Optimization Techniques for Machine Learning

AMLZC326 · #05 Vector Calculus I

Anshid Aboobacker

Motivation

- Gradient descent — the engine of machine learning — is just calculus applied iteratively
- Before we can descend, we must understand *what* we are descending: derivatives, Taylor approximations, and gradients
- This lecture builds the univariate and multivariate calculus tools that underpin every optimisation algorithm in the course

Learning Objectives

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

- Compute derivatives of univariate functions using the standard rules and the chain rule
- Use Taylor expansion to approximate a function locally by a polynomial
- Define and compute partial derivatives and the gradient $\nabla f(\boldsymbol{x})$
- Apply the directional derivative and the multivariate chain rule

Table of Contents

Why Derivatives Matter in ML

Core problem:

$$\min_{x \in \mathbb{R}^n} f(x)$$

Optimization algorithms need to know:

- How does f change locally?
- What is the best local linear model of f?

Central idea:

$$\text{Derivative} \implies \text{Best local linear approximation}$$

Examples:

- Linear regression: minimize squared loss
- Neural networks: minimize empirical risk

What is a Derivative?

Two nearby points in a curve: $(x, f(x)), \quad (x+h, f(x+h))$

Average rate of change: $\frac{f(x+h)-f(x)}{h}$

As $h \to 0$:

- Secant line $\to$ Tangent line
- Average slope $\to$ Instantaneous slope

$$\boxed{\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}}$$

Interpretations:

- Instantaneous rate of change
- Slope of tangent line

But this formula hides a deeper structure.

What is this limit really saying about f locally?

Derivative as Local Linear Approximation

Starting from $f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$,
multiply both sides by h: $f(x+h) - f(x) = f'(x)h + \varepsilon(h)h$,
where $\varepsilon(h) \to 0$ as $h \to 0$.

So we can rewrite: $f(x+h) = f(x) + f'(x)h + o(|h|)$.

Interpretation: $f'(x)h$ is the best local linear model of f.
The derivative tells us: $f(x+h) \approx f(x) + f'(x)h$ for small h.

Key idea:
A differentiable function behaves like a linear function locally.

This idea will generalize to $f : \mathbb{R}^n \to \mathbb{R}$ where the derivative becomes a linear map (the gradient).

Example: Derivative of x^n

Let $f(x) = x^n$

Using binomial expansion:

$$(x+h)^n = \sum_{i=0}^{n} \binom{n}{i} x^{n-i} h^i$$

$$\frac{f(x+h) - f(x)}{h} = \sum_{i=1}^{n} \binom{n}{i} x^{n-i} h^{i-1}$$

Taking limit as $h \to 0$:

$$\boxed{\frac{d}{dx} x^n = n x^{n-1}}$$

Differentiation Toolkit

Linearity

- Constant rule: $\frac{d}{dx}c = 0$
- Constant multiple rule: $\frac{d}{dx}\left[c\,u(x)\right] = c\,u'(x)$
- Sum/Difference rule: $\frac{d}{dx}\left[u(x) \pm v(x)\right] = u'(x) \pm v'(x)$

Product and Quotient

- Product rule: $\frac{d}{dx}\left[u(x)v(x)\right] = u(x)v'(x) + v(x)u'(x)$
- Quotient rule (for $v(x) \neq 0$): $\frac{d}{dx}\left(\frac{u(x)}{v(x)}\right) = \frac{v(x)u'(x) - u(x)v'(x)}{[v(x)]^2}$

Chain Rule

If $y = f(g(x))$, then $\frac{dy}{dx} = f'(g(x))\,g'(x)$

Common Derivatives

$\frac{d}{dx}x^n = nx^{n-1}$, $\frac{d}{dx}e^x = e^x$, $\frac{d}{dx}\log x = \frac{1}{x}$, $\frac{d}{dx}(\sin x) = \cos x$, $\frac{d}{dx}(\cos x) = -\sin x$

Chain Rule Example

Let

$$h(x) = (2x + 1)^4.$$

View as composition:

$$g(u) = u^4, \quad u(x) = 2x + 1.$$

Then

$$h'(x) = g'(u(x))\, u'(x) = 4(2x + 1)^3 \cdot 2 = 8(2x + 1)^3.$$

Key Insight: Deep models are compositions. Chain rule powers backpropagation.

