Linear Algebra

Vector Derivation

Department of Computer Engineering

Sharif University of Technology

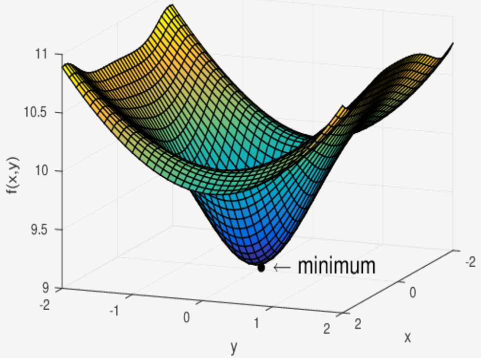
Maryam Ramezani maryam.ramezani@sharif.edu

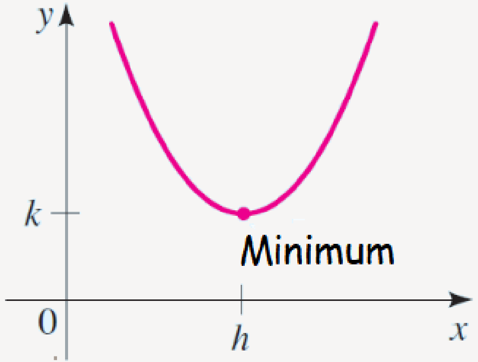
01

Introduction

Motivation

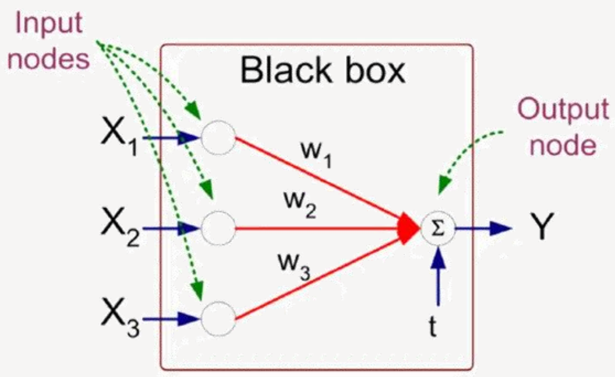
- Machine Learning training requires one to evaluate how one vector changes with respect to another?
- How output changes with respect to parameters?
- How do we find minimum of a scalar function?
- How do we find minimum of two variables?



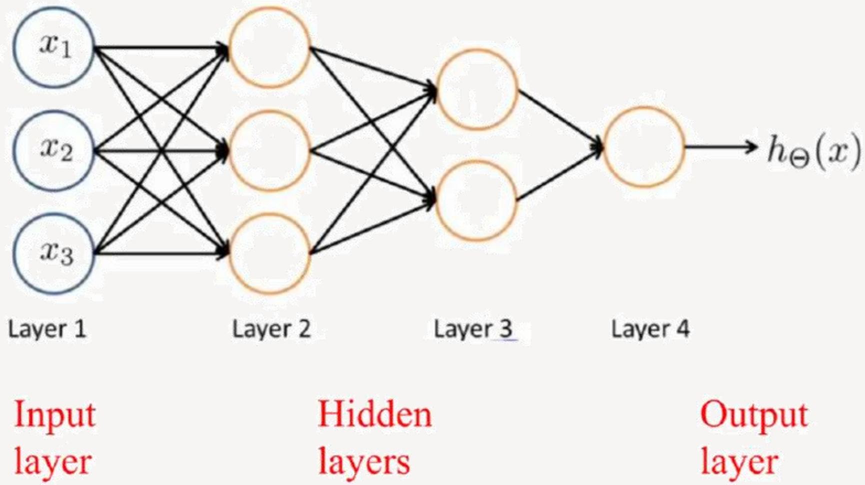


Neural Network

A single neuron

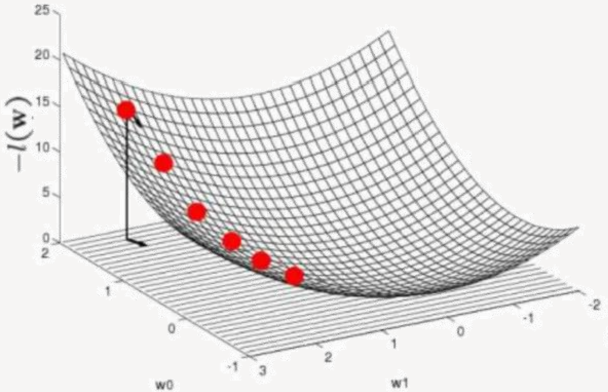


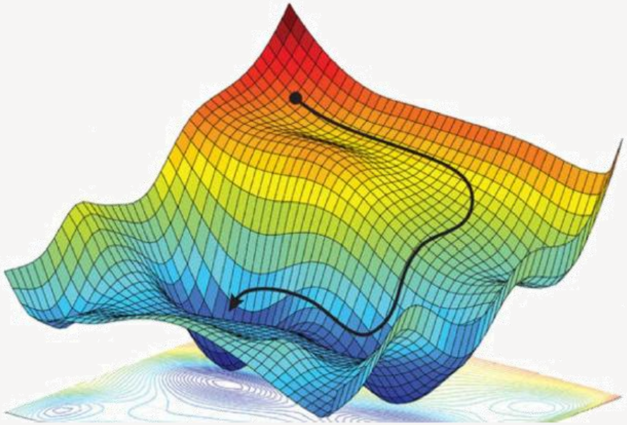
A neural network



ML Optimization

- Optimizing the weights of a neural network, or more generally the parameters of a machine learning model, can be an extremely complex task.
- Many tools have been developed for this purpose. The core of these tools relies on the use of "local information," such as derivatives (gradients) and similar methods.
- Here, the problem is to search for and find the optimal weights in a continuous space, which has an infinite number of potential candidates. Such a problem is also referred to as Continuous





Different Functions

- Scalar Function $f: \mathbb{R} \to \mathbb{R}$
- Scalar Field $f: \mathbb{R}^n \to \mathbb{R}$ or $f: \mathbb{R}^{n\times k} \to \mathbb{R}$ or $f: \mathbb{R} \to \mathbb{R}$
- Vector Field $f: \mathbb{R}^n \to \mathbb{R}^m$ or $f: \mathbb{R}^{n\times k} \to \mathbb{R}^m$ or $f: \mathbb{R} \to$
- Matrix Field $f: \mathbb{R}^n \to \mathbb{R}^{n\times m}$ or $f: \mathbb{R}^{n\times k} \to \mathbb{R}^{p\times m}$ or $f: \mathbb{R} \to \mathbb{R}^{p\times m}$
- Tensor Field $f: scalar, vector, matrix \to \mathbb{R}^{n\times m\times k}$

In higher dimensions, if we take the derivative of a scalar field, it will result in a scalar field (Gradient). If we take the derivative again, it will result in a matrix-valued function (Hessian).

