known as a *Jacobian* and is sometimes written as $\nabla_{\mathbf{x}} \mathbf{y}$ in other documents.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y} / \partial \mathbf{X}$ is a matrix where element (i, j) contains the derivative $\partial y / \partial x_{ij}$.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y} / \partial \mathbf{X}$ is a 3D tensor containing the derivatives $\partial y_i / \partial x_{jk}$.

Calculo matricial en capas lineales

Caso para $\mathbf{h}$

- Supongamos que la capa 3 recibe 2 entradas y produce 3. Dimensiones:

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

$\mathbf{h}_3 \in \mathbb{R}^{2 \times 1}$ - vector de entrada con componentes h_1, h_2

$\boldsymbol{\Omega}_3 \in \mathbb{R}^{3 \times 2}$ - matriz de pesos con entradas ω_{ij}

$\boldsymbol{\beta}_3 \in \mathbb{R}^{3 \times 1}$ - vector de sesgos con componentes $\beta_1, \beta_2, \beta_3$

$\mathbf{f}_3 \in \mathbb{R}^{3 \times 1}$ - vector de salida con componentes f_1, f_2, f_3

Calculo matricial en capas lineales

Caso para $\mathbf{h}$

- Supongamos que la capa 3 recibe 2 entradas y produce 3.

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

$$\begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \\ \omega_{31} & \omega_{32} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} \beta_1 + \omega_{11}h_1 + \omega_{12}h_2 \\ \beta_2 + \omega_{21}h_1 + \omega_{22}h_2 \\ \beta_3 + \omega_{31}h_1 + \omega_{32}h_2 \end{bmatrix}$$

$$f_i = \beta_i + \omega_{i1}h_1 + \omega_{i2}h_2 = \beta_i + \sum_j \omega_{ij}h_j$$

Calculo matricial en capas lineales

Caso para $\mathbf{h}$

- Supongamos que la capa 3 recibe 2 entradas y produce 3.

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

$$\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} = \begin{bmatrix} \frac{\partial f_1}{\partial h_1} & \frac{\partial f_2}{\partial h_1} & \frac{\partial f_3}{\partial h_1} \\ \frac{\partial f_1}{\partial h_2} & \frac{\partial f_2}{\partial h_2} & \frac{\partial f_3}{\partial h_2} \end{bmatrix}$$

$$\begin{aligned} \frac{\partial f_1}{\partial h_1} &= \omega_{11}, & \frac{\partial f_1}{\partial h_2} &= \omega_{12} \\ \frac{\partial f_2}{\partial h_1} &= \omega_{21}, & \frac{\partial f_2}{\partial h_2} &= \omega_{22} \\ \frac{\partial f_3}{\partial h_1} &= \omega_{31}, & \frac{\partial f_3}{\partial h_2} &= \omega_{32} \end{aligned}$$

Calculo matricial en capas lineales

Caso para $\mathbf{h}$

- Supongamos que la capa 3 recibe 2 entradas y produce 3. Dimensiones:

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

$$\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} = \begin{bmatrix} \frac{\partial f_1}{\partial h_1} & \frac{\partial f_2}{\partial h_1} & \frac{\partial f_3}{\partial h_1} \\ \frac{\partial f_1}{\partial h_2} & \frac{\partial f_2}{\partial h_2} & \frac{\partial f_3}{\partial h_2} \end{bmatrix} = \begin{bmatrix} \omega_{11} & \omega_{21} & \omega_{31} \\ \omega_{12} & \omega_{22} & \omega_{32} \end{bmatrix}$$

$$\boldsymbol{\Omega}_3 = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \\ \omega_{31} & \omega_{32} \end{bmatrix} \quad (3 \times 2), \quad \boldsymbol{\Omega}_3^T = \begin{bmatrix} \omega_{11} & \omega_{21} & \omega_{31} \\ \omega_{12} & \omega_{22} & \omega_{32} \end{bmatrix} \quad (2 \times 3)$$

Calculo matricial en capas lineales

Caso para **h**

Derivadas escalares:

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

Derivadas de la matriz:

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

Tareas para casa:

- Considera la función:

$$\mathbf{f} = \mathbf{B}\mathbf{a}$$

- Se puede escribir como:

$$f_i = \sum_j B_{ij} a_j$$

- Ahora calcula:

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

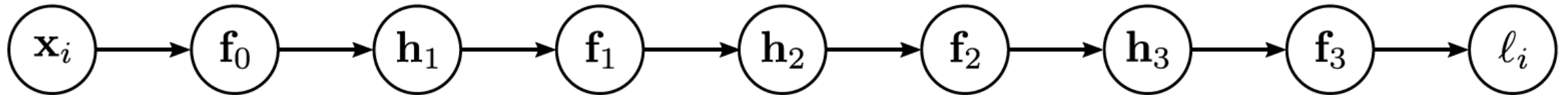
- Escribe la expresión final como una matriz

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

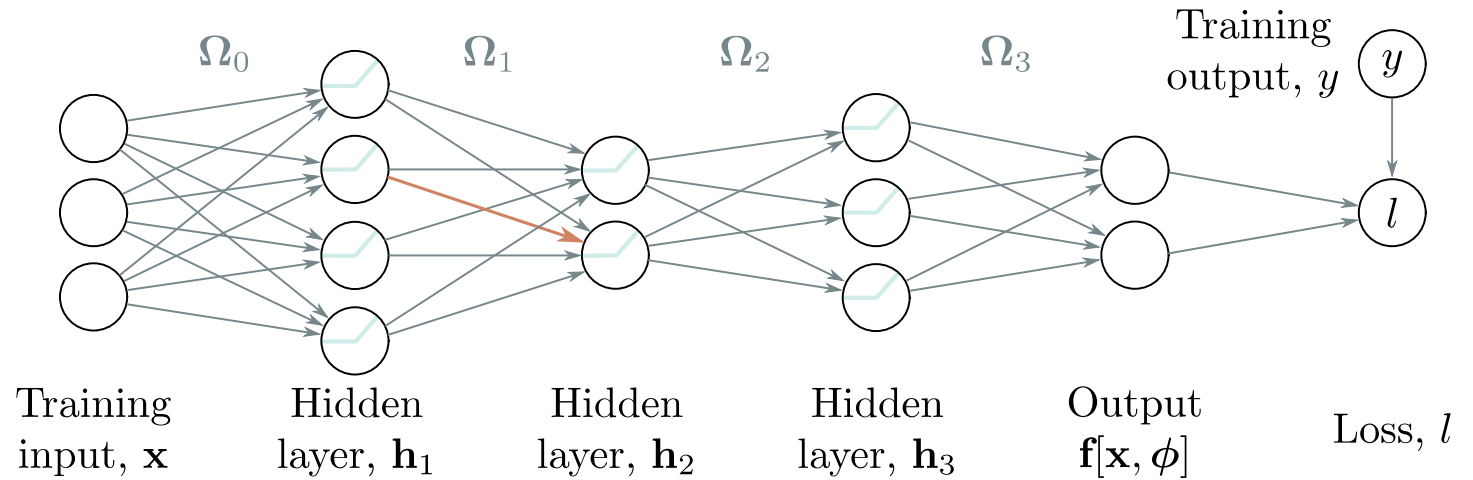
Gradientes

- La intuición de la retropropagación
- Modelo de juguete
- Matemáticas de base
- Paso adelante de retropropagación
- Paso hacia atrás por retropropagación
- Diferenciación algorítmica
- Código

