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.

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).

$$f(x) = \|x\|^2 = \sum_{i=1}^{d} x_i^2 = \begin{bmatrix} x_1^2 \\ \vdots \\ x_d^2 \end{bmatrix} \Rightarrow \nabla f(x) = \begin{bmatrix} 2x_1 \\ \vdots \\ 2x_d \end{bmatrix} = 2X$$

$$X \in \mathbb{R}^d$$

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).

$$f(x) = A^T x \quad , x \in \mathbb{R}^d \, , A^T \in \mathbb{R}^{n\times d}$$

$$J = \begin{bmatrix} \frac{df_1}{dx_1} & \cdots & \frac{df_1}{dx_d} \\ \vdots & & \vdots \\ \frac{df_n}{dx_1} & \cdots & \frac{df_n}{dx_n} \end{bmatrix}$$

$$(A_{1i}x_i)_1 = A_{1i}\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$$

$$\frac{df_1}{dx_1} = A_{1i}\sum_{i=1}^{d} x_i = \underline{A_{11}}$$

ההסבר הוא כאשר נגזור את $(A^Tx)_1$ לפי x_1 אז נקבל

A_{11} במקום ה-1 ובכל השאר - 0, כך גם לגבי שאר האיברים בהתאם

לכן:

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

ומקבל 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).

$$g(x) = A^T x \qquad \nabla g(x) = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ A_{n1} & & A_{nd} \end{pmatrix} = \boxed{A^T}$$

$$f(y) = \|y\|_2^2 = y^T y = \sum y_i^2$$

$$\nabla f(y) = 2Y$$

$$\nabla_x f(g(x)) = \left(\frac{dg}{dx}\right)^T \nabla_y f(y) = (A^T)^T \cdot 2y = A \cdot 2y$$

כאשר $y = g(x) = A^T x$ לכן

$$= A \cdot 2(A^T x) = \boxed{2AA^T x}$$

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.

$g(A) = A^T x \qquad \nabla g(A) = x$

$f(y) = \|y\|_2^2 = y^T y \qquad \nabla f(y) = 2y$

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

$y = g(A) = A^T x$ [illegible]

$$= x\,2(A^T x)^T = \boxed{2 x x^T A}$$

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:

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

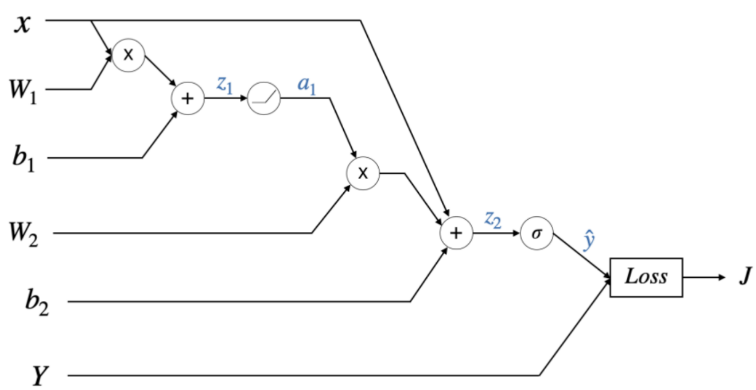

Figure 1: Neural architecture example

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?

7 7 7

0111 0111 0111

4 2 1

groups
me
public

w r x
1 1 1

chmod file-name 750

111 101 000

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

$(w_1 x)$

$$z_1 = (w_1 x + b_1) \longrightarrow a_1 = \max(0, w_1 x + b_1) \quad [\text{ReLU}]$$

$$z_2 = W_2(a_1) + b_2 + X$$

$$y = \hat{y} = \sigma(z_2) = \overset{\text{Sigmoid}}{\sigma}(W_2 \max(0, w_1 x + b_1) + b_2 + X)$$

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

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

לפי כלל השרשרת, נגזור "אחורה" =

$\boxed{\frac{dJ}{dW_2}}$ נתחיל עם

נגזרת פנימית.

