# Chapter 5　Elements of Calculus

## Definition (sequence in $\mathbb{R}$)

A sequence of in $\mathbb{R}$ is a set of numbers $\{x_1, x_2, \ldots, x_k, \ldots\}$, denoted by $\{x_k\}$ or $\{x_k\}_{k=1}^{\infty}$ or $\{x_k\}_{k\in\mathbb{N}}$.

★ A sequence $\{x_k\}$ is said to be $\begin{cases} \text{increasing}, & \text{if } x_k < x_{k+1} \text{ for all } k, \\ \text{nondecreasing}, & \text{if } x_k \leq x_{k+1} \text{ for all } k. \end{cases}$

★ Likewise, a sequence is decreasing/nonincreasing $\cdots$

★ nonincreasing/nondecreasing sequences are called monotone sequences.

## Definition (limit of sequence in $\mathbb{R}$)

An $x^* \in \mathbb{R}$ is the limit of the sequence $\{x_k\}$ if: $\forall \varepsilon > 0$, $\exists K > 0$ (may depend on $\varepsilon$) such that $|x_k - x^*| < \varepsilon$ for all $k > K$. Denote by $x_k \to x^*$ or $x^* = \lim_{k\to\infty} x_k$.

## Sequences and Limits

### Definition (sequence in $\mathbb{R}^n$)

A sequence in $\mathbb{R}^n$ is a set of vectors $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(k)}, \ldots\}$, denoted by $\{\boldsymbol{x}^{(k)}\}$ or $\{\boldsymbol{x}^{(k)}\}_{k=1}^{\infty}$.

★ for the sequences in $\mathbb{R}^n$, there is no terminologies,e.g., increasing/nonincreasing/... (reason: $\mathbb{R}^n$ has no order relation).

### Definition (limit of sequence in $\mathbb{R}^n$)

An $\boldsymbol{x}^* \in \mathbb{R}^n$ is the limit of the sequence $\{\boldsymbol{x}^{(k)}\}$ if: $\forall \varepsilon > 0$, $\exists K > 0$ (may depend on $\varepsilon$) such that $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| < \varepsilon$ for all $k > K$. Denote by $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ or $\boldsymbol{x}^* = \lim\limits_{k\to\infty} \boldsymbol{x}^{(k)}$.

### Theorem (let $\{\boldsymbol{x}^{(k)}\}$ be a sequence in $\mathbb{R}^n$)

- $\{\boldsymbol{x}^{(k)}\}$ *is called bounded if:* $\exists r > 0$ *such that* $\|\boldsymbol{x}^{(k)}\| \leq r$ *for all* $k$.
- *if* $\{\boldsymbol{x}^{(k)}\}$ *is convergent sequence, it is bounded and has unique limit.*

## Definition (supremum/infimum of a sequence $\{x_k\}$ in $\mathbb{R}$)

- if $\{x_k\}$ has an upper bound $\Longrightarrow$ it has a least upper bound (called the supremum), i.e., the smallest upper bound of $\{x_k\}$.
- if $\{x_k\}$ has a lower bound $\Longrightarrow$ it has a greatest lower bound (called the infimum), i.e., the largest lower bound of $\{x_k\}$.

## Theorem

*Every monotone bounded sequence in* $\mathbb{R}$ *is convergent.*

## Definition (subsequence of the sequence $\{\boldsymbol{x}^{(k)}\}$)

For an increasing sequence of natural numbers $\{m_k\}$. The sequence $\{\boldsymbol{x}^{(m_k)}\} = \{\boldsymbol{x}^{(m_1)}, \boldsymbol{x}^{(m_2)}, \ldots\}$ is called a subsequence of $\{\boldsymbol{x}^{(k)}\}$. i.e., $\{\boldsymbol{x}^{(m_k)}\}$ is obtained by neglecting some elements in $\{\boldsymbol{x}^{(k)}\}$.

# Sequences and Limits

## Theorem (convergent of subsequence)

*If a sequence $\{\boldsymbol{x}^{(k)}\}$ converges to $\boldsymbol{x}^*$, then any subsequence of $\{\boldsymbol{x}^{(k)}\}$ also converges to $\boldsymbol{x}^*$.*

## Theorem (Bolzano-Weierstrass theorem)

*Any bounded sequence has convergent subsequence.*

## Definition (continuity of $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$)

Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ be vector-valued function[a]. $\boldsymbol{f}$ is continuous at $\boldsymbol{x}_0 \in \mathbb{R}^n$

$\iff$ for any sequence $\{\boldsymbol{x}^{(k)}\}$ converges to $\boldsymbol{x}_0$, $\lim\limits_{k\to\infty} \boldsymbol{f}(\boldsymbol{x}^{(k)}) = \boldsymbol{f}(\boldsymbol{x}_0)$

$\iff$ $\lim\limits_{\boldsymbol{x}\to\boldsymbol{x}_0} \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_0)$.