Objetivo pase hacia adelante



El pase hacia delante



1. Escribir esto como una serie de cálculos intermedios

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

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

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

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

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

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

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

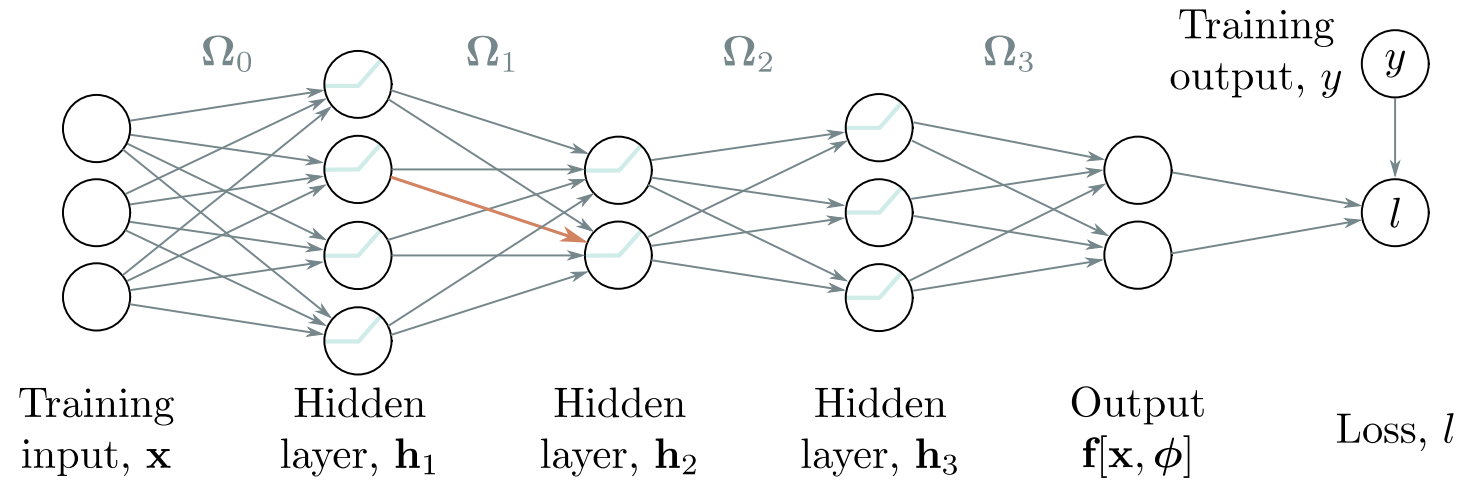
$$\ell_i = l[\mathbf{f}_3, y_i]$$

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

$$\mathbf{h}_k = \mathbf{a}[\mathbf{f}_{k-1}] \quad k \in \{1, 2, \dots, K\}$$

$$\mathbf{f}_k = \beta_k + \Omega_k \mathbf{h}_k. \quad k \in \{1, 2, \dots, K\}$$

El pase hacia delante



1. Escribir esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

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

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

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

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

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

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

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

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

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

$$\mathbf{h}_k = \mathbf{a}[\mathbf{f}_{k-1}]$$

$$\mathbf{f}_k = \beta_k + \Omega_k \mathbf{h}_k.$$

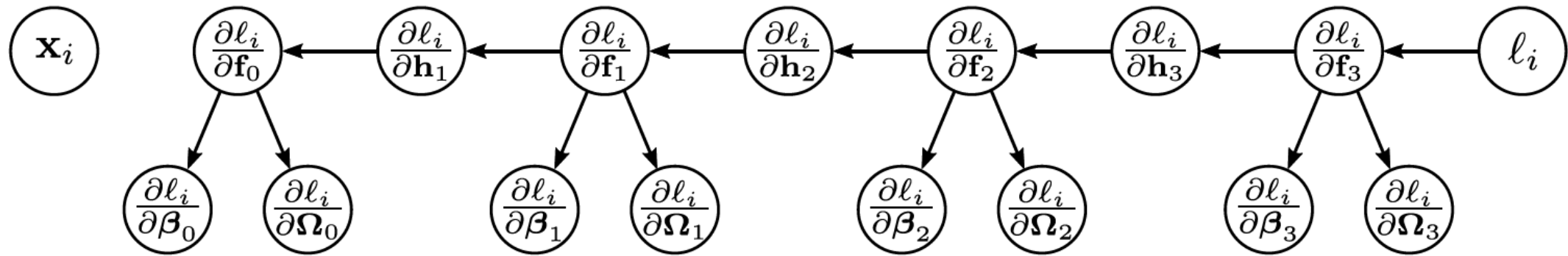
$$k \in \{1, 2, \dots, K\}$$

$$k \in \{1, 2, \dots, K\}$$

Gradientes

- La intuición de la retropropagación
- Modelo de juguete
- Matemáticas de fondo
- Paso adelante de retropropagación
- Paso hacia atrás por retropropagación
- Diferenciación algorítmica
- Código

Objetivo pase hacia atrás



El pase hacia atrás

Objetivos

$$\frac{\partial \ell_i}{\partial \mathbf{f}_k} \quad \frac{\partial \mathbf{f}_k}{\partial \mathbf{h}_k} \quad \frac{\partial \mathbf{h}_k}{\partial \mathbf{f}_{k-1}} \quad \frac{\partial \mathbf{f}_k}{\partial \boldsymbol{\beta}_k} \quad \frac{\partial \mathbf{f}_k}{\partial \boldsymbol{\Omega}_k}$$

1. Escribe esto como una serie de cálculos intermedios

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

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

2. Calcula estas cantidades intermedias

$$\mathbf{f}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1 \mathbf{h}_1 \quad \frac{\partial \ell_i}{\partial \mathbf{f}_2} = \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3}$$

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

$$\mathbf{f}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2 \mathbf{h}_2 \quad \frac{\partial \ell_i}{\partial \mathbf{f}_1} = \frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \left(\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$

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

3. Calcular derivadas de la función de pérdida con respecto a las cantidades intermedias

$$\mathbf{f}_3 = \boldsymbol{\beta}_3 + \boldsymbol{\Omega}_3 \mathbf{h}_3 \quad \frac{\partial \ell_i}{\partial \mathbf{f}_0} = \frac{\partial \mathbf{h}_1}{\partial \mathbf{f}_0} \frac{\partial \mathbf{f}_1}{\partial \mathbf{h}_1} \left(\frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3} \right)$$

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

El pase hacia atrás

Objetivos

$$\frac{\partial \ell_i}{\partial \mathbf{f}_k} \quad \frac{\partial \mathbf{f}_k}{\partial \mathbf{h}_k} \quad \frac{\partial \mathbf{h}_k}{\partial \mathbf{f}_{k-1}} \quad \frac{\partial \mathbf{f}_k}{\partial \beta_k} \quad \frac{\partial \mathbf{f}_k}{\partial \Omega_k}$$

