Week 2: Gradients

Partial Derivatives & Linear Regression

Machine Learning (MLVU)

February 13, 2025

Today's Plan

1. Review: Partial derivatives & the gradient
2. Practice: Differentiation rules
3. Application: Linear regression loss function
4. Gradient descent: One step by hand
5. Closed-form solution for linear regression
6. Compute optimal parameters on real data

Goal: Understand how gradients guide us toward optimal models.

Quick Recap: Partial Derivatives

Given a function of multiple variables, a **partial derivative** is the derivative with respect to *one* variable, treating all others as constants.

Example:

$$f(a, b) = 3a^2 + b^2 - ab + a$$

- $\frac{\partial f}{\partial a} = 6a - b + 1$ (treat b as constant)
- $\frac{\partial f}{\partial b} = 2b - a$ (treat a as constant)

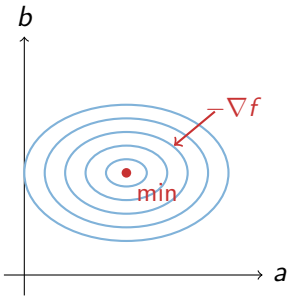
The **gradient** collects all partial derivatives into a vector:

$$\nabla f(a, b) = \left(\frac{\partial f}{\partial a}, \frac{\partial f}{\partial b} \right) = (6a - b + 1,\ 2b - a)$$

Gradient Intuition

Think of $f(a, b)$ as a landscape:

- The gradient at a point is an **arrow** pointing in the direction of **steepest ascent**
- A marble released at that point rolls in the **opposite** direction
- At a **minimum**, all partial derivatives are zero



Question 1: Finding the Minimum

Task: Find the values of a and b that minimize $f(a, b) = 3a^2 + b^2 - ab + a$.

Set the partial derivatives to zero:

$$6a - b + 1 = 0 \qquad\qquad 2b - a = 0$$

Your turn:

- Express a in terms of b from the first equation
- Express b in terms of a from the second equation
- Substitute and solve!

Take 2–3 minutes, then we'll solve it together.

Question 1: Solution

From the two equations:

$$a = \frac{b-1}{6} \qquad b = \frac{a}{2}$$

Substitute $a = \frac{b-1}{6}$ into $b = \frac{a}{2}$:

$$b = \frac{1}{2} \cdot \frac{b-1}{6} = \frac{b-1}{12}$$

$$b - \frac{b}{12} = -\frac{1}{12} \qquad \Longrightarrow \qquad \frac{11b}{12} = -\frac{1}{12}$$

$$\boxed{b = -\frac{1}{11}}$$

Then:

$$a = \frac{-\frac{1}{11} - 1}{6} = \frac{-\frac{12}{11}}{6} = \boxed{-\frac{2}{11}}$$

Differentiation Rules Cheat Sheet

Let c be a constant, $f(x)$ and $g(x)$ functions of x:

Rule	**Formula**
Constant	$\frac{\partial c}{\partial x} = 0$
Exponent	$\frac{\partial x^n}{\partial x} = nx^{n-1}$
Constant factor	$\frac{\partial (c \cdot f)}{\partial x} = c\,\frac{\partial f}{\partial x}$
Sum	$\frac{\partial (f+g)}{\partial x} = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial x}$
Chain rule	$\frac{\partial f(g(x))}{\partial x} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial x}$

Strategy: Match parts of your expression to the left-hand side of a rule, replace with the right-hand side, repeat until all ∂'s are gone.

Question 2: Chain Rule Practice

Task: Compute $\frac{\partial(x+y+z)^2}{\partial y}$ step by step.

Hints:

1. What rule applies to an "outer" square of an "inner" expression?
2. What is the derivative of the inner part $(x + y + z)$ with respect to y?
3. Remember: x and z are constants with respect to y.

Try it yourself first!