---

[a] A vector-valued function can be abbreviated by $\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{f}_1(\boldsymbol{x}) \\ \vdots \\ \boldsymbol{f}_m(\boldsymbol{x}) \end{bmatrix}$.

## Definition (lower semi-continuous)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be real-valued function. $f$ is lower semi-continuous (l.s.c) at $\boldsymbol{x}_0 \in \mathbb{R}^n$ if: $\liminf_{k\to\infty} f(\boldsymbol{x}_k) \geq f(\boldsymbol{x}_0)$, $\forall\{\boldsymbol{x}_k\} \subset \mathbb{R}^n$ satisfying $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}_0$.

★ E.g., $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ is continuous function.

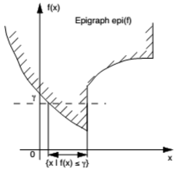
## Definition (closedness of real-valued function $f : \mathbb{R}^n \to \mathbb{R}$)

$f$ is closed $\iff$ dom$(f)$ is closed and $f$ is l.s.c over dom$(f)$.

## Theorem

*If $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued function, then the followings are equivalent:*

- $\{\boldsymbol{x} \mid f(\boldsymbol{x}) \leq \gamma\}$ *is closed for every scalar* $\gamma$.
- $f$ *is l.s.c at all* $\boldsymbol{x} \in \mathbb{R}^n$.
- $f$ *is closed function.*



**Definition (sequence in $\mathbb{R}^{n\times m}$)**

For a sequence $\{\boldsymbol{A}_k\} \subset \mathbb{R}^{n\times m}$, it converges to a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ if $\lim\limits_{k\to\infty} \|\boldsymbol{A}_k - \boldsymbol{A}\| = 0$.

**Lemma (Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$)**

$\lim\limits_{k\to\infty} \boldsymbol{A}^k = 0 \iff \rho(\boldsymbol{A}) = \max_{1\le i\le n} |\lambda_i(\boldsymbol{A})| < 1$ *(spectral radius).*

proof. (key point):

For any $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, it is similar to the Jordan form, i.e., $\exists \boldsymbol{T} \in \mathbb{R}^{n\times n}$ such that $\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \text{diag}[\boldsymbol{J}_1(\lambda_1), \boldsymbol{J}_2(\lambda_2), \cdots, \boldsymbol{J}_t(\lambda_t)] = \boldsymbol{J}$, where $\boldsymbol{J}$ is Jordan form and

$$\boldsymbol{J}_r(\lambda_r) = \begin{bmatrix} \lambda_r & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_r & 1 \\ & & & \lambda_r \end{bmatrix}.$$

$$\because\ [\boldsymbol{J}_r(\lambda)]^k = \begin{bmatrix} \lambda^k & c_k^1\lambda^{k-1} & \cdots & c_k^{k-r+1}\lambda^{k-r+1} \\ & \ddots & \ddots & \\ & & \lambda^k & c_k^1\lambda^{k-1} \\ & & & \lambda^k \end{bmatrix},$$

$$\Longrightarrow\ \lim_{k\to\infty} \boldsymbol{A}^k = \boldsymbol{T}^{-1}(\lim_{k\to\infty} \boldsymbol{J}^k)\boldsymbol{T} = 0 \iff |\lambda_i| < 1.$$

**Lemma (series of matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$)**

*The series* $\sum_{k=0}^{\infty} \boldsymbol{A}^k$ *converges if and only if* $\lim_{k\to\infty} \boldsymbol{A}^k = 0$. *Moreover,* $\sum_{k=0}^{\infty} \boldsymbol{A}^k = (I - \boldsymbol{A})^{-1}$. *[Proof on blackboard]*

**Definition (continuity of matrix-valued function $F : \mathbb{R}^r \to \mathbb{R}^{n\times n}$)**

$F$ is continuous at an $\boldsymbol{x}_0 \in \mathbb{R}^r$ if $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \|F(\boldsymbol{x}) - F(\boldsymbol{x}_0)\| = 0$.

