

INF721

2023/2



Aprendizado em Redes Neurais Profundas

A7: Backpropagation

Logística

Avisos

- ▶ Teste T2: Multilayer Perceptron na próxima aula!

Última aula

- ▶ Problemas linearmente não-separáveis
- ▶ Multilayer Perceptron (MLP)
- ▶ Forward Pass
- ▶ Funções de ativação

Plano de Aula

- ▶ Grafo computacional
- ▶ Backpropagation
 - ▶ Gradiente da Regressão Logística
 - ▶ Gradiente da MLP

Grafo Computacional

Um grafo dirigido que descreve as expressões matemáticas de uma RNA passo a passo:

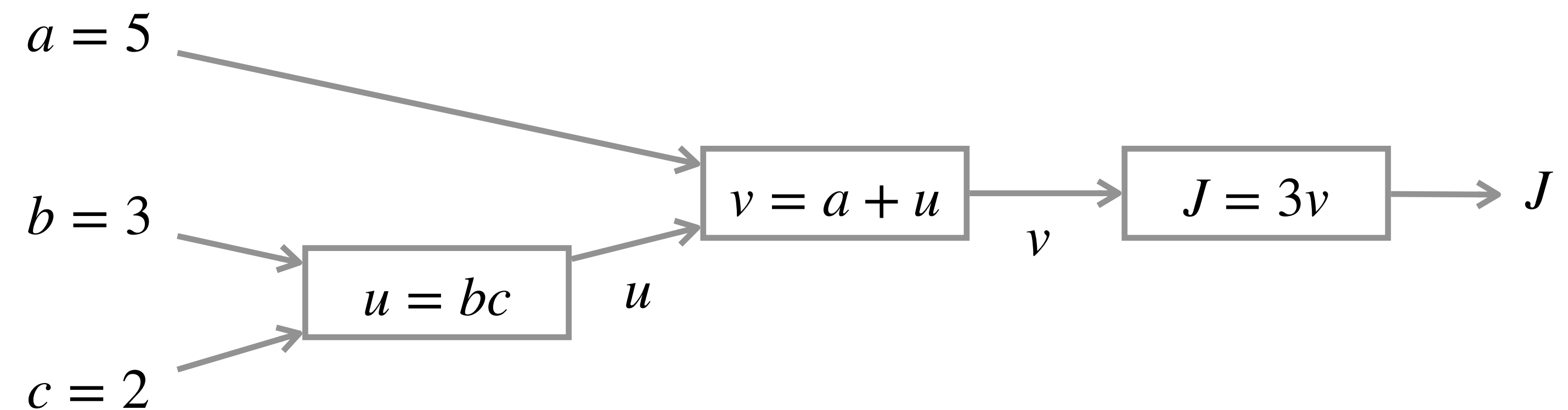
- ▶ Vértices representam operações
- ▶ Arestas representam entrada e saída

