Mathematics for Machine Learning

— Vector Calculus: Differentiation, Partial Differentiation & Gradients

Joseph Chuang-Chieh Lin

Department of Computer Science & Information Engineering,
Tamkang University

Fall 2023

Credits for the resource

- The slides are based on the textbooks:
  - *Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong: Mathematics for Machine Learning. Cambridge University Press. 2020.*
  - *Howard Anton, Chris Rorres, Anton Kaul: Elementary Linear Algebra. Wiley. 2019.*
- We could partially refer to the monograph:
  Francesco Orabona: A Modern Introduction to Online Learning. https://arxiv.org/abs/1912.13213

Outline

Motivations

- Machine learning algorithms that optimize an objective function w.r.t. a set of model parameters.
- Examples:
  - Curve-fitting.
  - Neural networks (parameters as weights & biases of layers, repeatedly application of chain rule, etc.)
  - Gaussian mixture models (maximizing the likelihood of the model).
- We focus on functions.
  - $f : \mathbb{R}^D \mapsto \mathbb{R}$ (i.e., $\mathbf{x} \mapsto f(\mathbf{x})$).

Example

Get used to

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^2.$$

Example

Get used to

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^2.$$

$$\mathbf{x} \;\mapsto\; x_1^2 + x_2^2.$$

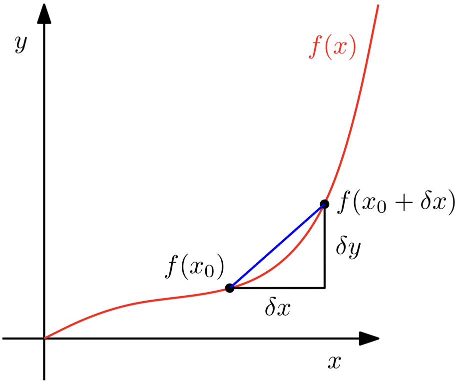
Outline

Derivative

Consider a univariate function $y = f(x)$, $x, y \in \mathbb{R}$.

Difference Quotient

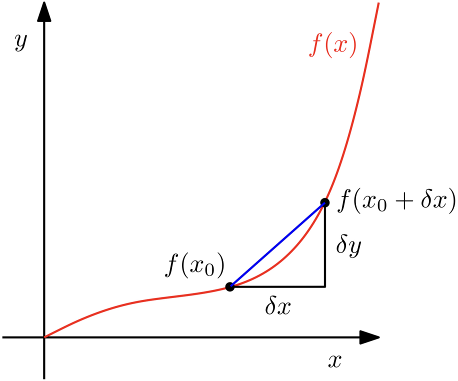
$$\frac{\delta y}{\delta x} := \frac{f(x + \delta x) - f(x)}{\delta x}.$$



Derivative

For $h > 0$, the derivative of f at x:

$$\frac{\mathrm{d}f}{\mathrm{d}x} := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$



Example

Derivative of a Polynomial

Given $f(x) = x^n$.

$$\begin{aligned} \frac{\mathrm{d}f}{\mathrm{d}x} &= \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \\ &= \lim_{h \to 0} \frac{(x+h)^n - x^n}{h} \\ &= \lim_{h \to 0} \frac{\sum_{i=0}^{n} \binom{n}{i} x^{n-i} h^i - x^n}{h} \end{aligned}$$

Note that $x^n = \binom{n}{0} x^{n-0} h^0$.

Example

Derivative of a Polynomial

Given $f(x) = x^n$.

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \lim_{h \to 0} \frac{\sum_{i=1}^{n} \binom{n}{i} x^{n-i} h^i}{h}$$

Example

Derivative of a Polynomial

Given $f(x) = x^n$.

$$
\begin{aligned}
\frac{\mathrm{d}f}{\mathrm{d}x} &= \lim_{h \to 0} \frac{\sum_{i=1}^{n} \binom{n}{i} x^{n-i} h^i}{h} \\
&= \lim_{h \to 0} \sum_{i=1}^{n} \binom{n}{i} x^{n-i} h^{i-1} \\
&= \lim_{h \to 0} \binom{n}{1} x^{n-1} + \lim_{h \to 0} \sum_{i=2}^{n} \binom{n}{i} x^{n-i} h^{i-1} \\
&= nx^{n-1} + 0.
\end{aligned}
$$