Overview

Types of matrix derivative

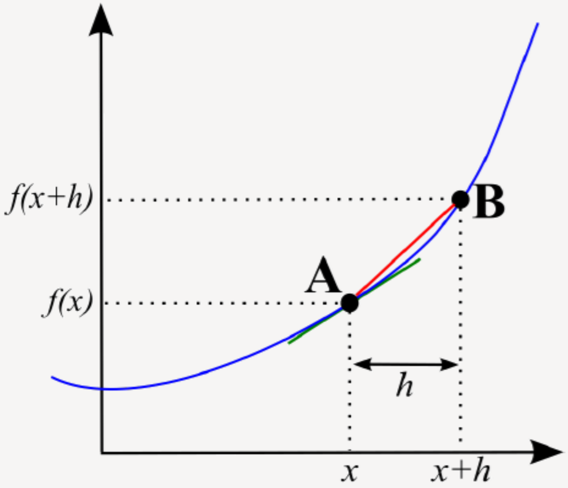
Types	Scalar	Vector	Matrix
Scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{Y}}{\partial x}$
Vector	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
Matrix	$\frac{\partial y}{\partial \mathbf{X}}$	Tensor! (Optional part of this course)	

02

Scalar function Derivation

Overview

$$\boxed{f'(x) \quad \text{or} \quad \frac{\mathrm{d}f}{\mathrm{d}x}(x) \;=\; \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}}$$



- A derivative, which itself is a function $f : \mathbb{R} \rightarrow \mathbb{R}$, stores local/instantaneous information about changes in the function.
- Note that the derivative may not be defined at certain points (or anywhere at all). Functions that are differentiable throughout their domain are referred to as differentiable.

Simple Rules

1. Constant Rule : $\frac{d}{dx}(c) = 0$

2. Constant Multiple Rule : $\frac{d}{dx}[cf(x)] = cf'(x)$

3. Power Rule : $\frac{d}{dx}(x^n) = nx^{n-1}$

4. Sum Rule : $\frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x)$

5. Difference Rule : $\frac{d}{dx}[f(x) - g(x)] = f'(x) - g'(x)$

6. Product Rule : $\frac{d}{dx}[f(x)g(x)] = f(x)g'(x) + g(x)f'(x)$

7. Quotient Rule : $\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)f'(x) - f(x)g'(x)}{[g(x)]^2}$

8. Chain Rule : $\frac{d}{dx}f[g(x)] = f'[g(x)]g'(x)$

Well-Behaved Functions in Differentiation

A function is considered well-behaved if it satisfies these criteria:

- Continuity: The function is continuous across its domain (no jumps or breaks).
- Differentiability: The function is differentiable at every point in its domain (no sharp corners).
- Smoothness: The derivative is also continuous, ensuring smooth transitions.

Well-Behaved Functions in Differentiation

Examples of Well-Behaved Functions

- Polynomials: $f(x) = x^2$, $f(x) = 3x^3 + 2x - 5$
- Trigonometric: $f(x) = \sin(x)$, $f(x) = \cos(x)$
- Exponential: $f(x) = e^x$
- Logarithmic (defined domain): $f(x) = \ln(x)$, $x > 0$

Non-Well-Behaved Functions

- Discontinuous: $f(x) = \frac{1}{x}$ at $x = 0$
- Sharp Points: $f(x) = |x|$ at $x = 0$
- Oscillatory: $f(x) = x\sin(1/x)$ at $x = 0$

Interpretation of First and Second Derivatives

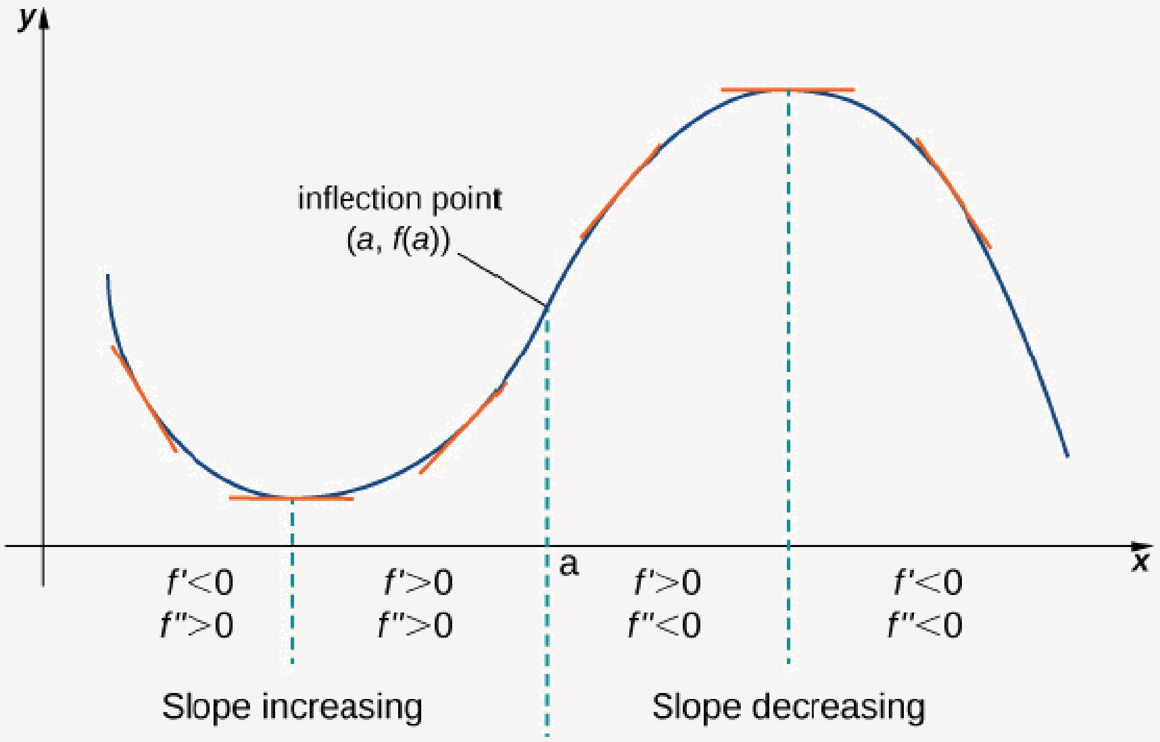
Assume f is a function that is at least twice differentiable, meaning f and f' are both differentiable.

Points where $f'(x) = 0$ are called stable points of f. Note that a function may have no stable points, a finite number of stable points, or an infinite number of them!

At a stable point x^* for f:

- If $f''(x^*) > 0$, the point is a local minimum.
- If $f''(x^*) < 0$, the point is a local maximum.
- If $f''(x^*) = 0$, we cannot determine the nature of the point based solely on the second derivative and must analyze higher-order derivatives.

Interpretation of First and Second Derivatives



Taylor series: Estimating a Function with a Polynomial

Assume that f is a well-behaved function, meaning it is infinitely differentiable (this is a very strong condition but can sometimes be relaxed). Also, assume that $x_0 \in \mathbb{R}$ is a fixed and desired point on the real number line.