Question 2: Solution

$$\frac{\partial(x+y+z)^2}{\partial y} = \underbrace{\frac{\partial(x+y+z)^2}{\partial(x+y+z)}}_{\text{outer derivative}} \cdot \underbrace{\frac{\partial(x+y+z)}{\partial y}}_{\text{inner derivative}} \qquad \text{(chain rule)}$$

$$= 2(x+y+z) \cdot \frac{\partial(x+y+z)}{\partial y} \qquad \text{(exponent rule)}$$

$$= 2(x+y+z) \cdot \left(\underbrace{\frac{\partial x}{\partial y}}_{0} + \underbrace{\frac{\partial y}{\partial y}}_{1} + \underbrace{\frac{\partial z}{\partial y}}_{0} \right) \qquad \text{(sum rule)}$$

$$= \boxed{2(x+y+z)}$$

Linear Regression: The Setup

Scenario: Predicting a baby's height from age.

Age a_i (months)	**Height** h_i (cm)
0	30
2	40
4	50

Model: $f_{s,b}(a) = sa + b$ (slope s, intercept b)

Loss function (sum of squared errors):

$$\text{loss}(s, b) = \frac{1}{2} \sum_i \left(s\, a_i + b - h_i\right)^2$$

The $\frac{1}{2}$ is a convenience factor — it cancels with the 2 from differentiation.

Question 3: Computing the Loss

Task: For $s = 1$ and $b = 0$, compute $\text{loss}(1, 0)$ using the data.

$$\text{loss}(1,0) = \frac{1}{2}\sum_i (1 \cdot a_i + 0 - h_i)^2 = \frac{1}{2}\sum_i (a_i - h_i)^2$$

Your turn: Plug in each data point:

i	a_i	h_i	$(a_i - h_i)^2 = ?$
1	0	30	
2	2	40	
3	4	50	

What does this loss value tell us about $s = 1, b = 0$?

Question 3: Solution

$$\text{loss}(1, 0) = \frac{1}{2}\Big[(0-30)^2 + (2-40)^2 + (4-50)^2\Big]$$
$$= \frac{1}{2}\Big[900 + 1444 + 2116\Big] = \frac{4460}{2} = \boxed{2230}$$

Interpretation: This is a *huge* loss. The model $f(a) = a$ predicts heights of 0, 2, and 4 cm — way off from the actual 30, 40, 50 cm.

We need to find better values of s and b!

Question 4: Why Square the Errors?

Discussion: The residual $f(a_i) - h_i$ measures the error per example. Why not just sum the residuals directly? Why square first?

Think about this:

- What happens if one prediction is too high by 100 and another is too low by 100?
- Could you use absolute values instead? What's different?

Discuss with your neighbor for 1 minute.

Question 4: Solution

Problem with plain summation:
Positive and negative errors **cancel out**. A model that's wildly wrong in both directions could appear perfect.

Example: errors of $+100$ and -100 sum to 0, suggesting no error at all!

Why squaring specifically (vs. absolute value)?

- Squaring **penalizes large errors more** than absolute value
- It's **differentiable everywhere** (absolute value has a kink at 0)
- Under the assumption of Gaussian noise, minimizing squared errors is equivalent to **maximum likelihood estimation** (more on this in the probability lectures)

Question 5: Derivatives of the Loss

Task: Find $\frac{\partial \text{loss}}{\partial s}$ and $\frac{\partial \text{loss}}{\partial b}$ for $\text{loss}(s, b) = \frac{1}{2}\sum_i (sa_i + b - h_i)^2$.

Strategy for $\frac{\partial \text{loss}}{\partial s}$:

1. Apply the **chain rule**: outer function is $(\cdot)^2$, inner is $(sa_i + b - h_i)$
2. Derivative of outer: $2(\cdot)$, times derivative of inner w.r.t. s
3. What is $\frac{\partial (sa_i + b - h_i)}{\partial s}$?