$$J(a, b, c) = 3(a + bc)$$

$$u = bc$$

$$v = a + u$$

$$J = 3v$$



Grafo Computacional e RNAs

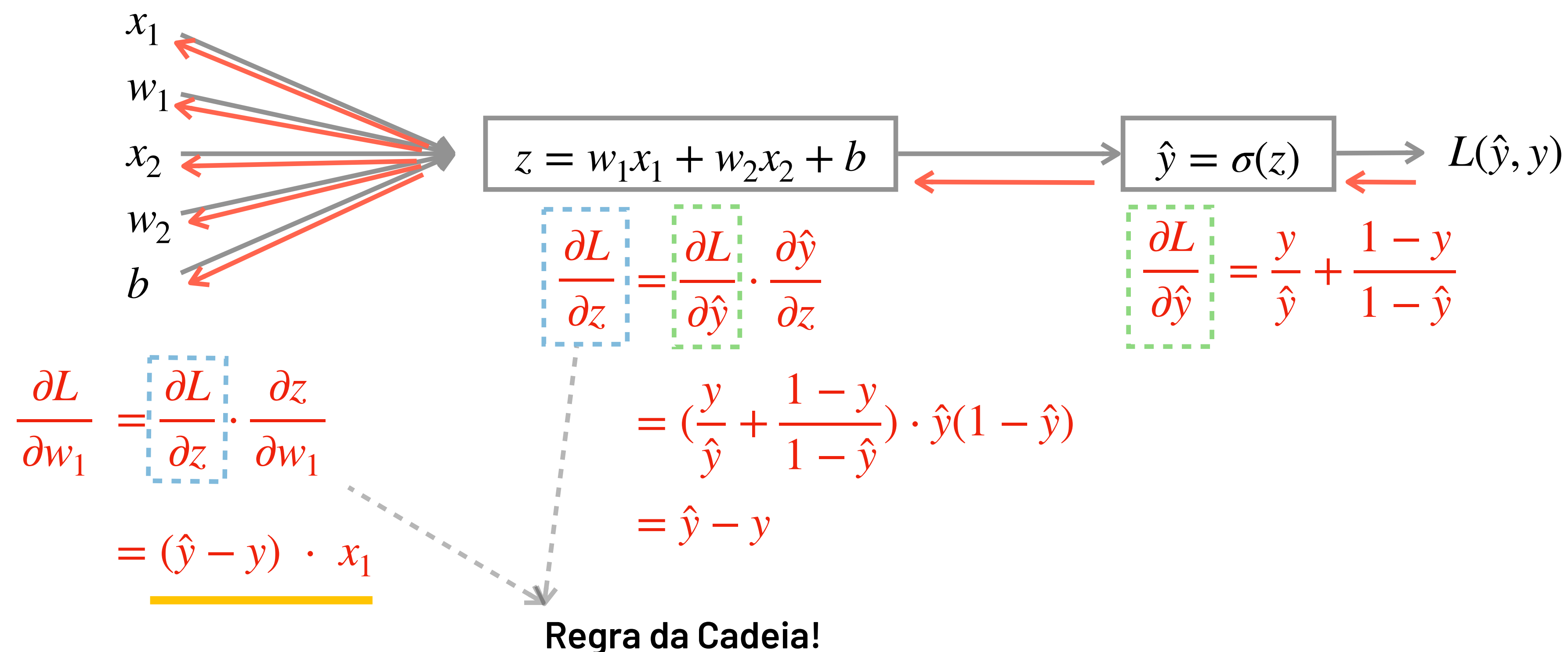
Grafos computacionais nos ajudam a calcular o gradiente de uma função de perda com relação aos pesos de uma RNA

Regressão Logística

$$z = \mathbf{w} \cdot \mathbf{x} + b$$

$$\hat{y} = h(\mathbf{x}) = \frac{1}{1 + e^{-z}}$$

$$L(\hat{y}, y) = -y \log \hat{y} + (1 - y) \log (1 - \hat{y})$$



Retropropagação (Backprop)

