Exercise 5: Manual Differentiation

Advanced Deep Learning for Physics (IN2298)

Andrea Scaioli

Summer Semester 2025

Exercise 1: One-step backpropagation

Given:

$$x_{n+1} = \begin{bmatrix} x_{n,1}^2 + x_{n,2} \\ -x_{n,1} + \dfrac{x_{n,2}}{2} \\ -x_{n,2}^2 + x_{n,3} \end{bmatrix} + \begin{bmatrix} x_{n,1}\theta_1 \\ x_{n,2}\theta_2 \\ x_{n,3}\theta_3 \end{bmatrix},$$

$$x_0 = \begin{bmatrix} 2 \\ -1 \\ 3 \end{bmatrix}, \quad \theta = \begin{bmatrix} -1 \\ 4 \\ 2 \end{bmatrix}, \quad L = \|x_1\|_1$$

Step 1: Compute x_1

$$x_{1,1} = 2^2 + (-1) + 2(-1) = 4 - 1 - 2 = 1$$

$$x_{1,2} = -2 + \frac{-1}{2} + (-1)(4) = -2 - 0.5 - 4 = -6.5$$

$$x_{1,3} = -(-1)^2 + 3 + 3(2) = -1 + 3 + 6 = 8$$

$$x_1 = \begin{bmatrix} 1 \\ -6.5 \\ 8 \end{bmatrix}, \quad L = |1| + |-6.5| + |8| = 15.5$$

Step 2: Backpropagation

$$\frac{\partial L}{\partial x_{1,1}} = 1, \quad \frac{\partial L}{\partial x_{1,2}} = -1, \quad \frac{\partial L}{\partial x_{1,3}} = 1$$

$$\frac{\partial x_{1,1}}{\partial \theta_1} = x_{0,1} = 2 \Rightarrow \frac{\partial L}{\partial \theta_1} = 1 \cdot 2 = 2$$

$$\frac{\partial x_{1,2}}{\partial \theta_2} = x_{0,2} = -1 \Rightarrow \frac{\partial L}{\partial \theta_2} = -1 \cdot (-1) = 1$$

$$\frac{\partial x_{1,3}}{\partial \theta_3} = x_{0,3} = 3 \Rightarrow \frac{\partial L}{\partial \theta_3} = 1 \cdot 3 = 3$$

Final result:

$$\nabla_\theta L = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$$

Exercise 2: Multi-step control

Given:

$$x_0 = \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix}, \quad \theta = \begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix}, \quad L = \|x_2\|_1$$

Step 1: Compute x_1

$$x_1 = \begin{bmatrix} (-1)^2 + 1 + (-1)(3) \\ 1 + 0.5 + 1(-1) \\ -1 + 2 + 2(1) \end{bmatrix} = \begin{bmatrix} 1 - 3 = -1 \\ 1.5 - 1 = 0.5 \\ 1 + 2 = 3 \end{bmatrix}$$

Step 2: Compute x_2

$$x_2 = \begin{bmatrix} (-1)^2 + 0.5 + (-1)(3) \\ 1 + 0.25 + 0.5(-1) \\ -0.25 + 3 + 3(1) \end{bmatrix} = \begin{bmatrix} 1.5 - 3 = -1.5 \\ 1.25 - 0.5 = 0.75 \\ -0.25 + 3 + 3 = 5.75 \end{bmatrix}$$
$$L = |-1.5| + |0.75| + |5.75| = 8.0$$

Step 3: Backpropagation

$$\frac{\partial L}{\partial x_2} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$$

Using the chain rule, we compute gradients w.r.t. x_1, then to θ:

$$\frac{\partial L}{\partial x_{1,1}} = (-1)(2x_{1,1} + \theta_1) + (1)(-1) = (-1)(-2 + 3) + (-1) = -2$$
$$\frac{\partial L}{\partial x_{1,2}} = (-1)(1) + (1)(0.5 + \theta_2) + (1)(-2x_{1,2}) = -1 + (-0.5) - 1 = -2.5$$
$$\frac{\partial L}{\partial x_{1,3}} = (1)(1 + \theta_3) = 2$$

Now use:

$$\frac{\partial x_{1,1}}{\partial \theta_1} = x_{0,1} = -1 \Rightarrow \frac{\partial L}{\partial \theta_1} = (-2)(-1) + (-1)(-1) = 3$$
$$\frac{\partial x_{1,2}}{\partial \theta_2} = x_{0,2} = 1 \Rightarrow \frac{\partial L}{\partial \theta_2} = (-2.5)(1) + (1)(0.5) = -2$$
$$\frac{\partial x_{1,3}}{\partial \theta_3} = x_{0,3} = 2 \Rightarrow \frac{\partial L}{\partial \theta_3} = (2)(2) + (1)(3) = 7$$

Final result:

$$\nabla_\theta L = \begin{bmatrix} 3 \\ -2 \\ 7 \end{bmatrix}$$

Exercise 3: Forward mode autodiff

We re-solve Exercise 1 using forward-mode differentiation.

Recall:

$$x_0 = \begin{bmatrix} 2 \\ -1 \\ 3 \end{bmatrix}, \quad \theta = \begin{bmatrix} -1 \\ 4 \\ 2 \end{bmatrix}, \quad x_1 = \begin{bmatrix} 1 \\ -6.5 \\ 8 \end{bmatrix}, \quad L = 15.5$$

Compute directional derivatives:

$$\frac{\partial L}{\partial \theta_1} = \frac{\partial x_{1,1}}{\partial \theta_1} \cdot \operatorname{sign}(x_{1,1}) = x_{0,1} \cdot 1 = 2$$
$$\frac{\partial L}{\partial \theta_2} = x_{0,2} \cdot \operatorname{sign}(x_{1,2}) = (-1)(-1) = 1$$
$$\frac{\partial L}{\partial \theta_3} = x_{0,3} \cdot \operatorname{sign}(x_{1,3}) = 3 \cdot 1 = 3$$

Final result:

$$\nabla_\theta L = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$$

Forward vs Backward mode:

- **Forward mode**: Efficient when few inputs and many outputs; computes $\frac{\mathrm{d}y}{\mathrm{d}x}$ one input at a time.
- **Backward mode (backpropagation)**: Efficient when many inputs and one output (common in deep learning).
- In deep learning, we compute gradients of a scalar loss w.r.t. millions of parameters: *backward mode is preferred. Nevertheless sometimes forward propagation of gradients is competitive or even better than the backpropagation..*