

Exercises

§13.4: The Chain Rule and Backpropagation

Python: There is an optional starter script that supports Python-based problem(s) on this sheet. Use it as scaffolding: read the comments, complete the lines marked **TODO**, and compare the output with the mathematics requested in the problem(s).

Problem 1 For the scalar ReLU network $a = wx + b$, $h = \text{ReLU}(a)$, $\hat{y} = vh + c$, $L = \frac{1}{2}(\hat{y} - y)^2$, use $x = 3$, $w = 2$, $b = -1$, $v = -2$, $c = 1$, $y = -8$. Compute the forward pass and all four parameter derivatives.

Problem 2 Starting from $a_i = \sum_j W_{ij}x_j + b_i$ and $\delta_i = \partial L / \partial a_i$, derive $\nabla_w L = \delta \mathbf{x}^T$ and $\nabla_b L = \delta$.

Problem 3 Using the Lesson 1.32 network and target $y = 7/4$, verify $\nabla_v L = (-1, -3/2)^T$, $\partial L / \partial c = -1$, $\nabla_w L = \begin{bmatrix} -4 & -2 \\ 2 & 1 \end{bmatrix}$, and $\nabla_b L = (-2, 1)^T$.

Problem 4 For a two-hidden-layer network, write formulas for δ_2 , $\nabla_{w_2} L$, $\nabla_{b_2} L$, δ_1 , $\nabla_{w_1} L$, and $\nabla_{b_1} L$.

Problem 5 For sigmoid binary classification with cross-entropy, show that the output delta is $p - y$. If $f = \mathbf{v}^T \mathbf{h} + c$, derive $\nabla_{\mathbf{v}} L$ and $\partial L / \partial c$.

Problem 6 Let $H_{\ell-1} \in \mathbb{R}^{20 \times 5}$, $W_{\ell} \in \mathbb{R}^{8 \times 5}$, and $\Delta_{\ell} \in \mathbb{R}^{20 \times 8}$. Verify the dimensions of $\Delta_{\ell}^T H_{\ell-1}$, $\nabla_{W_{\ell}} g$, $\Delta_{\ell} W_{\ell}$, and $\Delta_{\ell-1}$.

Problem 7 Implement the one-hidden-layer example in Python. Use centered finite differences with $\varepsilon = 10^{-6}$ to check every entry of W , $\mathbf{b}$, $\mathbf{v}$, and c . Report the maximum absolute discrepancy.

Problem 8 Explain why backpropagation is computationally preferable to estimating every derivative with finite differences. Discuss cached forward quantities, reuse of downstream derivatives, the roughly $2D$ extra evaluations needed for D parameters, and reverse-mode automatic differentiation.