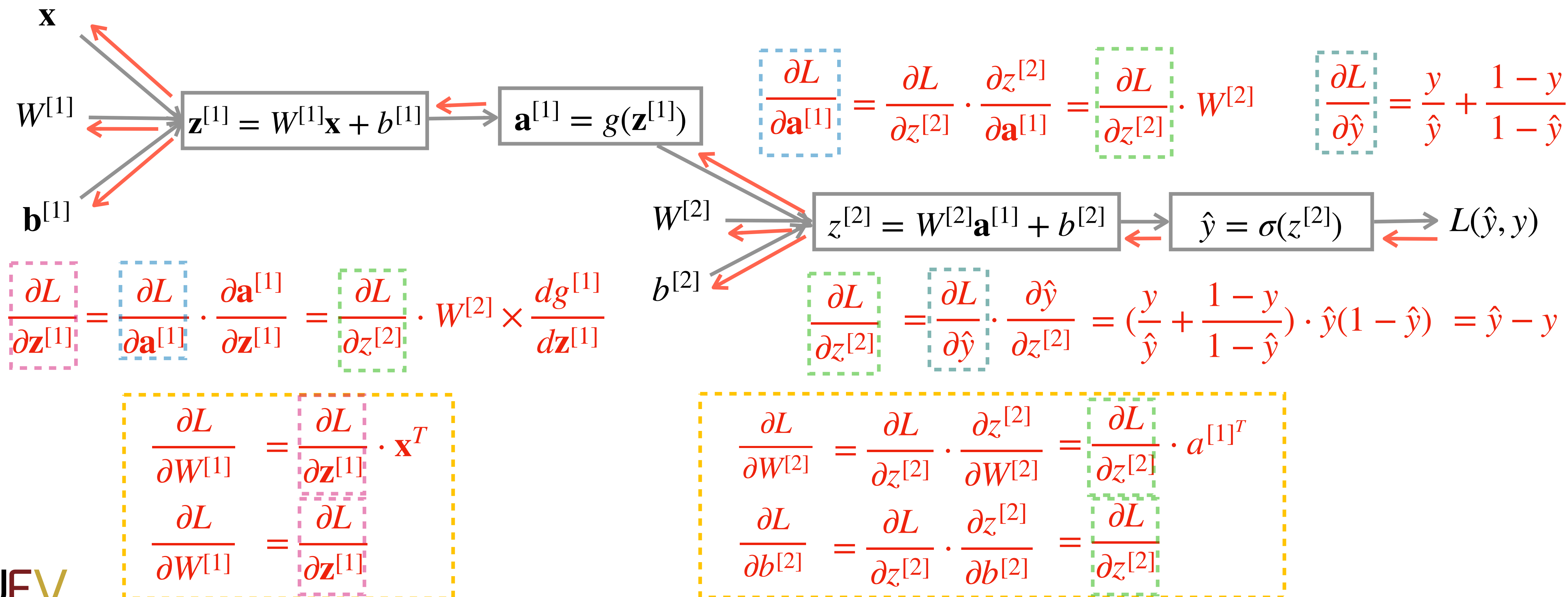
Calcular as derivadas parciais da função de perda com relação aos pesos $W^{[l]}$ e $b^{[l]}$ para todas as camadas l de trás pra frente com a regra da cadeia.

MLP (2 camadas)

$$\mathbf{z}^{[1]} = W^{[1]}\mathbf{x} + \mathbf{b}^{[1]} \quad z^{[2]} = W^{[2]}\mathbf{a}^{[1]} + b^{[2]}$$

$$\mathbf{a}^{[1]} = g^{[1]}(\mathbf{z}^{[1]}) \quad \hat{y} = \sigma(z^{[2]})$$

$$L(\hat{y}, y) = -y \log \hat{y} + (1 - y) \log (1 - \hat{y})$$



Retropropagação (Backprop)

MLP (2 camadas)

$$\mathbf{z}^{[1]} = W^{[1]}\mathbf{x} + \mathbf{b}^{[1]}$$

$$\mathbf{a}^{[1]} = g^{[1]}(\mathbf{z}^{[1]})$$

$$z^{[2]} = W^{[2]}\mathbf{a}^{[1]} + b^{[2]}$$

$$\hat{y} = \sigma(z^{[2]})$$

Função de Perda

$$L(h) = -\frac{1}{n} \sum_{i=1}^n (y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i))$$

Inicialização do backpropagation

Derivada parcial da ativação da camada de saída [2]

$$\frac{\partial L}{\partial \hat{y}} = \frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}$$

$$\frac{\partial L}{\partial z^{[2]}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{[2]}} = \frac{\partial L}{\partial \hat{y}} \cdot \hat{y}(1 - \hat{y}) = \hat{y} - y$$

Derivadas parciais da parte linear da camada de saída [2]

$$\frac{\partial L}{\partial W^{[2]}} = \frac{\partial L}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial W^{[2]}} = \frac{\partial L}{\partial z^{[2]}} \cdot a^{[1]T}$$

$$\frac{\partial L}{\partial b^{[2]}} = \frac{\partial L}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial b^{[2]}} = \frac{\partial L}{\partial z^{[2]}}$$

$$\frac{\partial L}{\partial \mathbf{a}^{[1]}} = \frac{\partial L}{\partial z^{[2]}} \cdot W^{[2]} \quad \text{Derivada parcial da ativação da camada escondida [1]}$$

$$\frac{\partial L}{\partial \mathbf{z}^{[1]}} = \frac{\partial L}{\partial \mathbf{a}^{[1]}} \cdot \frac{\partial \mathbf{a}^{[1]}}{\partial \mathbf{z}^{[1]}} = \frac{\partial L}{\partial \mathbf{a}^{[1]}} \cdot \frac{dg^{[1]}}{dz^{[1]}}$$