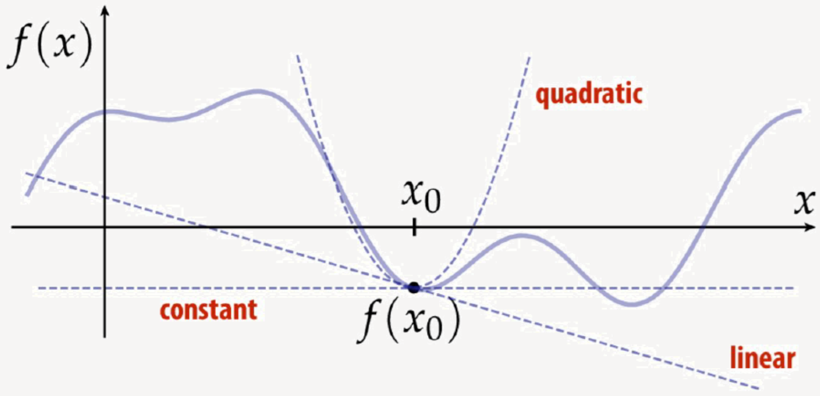
Under these conditions, and for some x (sometimes for all $x \in \mathbb{R}$), we have:

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

The Taylor series of $f(x)$, even for points far away from x, provides an approximation of $f(x)$ based on the local information at the point x_0.

Estimating a Function with a Polynomial

$$f(x) = \overset{\text{constant}}{f(x_0)} + \overset{\text{linear}}{f'(x_0)(x - x_0)} + \overset{\text{quadratic}}{\frac{(x-x_0)^2}{2!} f''(x_0)} + \cdots$$



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

Taylor series Example

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots$$

$$x^3 - 3x + 1 = x^3 - 3x + 1$$

Taylor series Example

We consider the polynomial

$$f(x) = x^4 \tag{5.9}$$

and seek the Taylor polynomial T_6, evaluated at $x_0 = 1$. We start by computing the coefficients $f^{(k)}(1)$ for $k = 0, \ldots, 6$:

$$f(1) = 1 \tag{5.10}$$

$$f'(1) = 4 \tag{5.11}$$

$$f''(1) = 12 \tag{5.12}$$

$$f^{(3)}(1) = 24 \tag{5.13}$$

$$f^{(4)}(1) = 24 \tag{5.14}$$

$$f^{(5)}(1) = 0 \tag{5.15}$$

$$f^{(6)}(1) = 0 \tag{5.16}$$

Therefore, the desired Taylor polynomial is

$$T_6(x) = \sum_{k=0}^{6} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{5.17a}$$

$$= 1 + 4(x-1) + 6(x-1)^2 + 4(x-1)^3 + (x-1)^4 + 0\,. \tag{5.17b}$$

Multiplying out and re-arranging yields

$$\begin{aligned} T_6(x) &= (1 - 4 + 6 - 4 + 1) + x(4 - 12 + 12 - 4) \\ &\quad + x^2(6 - 12 + 6) + x^3(4 - 4) + x^4 && (5.18a) \\ &= x^4 = f(x)\,, && (5.18b) \end{aligned}$$

i.e., we obtain an exact representation of the original function.

03

Scalar Field Derivation

Scalar with respect to scalar

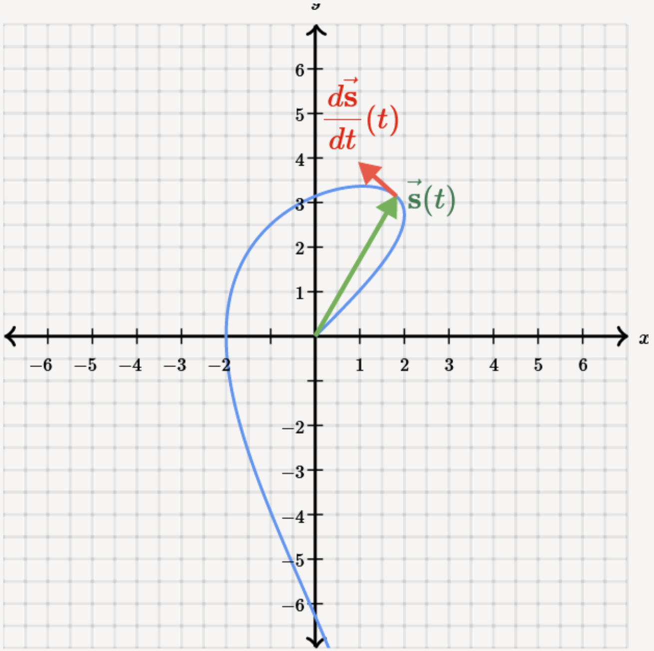
Vector-Valued Function

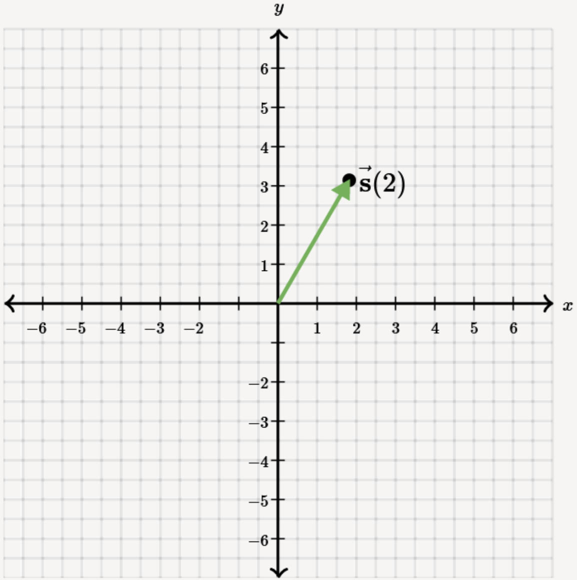
https://youtu.be/GkB4vW16QH

Directional Derivative

$$\vec{\mathbf{s}}(t) = \begin{bmatrix} 2\sin(t) \\ \\ 2\cos(t/3)t \end{bmatrix}$$