Then for $\frac{\partial \text{loss}}{\partial b}$:

Same chain rule approach — what changes?

Work through this carefully. Take 3–4 minutes.

Question 5: Solution — $\frac{\partial \text{loss}}{\partial s}$

$$\frac{\partial \text{loss}}{\partial s} = \frac{1}{2} \sum_i \frac{\partial (s a_i + b - h_i)^2}{\partial (s a_i + b - h_i)} \cdot \frac{\partial (s a_i + b - h_i)}{\partial s}$$

$$= \frac{1}{2} \sum_i 2(a_i s + b - h_i) \cdot a_i$$

$$= \sum_i (a_i s + b - h_i)\, a_i$$

$$= \sum_i (a_i^2\, s + a_i\, b - a_i\, h_i)$$

$$\boxed{\frac{\partial \text{loss}}{\partial s} = s \sum_i a_i^2 + b \sum_i a_i - \sum_i a_i h_i}$$

Question 5: Solution — $\frac{\partial \text{loss}}{\partial b}$

$$\frac{\partial \text{loss}}{\partial b} = \frac{1}{2} \sum_i 2(a_i s + b - h_i) \cdot \underbrace{\frac{\partial (s a_i + b - h_i)}{\partial b}}_{=1}$$

$$= \sum_i (a_i s + b - h_i)$$

$$\boxed{\frac{\partial \text{loss}}{\partial b} = s \sum_i a_i + b\, n - \sum_i h_i}$$

The gradient:

$$\nabla \text{loss}(s, b) = \left(s \sum_i a_i^2 + b \sum_i a_i - \sum_i a_i h_i, \;\; s \sum_i a_i + bn - \sum_i h_i \right)$$

Question 6: One Step of Gradient Descent

Task: Starting at $s = 1,\ b = 0$ with learning rate $\eta = 0.01$, perform **one step** of gradient descent.

Recall the update rule:

$$\begin{pmatrix} s^{\text{new}} \\ b^{\text{new}} \end{pmatrix} = \begin{pmatrix} s \\ b \end{pmatrix} - \eta \nabla \text{loss}(s, b)$$

Steps:

1. Compute $\sum a_i^2$, $\sum a_i$, $\sum a_i h_i$, $\sum h_i$ from the data
2. Plug $s = 1, b = 0$ into the gradient
3. Apply the update rule

Data: $(0, 30)$, $(2, 40)$, $(4, 50)$. *Take 2–3 minutes.*

Question 6: Solution

Data statistics:

$$\sum a_i^2 = 0 + 4 + 16 = 20, \quad \sum a_i = 6, \quad \sum a_i h_i = 0 + 80 + 200 = 280, \quad \sum h_i = 120$$

Gradient at $s = 1, b = 0$:

$$\nabla \text{loss} = \big(1 \cdot 20 + 0 \cdot 6 - 280, \;\; 1 \cdot 6 + 0 \cdot 3 - 120\big) = (-260, \; -114)$$

Update:

$$\begin{pmatrix} s^{\text{new}} \\ b^{\text{new}} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - 0.01 \begin{pmatrix} -260 \\ -114 \end{pmatrix} = \boxed{\begin{pmatrix} 3.6 \\ 1.14 \end{pmatrix}}$$

The negative gradient told us to **increase both** s **and** b — which makes sense, since $s = 1, b = 0$ was far too low.

Question 7: Closed-Form Solution (Part 1)

Task: Set both derivatives to zero and solve for s and b.
Express s so it does **not** depend on b.

Notation for data statistics:

$$\bar{a} = \frac{1}{n}\sum a_i, \quad \bar{h} = \frac{1}{n}\sum h_i, \quad \overline{a^2} = \frac{1}{n}\sum a_i^2, \quad \overline{ah} = \frac{1}{n}\sum a_i h_i$$

Approach:

1. From $\frac{\partial \text{loss}}{\partial b} = 0$, express b in terms of s, $\bar{a}$, $\bar{h}$
2. From $\frac{\partial \text{loss}}{\partial s} = 0$, express s in terms of b, $\bar{a}$, $\overline{a^2}$, $\overline{ah}$
3. Substitute to eliminate b

This is the hardest part — take 5 minutes.

