PS1: Part 1

Ibrahim Bashir

52025 Fall 2025
November 12, 2025

Instructions

- Please answer the questions below.
- Submit full answers with complete work in a PDF file into the relevant submission box in Moodle.
- You don't have to type your answers, but please make sure they are legible and clear.

Preliminaries

- The function $f : \mathbb{R}^d \to \mathbb{R}$ maps a d-dimensional vector to a scalar.
- The column vector $\nabla_x f(x)$ is the gradient of $f(x)$ with partial derivatives:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_d}(x) \end{bmatrix}$$

- The Jacobian $\frac{\partial f}{\partial x} \in \mathbb{R}^{n \times m}$ is a matrix where each element (i, j) is given by $\frac{\partial f_j}{\partial x_i}$.
- Multivariate chain rule: see here.
- A useful guide on neural network gradients.
- This is a very intuitive explanation of gradients in deep neural networks.

A (50 pts)

Answer the following questions[1]:

1. Let $x \in \mathbb{R}^d$, and $f(x) = \|x\|_2^2 = x^\top x$. Compute the gradient $\nabla f(x)$ (gradient of the ℓ_2 norm).
2. Let $f(x) = A^\top x \in \mathbb{R}^n$, for $A \in \mathbb{R}^{d \times n}$. Compute the Jacobian of f with respect to x (Jacobian of a linear map).
3. Let $g(x) = A^\top x \in \mathbb{R}^n$ and $f(y) = \|y\|_2^2$. Compute the gradient of $f(g(x))$ with respect to x (hint: use the chain rule).
4. Let $g(A) = A^\top x \in \mathbb{R}^n$ and $f(y) = \|y\|_2^2$. Compute the gradient of $f(g(A))$ with respect to A.

[1] Based on Berkeley's CS182 course.

B (50 pts)

Figure 1 portrays a basic neural network architecture schema with weights, biases, activation functions, and loss components. The loss is defined as:

$$\text{Loss} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

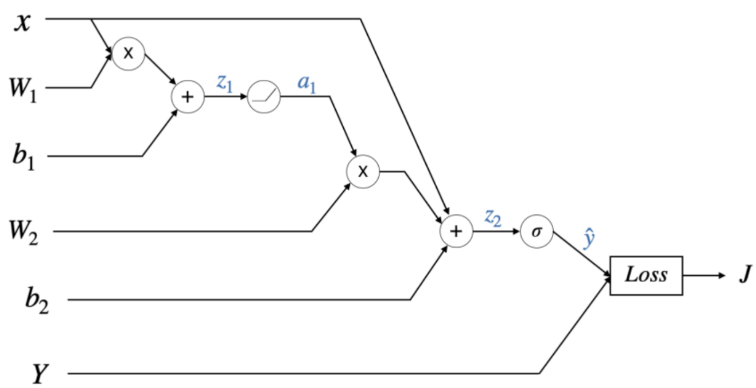

Figure 1: Neural architecture example

1. Express $\hat{y}$ as a function of x, W_1, b_1, W_2, b_2.
2. Compute the gradients $\frac{\partial J}{\partial W_2}$ and $\frac{\partial J}{\partial b_2}$.
3. Compute the gradients $\frac{\partial J}{\partial W_1}$, $\frac{\partial J}{\partial b_1}$, and $\frac{\partial J}{\partial x}$.
4. What intermediate variables do we need to cache in the above calculations?