Finding Extrema: Where Can Optima Occur?

Local Extremum

- Local minimum at x_0 if $f(x) \geq f(x_0)$ for all x near x_0
- Local maximum at x_0 if $f(x) \leq f(x_0)$ for all x near x_0

First-Order Necessary Condition If f has a local extremum at an interior point x_0 and $f'(x_0)$ exists, then $f'(x_0) = 0$.

Critical Points
Interior points where $f'(x) = 0$ or $f'(x)$ does not exist.

Important: $f'(x) = 0$ is necessary, not sufficient.
Example: $f(x) = x^3$ has $f'(0) = 0$ but no local extremum.

Optimal points occur where the derivative vanishes. This idea generalizes to $\mathbb{R}^n$ using the **gradient**.

Classifying Critical Points

Suppose $f'(x_0) = 0$ and $f''(x_0)$ exists.

Second Derivative Test

- If $f''(x_0) > 0$, then f has a local minimum at x_0
- If $f''(x_0) < 0$, then f has a local maximum at x_0
- If $f''(x_0) = 0$, the test is inconclusive

Optimization View:
Second derivative measures curvature.

- Positive curvature($f''(x_0) > 0$) $\Rightarrow$ convex locally
- Negative curvature($f''(x_0) < 0$) $\Rightarrow$ concave locally

This idea generalizes to the Hessian in $\mathbb{R}^n$.

Table of Contents

Why Taylor Expansion?

Suppose we know:

$$f(x_0), \quad f'(x_0), \quad f''(x_0), \ldots$$

Question:
How well can we approximate $f(x)$ near x_0?

From differentiability:

$$f(x_0 + h) = f(x_0) + f'(x_0)h + o(|h|)$$

Taylor expansion systematically improves this local model.

Taylor Approximation

First-Order Taylor Approximation: $f(x_0 + h) \approx f(x_0) + f'(x_0)h$

Interpretation: Locally, a smooth function behaves like a line.

Used in:

- Gradient descent
- Linearization of nonlinear models

Second-Order Approximation

Including curvature: $f(x_0 + h) \approx f(x_0) + f'(x_0)h + \frac{1}{2}f''(x_0)h^2$

Interpretation:

- First derivative $\rightarrow$ slope
- Second derivative $\rightarrow$ curvature

Used in Newton's method.

Question: Can we improve this approximation?

Taylor Polynomial

Taylor Polynomial of Degree k at x_0

$$T_k(x) = \sum_{n=0}^{k} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

- Built from derivatives at x_0
- Higher degree $\Rightarrow$ better local approximation
- Exact for polynomials of degree $\leq k$

Define the remainder(error):

$$R_k(x) = f(x) - T_k(x)$$

Taylor's Theorem

Let $f : \mathbb{R} \to \mathbb{R}$ be $(k+1)$ times differentiable on an open interval containing x_0 and x, and suppose $f^{(k)}$ is continuous on the closed interval between x_0 and x.
Then

$$f(x) = T_k(x) + R_k(x),$$

where

$$R_k(x) = \frac{f^{(k+1)}(\xi)}{(k+1)!}(x - x_0)^{k+1}$$

for some ξ between x_0 and x.

Meaning: Taylor approximation comes with a precise error term.

Understanding the Error Term

$$R_k(x) = \frac{f^{(k+1)}(\xi)}{(k+1)!}(x - x_0)^{k+1}$$

- Error depends on:
  - $(k+1)$-th derivative
  - Distance from expansion point
- If $\left|f^{(k+1)}(t)\right| \leq M$ on the interval, then

$$|R_k(x)| \leq \frac{M}{(k+1)!}|x - x_0|^{k+1}$$

- Error shrinks rapidly as k increases

Taylor Series

If all derivatives exist,

$$T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

- Infinite polynomial expansion
- Special case $x_0 = 0$: **Maclaurin series**
- If $f(x) = T_\infty(x)$ near x_0, then f is analytic

Example: $f(x) = x^4$

- Compute derivatives at $x_0 = 1$:
  $f(1) = 1$; $f'(1) = 4$; $f''(1) = 12$; $f^{(3)}(1) = 24$; $f^{(4)}(1) = 24$
- Since $f^{(5)}(x) = 0$, all higher derivatives vanish, hence the remainder $R_4(x) = 0$
- Compute Taylor series at $x_0 = 1$:
  $T_4(x) = 1 + 4(x-1) + 6(x-1)^2 + 4(x-1)^3 + (x-1)^4$
- Hence, $T_n(x) = x^4 \quad$ for all $n \geq 4$
- In general, the Taylor polynomial of degree $\geq m$ exactly reproduces any polynomial of degree m.