Question 7: Solution — Solving for b

From $\frac{\partial \text{loss}}{\partial b} = 0$:

$$s \sum_i a_i + bn - \sum_i h_i = 0$$

$$bn = \sum_i h_i - s \sum_i a_i$$

$$\boxed{b = \bar{h} - s\,\bar{a}} \tag{2}$$

From $\frac{\partial \text{loss}}{\partial s} = 0$, dividing through by n:

$$\boxed{s = -b\,\frac{\bar{a}}{\overline{a^2}} + \frac{\overline{ah}}{\overline{a^2}}} \tag{1}$$

Question 7: Solution — Eliminating b

Substitute $b = \bar{h} - s\bar{a}$ into equation (1):

$$s = -(\bar{h} - s\bar{a})\,\frac{\bar{a}}{\overline{a^2}} + \frac{\overline{ah}}{\overline{a^2}}$$

$$s = -\frac{\bar{h}\,\bar{a}}{\overline{a^2}} + s\,\frac{\bar{a}^2}{\overline{a^2}} + \frac{\overline{ah}}{\overline{a^2}}$$

$$s\left(1 - \frac{\bar{a}^2}{\overline{a^2}}\right) = \frac{\overline{ah} - \bar{a}\,\bar{h}}{\overline{a^2}}$$

$$\boxed{s = \frac{\overline{ah} - \bar{a}\,\bar{h}}{\overline{a^2} - \bar{a}^2}} \qquad b = \bar{h} - s\,\bar{a}$$

Both parameters expressed purely in terms of **data statistics**!

Question 8: Computing the Optimal Parameters

Task: Plug in the data and compute the optimal s and b.
What do these parameters tell us about the baby?

Data: $(0, 30),\ (2, 40),\ (4, 50)$

Compute these statistics:

$\bar{a} = ?$ $\quad \bar{h} = ?$

$\overline{a^2} = ?$ $\quad \overline{ah} = ?$

Then plug into:

$$s = \frac{\overline{ah} - \bar{a}\,\bar{h}}{\overline{a^2} - \bar{a}^2} \qquad b = \bar{h} - s\,\bar{a}$$

Take 2 minutes, then we'll check together.

Question 8: Solution

Statistics:

$$\bar{a} = \frac{0+2+4}{3} = 2, \quad \bar{h} = \frac{30+40+50}{3} = 40$$

$$\overline{a^2} = \frac{0+4+16}{3} = \frac{20}{3}, \quad \overline{ah} = \frac{0+80+200}{3} = \frac{280}{3}$$

Optimal slope:

$$s = \frac{\frac{280}{3} - 2 \cdot 40}{\frac{20}{3} - 4} = \frac{\frac{40}{3}}{\frac{8}{3}} = \boxed{5}$$

Optimal intercept:

$$b = 40 - 5 \cdot 2 = \boxed{30}$$

Interpretation: The baby was **30 cm at birth** and grows approximately **5 cm per month**.

Summary

1. **Partial derivatives** — differentiate w.r.t. one variable, hold others constant
2. **Gradient** — vector of all partial derivatives; points uphill
3. **Loss function** — measures how bad a model is; we want to minimize it
4. **Gradient descent** — iteratively step in the direction of $-\nabla \text{loss}$
5. **Closed-form solution** — when possible, set $\nabla \text{loss} = 0$ and solve directly

Key formula for linear regression:

$$s = \frac{\overline{ah} - \bar{a}\,\bar{h}}{\overline{a^2} - \bar{a}^2} \qquad b = \bar{h} - s\,\bar{a}$$

Questions?

Next week: Probability & Maximum Likelihood