Optimization Techniques for Machine Learning

AMLZC326 · #06 Vector Calculus II

Anshid Aboobacker

Motivation

- We can now compute gradients (CH05). To classify critical points and design second-order optimisers, we need the **Hessian**.
- For multi-output functions such as neural network layers, we need the **Jacobian**.
- This lecture extends the calculus toolkit to second derivatives, curvature, and vector-valued functions.

Learning Objectives

By the end of this lecture you should be able to:

- Construct the second-order Taylor approximation for a multivariate function
- Define and compute the Hessian matrix; interpret its definiteness geometrically
- Classify critical points (minimum / maximum / saddle) using the Hessian
- Define the Jacobian for vector-valued functions and apply the multivariate chain rule

Table of Contents

Taylor Expansion in Multiple Variables

In optimization we often minimize functions of many variables:

$$f : \mathbb{R}^n \to \mathbb{R}.$$

Optimization algorithms need a **local model** of f near a point x.
In one variable we used: $f(x+h) \approx f(x) + f'(x)h$.

Question:
How do we approximate a nonlinear function

$$f(x_1, x_2, \ldots, x_n)$$

near a point x?

The answer is a **multivariable Taylor expansion**.

First-Order Taylor Approximation

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable.
For a small displacement $h \in \mathbb{R}^n$,

$$f(x + h) \approx f(x) + \nabla f(x) \cdot h.$$

Recall:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}.$$

Thus,

$$\nabla f(x)h = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} h_i.$$

Interpretation

- The gradient defines the **best local linear approximation**.
- This generalizes the tangent line to a **tangent plane** (or tangent hyperplane in $\mathbb{R}^n$).

Second-Order Taylor Approximation

To capture **curvature**, we include second derivatives.

For small h,

$$f(x+h) = f(x) + \nabla f(x)h + \frac{1}{2}h^T H(x)h + o(\|h\|^2).$$

Hessian matrix

$$H(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

- Gradient $\rightarrow$ first-order changes
- Hessian $\rightarrow$ curvature of the function

Geometric Interpretation

For $f(x, y)$ the graph is a surface in $\mathbb{R}^3$.

First-order Taylor approximation $f(x + h) \approx f(x) + \nabla f(x)h$ gives the **tangent plane**.

Second-order Taylor approximation

$$f(x + h) \approx f(x) + \nabla f(x)h + \frac{1}{2}h^T H(x)h$$

captures the **local curvature** of the surface.

Optimization insight

- Gradient descent uses the first-order model
- Newton's method uses the second-order model

Taylor expansion provides the local structure used by optimization algorithms.

Table of Contents

Second Derivatives in Multiple Variables

For $f : \mathbb{R}^n \to \mathbb{R}$ we defined first derivatives $\frac{\partial f}{\partial x_i}$.

To study **curvature**, we consider **second partial derivatives**.

$$\frac{\partial^2 f}{\partial x_i^2}, \qquad \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

These describe how the first derivatives change.

Example

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_2}\right).$$

Collecting all second derivatives leads to an important object: the **Hessian matrix**.

Hessian Matrix

For $f : \mathbb{R}^n \to \mathbb{R}$ the **Hessian matrix** is

$$H(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Properties

- If second derivatives are continuous, then

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

- Hence the Hessian is **symmetric**.

The Hessian captures the **second-order structure** of f.

Geometric Meaning of the Hessian

For $f(x, y)$ the graph is a surface in $\mathbb{R}^3$.

Recall the Taylor approximation

$$f(x + h) = f(x) + \nabla f(x)h + \frac{1}{2}h^T H(x)h.$$

Interpretation

- Gradient $\rightarrow$ slope of tangent plane
- Hessian $\rightarrow$ curvature of the surface

Example

Let $f(x, y) = x^2 + y^2$.

$$\nabla f = \begin{bmatrix} 2x & 2y \end{bmatrix}, \qquad H = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

The Hessian shows that the surface is a **bowl-shaped quadratic**.

Table of Contents

Critical Points in Multiple Variables

In one variable we saw: $f'(x_0) = 0$ is a necessary condition for a local extremum.

For a function $f : \mathbb{R}^n \to \mathbb{R}$, the derivative is the gradient.

First–Order Necessary Condition
If f has a local extremum at an interior point x_0 and f is differentiable, then $\nabla f(x_0) = 0$.

Critical (stationary) Points: Points where $\nabla f(x) = 0$

As in one variable, this condition is **necessary but not sufficient**.

Classifying Critical Points Using the Hessian

Suppose x_0 is a critical point: $\nabla f(x_0) = 0$.
To determine its nature we examine the Hessian $H(x_0)$.