Example: $f(x) = \sin x + \cos x$

Maclaurin expansions ($x_0 = 0$):

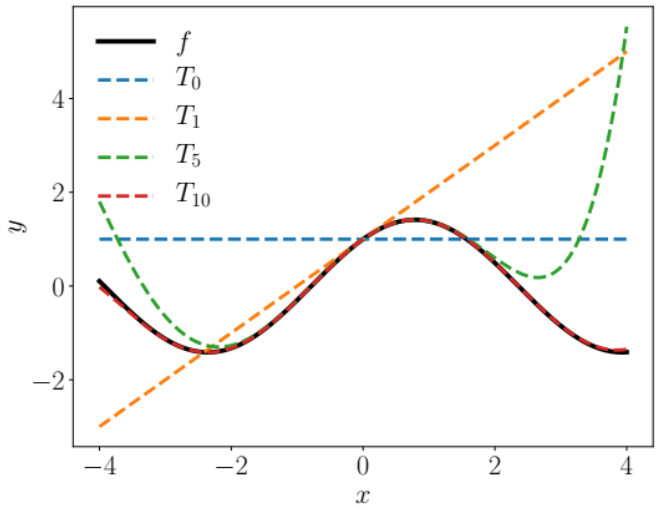
$$\cos x = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k} \qquad \sin x = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1}$$

$$f(x) = \sin x + \cos x$$

- Derivatives are bounded by 1
- Error bound:

$$|R_k(x)| \leq \frac{|x|^{k+1}}{(k+1)!}$$

- Higher order $\Rightarrow$ better approximation near 0
- Taylor series allows us to approximate nonlinear functions by polynomials.



Taylor Expansion and Optimization

Using $x = a + h$:

$$f(a+h) = f(a) + f'(a)h + \frac{1}{2}f''(\xi)h^2$$

Implications:

- Gradient descent uses first-order model
- Newton's method uses second-order model
- Convergence depends on curvature

Taylor expansion is the foundation of optimization algorithms.

Table of Contents

From One Variable to Many

Previously:

$$f : \mathbb{R} \to \mathbb{R}$$

Now:

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

Key Question:
What does differentiability mean in higher dimensions?

We say f is differentiable at x if:

$$f(x+h) = f(x) + L(h) + o(\|h\|)$$

where L is a linear map.

Directional Derivative

Given a direction $v \in \mathbb{R}^n$, $\quad D_v f(x) = \lim_{t \to 0} \dfrac{f(x+tv) - f(x)}{t}$.

If f is differentiable: $D_v f(x) = L(v)$.
So the derivative tells us: How fast does f change in any direction?

Partial Derivatives

Choose direction e_i(coordinate axis) $\quad \dfrac{\partial f}{\partial x_i} = D_{e_i} f(x)$.

So partial derivatives are: Directional derivatives along coordinate directions.
Existence of partial derivatives does NOT guarantee differentiability.

Partial Derivatives

For $f : \mathbb{R}^n \to \mathbb{R}$,

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x)}{h}$$

- Vary only x_i
- Keep other variables fixed
- Ordinary derivative in one coordinate direction

Geometric Meaning

If $y = f(x_1, x_2)$ defines a surface in $\mathbb{R}^3$:

$\frac{\partial f}{\partial x_1}$ = slope along x_1-direction $\quad$ $\frac{\partial f}{\partial x_2}$ = slope along x_2-direction

Each partial derivative measures change along a coordinate axis.

But real change can occur in any direction.

Constructing the Derivative

If f is differentiable, then, $\quad L(h) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} h_i.$

This linear map can be written as: $\quad L(h) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} h.$

This row vector is called the gradient.