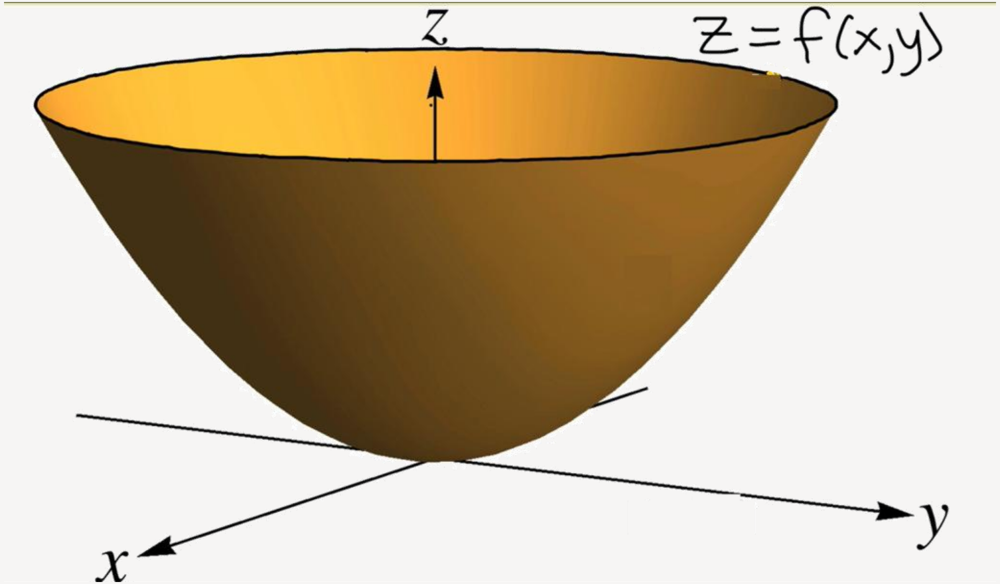
$$\vec{\mathbf{s}}(2) = \begin{bmatrix} 2\sin(2) \\ \\ 2\cos(2/3)\cdot 2 \end{bmatrix} \approx \begin{bmatrix} 1.819 \\ \\ 3.144 \end{bmatrix}$$