Derivadas parciais da parte linear da camada escondida [1]

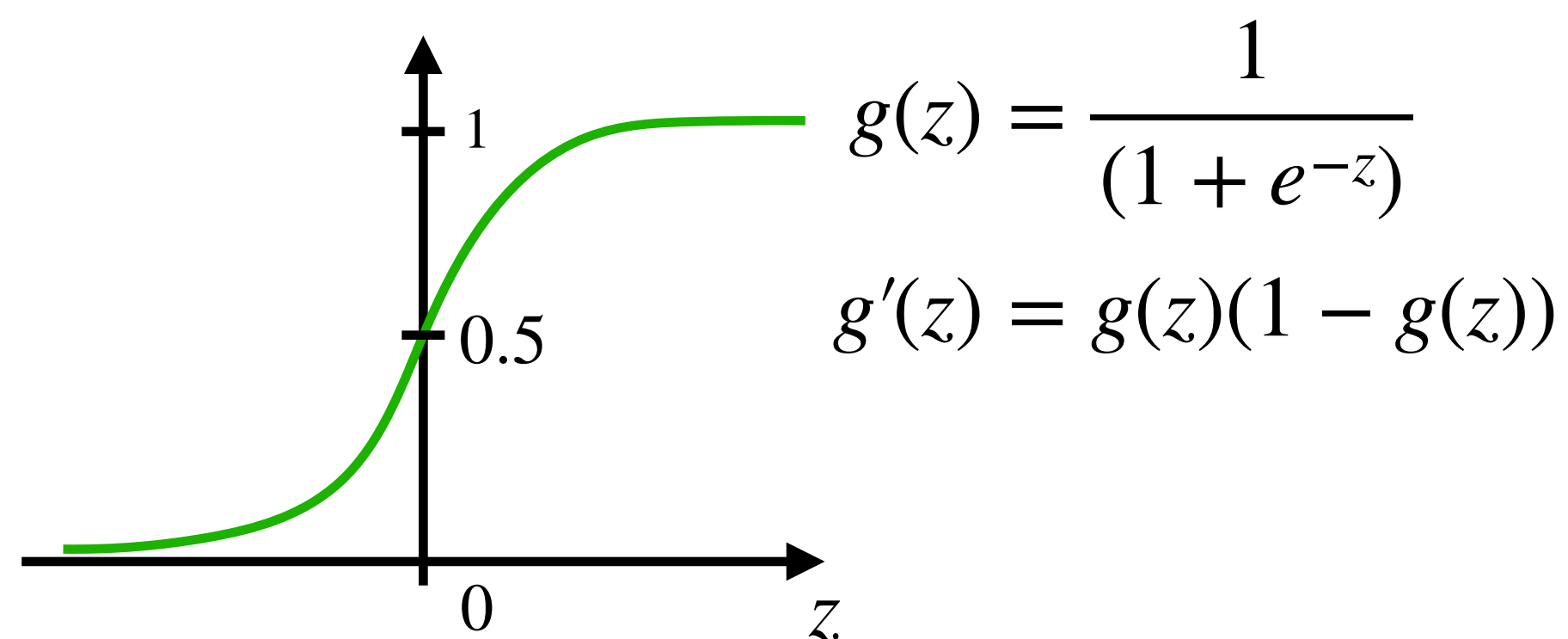
$$\frac{\partial L}{\partial W^{[1]}} = \frac{\partial L}{\partial \mathbf{z}^{[1]}} \cdot \mathbf{x}^T$$

