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:

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

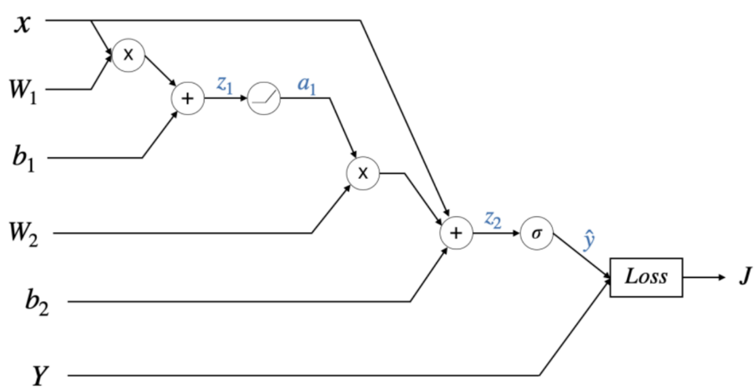

Figure 1: Neural architecture example

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?

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).

$$\rightarrow \nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix} = \begin{pmatrix} \frac{\partial \sum x_i^2}{\partial x_1} \\ \vdots \\ \frac{\partial \sum x_i^2}{\partial x_d} \end{pmatrix} = \begin{pmatrix} 2x_1 \\ \vdots \\ 2x_d \end{pmatrix} = 2x$$

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).

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \vdots & & \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_d} \end{pmatrix}$$

$$= \begin{pmatrix} A_{11} & \cdots & A_{1d} \\ & & \\ A_{n1} & \cdots & A_{nd} \end{pmatrix} = A^T$$

ההסבר הוא שכל רכיב של f באינדקס i הוא $(A^Tx)_i$

ולכן זה שקול ל-f: $A_i x$ ואז אם נגזור לפי x_k מקבלים

יוצא A_{ik}

בסופו אפשר לראות שהמטריצה היא בדיוק A^T

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).

$$\nabla f(g(x)) = \left(\frac{\partial g}{\partial x}\right)^T \frac{\partial f}{\partial y} = (A^T)^T 2y$$

אבל מכיון ש $y=g(x)=A^Tx$ לכן:

$$= 2AA^Tx$$

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.

$$\to \nabla f(g(A)) = \frac{\partial g}{\partial A} \cdot \nabla_y f^T = x(2y)^T$$

ומכיון שאנו נוגעים ש $y=g(A)=A^Tx$ לכן

$$= 2xx^TA$$

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:

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

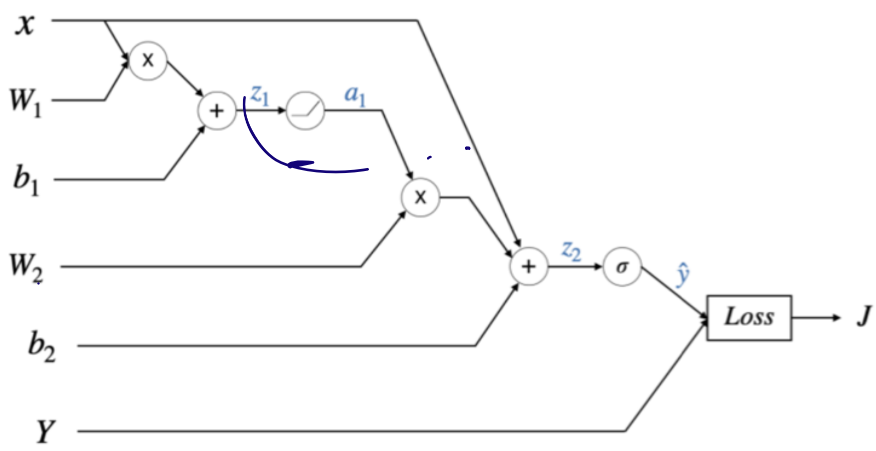

Figure 1: Neural architecture example

1. Express $\hat{y}$ as a function of x, W_1, b_1, W_2, b_2.

$\to z_1 = W_1 x + b_1$

$a_1 = \max(0, z_1) = \max(0, W_1 x + b_1)$

$z_2 = W_2 a_1 + b_2 + x = W_2 \max(0, W_1 x + b_1) + b_2 + x$

$\hat{y} = \sigma(z_2) = \sigma(W_2 \max(0, W_1 x + b_1) + b_2 + x$

2. Compute the gradients $\frac{\partial J}{\partial W_2}$ and $\frac{\partial J}{\partial b_2}$.

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

$$\frac{\partial \hat{y}}{\partial z_2} = \sigma'(z_2) = \left(\frac{1}{1+e^{-z_2}}\right)' = \frac{e^{-z_2}}{1+e^{-z_2}} = \sigma(z_2)(1-\sigma(z_2))$$

$$= \hat{y}(1-\hat{y})$$

$$\frac{\partial z_2}{\partial W_2} = a_1 = \max(0, W_1x+b_1)$$

$$\frac{\partial z_2}{\partial b_2} = 1$$

$$\frac{\partial J}{\partial W_2} = \frac{\partial J}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_2} \cdot \frac{\partial z_2}{\partial W_2} = \left(-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\right)(\hat{y}(1-\hat{y})) a_1$$

$$= (-y(1-\hat{y}) + (1-y)\hat{y}) a_1$$

$$= (\hat{y}-y) a_1$$

$$\frac{\partial J}{\partial b_2} = \frac{\partial J}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_2} \cdot \frac{\partial z_2}{\partial b_2} = \left(-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\right)(\hat{y}(1-\hat{y})) \cdot 1$$

$$= \hat{y}-y$$

3. Compute the gradients $\frac{\partial J}{\partial W_1}$, $\frac{\partial J}{\partial b_1}$, and $\frac{\partial J}{\partial x}$.

$$\frac{\partial z_2}{\partial a_1} = W_2$$

$$\frac{\partial a_1}{\partial z_1} = \mathbb{1}_{\{z_1 \geq 0\}}$$

$$\frac{\partial z_1}{\partial W_1} = x \quad ; \quad \frac{\partial z_1}{\partial x} = W_1 \quad ; \quad \frac{\partial z_1}{\partial b_1} = 1$$

$$\Rightarrow \frac{\partial J}{\partial W_1} = \frac{\partial J}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial W_1} = (\hat{y}-y) W_2 \mathbb{1}_{\{z_1 \geq 0\}} x$$

$$\Rightarrow \frac{\partial J}{\partial b_1} = \frac{\partial J}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial b_1} = (\hat{y}-y) W_2 \mathbb{1}_{\{z_1 \geq 0}$$

$$\Rightarrow \frac{\partial J}{\partial x} = \frac{\partial J}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial x} = (\hat{y}-y) W_2 \mathbb{1}_{\{z_1 \geq 0\}} W_1$$

4. What intermediate variables do we need to cache in the above calculations?

We need to cache all intermediate variables

$\rightarrow$ cache $= \{z_1, a_1, z_2, \hat{y}\}$

because we need them many times to calculate derivatives for all others leafs