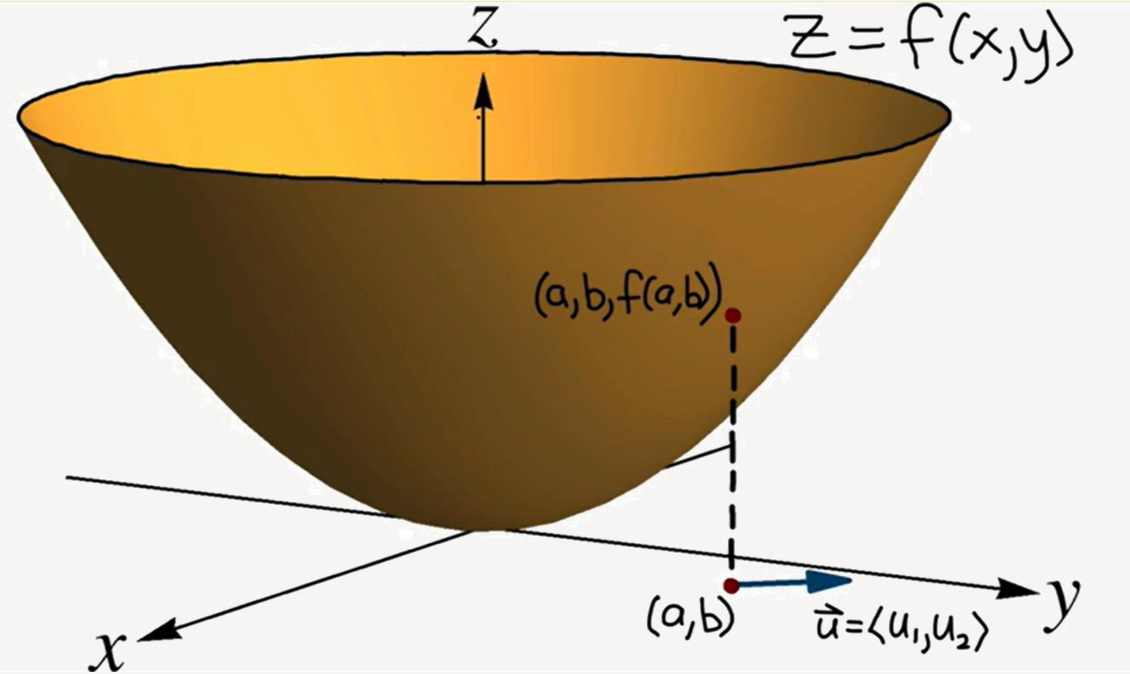
Directional Derivative

- Example $D_{\vec{u}} f(a,b) = \vec{\nabla} f(a,b) \cdot \vec{u}.$



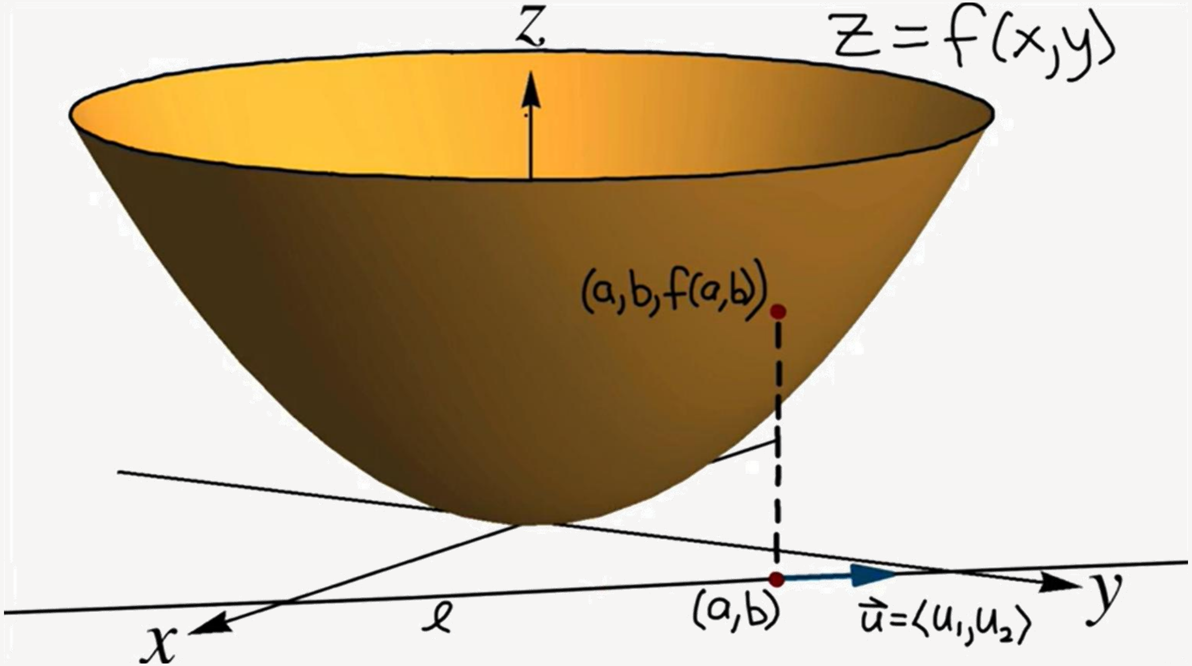
Directional Derivative

- Example $D_{\vec{u}} f(a,b) = \vec{\nabla} f(a,b) \cdot \vec{u}.$



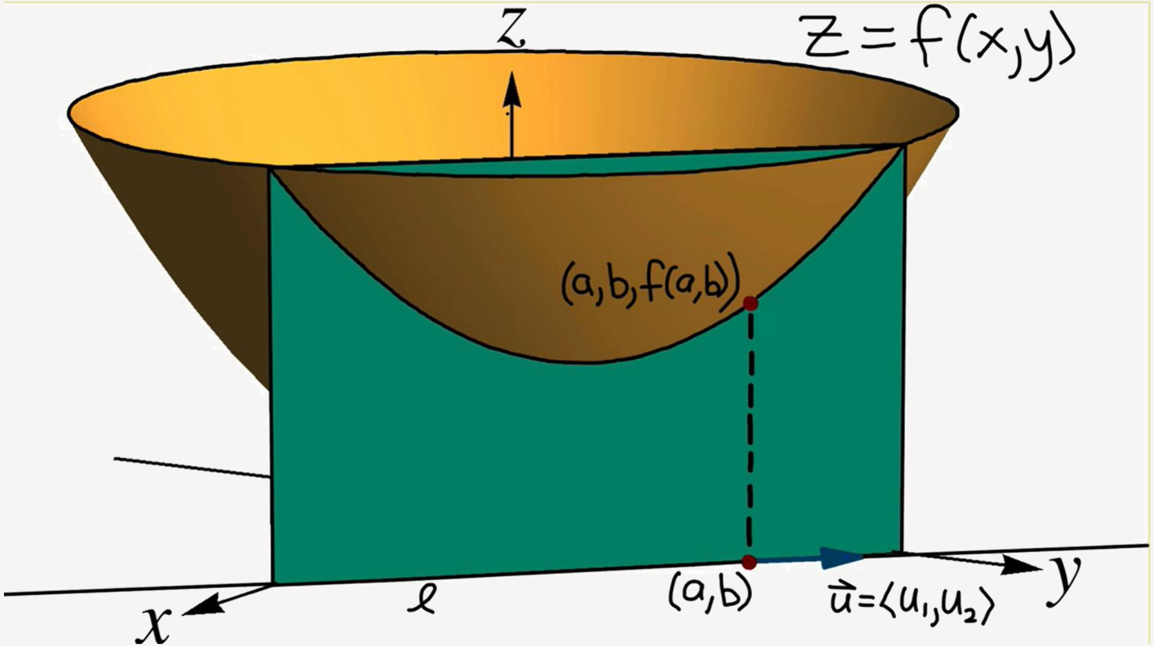
Directional Derivative

- Example $D_{\vec{u}} f(a,b) = \vec{\nabla} f(a,b) \cdot \vec{u}.$



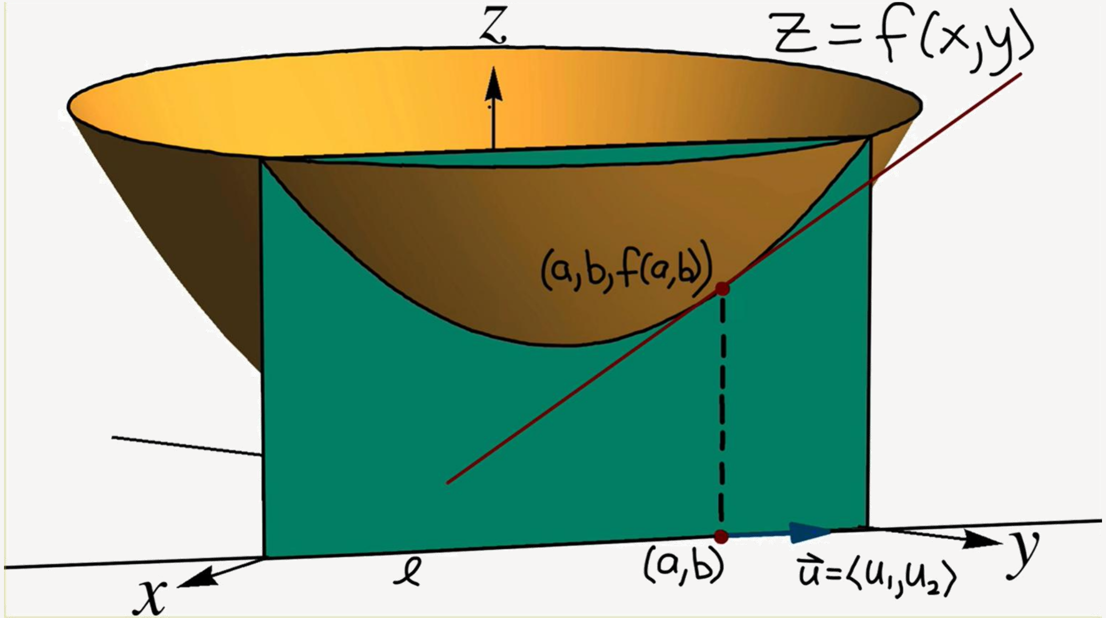
Directional Derivative

- Example $D_{\vec{u}} f(a,b) = \vec{\nabla} f(a,b) \cdot \vec{u}$.



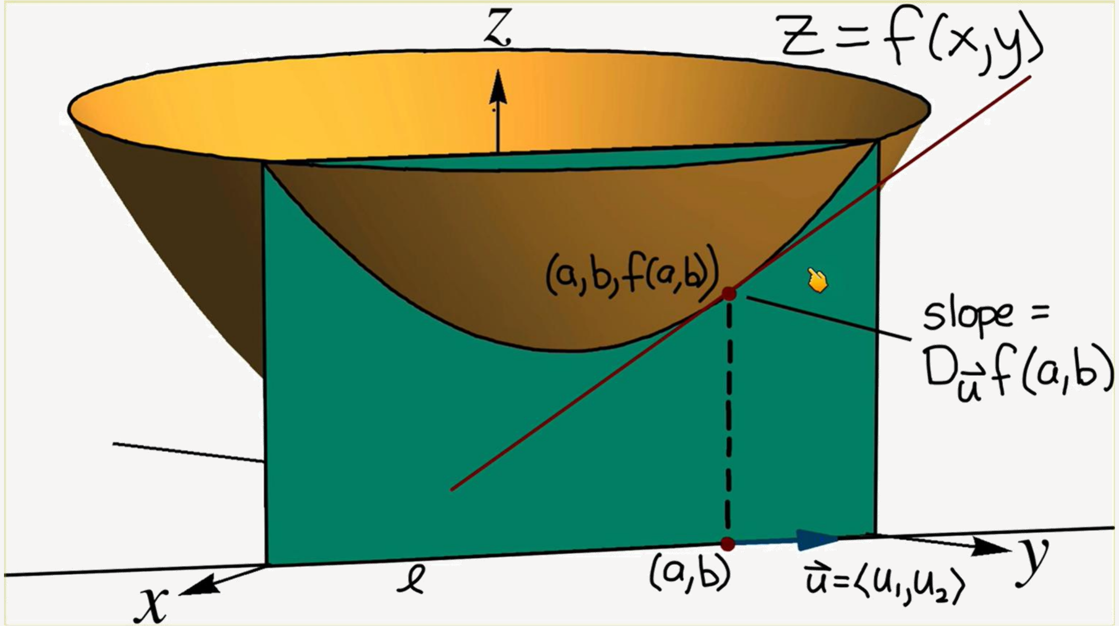
Directional Derivative

- Example $D_{\vec{u}} f(a,b) = \vec{\nabla} f(a,b) \cdot \vec{u}.$



Directional Derivative

- Example $D_{\vec{u}} f(a,b) = \vec{\nabla} f(a,b) \cdot \vec{u}$.