1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

3. Calcular derivadas de la función de pérdida con respecto a las cantidades intermedias

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

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

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

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

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

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

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

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

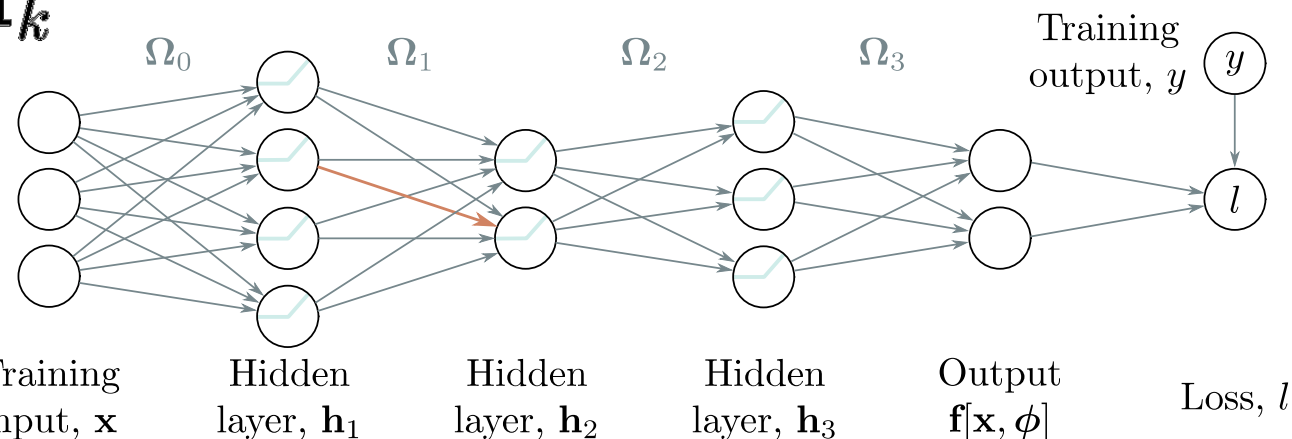
$$\frac{\partial \ell_i}{\partial \mathbf{f}_2} = \frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3}$$

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

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

El pase hacia atrás

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



1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

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

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

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

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

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

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

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

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

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

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

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

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

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

El pase hacia atrás $\frac{\partial \ell_i}{\partial \mathbf{f}_k}$
Regresión

$$\ell_i = \frac{1}{2} \|\mathbf{f}_3 - \mathbf{y}_i\|^2 = \frac{1}{2} \sum_j (f_{3,j} - y_{i,j})^2 \quad \frac{\partial \ell_i}{\partial f_{3,j}} = f_{3,j} - y_{i,j}$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_3} = \mathbf{f}_3 - \mathbf{y}_i \in \mathbb{R}^{D_{f_3} \times 1}$$

El pase hacia atrás $\frac{\partial \ell_i}{\partial \mathbf{f}_k}$

Clasificación

Cross entropy con soft max

$$\text{softmax} [f_3]_j = \frac{\exp (f_{3,j})}{\sum_{k=1}^{D_{f_3}} \exp (f_{3,k})}, \quad j = 1, \dots, D_{f_3}$$

$$\mathbf{p}_i = \text{softmax} [\mathbf{f}_3] \in \mathbb{R}^{D_{f_3} \times 1}, \quad \sum_j p_{i,j} = 1$$

$$\ell_i = 1 [\mathbf{f}_3, y_i] = -\log \left(\text{softmax} [\mathbf{f}_3]_{y_i} \right) = -\log p_{i,y_i} = -f_{3,y_i} + \log \sum_k \exp (f_{3,k})$$

$$\frac{\partial \ell_i}{\partial \mathbf{f}_3} = \text{softmax} [\mathbf{f}_3] - \mathbf{1}_{y_i} = \mathbf{p}_i - \mathbf{1}_{y_i} \in \mathbb{R}^{D_{f_3} \times 1} \quad \frac{\partial \ell_i}{\partial f_{3,j}} = p_{i,j} - \mathbb{I} [j = y_i]$$

El pase hacia atrás $\frac{\partial \ell_i}{\partial \mathbf{f}_k}$

Clasificación

Cross-entropy binaria con sigmoide $\text{sig}[f_3] = \frac{1}{1 + \exp(-f_3)} \in (0, 1)$

$$\ell_i = -[y_i \log \sigma(f_3) + (1 - y_i) \log (1 - \sigma(f_3))]$$

$$\frac{\partial \ell_i}{\partial f_3} = \sigma(f_3) - y_i \qquad \frac{\partial \ell_i}{\partial f_3} = \text{sig}[f_3] - y_i = p_i - y_i \in \mathbb{R}$$

El pase hacia atrás

Objetivos

$$\frac{\partial \ell_i}{\partial \mathbf{f}_k} \quad \frac{\partial \mathbf{f}_k}{\partial \mathbf{h}_k} \quad \frac{\partial \mathbf{h}_k}{\partial \mathbf{f}_{k-1}} \quad \frac{\partial \mathbf{f}_k}{\partial \beta_k} \quad \frac{\partial \mathbf{f}_k}{\partial \Omega_k}$$

1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

3. Calcular derivadas de la función de pérdida con respecto a las cantidades intermedias

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

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

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

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

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

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

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

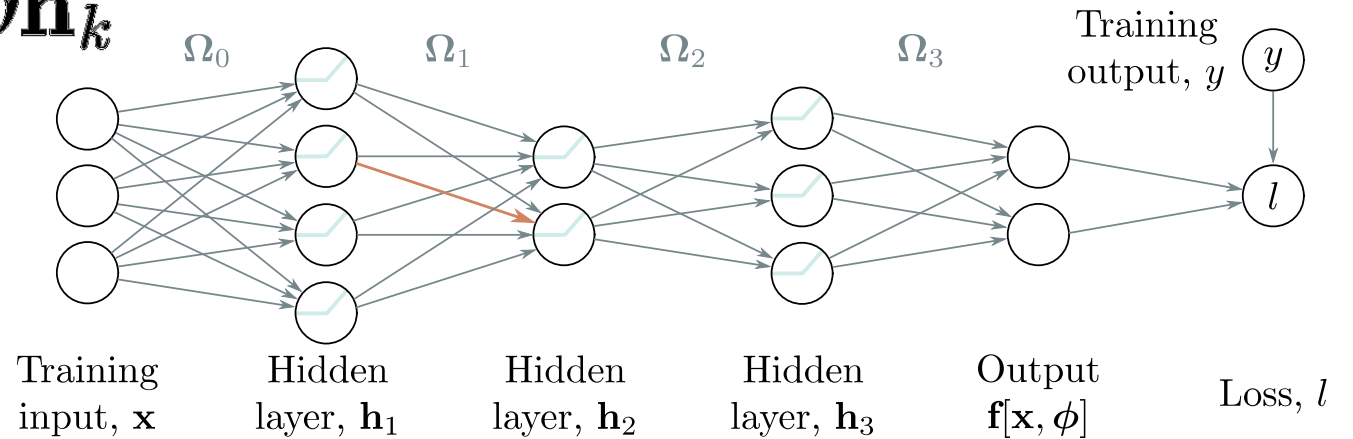
$$\ell_i = l[\mathbf{f}_3, y_i]$$

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

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

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

El pase hacia atrás $\frac{\partial \mathbf{f}_k}{\partial \mathbf{h}_k}$



1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

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

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

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

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

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

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

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

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

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

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

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

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

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

Tenemos parte del camino hecho $\frac{\partial \mathbf{f}_k}{\partial \mathbf{h}_k}$

