Backpropagation Handout

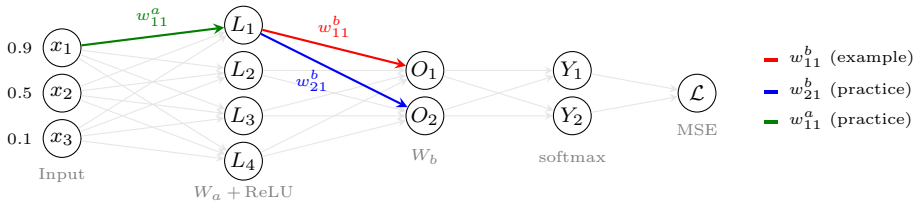
Step-by-Step Feedforward & Backpropagation

Your Tasks: **Forward:** compute all intermediate values H, L, O, Y, $\mathcal{L}$ (Steps 1–5). **Backward:** compute $\partial\mathcal{L}/\partial w^b_{21}$ and $\partial\mathcal{L}/\partial w^a_{11}$.

Setup (from `BackProp.py`)

Input: $x = (0.9,\ 0.5,\ 0.1)^T$ **Target:** $y = (1,\ 0)^T$

W_a: 4×3, no bias, all entries $= 0.1$ W_b: 2×4, no bias, all entries $= 0.1$



Forward Pass

Step 1: $H = W_a \cdot x$ (hidden pre-activation)

$$H = \underbrace{\begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix}}_{W_a} \cdot \underbrace{\begin{pmatrix} 0.9 \\ 0.5 \\ 0.1 \end{pmatrix}}_{x} = \begin{pmatrix} \_\_\_\_\_ \\ \_\_\_\_\_ \\ \_\_\_\_\_ \\ \_\_\_\_\_ \end{pmatrix} = \begin{pmatrix} \_\_\_\_\_ \\ \_\_\_\_\_ \\ \_\_\_\_\_ \\ \_\_\_\_\_ \end{pmatrix}$$

Step 2: $L = \mathrm{ReLU}(H)$ $\mathrm{ReLU}(z) = \max(0, z)$

$$L = \mathrm{ReLU}\begin{pmatrix} \_\_\_\_\_ \\ \_\_\_\_\_ \\ \_\_\_\_\_ \\ \_\_\_\_\_ \end{pmatrix} = \begin{pmatrix} \_\_\_\_\_ \\ \_\_\_\_\_ \\ \_\_\_\_\_ \\ \_\_\_\_\_ \end{pmatrix}$$

Step 3: $O = W_b \cdot L$ (output pre-activation)

$$O = \underbrace{\begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}}_{W_b} \cdot \underbrace{\begin{pmatrix} 0.15 \\ 0.15 \\ 0.15 \\ 0.15 \end{pmatrix}}_{L} = \begin{pmatrix} \_\_\_\_\_ \\ \_\_\_\_\_ \end{pmatrix} = \begin{pmatrix} \_\_\_\_\_ \\ \_\_\_\_\_ \end{pmatrix}$$

Step 4: $Y = \mathrm{softmax}(O)$ $Y_i = \dfrac{e^{O_i}}{\sum_j e^{O_j}}$

$$Y_1 = \frac{e^{\_\_\_\_}}{e^{\_\_\_\_} + e^{\_\_\_\_}} = \_\_\_\_\_ \qquad Y_2 = \_\_\_\_\_ \qquad \Rightarrow \quad Y = \begin{pmatrix} \_\_\_\_\_ \\ \_\_\_\_\_ \end{pmatrix}$$

Step 5: MSE Loss $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i - Y_i)^2$, $N=2$

$$\mathcal{L} = \tfrac{1}{2}\left[(\_\_\_ - \_\_\_)^2 + (\_\_\_ - \_\_\_)^2\right] = \_\_\_\_\_ = \_\_\_\_\_$$

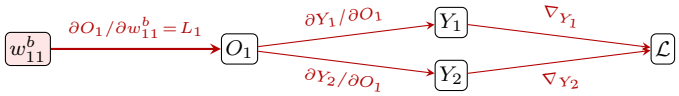
Backward Pass — Per-Parameter Gradient Paths

We trace the **gradient path** backward from $\mathcal{L}$ to each parameter, multiplying local derivatives along the way (chain rule). At branches, we **sum** contributions from all paths.