Esse termo depende da escolha de função de ativação

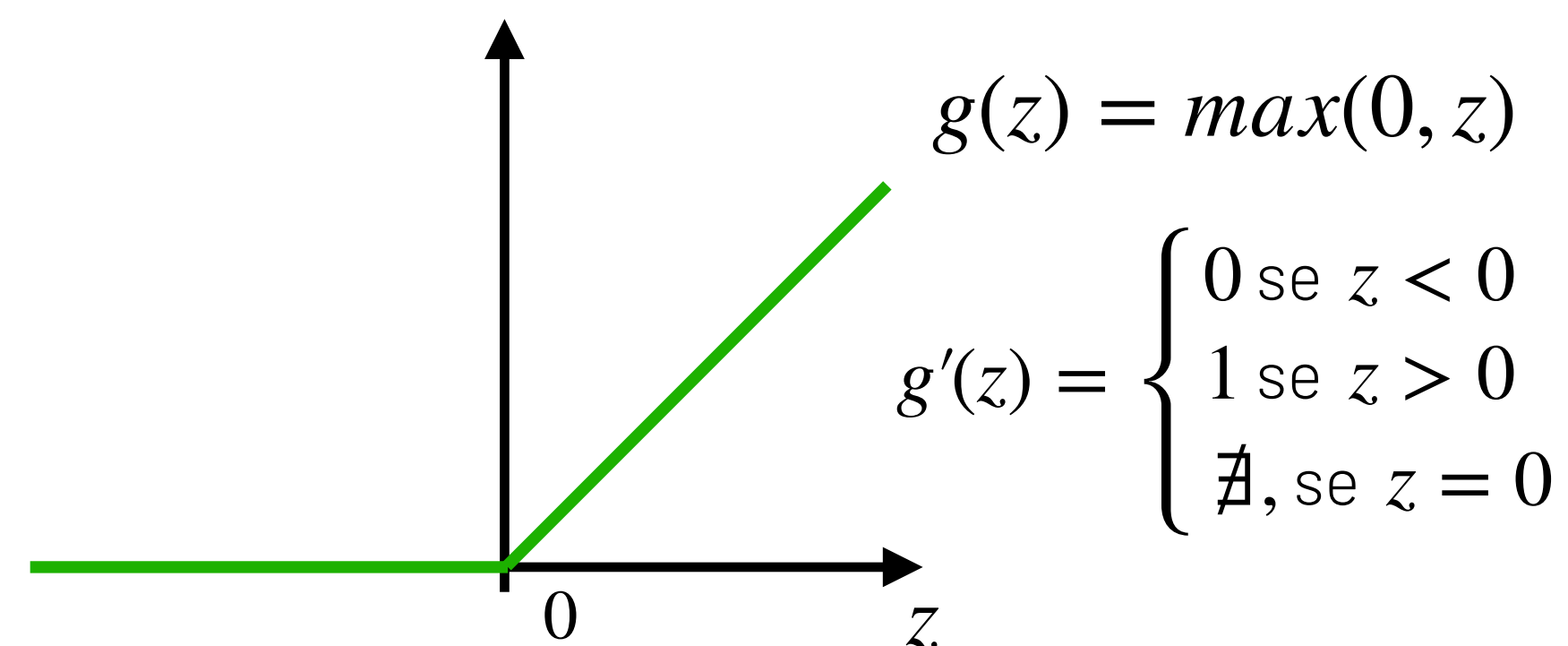
$$\frac{\partial L}{\partial b^{[1]}} = \frac{\partial L}{\partial \mathbf{z}^{[1]}}$$