Taylor Series

The Taylor polynomial of degree n of $f : \mathbb{R} \mapsto \mathbb{R}$ at x_0 is:

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

Taylor Series

For a function $f : \mathbb{R} \mapsto \mathbb{R}, f \in \mathcal{C}^{\infty}$, the Taylor series f at x_0 is:

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

Taylor Series

For a function $f : \mathbb{R} \mapsto \mathbb{R}, f \in \mathcal{C}^{\infty}$, the Taylor series f at x_0 is:

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

For $x_0 = 0$, it is the *Maclaurin series*.

Taylor Series

For a function $f : \mathbb{R} \mapsto \mathbb{R}, f \in \mathcal{C}^{\infty}$, the Taylor series f at x_0 is:

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

For $x_0 = 0$, it is the *Maclaurin series*.

f is analytic: $f(x) = T_{\infty}(x)$.

Example

Example

$f(x) = x^4$. Seek the Taylor polynomial T_6 evaluated at $x_0 = 1$.

Check if $T_6(x) = f(x)$.

$$f'(x) =$$

$$f'(x) = 4x^3,$$

$f'(x) = 4x^3, f''(x) =$

$f'(x) = 4x^3, f''(x) = 12x^2,$

$f'(x) = 4x^3, f''(x) = 12x^2, f^{(3)}(x) = 24x, f^{(4)}(x) = 24,$
$f^{(5)}(x) = f^{(6)}(x) = 0.$

$f'(x) = 4x^3, f''(x) = 12x^2, f^{(3)}(x) = 24x, f^{(4)}(x) = 24,$
$f^{(5)}(x) = f^{(6)}(x) = 0.$

$$\begin{aligned} T_6(x) &= \sum_{k=0}^{6} \frac{f^k(x_0)}{k!}(x - x_0)^k \\ &= 1 + 4(x-1) + 6(x-1)^2 + 4(x-1)^3 + (x-1)^4 + 0 \\ &= x^4. \end{aligned}$$

Example

Example

Given $f(x) = \sin(x) + \cos(x)$. We know $f(x) \in \mathcal{C}^{\infty}$. Seek the Taylor series $T_{\infty}(x)$ evaluated at $x_0 = 0$.

Check if $T_{\infty}(x) = f(x)$.

Example

Example

Given $f(x) = \sin(x) + \cos(x)$. We know $f(x) \in \mathcal{C}^{\infty}$. Seek the Taylor series $T_{\infty}(x)$ evaluated at $x_0 = 0$.

Check if $T_{\infty}(x) = f(x)$.

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

$$f(0) = \sin(0) + \cos(0) = 1$$

$$
\begin{aligned}
f(0) &= \sin(0) + \cos(0) = 1 \\
f'(0) &= \cos(0) - \sin(0) = 1
\end{aligned}
$$

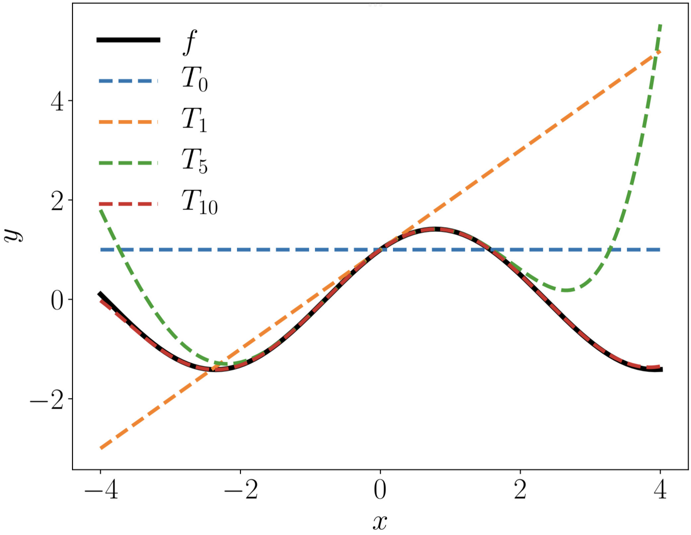
$$
\begin{aligned}
f(0) &= \sin(0) + \cos(0) = 1 \\
f'(0) &= \cos(0) - \sin(0) = 1 \\
f''(0) &= -\sin(0) - \cos(0) = -1
\end{aligned}
$$