Directional Derivative

$$f: \mathbb{R}^n \to \mathbb{R} \qquad v = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$D_v f = \mathrm{v}.\nabla f$$

$$\nabla_{\vec{\mathbf{v}}} f(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\vec{\mathbf{v}}) - f(\mathbf{x})}{h||\vec{\mathbf{v}}||}$$

Scalar with respect to vector

Definition 5.5 (Partial Derivative). For a function $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{x} \mapsto f(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$ of n variables $x_1, \ldots, x_n$ we define the *partial derivatives* as

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(\boldsymbol{x})}{h} \\ &\vdots \\ \frac{\partial f}{\partial x_n} &= \lim_{h \to 0} \frac{f(x_1, \ldots, x_{n-1}, x_n + h) - f(\boldsymbol{x})}{h} \end{aligned} \tag{5.39}$$

and collect them in the row vector

$$\nabla_{\boldsymbol{x}} f = \operatorname{grad} f = \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} & \frac{\partial f(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1 \times n}, \tag{5.40}$$

The row vector in (5.40) is called the *gradient* of f or the *Jacobian*

Note!

Example

- $\dfrac{\partial(\boldsymbol{x}^T\boldsymbol{a})}{\partial \boldsymbol{x}} = \boldsymbol{a}^T$

Remark (Gradient as a Row Vector). It is not uncommon in the literature to define the gradient vector as a column vector, following the convention that vectors are generally column vectors. The reason why we define the gradient vector as a row vector is twofold: First, we can consistently generalize the gradient to vector-valued functions $f : \mathbb{R}^n \to \mathbb{R}^m$ (then the gradient becomes a matrix). Second, we can immediately apply the multi-variate chain rule without paying attention to the dimension of the gradient.

Rules

Product rule: $$\frac{\partial}{\partial \boldsymbol{x}}\left(f(\boldsymbol{x})g(\boldsymbol{x})\right) = \frac{\partial f}{\partial \boldsymbol{x}} g(\boldsymbol{x}) + f(\boldsymbol{x}) \frac{\partial g}{\partial \boldsymbol{x}}$$

Sum rule: $$\frac{\partial}{\partial \boldsymbol{x}}\left(f(\boldsymbol{x}) + g(\boldsymbol{x})\right) = \frac{\partial f}{\partial \boldsymbol{x}} + \frac{\partial g}{\partial \boldsymbol{x}}$$

Chain rule: $$\frac{\partial}{\partial \boldsymbol{x}}(g \circ f)(\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}}\left(g(f(\boldsymbol{x}))\right) = \frac{\partial g}{\partial f} \frac{\partial f}{\partial \boldsymbol{x}}$$

Chain Rule

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

- **Example 1:**

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

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}t} &= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} \\ &= 2\sin t\frac{\partial \sin t}{\partial t} + 2\frac{\partial \cos t}{\partial t} \\ &= 2\sin t\cos t - 2\sin t = 2\sin t(\cos t - 1)\end{aligned}$$

is the corresponding derivative of f with respect to t.

Chain Rule

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

- **Example 2:**

If $f(x_1, x_2)$ is a function of x_1 and x_2, where $x_1(s,t)$ and $x_2(s,t)$ are themselves functions of two variables s and t, the chain rule yields the partial derivatives

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s},$$
$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t},$$

and the gradient is obtained by the matrix multiplication

$$\frac{\mathrm{d}f}{\mathrm{d}(s,t)} = \frac{\partial f}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial (s,t)} = \underbrace{\begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix}}_{=\frac{\partial f}{\partial \boldsymbol{x}}} \underbrace{\begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix}}_{=\frac{\partial \boldsymbol{x}}{\partial (s,t)}}.$$

Scalar with respect to matrix

The derivative of a scalar y by a matrix $X \in \mathbb{R}^{m\times n}$ is given by:

$$\frac{\partial y}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial X_{11}} & \frac{\partial y}{\partial X_{21}} & \cdots & \frac{\partial y}{\partial X_{m1}} \\ \frac{\partial y}{\partial X_{12}} & \frac{\partial y}{\partial X_{22}} & \cdots & \frac{\partial y}{\partial X_{m2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial X_{1n}} & \frac{\partial y}{\partial X_{2n}} & \cdots & \frac{\partial y}{\partial X_{mn}} \end{bmatrix}$$

04

Vector Field Derivation

Vector with respect to vector

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

For a function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $\boldsymbol{x} = [x_1, \ldots, x_n]^\top \in \mathbb{R}^n$, the corresponding vector of function values is given as

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^m .$$

The differentiation rules for every f_i are exactly the ones we discussed in section 03

Why this happen??

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

Vector with respect to vector

$$\frac{\partial \boldsymbol{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} = \begin{bmatrix} \lim_{h\to 0} \frac{f_1(x_1,\ldots,x_{i-1},x_i+h,x_{i+1},\ldots x_n)-f_1(\boldsymbol{x})}{h} \\ \vdots \\ \lim_{h\to 0} \frac{f_m(x_1,\ldots,x_{i-1},x_i+h,x_{i+1},\ldots x_n)-f_m(\boldsymbol{x})}{h} \end{bmatrix} \in \mathbb{R}^m$$

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

$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

Jacobian Matrix

Vector with respect to scalar

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

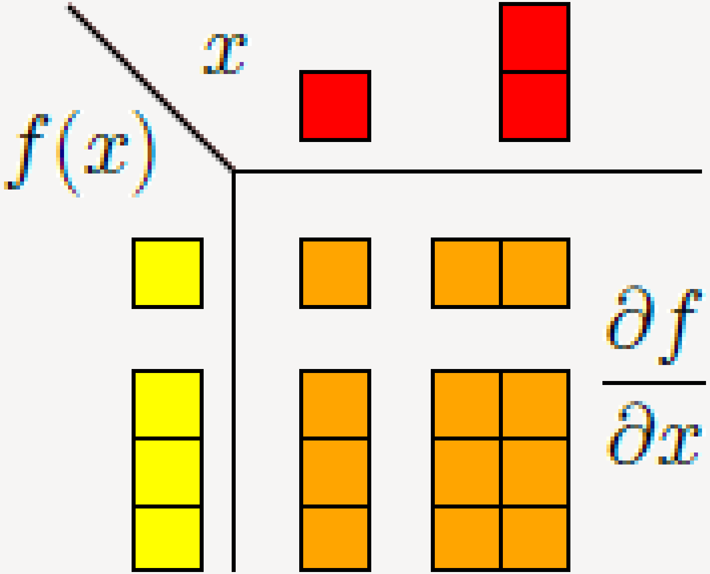
For a function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $\boldsymbol{x} = [x_1, \ldots, x_n]^\top \in \mathbb{R}^n$, the corresponding vector of function values is given as

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^m .$$

The differentiation rules for every f_i are exactly the ones we discussed in section 03

- If $x \in \mathbb{R}$ is a scalar, then it is a column vector $= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x} \\ \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x} \end{bmatrix}$

Dimensionality of (partial) derivatives



If $f : \mathbb{R} \to \mathbb{R}$ the gradient is simply a scalar (top-left entry). For $f : \mathbb{R}^D \to \mathbb{R}$ the gradient is a $1 \times D$ row vector (top-right entry). For $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^E$, the gradient is an $E \times 1$ column vector, and for $\boldsymbol{f} : \mathbb{R}^D \to \mathbb{R}^E$ the gradient is an $E \times D$ matrix.