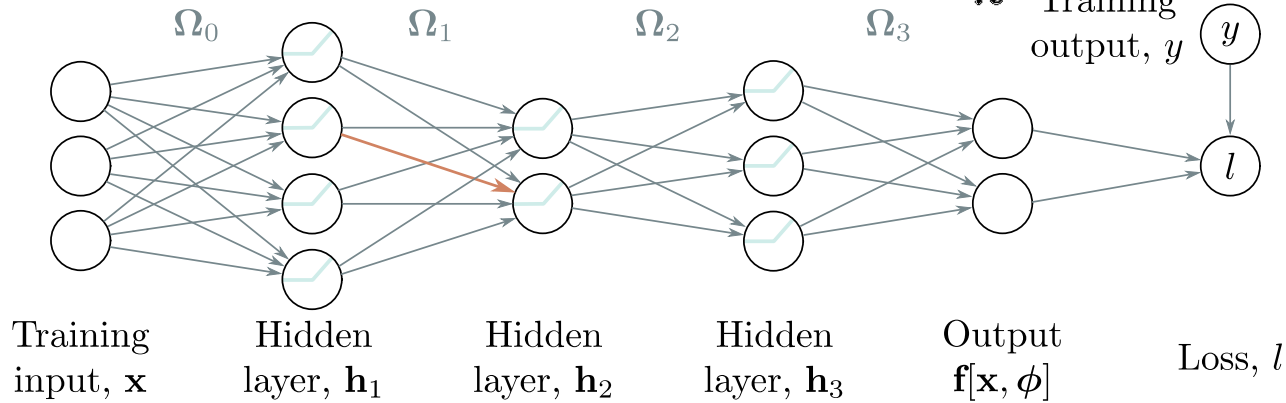
- Porque:

$$\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} = \frac{\partial}{\partial \mathbf{h}_3} (\beta_3 + \mathbf{\Omega}_3 \mathbf{h}_3) = \mathbf{\Omega}_3^T$$

- Recordemos similar a:

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

El pase hacia atrás $\frac{\partial \mathbf{f}_k}{\partial \mathbf{h}_k}$



1. Escribe esto como una serie de cálculos intermedios

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

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

2. Calcula estas cantidades intermedias

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

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

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

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

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

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

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

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

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

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

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

$$\frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} = \frac{\partial}{\partial \mathbf{h}_3} (\beta_3 + \Omega_3 \mathbf{h}_3) = \Omega_3^T$$

El pase hacia atrás

Objetivos

$$\frac{\partial \ell_i}{\partial \mathbf{f}_k} \quad \frac{\partial \mathbf{f}_k}{\partial \mathbf{h}_k} \quad \frac{\partial \mathbf{h}_k}{\partial \mathbf{f}_{k-1}} \quad \frac{\partial \mathbf{f}_k}{\partial \beta_k} \quad \frac{\partial \mathbf{f}_k}{\partial \Omega_k}$$