Derivative formulas: MSE: $\frac{\partial \mathcal{L}}{\partial Y_i} = \frac{2}{N}(Y_i - y_i)$ Softmax: $\frac{\partial Y_i}{\partial O_j} = Y_i(\delta_{ij} - Y_j)$ Linear: $\frac{\partial O_i}{\partial w_{ij}^b} = L_j$, $\frac{\partial O_i}{\partial L_j} = w_{ij}^b$, $\frac{\partial H_i}{\partial w_{ij}^a} = x_j$

ReLU: $\frac{\partial L_i}{\partial H_i} = \mathbf{1}_{H_i > 0}$

Example: $\dfrac{\partial \mathcal{L}}{\partial w_{11}^b}$ (red path: w_{11}^b connects L_1 to O_1)



Derivatives along path: $\frac{\partial \mathcal{L}}{\partial Y_k} = (Y_k - y_k)$, $\frac{\partial Y_k}{\partial O_1} = Y_k(\delta_{k1} - Y_1)$, $\frac{\partial O_1}{\partial w_{11}^b} = L_1$

Values (used by all paths below): $\nabla_{Y_1} = Y_1 - y_1 =$ _____, $\nabla_{Y_2} = Y_2 - y_2 =$ _____

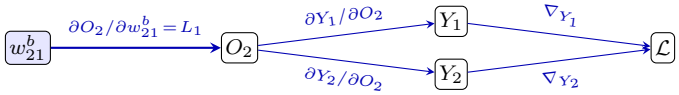
$\frac{\partial Y_1}{\partial O_1} =$ _____, $\frac{\partial Y_2}{\partial O_1} =$ _____, $\frac{\partial Y_1}{\partial O_2} =$ _____, $\frac{\partial Y_2}{\partial O_2} =$ _____

$$\frac{\partial \mathcal{L}}{\partial w_{11}^b} = \underbrace{[\_\_\_\_\_]}_{\nabla_{O_1}} \cdot \underbrace{\_\_\_\_\_}_{= L_1}$$

$$= [\_\_\_\_\_] \cdot \_\_\_\_\_$$

$$= (\_\_\_\_\_)(\_\_\_\_\_) = \_\_\_\_\_$$

Practice: $\dfrac{\partial \mathcal{L}}{\partial w_{21}^b}$ (blue path: w_{21}^b connects L_1 to O_2)



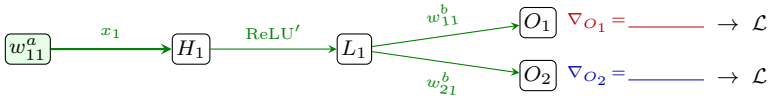
Derivatives along path: $\frac{\partial \mathcal{L}}{\partial Y_k} = (Y_k - y_k)$, $\frac{\partial Y_k}{\partial O_2} = Y_k(\delta_{k2} - Y_2)$, $\frac{\partial O_2}{\partial w_{21}^b} = L_1$

$$\frac{\partial \mathcal{L}}{\partial w_{21}^b} = \underbrace{\left[\nabla_{Y_1} \cdot \frac{\partial Y_1}{\partial O_2} + \nabla_{Y_2} \cdot \frac{\partial Y_2}{\partial O_2}\right]}_{\nabla_{O_2}} \cdot \underbrace{\frac{\partial O_2}{\partial w_{21}^b}}_{= L_1}$$

$$= [\ \_\_\_\_\_\ ] \cdot \_\_\_\_\_ = \_\_\_\_\_$$

Practice: $\dfrac{\partial \mathcal{L}}{\partial w_{11}^a}$ (green path: w_{11}^a connects x_1 to H_1; longer chain through hidden layer)

Derivatives along path: $\frac{\partial H_1}{\partial w_{11}^a} = x_1$, $\frac{\partial L_1}{\partial H_1} = \text{ReLU}'(H_1) = \mathbf{1}_{H_1 > 0}$, $\frac{\partial O_j}{\partial L_1} = w_{j1}^b$ (L_1 feeds both O_1, O_2 — sum their contributions)



$$\frac{\partial \mathcal{L}}{\partial w_{11}^a} = \underbrace{[\_\_\_\_\_]}_{\nabla_{L_1}} \cdot \underbrace{\text{ReLU}'(H_1)}_{= 1} \cdot \underbrace{x_1}_{= 0.9}$$

$$= [\ \_\_\_\_\_\ ] \cdot \_\_\_\_ \cdot \_\_\_\_ = \_\_\_\_\_$$

Key Insight — Initialization Problem: Based on your computation above, what is the gradient of w_a? What does this tell you about weight Initialization?