Derivadas das funções de ativação

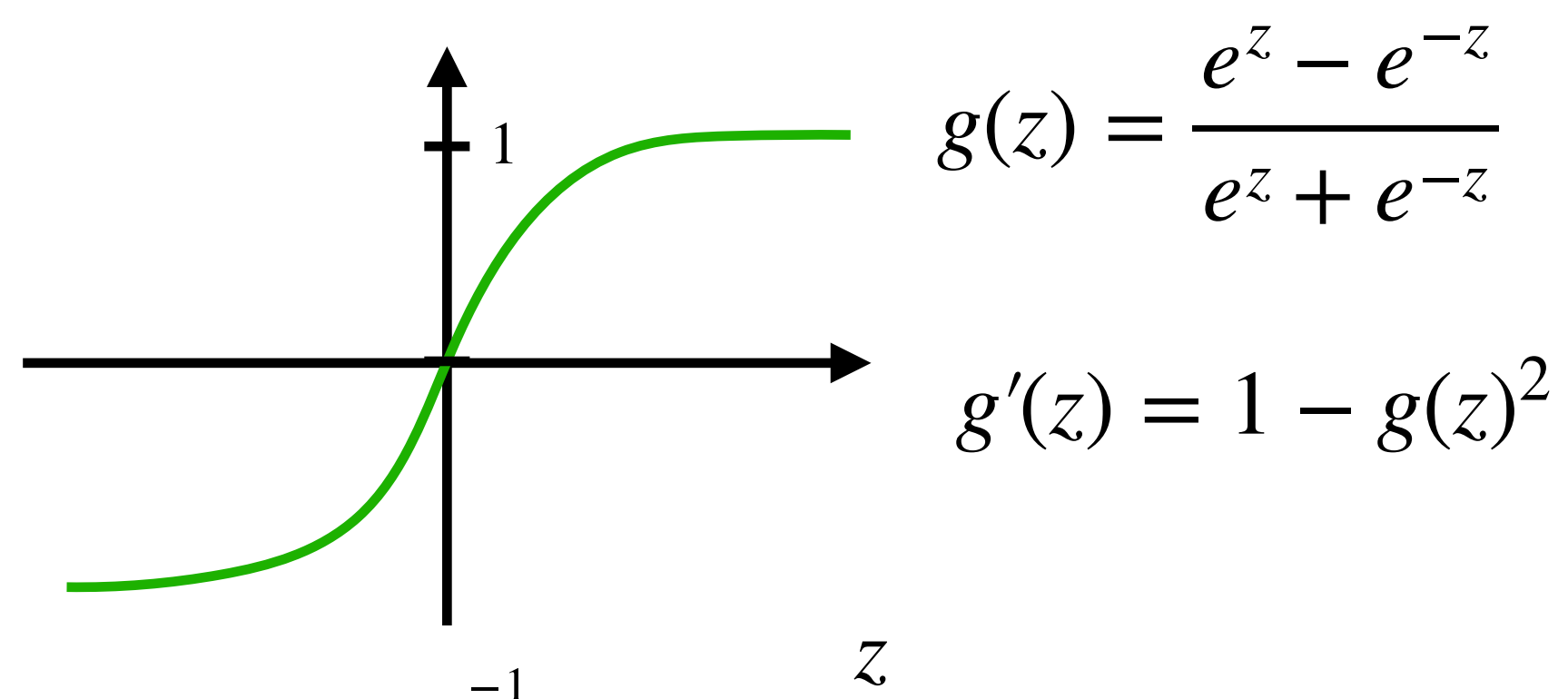
Logística (sigmoide)