1. Escribe esto como una serie de cálculos intermedios

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

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

2. Calcula estas cantidades intermedias

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

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

3. Calcular derivadas de la función de pérdida con respecto a las cantidades intermedias

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

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

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

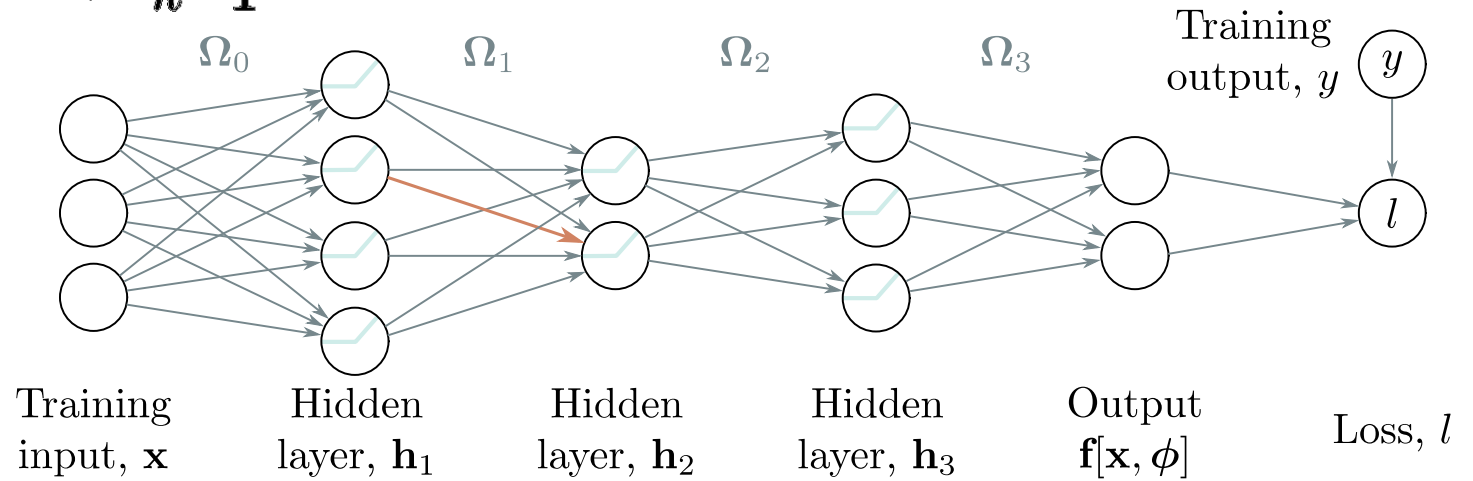
$$\ell_i = l[\mathbf{f}_3, y_i]$$

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

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

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

El pase hacia atrás $\frac{\partial \mathbf{h}_k}{\partial \mathbf{f}_{k-1}}$



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

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

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

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

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

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

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

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

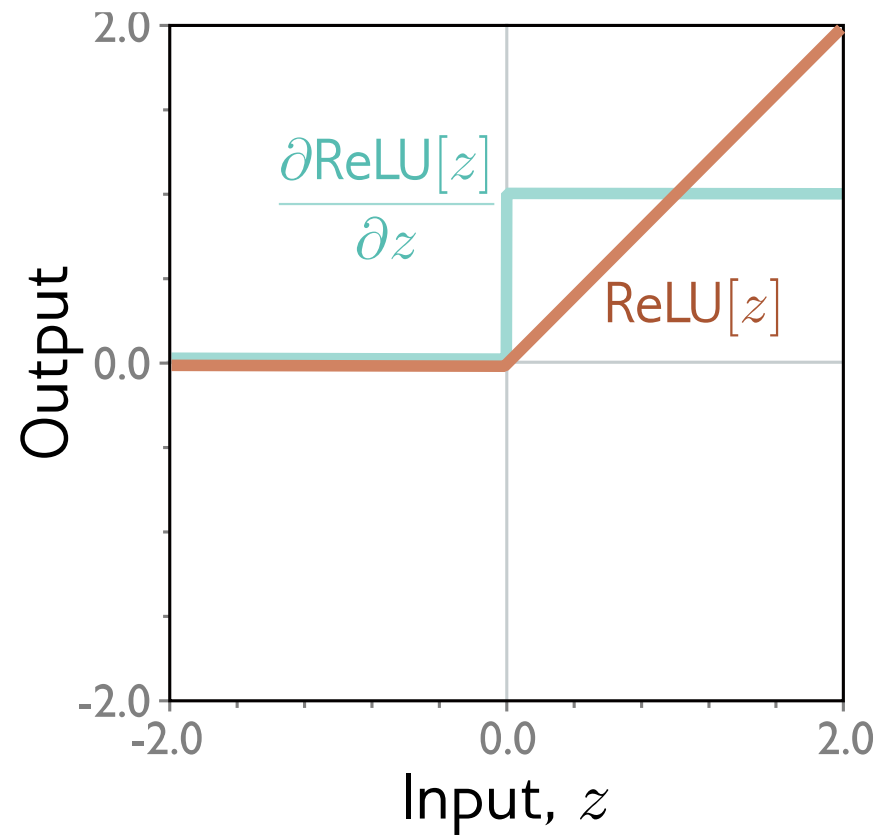
$$\frac{\partial \ell_i}{\partial \mathbf{f}_1} = \frac{\partial \mathbf{h}_2}{\partial \mathbf{f}_1} \frac{\partial \mathbf{f}_2}{\partial \mathbf{h}_2} \left(\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} \frac{\partial \mathbf{f}_3}{\partial \mathbf{h}_3} \frac{\partial \ell_i}{\partial \mathbf{f}_3}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

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

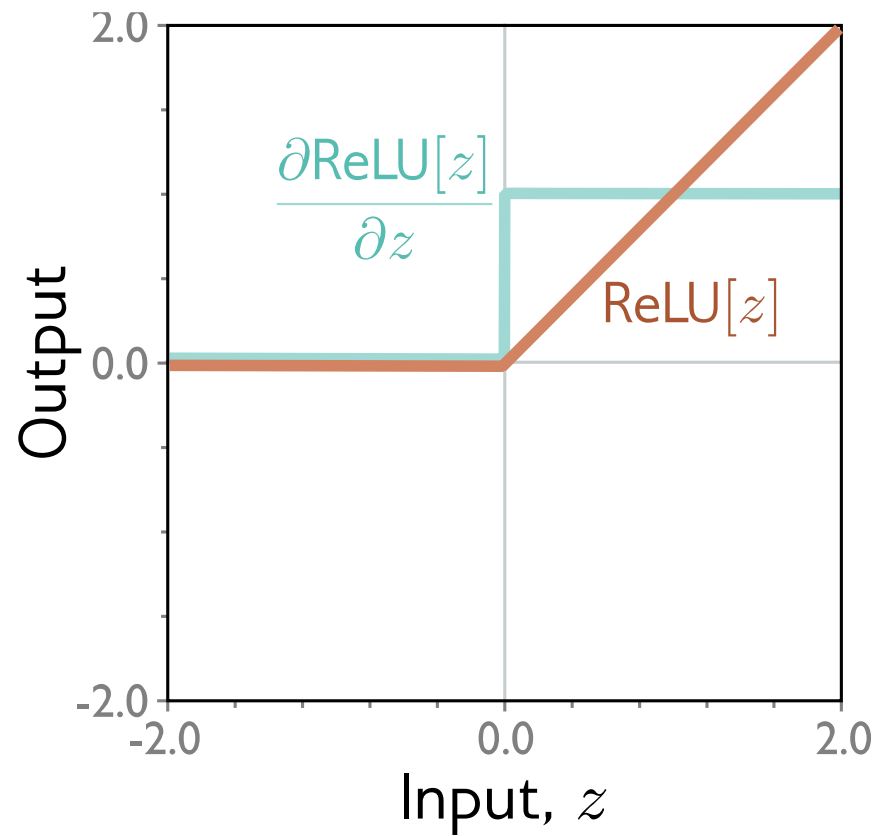
$$\ell_i = l[\mathbf{f}_3, y_i]$$

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

Derivada de ReLU



Derivada de ReLU



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

"Función de indicador"

Derivada de RELU

1. Considera:

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

donde:

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

2. Podríamos escribir equivalentemente:

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

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

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

Derivada de RELU

4. De forma equivalente, podemos multiplicar puntualmente por la diagonal $\mathbb{I}[\mathbf{b} > 0] \odot$

$$\frac{\partial \mathbf{a}}{\partial \mathbf{b}} = \text{diag}(\mathbb{I}[\mathbf{b} > 0]) = \begin{bmatrix} \mathbb{I}[b_1 > 0] & 0 & 0 \\ 0 & \mathbb{I}[b_2 > 0] & 0 \\ 0 & 0 & \mathbb{I}[b_3 > 0] \end{bmatrix} \in \mathbb{R}^{D \times D}$$

$$\mathbb{I}[\mathbf{b} > 0] = \begin{bmatrix} \mathbb{I}[b_1 > 0] \\ \mathbb{I}[b_2 > 0] \\ \mathbb{I}[b_3 > 0] \end{bmatrix} \in \mathbb{R}^{D \times 1}$$

Derivada de RELU

4. De forma equivalente, podemos multiplicar puntualmente por la diagonal $\mathbb{I}[\mathbf{b} > 0] \odot$

Para cualquier matriz diagonal $\mathbf{D} = \text{diag}(\mathbf{d})$ y vector $\mathbf{v}$ tenemos

$$\mathbf{D}\mathbf{v} = \mathbf{d} \odot \mathbf{v}$$

Por qué esto es relevante?

$$\begin{bmatrix} \mathbb{I}[b_1 > 0] & 0 & 0 \\ 0 & \mathbb{I}[b_2 > 0] & 0 \\ 0 & 0 & \mathbb{I}[b_3 > 0] \end{bmatrix}$$

Derivada de RELU

4. De forma equivalente, podemos multiplicar puntualmente por la diagonal $\mathbb{I}[\mathbf{b} > 0] \odot$

Para cualquier matriz diagonal $\mathbf{D} = \text{diag}(\mathbf{d})$ y vector $\mathbf{v}$ tenemos

$$\mathbf{D}\mathbf{v} = \mathbf{d} \odot \mathbf{v}$$

Por qué esto es relevante?

Una matriz diagonal $\mathbf{D} = \text{diag}(\mathbf{d})$ “cuesta” $O(D^2)$ en almacenamiento

La operación $\mathbf{D}\mathbf{v}$ “cuesta” $\sim O(D^2)$ operaciones de CPU/GPU.