The Gradient

For $f : \mathbb{R}^n \to \mathbb{R}$, $\quad \nabla_x f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1 \times n}$

It represents the linear map: $\quad Df(x)[h] = \nabla f(x)\, h.$

- Collects all partial derivatives
- Generalizes the derivative
- Row vector encoding complete first-order changes

Gradient and Steepest Ascent

Directional derivative: $D_v f(x) = \nabla f(x) \cdot v$.

By Cauchy–Schwarz, $|D_v f(x)| \leq \|\nabla f(x)\| \, \|v\|$.
If $\|v\| = 1$, the maximum occurs when $v = \nabla f(x)/\|\nabla f(x)\|$.
Thus, $\max_{\|v\|=1} D_v f(x) = \|\nabla f(x)\|$.

$\Longrightarrow$ The gradient points in the direction of steepest ascent.

Geometric interpretation
Level set: $f(x) = c$.
At a point on the level set, $\nabla f(x) \perp$ level set.
The gradient is normal to contour lines, so optimization moves orthogonally across level sets.

Gradient generalizes the one-variable condition $f'(x) = 0$ to higher dimensions: $f'(x) = 0 \quad \text{to} \quad \nabla f(x) = 0$.
In multiple variables, critical points occur where the gradient vanishes.

Examples: Computing Gradients

Example 1 $f(x, y) = (x + 2y^3)^2$

$$\frac{\partial f}{\partial x} = 2(x + 2y^3), \qquad \frac{\partial f}{\partial y} = 12(x + 2y^3)y^2.$$

$$\nabla f = \begin{bmatrix} 2(x + 2y^3) & 12(x + 2y^3)y^2 \end{bmatrix}.$$

Each partial derivative uses the ordinary chain rule.

Example 2 $f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3$

$$\frac{\partial f}{\partial x_1} = 2x_1 x_2 + x_2^3, \qquad \frac{\partial f}{\partial x_2} = x_1^2 + 3x_1 x_2^2.$$

$$\nabla f = \begin{bmatrix} 2x_1 x_2 + x_2^3 & x_1^2 + 3x_1 x_2^2 \end{bmatrix}.$$

Key Idea: Gradient computation reduces to applying single-variable rules (chain + product) coordinatewise.

Table of Contents

Chain Rule: Scalar Output, Vector Input

$$f : \mathbb{R}^n \to \mathbb{R}, \qquad x(t) \in \mathbb{R}^n.$$

We compute the derivative of f along the path $x(t)$.

Componentwise,

$$\frac{df}{dt} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dx_i}{dt}.$$

Matrix form:

$$\frac{df}{dt} = \underbrace{\nabla f(x)}_{\in \mathbb{R}^{1\times n}} \underbrace{\frac{dx}{dt}}_{\in \mathbb{R}^{n\times 1}} \in \mathbb{R}.$$

Interpretation:

$$\frac{df}{dt} = \nabla f(x) \cdot \dot{x}(t).$$

The derivative along a curve is the gradient dotted with the velocity.

Example

Let

$$f(x_1, x_2) = x_1^2 + 2x_2, \qquad x_1 = \sin t, \quad x_2 = \cos t.$$

Then

$$\nabla f = \begin{bmatrix} 2x_1 & 2 \end{bmatrix}, \qquad \frac{dx}{dt} = \begin{bmatrix} \cos t \\ -\sin t \end{bmatrix}.$$

Thus

$$\frac{df}{dt} = 2x_1 \cos t - 2 \sin t = 2 \sin t \cos t - 2 \sin t = 2 \sin t (\cos t - 1).$$

Key Takeaways

- A derivative $f'(x)$ is the slope of the best local linear approximation to f at x
- Taylor expansion: $f(x+h) \approx f(x) + f'(x)h + \frac{1}{2}f''(x)h^2$ (higher-order terms for greater accuracy)
- The gradient $\nabla f(\boldsymbol{x}) \in \mathbb{R}^n$ collects all partial derivatives and points in the direction of steepest ascent
- Directional derivative $D_{\boldsymbol{u}} f(\boldsymbol{x}) = \nabla f^\top \boldsymbol{u} = \|\nabla f\| \cos\theta$ — gradient descent sets $\boldsymbol{u} = -\nabla f / \|\nabla f\|$

Thank you :)