Hessian Matrix

Suppose that $f(x) : \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in $\mathbb{R}^n$ and returns a real number. Then the Hessian matrix with respect to x, written $\nabla_x^2 f(x)$ or simply as H is the $n \times n$ matrix of partial derivatives,

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

In other words, $\nabla_x^2 f(x) \in \mathbb{R}^{n \times n}$, with

$$(\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

Note that the Hessian is always symmetric, since

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

05

Matrix Field Derivation

Matrix with respect to scalar

The derivative of a matrix $Y \in \mathbb{R}^{m \times n}$ by a scalar x is given by:

$$\frac{\partial Y}{\partial x} = \begin{bmatrix} \frac{\partial Y_{11}}{\partial x} & \frac{\partial Y_{12}}{\partial x} & \cdots & \frac{\partial Y_{1n}}{\partial x} \\ \frac{\partial Y_{21}}{\partial x} & \frac{\partial Y_{22}}{\partial x} & \cdots & \frac{\partial Y_{2n}}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Y_{m1}}{\partial x} & \frac{\partial Y_{m2}}{\partial x} & \cdots & \frac{\partial Y_{mn}}{\partial x} \end{bmatrix}$$

06

Beautiful Examples!

Important note on product Rule

Product rule: $$\frac{\partial}{\partial \boldsymbol{x}}(f(\boldsymbol{x})g(\boldsymbol{x})) = \frac{\partial f}{\partial \boldsymbol{x}} g(\boldsymbol{x}) + f(\boldsymbol{x})\frac{\partial g}{\partial \boldsymbol{x}}$$

Note. Please pay attention to following example!

- $$\frac{\partial(x^T y)}{\partial z} = x^T \frac{\partial(y)}{\partial z} + y^T \frac{\partial(x)}{\partial z}$$

- if x and y be vectors which elements are function of vector z

Let's practice

- $\frac{\partial(\boldsymbol{u}(\boldsymbol{x})+\boldsymbol{v}(\boldsymbol{x}))}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{u}(\boldsymbol{x})}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{v}(\boldsymbol{x})}{\partial \boldsymbol{x}}$
- $\frac{\partial(\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{A}$
- $\frac{\partial(\boldsymbol{x}^T\boldsymbol{a})}{\partial \boldsymbol{x}} = \boldsymbol{a}^T$
- $\frac{\partial(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{x}^T(\boldsymbol{A} + \boldsymbol{A}^T)$
- $\frac{\partial(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\boldsymbol{x}^T\boldsymbol{A}$ if $\boldsymbol{A}$ is symmetric

Hint!

$$A\vec{x} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_1 x_1 + a_2 x_2 \\ a_3 x_1 + a_4 x_2 \end{bmatrix}$$

$$\frac{dA\vec{x}}{dx} = \begin{bmatrix} \frac{\partial(a_1 x_1 + a_2 x_2)}{\partial x_1} & \frac{\partial(a_1 x_1 + a_2 x_2)}{\partial x_2} \\ \frac{\partial(a_3 x_1 + a_4 x_2)}{\partial x_1} & \frac{\partial(a_3 x_1 + a_4 x_2)}{\partial x_2} \end{bmatrix}$$

$$= \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} = A$$

Let's practice

- $\frac{\partial (A(t))^{-1}}{\partial t} = -A(t)^{-1}\,\frac{\partial (A(t))}{\partial t}\,A(t)^{-1}$
- $\frac{\partial \det(A)}{\partial A} = \det(A)\,A^{-1}$
- $\frac{\partial \ln(\det(A))}{\partial A} = A^{-1}$
- $\frac{\partial \det(A(t))}{\partial t} = \det(A)\,trace(A^{-1}\,\frac{\partial (A(t))}{\partial t})$
- $\frac{\partial trace(BA^{-1})}{\partial A} = -A^{-1}BA^{-1}$
- $\frac{\partial (y^T Ax)}{\partial A} = xy^T$
- $\frac{\partial (x^T Ax)}{\partial A} = xx^T$

Review

Given $A = [a_{ij}]$, the (i, j)-cofactor of A is the number C_{ij} given by

$$C_{ij} = (-1)^{i+j} \det(A_{ij})$$

Then

$$\det(A) = a_{11}C_{11} + a_{12}C_{12} + \cdots + a_{1n}C_{1n}$$

Which is a cofactor expansion across the first row of A.

$$A^{-1} = \frac{1}{|A|}\begin{bmatrix} C_{11} & C_{21} & \cdots & C_{n1} \\ C_{12} & C_{22} & \cdots & C_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{1n} & C_{2n} & \cdots & C_{nn} \end{bmatrix} = A^{-1} = \frac{1}{|A|}\, adj\ A$$

$$adj\ (A) = C^T$$

The matrix of cofactors is called the adjugate (or classical adjoint) of A, denoted by $\operatorname{adj} A$.