Derivada de RELU

4. De forma equivalente, podemos multiplicar puntualmente por la diagonal $\mathbb{I}[\mathbf{b} > 0] \odot$

Para cualquier matriz diagonal $\mathbf{D} = \text{diag}(\mathbf{d})$ y vector $\mathbf{v}$ tenemos

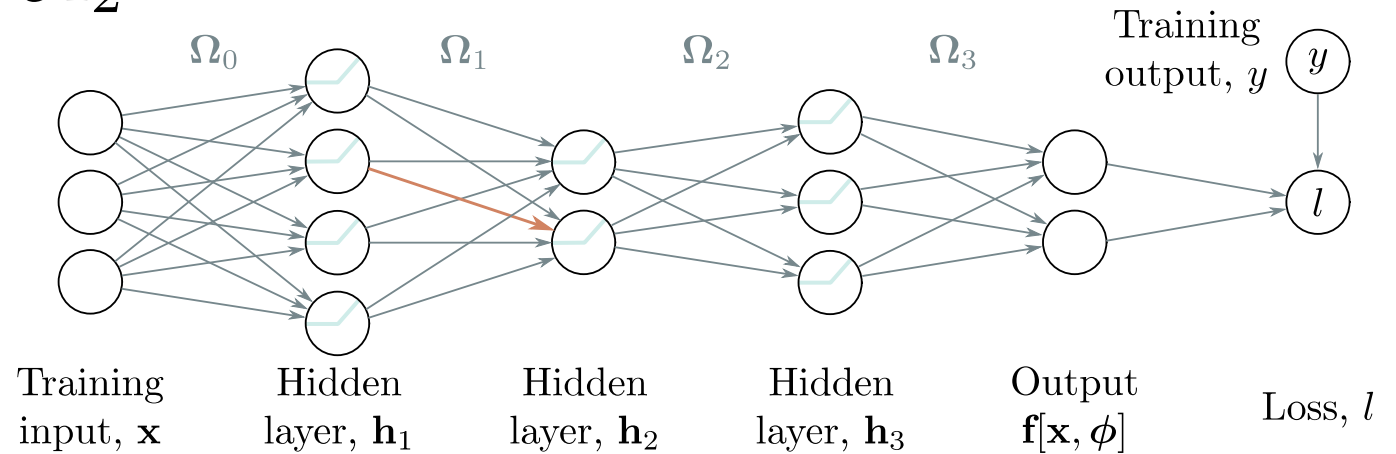
$$\mathbf{D}\mathbf{v} = \mathbf{d} \odot \mathbf{v}$$

Por qué esto es relevante?

Una vector $\mathbf{d}$ “cuesta” $O(D)$ en almacenamiento

La operación $\mathbf{d} \odot \mathbf{v}$ “cuesta” $\sim O(D)$ operaciones de CPU/GPU.

El pase hacia atrás $\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2}$



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

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

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

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

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

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

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

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

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

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

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

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

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

$$\frac{\partial \mathbf{h}_3}{\partial \mathbf{f}_2} = \mathbb{I}[\mathbf{f}_2 > 0]$$

El pase hacia atrás

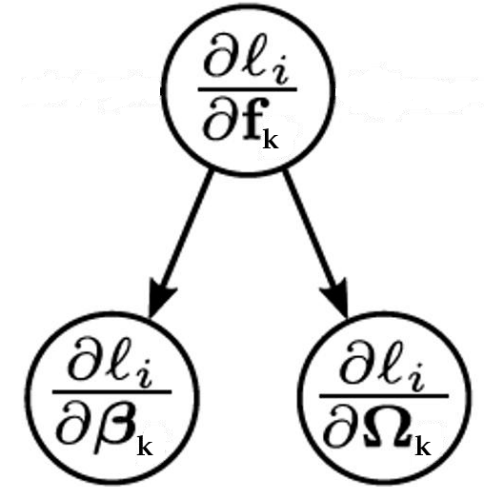
Objetivos

1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

3. Calcular derivadas de la función de pérdida con respecto a las cantidades intermedias

$$\frac{\partial \ell_i}{\partial \mathbf{f}_k} \quad \frac{\partial \mathbf{f}_k}{\partial \mathbf{h}_k} \quad \frac{\partial \mathbf{h}_k}{\partial \mathbf{f}_{k-1}} \quad \frac{\partial \mathbf{f}_k}{\partial \boldsymbol{\beta}_k} \quad \frac{\partial \mathbf{f}_k}{\partial \boldsymbol{\Omega}_k}$$



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

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

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

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

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

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

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

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

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

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

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

El pase hacia atrás $\frac{\partial \ell_i}{\partial \beta_k}$

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

4. Tomar derivadas respecto a parámetros

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

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

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

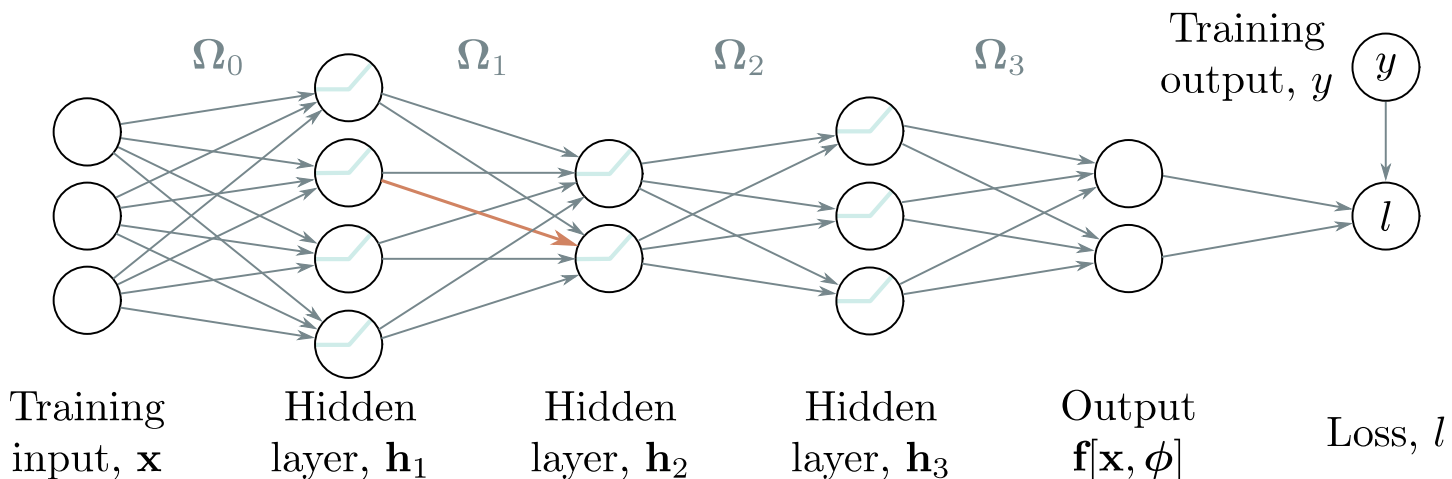
$$\mathbf{h}_2 = \mathbf{a}[\mathbf{f}_1]$$

