

## Redes profundas

25 de febrero de 2023



# Problema

- 1 Problema
- 2 Soluciones candidatas
- 3 Nota final

# Definición

## Definición (Desvanecimiento del gradiente (*Vanishing gradient*))

Para funciones de activación, cuya derivada se acerca a cero, conforme se agregan capas a la red neuronal esta se vuelve más y más difícil de entrenar.

¿Por qué?

$$W_{t+1} = W_t - \alpha \nabla_W J \quad (1)$$

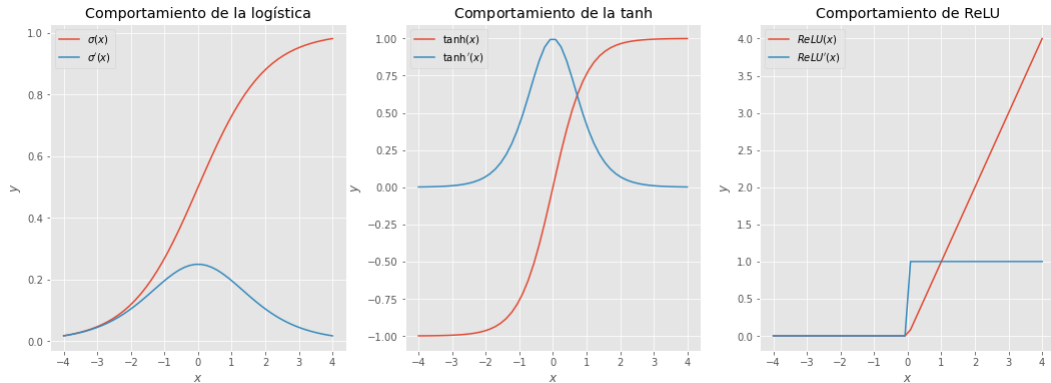
$$\nabla_W J \propto \Delta^{(l)} \quad (2)$$

$$\Delta^{(l-1)} = \Delta^{(l)} \left( W^{(l-1)} \right)^T \circ g'(Z^{(l-1)}) \quad (3)$$

$$\Delta^{(1)} \propto g'(Z^{(0)}) \circ g'(Z^{(1)}) \circ \dots \circ g'(Z^{(L-1)}) \quad (4)$$

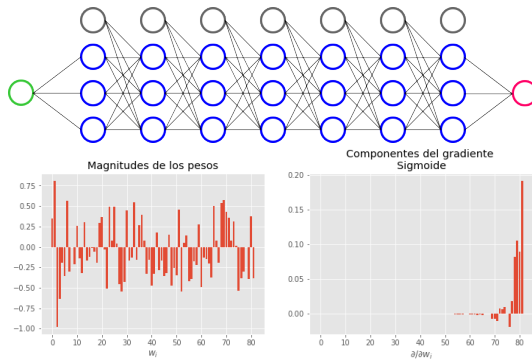
siendo todos estos números  $g'(z) \ll 1$ .

# Funciones de activación y sus derivadas



**Figura:** Todas estas funciones de activación tienen regiones donde su derivada es cero o cercana a cero.

# Componentes del gradiente



**Figura:** **Izquierda:** Magnitudes de los pesos como punto de comparación. **Derecha:** Las primeras componentes del gradiente, a la izquierda, tiene magnitudes mucho más pequeñas que las componentes más cercanas a la salida, a la derecha.

## Soluciones candidatas

- 1 Problema
- 2 Soluciones candidatas
- 3 Nota final

# Soluciones candidatas

- Usar ReLU, dependiendo del problema.
- Normalización por lotes.
- Redes con capas residuales: éstas se conectan directamente a capas más adelante.

$$A^{(l)} = g(A^{(l-1)}W^{(l-1)} + B^{(l-1)} + A^{(l-2)}W^{(l-2,l)})$$

- Valores de activación de capas más atrás tienen un efecto más significativo en capas posteriores.
- La matriz  $W^{(l-2,l)}$  puede entrenarse como las demás o puede ser tener valores fijos, como ser la matriz identidad.

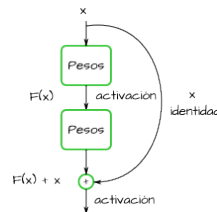


Figura: Salto de capa (*Skip layer*)

## Nota final

- 1 Problema
- 2 Soluciones candidatas
- 3 Nota final




# Explosión del gradiente

- Es un poco menos común, pero la misma fórmula que provoca el devanecimiento puede provocar una explosión si las derivadas tienen componentes con valores grandes:

$$\Delta^{(1)} \propto g'(Z^{(0)}) \circ g'(Z^{(1)}) \circ \dots \circ g'(Z^{(L-1)}) \quad (5)$$

# Referencias I

 *The Vanishing Gradient Problem The Problem, Its Causes, Its Significance, and Its Solutions*, Towards data science 2019,

<https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484>

 *Residual neural network*, [https://en.wikipedia.org/wiki/Residual\\_neural\\_network](https://en.wikipedia.org/wiki/Residual_neural_network)

 *Gentle Introduction to the Adam Optimization Algorithm for Deep Learning*

<https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/>

 *Code Adam Optimization Algorithm From Scratch*

<https://machinelearningmastery.com/adam-optimization-from-scratch/>

# Licencia

Creative Commons  
Atribución-No Comercial-Compartir Igual