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

(1)
$$\frac{dJ}{d\hat{y}} = -\frac{y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})}$$

(2)
$$\frac{d\hat{y}}{dz_2} = \frac{d\sigma(z_2)}{dz_2} = \frac{d}{z_2}\left(\frac{1}{1+e^{-z_2}}\right)' = \frac{-e^{-z}}{(1+e^{-z})^2}$$

$$= \frac{1}{(1+e^{-z})} \cdot \frac{-e^{-z}}{(1+e^{-z})} = \sigma(z_2) \cdot \frac{-e^{-z}}{(1+e^{-z})}$$

$$= \sigma(z_2)\,(1-\sigma(z_2)) = \hat{y}(1-\hat{y})$$

$$\boxed{\frac{dJ}{dz_2} = \hat{y}(1-\hat{y})}$$

$$\frac{dz^2}{dw_2} = \frac{d}{dw_2}\left(w_2(a_1)+b_2+x\right) = a_1$$

: chain rule

$$\frac{dJ}{d\hat{y}} \cdot \frac{d\hat{y}}{dz_2} \cdot \frac{dz^2}{dw_2} = \left(\frac{y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})}\right)\left[\hat{y}(1-\hat{y})\right] \cdot a_1$$

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

$$\boxed{\frac{dJ}{db_2}}$$

$$\frac{dJ}{db_2} = \frac{dJ}{d\hat{y}} \cdot \frac{d\hat{y}}{dz_2} \cdot \frac{dz^2}{db^2} = \left(\frac{y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})}\right)\left[\hat{y}(1-\hat{y})\right] \cdot 1$$

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

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

$$\frac{dJ}{dW_1} = \frac{dJ}{dz_2} \cdot \frac{dz_2}{da_1} \cdot \frac{da_1}{dz_1} \cdot \frac{dz_1}{dW_1}$$

$$\frac{dJ}{dz_2} = \hat{y} - y$$

$$\frac{dz_2}{da_1} = \left((W_2 a_1 + b_1) + b_2\right)' = W_2$$

$$\frac{da_1}{dz_1} = \left[\max(0, W_1 x + b_1)\right]' = \frac{d\,\mathrm{ReLu}(z_1)}{dz_1} = 1_{\{z_1 > 0\}}$$

$$\frac{dz_1}{dW_1} = (W_1 x + b_1)' = X^T$$

חיבור הכל ימינה!

$$\frac{dJ}{dW_1} = \frac{dJ}{dz_2} \cdot \frac{dz_2}{da_1} \cdot \frac{da_1}{dz_1} \cdot \frac{dz_1}{dW_1} = \boxed{\left[(\hat{y} - y) \cdot W_2 \cdot 1_{\{z_1 > 0\}}\right] \cdot X^T}$$

$$\frac{dz_1}{db_1} = 1$$

$$\boxed{\frac{dJ}{db_1}}$$

כעת

$$\frac{dJ}{db_1} = \frac{dJ}{dW_1} = \frac{dJ}{dz_2} \cdot \frac{dz_2}{da_1} \cdot \frac{da_1}{dz_1} \cdot \frac{dz_1}{db_1} \quad \boxed{\left[(\hat{y}-y) \cdot W_2 \cdot 1_{\{z_1>0\}}\right]}$$

$$\frac{dz_1}{dx} = W_1 \qquad\qquad \frac{dJ}{dx}$$

$$\frac{dJ}{dx} = \frac{dJ}{dW_1} = \frac{dJ}{dz_2} \cdot \frac{dz_2}{da_1} \cdot \frac{da_1}{dz_1} \cdot \frac{dz_1}{dx} = \left[(\hat{y}-y) \cdot W_2 \cdot 1_{\{z_1>0\}}\right] \cdot W_1$$

We need to cache all intermadiate Variables

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

becuse we need many times to calculate

derivatives for all other leafs.