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

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

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

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



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

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

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

El pase hacia atrás $\frac{\partial \ell_i}{\partial \Omega_k}$

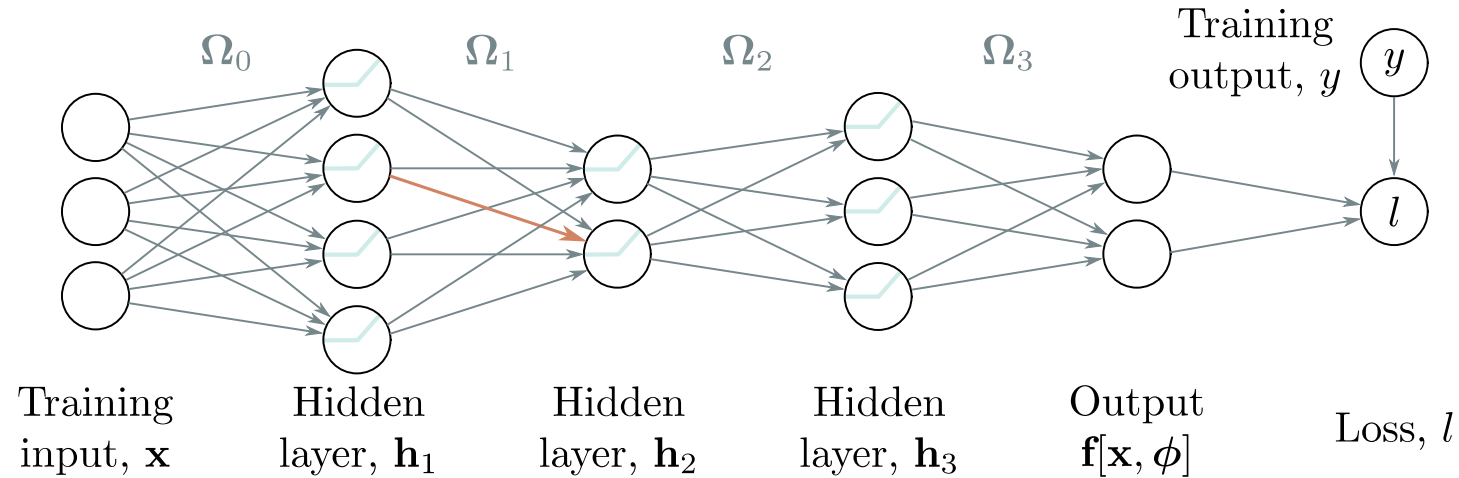
1. Escribe esto como una serie de cálculos intermedios

2. Calcula estas cantidades intermedias

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

4. Tomar derivadas respecto a parámetros

$$\begin{aligned} \mathbf{f}_0 &= \beta_0 + \Omega_0 \mathbf{x}_i \\ \mathbf{h}_1 &= \mathbf{a}[\mathbf{f}_0] \\ \mathbf{f}_1 &= \beta_1 + \Omega_1 \mathbf{h}_1 \\ \mathbf{h}_2 &= \mathbf{a}[\mathbf{f}_1] \\ \mathbf{f}_2 &= \beta_2 + \Omega_2 \mathbf{h}_2 \\ \mathbf{h}_3 &= \mathbf{a}[\mathbf{f}_2] \\ \mathbf{f}_3 &= \beta_3 + \Omega_3 \mathbf{h}_3 \\ \ell_i &= l[\mathbf{f}_3, y_i] \end{aligned}$$



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

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:

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

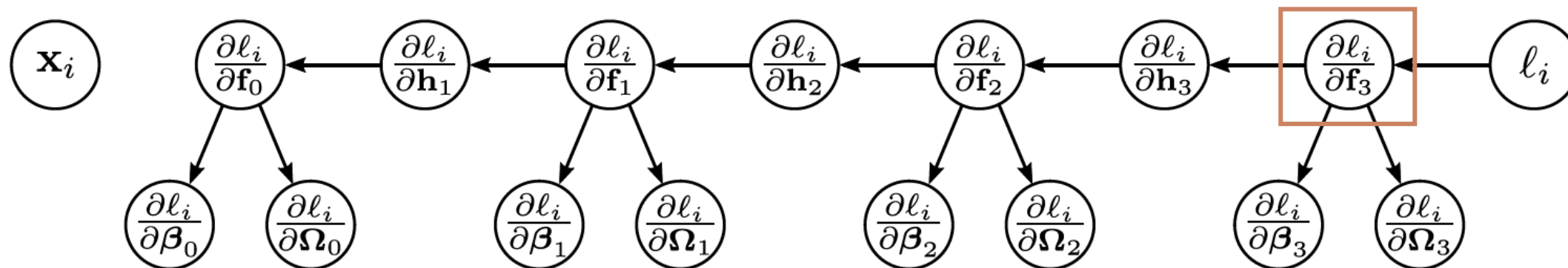
where $\odot$ denotes pointwise multiplication and $\mathbb{I}[\mathbf{f}_{k-1} > 0]$ is a vector containing ones where $\mathbf{f}_{k-1}$ is greater than zero and zeros elsewhere.

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:

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

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.

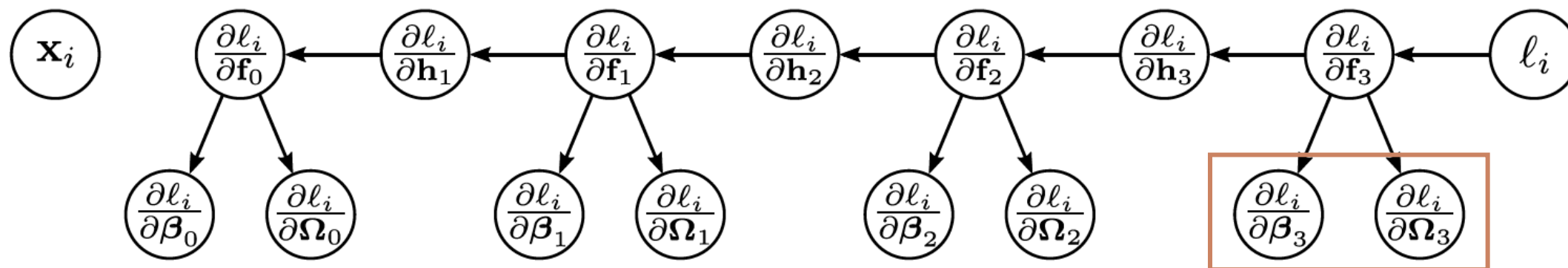


Resumen del backprop

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:

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

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.