Second–Order Test

- If $H(x_0)$ is **positive definite** $\Rightarrow$ x_0 is a local minimum
- If $H(x_0)$ is **negative definite** $\Rightarrow$ x_0 is a local maximum
- If $H(x_0)$ has both positive and negative eigenvalues $\Rightarrow$ saddle point

Interpretation

- Positive definite Hessian $\rightarrow$ locally bowl-shaped
- Negative definite Hessian $\rightarrow$ locally upside-down bowl
- Indefinite Hessian $\rightarrow$ surface bends in opposite directions

Examples

Example: $f(x, y) = x^2 + y^2$

$$\nabla f = \begin{bmatrix} 2x & 2y \end{bmatrix} \implies \text{Critical point: } (0, 0)$$

Hessian:

$$H = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Positive definite $\Rightarrow$ local minimum.

Example: $f(x, y) = x^2 - y^2$

$$H = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

Indefinite $\Rightarrow$ saddle point.

Table of Contents

Vector-Valued Functions

So far we considered functions $f : \mathbb{R}^n \to \mathbb{R}$.

Now consider functions with **vector output** $f : \mathbb{R}^n \to \mathbb{R}^m$.

Such a function can be written as

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{bmatrix}.$$

Each component $f_i(x)$ is a scalar function.

To describe how f changes locally we generalize the gradient.

Jacobian Matrix

For $f : \mathbb{R}^n \to \mathbb{R}^m$ the derivative is the **Jacobian** $J_f(x) \in \mathbb{R}^{m \times n}$.

$$J_f(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Interpretation

The Jacobian represents the **linear map**

$$Df(x) : \mathbb{R}^n \to \mathbb{R}^m$$

that approximates the function locally and is the matrix representation of derivative.

It generalizes the gradient to vector-valued functions.

Example

$$f(x, y) = \begin{bmatrix} x^2 + y \\ xy \end{bmatrix}.$$

Compute partial derivatives.

$$\frac{\partial f_1}{\partial x} = 2x \qquad \frac{\partial f_1}{\partial y} = 1$$

$$\frac{\partial f_2}{\partial x} = y \qquad \frac{\partial f_2}{\partial y} = x$$

Thus the Jacobian is

$$J_f(x, y) = \begin{bmatrix} 2x & 1 \\ y & x \end{bmatrix}.$$

Each row corresponds to the gradient of one component function.

Example

Let: $f(x) = Ax$ with $A \in \mathbb{R}^{m \times n}$

Compute:

$$\frac{\partial f_i}{\partial x_j} = A_{ij}$$

Therefore:

$$\frac{df}{dx} = J_f(x) = A$$

Jacobian of linear function = matrix itself.

Table of Contents

Chain Rule for Vector Functions

Consider the functions $g : \mathbb{R}^n \to \mathbb{R}^m, \quad f : \mathbb{R}^m \to \mathbb{R}^k$.

Define the composite function $h(x) = f(g(x))$.

Question
How do we compute the derivative of h?

The derivative of a vector-valued function is the Jacobian.

Multivariable Chain Rule

$$J_{f \circ g}(x) = J_f(g(x))\, J_g(x)$$

Thus the derivative of a composition is the product of the Jacobians. Note the order: first apply J_g, then J_f.

Example

$$g(x,y) = \begin{bmatrix} x^2 + y \\ xy \end{bmatrix}, \qquad f(u,v) = u + v.$$

First compute Jacobians.

$$J_g(x,y) = \begin{bmatrix} 2x & 1 \\ y & x \end{bmatrix} \qquad J_f(u,v) = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

Using the chain rule: $J_{f\circ g}(x,y) = J_f(g(x,y))J_g(x,y)$.

$$= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 2x & 1 \\ y & x \end{bmatrix}$$

$$= \begin{bmatrix} 2x + y & 1 + x \end{bmatrix}.$$

Key Takeaways

- The Hessian $H_{ij} = \partial^2 f / \partial x_i \partial x_j$ is symmetric for smooth f; it encodes curvature in every direction
- At a critical point ($\nabla f = \mathbf{0}$): $H \succ 0 \Rightarrow$ minimum, $H \prec 0 \Rightarrow$ maximum, indefinite $\Rightarrow$ saddle point
- Jacobian $[J_f]_{ij} = \partial f_i / \partial x_j$ generalises the derivative to $f : \mathbb{R}^n \to \mathbb{R}^m$
- Multivariate chain rule: $J_{f \circ g}(\boldsymbol{x}) = J_f(g(\boldsymbol{x})) \cdot J_g(\boldsymbol{x})$ — this is the foundation of back-propagation

Thank you :)