$$\begin{aligned} f(0) &= \sin(0)+\cos(0)=1 \\ f'(0) &= \cos(0)-\sin(0)=1 \\ f''(0) &= -\sin(0)-\cos(0)=-1 \\ f^{(3)}(0) &= -\cos(0)+\sin(0)=-1 \end{aligned}$$

$$\begin{aligned} f(0) &= \sin(0) + \cos(0) = 1 \\ f'(0) &= \cos(0) - \sin(0) = 1 \\ f''(0) &= -\sin(0) - \cos(0) = -1 \\ f^{(3)}(0) &= -\cos(0) + \sin(0) = -1 \\ f^{(4)}(0) &= \sin(0) + \cos(0) = 1 \end{aligned}$$

$$
\begin{aligned}
f(0) &= \sin(0) + \cos(0) = 1 \\
f'(0) &= \cos(0) - \sin(0) = 1 \\
f''(0) &= -\sin(0) - \cos(0) = -1 \\
f^{(3)}(0) &= -\cos(0) + \sin(0) = -1 \\
f^{(4)}(0) &= \sin(0) + \cos(0) = 1 \\
&\vdots
\end{aligned}
$$

$$
\begin{aligned}
T_\infty(x) &= \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_o)^k \\
&= 1 + x - \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 - \cdots \\
&= \cos(x) + \sin(x).
\end{aligned}
$$


f
T0
T1
T5
T10
4
2
0
−2
y
−4
−2
0
2
4
x

Differentiation Rules

- $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$.
- $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$.
- $(f(x) + g(x))' = f'(x) + g'(x)$.
- $(g(f(x)))' = (g \circ f)'(x) = g'(f(x))f'(x)$.
  - Chain rule.

- **Example:** Compute $h'(x)$ where $h(x) = (2x+1)^4$.

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

- $h(x) = (2x+1)^4$.
- Let $f(x) = 2x + 1$, $g(f) = f^4$.

- $h(x) = (2x+1)^4$.
- Let $f(x) = 2x+1$, $g(f) = f^4$.
- $f'(x) = 2$,

- $h(x) = (2x + 1)^4$.
- Let $f(x) = 2x + 1$, $g(f) = f^4$.
- $f'(x) = 2$, $g'(f) = 4f^3$.

- $h(x) = (2x+1)^4$.
- Let $f(x) = 2x+1$, $g(f) = f^4$.
- $f'(x) = 2$, $g'(f) = 4f^3$.
- $h'(x) = g'(f)f'(x) =$

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

Outline

Motivation

- We consider a more general case: $f : \mathbb{R}^n \mapsto \mathbb{R}$.
  - The derivative to functions of several variables $\Rightarrow$ gradient.

Partial Derivative

Partial Derivative

For a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $x \in \mathbb{R}^n$ of n variables $x_1, \ldots, x_n$, the partial derivatives are:

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(\mathbf{x})}{h} \\ &\vdots \\ \frac{\partial f}{\partial x_n} &= \lim_{h \to 0} \frac{f(x_1, \ldots, x_{n-1}, x_n + h) - f(\mathbf{x})}{h} \end{aligned}$$

We collect them in the row vector:

$$\nabla_{\mathbf{x}} f = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

Partial Derivative

Partial Derivative

For a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $x \in \mathbb{R}^n$ of n variables $x_1, \ldots, x_n$, the partial derivatives are:

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(\mathbf{x})}{h} \\ &\vdots \\ \frac{\partial f}{\partial x_n} &= \lim_{h \to 0} \frac{f(x_1, \ldots, x_{n-1}, x_n + h) - f(\mathbf{x})}{h} \end{aligned}$$

We collect them in the row vector:

$$\nabla_{\mathbf{x}} f = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}} = \left[\frac{\partial f(\mathbf{x})}{\partial x_1} \; \frac{\partial f(\mathbf{x})}{\partial x_2} \; \cdots \; \frac{\partial f(\mathbf{x})}{\partial x_n} \right] \in \mathbb{R}^{1 \times n},$$

where $\mathbf{x} = [x_1, \ldots, x_n]^\top$.