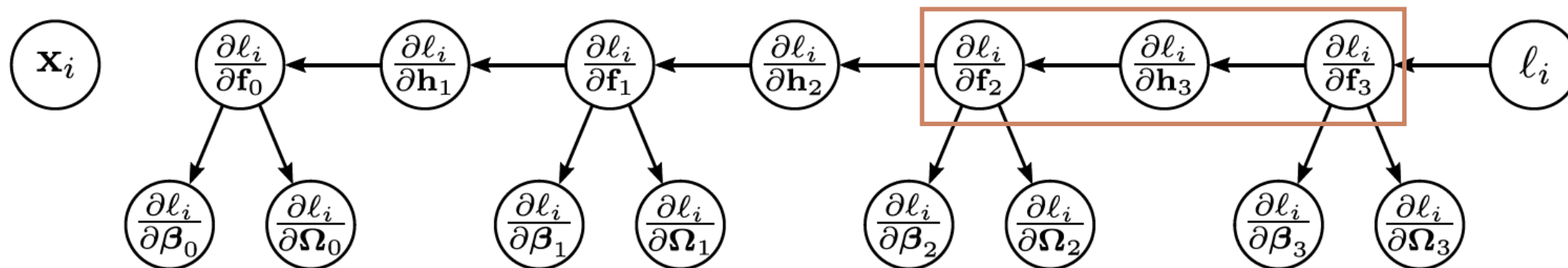


Resumen del backprop

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:

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

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:

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

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.

$$\underbrace{\text{diag}(\mathbb{I}[\mathbf{f}_{k-1} > 0])}_{D_{k-1} \times D_{k-1}} \cdot \underbrace{\left(\boldsymbol{\Omega}_k^T \frac{\partial\ell_i}{\partial\mathbf{f}_k} \right)}_{D_{k-1} \times 1} = \underbrace{\mathbb{I}[\mathbf{f}_{k-1} > 0]}_{D_{k-1} \times 1} \odot \underbrace{\left(\boldsymbol{\Omega}_k^T \frac{\partial\ell_i}{\partial\mathbf{f}_k} \right)}_{D_{k-1} \times 1}$$

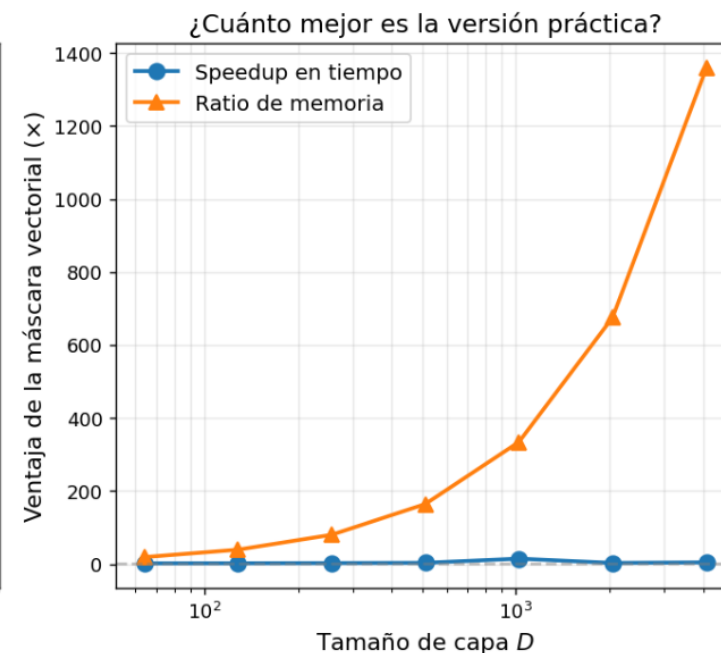
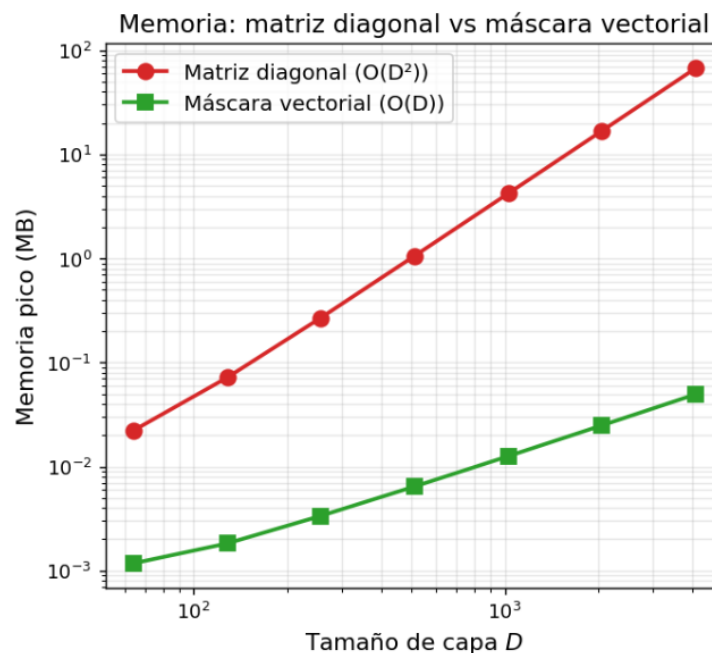
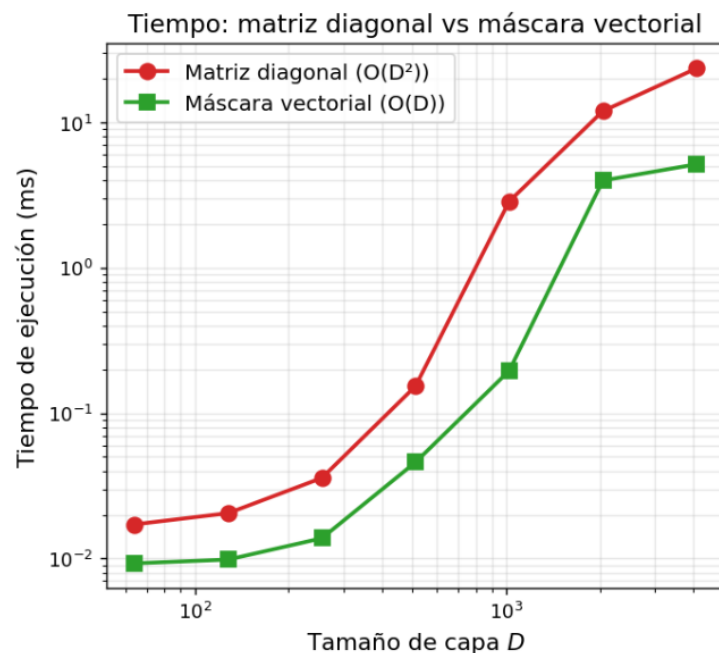
Derivada de RELU

Por qué esto es relevante?

$$\underbrace{\text{diag}(\mathbb{I}[\mathbf{f}_{k-1} > 0])}_{D_{k-1} \times D_{k-1}} \cdot \underbrace{\left(\Omega_k^T \frac{\partial \ell_i}{\partial \mathbf{f}_k}\right)}_{D_{k-1} \times 1} = \underbrace{\mathbb{I}[\mathbf{f}_{k-1} > 0]}_{D_{k-1} \times 1} \odot \underbrace{\left(\Omega_k^T \frac{\partial \ell_i}{\partial \mathbf{f}_k}\right)}_{D_{k-1} \times 1}$$

Una vector $\mathbf{d}$ “cuesta” $O(D)$ en almacenamiento

La operación $\mathbf{d} \odot \mathbf{v}$ “cuesta” $\sim O(D)$ operaciones de CPU/GPU.



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:

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

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 umbralización
- Uso exhaustivo 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 retropropagación
- Modelo de juguete
- Matemáticas de fondo
- Paso adelante de retropropagación
- Paso hacia atrás por retropropagación
- 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. Sequential
 - Puede calcular la cadena de derivadas
- Funciona con ramas siempre que siga siendo un grafo acíclico

Gradientes

- La intuición de la retropropagación
- Modelo de juguete
- Matemáticas de fondo
- Paso adelante de retropropagación
- Paso hacia atrás por retropropagación
- 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()
```



Comentarios