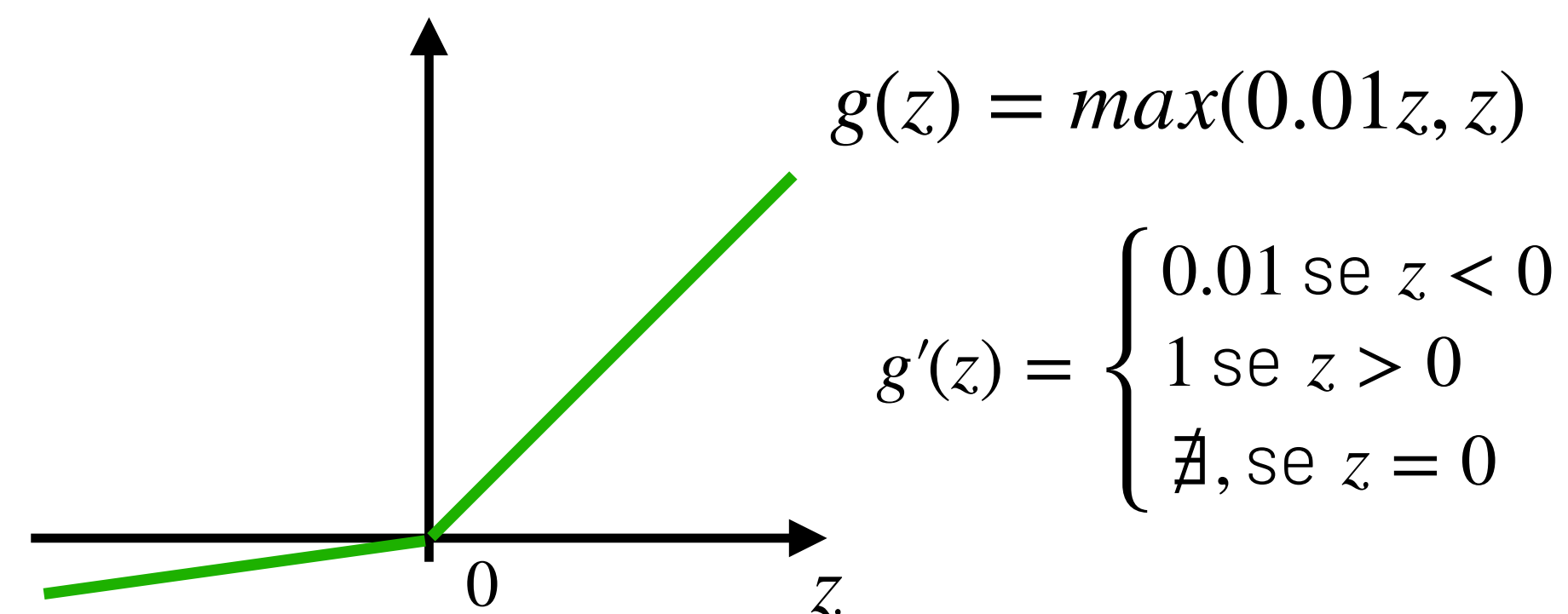
Unidade Linear Retificada (ReLU)



Tangente Hiperbólica



Leaky ReLU



Retropropagação (Backprop)

MLP (L camadas)

$$Z^{[l]} = W^{[l]}A^{[l-1]} + \mathbf{b}^{[l]}$$

$$\mathbf{A}^{[l]} = g^{[l]}(Z^{[l]})$$

$$A^{[0]} = X$$

$$A^{[L]} = \hat{Y}$$

Função de Perda

$$L(h) = -\frac{1}{n} \sum_{i=1}^n (y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i))$$

$$dA^{[L]} = \frac{Y}{A^{[L]}} + \frac{1 - Y}{1 - A^{[L]}} \quad \text{Inicialização do backpropagation}$$

Derivada parcial da ativação da camada de saída [L]

$$dZ^{[L]} = dA^{[L]} \times \frac{dg^{[L]}}{dZ^{[L]}}$$

$$dW^{[L]} = \frac{1}{n} dZ^{[L]} A^{[L-1]T}$$

Derivadas parciais da parte linear da camada de saída [L]

$$db^{[L]} = \frac{1}{n} \sum_{i=1}^n dZ^{[L](i)}$$

$$dA^{[L-1]} = W^{[L]T} dZ^{[L]} \quad \text{Derivada parcial da ativação da camada escondida [L-1]}$$

$$dZ^{[L-l]} = dA^{[L-l]} \times \frac{dg^{[L-l]}}{dZ^{[L-l]}} \quad \text{Esse termo depende da escolha de função de ativação}$$

$$dW^{[L-l]} = \frac{1}{n} dZ^{[L-l]} A^{[L-l-1]T}$$

$$db^{[L-l]} = \frac{1}{n} \sum_{i=1}^n dZ^{[L-l](i)}$$

Para $l = [1, L]$, derivadas parciais da parte linear da camada de saída $[L - l]$

$$dA^{[L-l-1]} = W^{[L-l]T} dZ^{[L-l]} \quad \text{Derivada parcial da ativação da camada escondida [L - l - 1]}$$

Próxima aula

A7: MLP em Numpy

Aula prática sobre implementação de redes neurais profundas com Numpy.