Examples

Example

Given $f(x, y) = (x + 2y^3)^2$, compute $\frac{\partial f(x,y)}{\partial x}$ and $\frac{\partial f(x,y)}{\partial y}$.

Example

Given $f(x, y) = x_1^2 x_2 + x_1 x_2^3 \in \mathbb{R}$, compute $\frac{\partial f(x,y)}{\partial x}$, $\frac{\partial f(x,y)}{\partial y}$ and $\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}}$.

Basic Partial Differentiation Rules

- $\frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x})g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}} g(x) + f(x) \frac{\partial g}{\partial \mathbf{x}}$.
- $\frac{\partial}{\partial \mathbf{x}}(f(x) + g(x)) = \frac{\partial f}{\partial \mathbf{x}} + \frac{\partial g}{\partial \mathbf{x}}$.
- $\frac{\partial}{\partial \mathbf{x}}(g \circ f)(\mathbf{x}) = \frac{\partial g}{\partial \mathbf{x}}(g(f(\mathbf{x}))) = \frac{\partial g}{\partial f} \frac{\partial f}{\partial \mathbf{x}}$.
  - Chain rule.

Chain Rule (Partial Differentiation)

- Consider a function $f : \mathbb{R}^2 \mapsto \mathbb{R}$ of two variables x_1, x_2.
  - $x_1(t), x_2(t) : \mathbb{R} \mapsto \mathbb{R}$.

Then,

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}.$$

Here 'd' denotes the gradient and '∂' denotes partial derivatives.

- **Note:** Here the 't' in dt is in $\mathbb{R}^1$.
- Trick: View $[x_1, x_2]^\top$ as $\mathbf{x} \in \mathbb{R}^2$.
  $\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}}$: $\mathbb{R}$ w.r.t. $\mathbb{R}^2$.
  $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}$: $\mathbb{R}^2$ w.r.t. $\mathbb{R}$.

Example

Example

Consider $f(x_1, x_2) = x_1^2 + 2x_2$, where $x_1 = \sin t$ and $x_2 = \cos t$. Calculate

$$\frac{\mathrm{d}f}{\mathrm{d}t} = ?$$

What if $x_1, x_2 : \mathbb{R}^2 \mapsto \mathbb{R}$?

- Again, consider a function $f : \mathbb{R}^2 \mapsto \mathbb{R}$ of two variables x_1, x_2. However,

What if $x_1, x_2 : \mathbb{R}^2 \mapsto \mathbb{R}$?

- Again, consider a function $f : \mathbb{R}^2 \mapsto \mathbb{R}$ of two variables x_1, x_2. However,
  - $x_1(s, t), x_2(s, t) : \mathbb{R}^2 \mapsto \mathbb{R}$.

  Then,

$$\begin{aligned} \frac{\partial f}{\partial s} &= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s}, \\ \frac{\partial f}{\partial t} &= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}, \end{aligned}$$

  - Trick: View $[x_1, x_2]^\top$ as $\mathbf{x} \in \mathbb{R}^2$ and $[s, t]^\top$ as $\boldsymbol{\theta} \in \mathbb{R}^2$.
    $\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}}$: $\mathbb{R}$ w.r.t. $\mathbb{R}^2$.
    $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\boldsymbol{\theta}}$: $\mathbb{R}^2$ w.r.t. $\mathbb{R}^2$.

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}} = \frac{\partial f}{\partial \mathbf{x}}\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}} =$$

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}} = \frac{\partial f}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix}$$

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}} = \frac{\partial f}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix}.$$

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}} = \frac{\partial f}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix}.$$

Somehow we can see why the gradient is defined as a row vector.

Heads up

We will see that

- $f : \mathbb{R}^D \mapsto \mathbb{R}$: the gradient is a $1 \times D$ row vector.
- $\mathbf{f} : \mathbb{R} \mapsto \mathbb{R}^E$: the gradient is a $E \times 1$ column vector.
- $\mathbf{f} : \mathbb{R}^D \mapsto \mathbb{R}^E$: the gradient is a $E \times D$ matrix.

Discussions