49

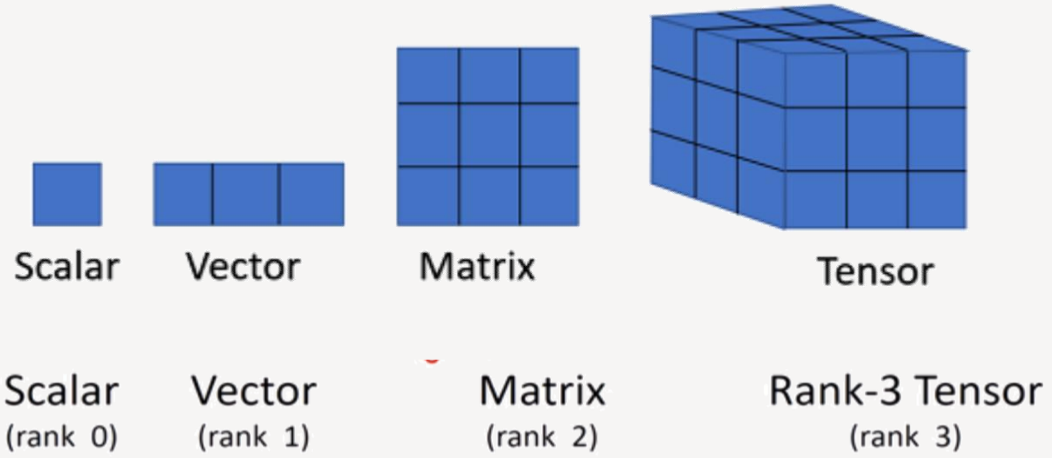
07

Tensors

Tensor

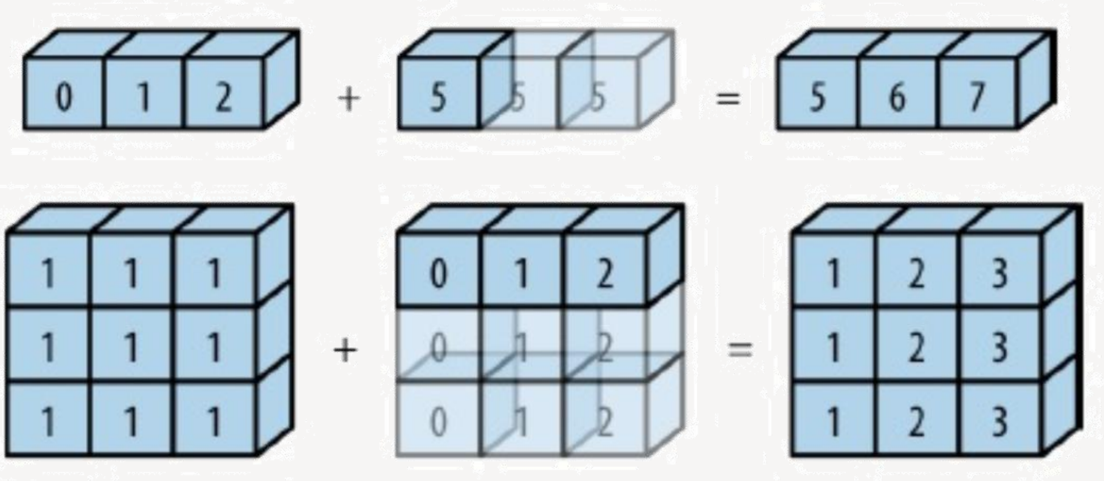
- Multi-dimensional array of numbers

```
w = torch.empty(3)
x = torch.empty(2, 3)
y = torch.empty(2, 3, 4)
z = torch.empty(2, 3, 2, 4)
```



Tensors Addition

- Adding tensors with same size
- Adding scalar to tensor
- Adding tensors with different size: if broacastable

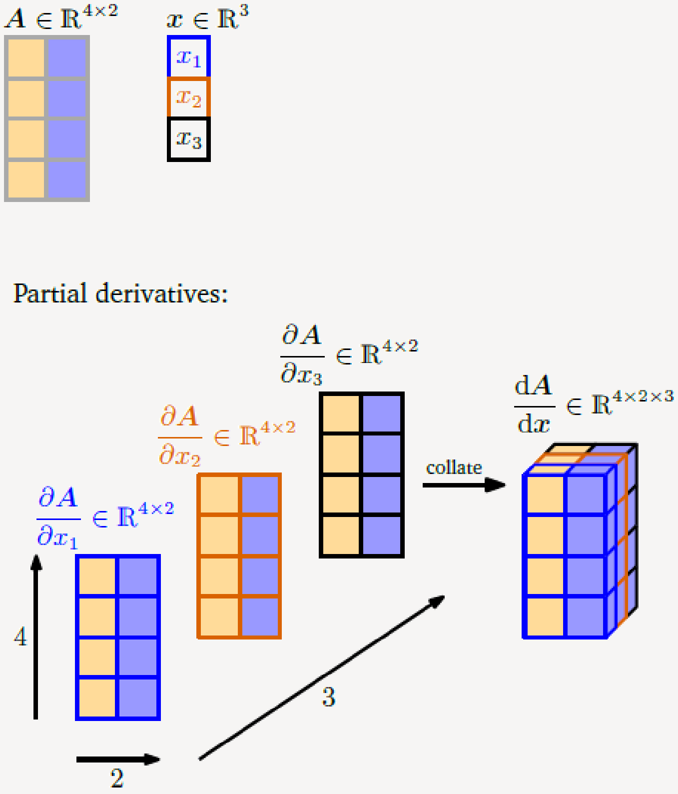


Tensors Product

$$\overbrace{(m \times \underbrace{n) \cdot (n}_{\text{product is defined}} \times k)}^{} = (m \times k)$$

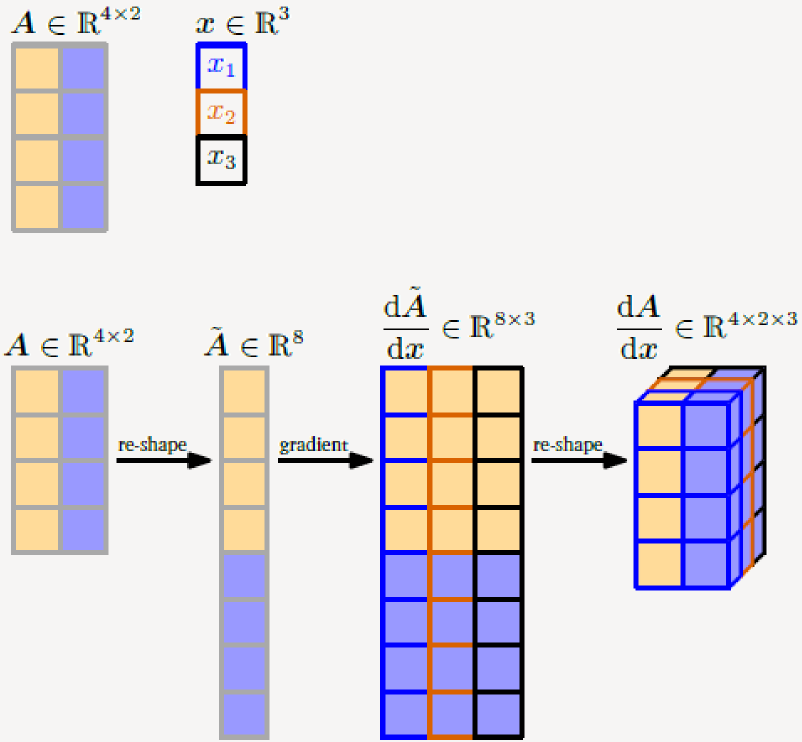
Matrix with respect to vector

- Approach 1



Matrix with respect to vector

- Approach 2



References

- https://explained.ai/matrix-calculus/
- https://paulklein.ca/newsite/teaching/matrix%20calculus.pdf
- https://web.stanford.edu/~jduchi/projects/matrix_prop.pdf
- https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf
- https://www.kamperh.com/notes/kamper_matrixcalculus13.pdf