## Lemma (property of continuous matrix-valued function)

*Let $F : \mathbb{R}^r \to \mathbb{R}^{n\times n}$ be continuous at $\boldsymbol{x}_0 \in \mathbb{R}^r$. If $F(\boldsymbol{x}_0)$ is invertible, then $\exists \varepsilon > 0$ such that $F(\boldsymbol{x})$ is invertible for all $\boldsymbol{x} \in N(\boldsymbol{x}_0)$. Moreover, $[F(\cdot)]^{-1}$ is continuous at $\boldsymbol{x}_0$.*

proof. By defining $K(\boldsymbol{x}) = F(\boldsymbol{x}_0)^{-1}[F(\boldsymbol{x}_0) - F(\boldsymbol{x})]$, we have

$$F(\boldsymbol{x}) = F(\boldsymbol{x}_0) - F(\boldsymbol{x}_0) + F(\boldsymbol{x}) = F(\boldsymbol{x}_0)[I - K(\boldsymbol{x})], \tag{1}$$

Thus, $\|K(\boldsymbol{x})\| \le \underbrace{\|F(\boldsymbol{x}_0)^{-1}\|}_{\text{bounded}} \cdot \underbrace{\|F(\boldsymbol{x}_0) - F(\boldsymbol{x})\|}_{\text{continuity}}$

$\therefore \quad \lim\limits_{\boldsymbol{x}\to\boldsymbol{x}_0} \|K(\boldsymbol{x})\| = 0 \Longrightarrow K(\boldsymbol{x})$ is continuous $\xRightarrow{\text{by (1)}}$ $F$ is continuous.

$\because \quad \|K(\boldsymbol{x})\| < \theta < 1$ as $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \varepsilon \xRightarrow{\text{by Lemma}} I - K(\boldsymbol{x})$ is invertible

$\therefore \xRightarrow{\text{by (1)}} F(\boldsymbol{x})$ is invertible.

## Definition (affine function)

A function $\boldsymbol{a}: \mathbb{R}^n \to \mathbb{R}^m$ is affine if there exists a linear function $\boldsymbol{l}: \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{b} \in \mathbb{R}^m$ such that $\boldsymbol{a}(\boldsymbol{x}) = \boldsymbol{l}(\boldsymbol{x}) + \boldsymbol{b}$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

How to approximate vector-valued $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ at $\boldsymbol{x}_0 \in \mathbb{R}^n$ by an affine function $\boldsymbol{a}(\boldsymbol{x}) = \boldsymbol{l}(\boldsymbol{x}) + \boldsymbol{b}$?

$$\text{natural condition:} \quad \boldsymbol{a}(\boldsymbol{x}_0) = \boldsymbol{f}(\boldsymbol{x}_0). \tag{2}$$

$$\because \; \boldsymbol{a}(\boldsymbol{x}) = \boldsymbol{l}(\boldsymbol{x}) + \boldsymbol{b} \xRightarrow{(2)} \boldsymbol{b} = \boldsymbol{f}(\boldsymbol{x}_0) - \boldsymbol{l}(\boldsymbol{x}_0). \tag{3}$$

Thus, $\boldsymbol{a}(\boldsymbol{x}) = \boldsymbol{l}(\boldsymbol{x}) + \boldsymbol{b} \overset{(3)}{=\!=} \boldsymbol{f}(\boldsymbol{x}_0) + \boldsymbol{l}(x) - \boldsymbol{l}(\boldsymbol{x}_0) = \boldsymbol{f}(\boldsymbol{x}_0) + \boldsymbol{l}(\boldsymbol{x} - \boldsymbol{x}_0)$.

# Differentiability

**Definition (differentiability of a function $\boldsymbol{f} : \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$)**

$\boldsymbol{f}$ is said to be differentiable at $\boldsymbol{x}_0 \in \Omega$ if there exists an affine function $\boldsymbol{a}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_0) + \boldsymbol{l}(\boldsymbol{x} - \boldsymbol{x}_0)$ such that

$$\lim_{\substack{\boldsymbol{x}\to\boldsymbol{x}_0\\ \boldsymbol{x}\in\Omega}} \frac{\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{a}(\boldsymbol{x})\|}{\|\boldsymbol{x} - \boldsymbol{x}_0\|} = \lim_{\substack{\boldsymbol{x}\to\boldsymbol{x}_0\\ \boldsymbol{x}\in\Omega}} \frac{\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_0) - \boldsymbol{l}(\boldsymbol{x} - \boldsymbol{x}_0)\|}{\|\boldsymbol{x} - \boldsymbol{x}_0\|} = 0. \tag{4}$$

If the linear function $\boldsymbol{l}$ is unique, it is called the derivative of $\boldsymbol{f}$ at $\boldsymbol{x}_0$.

**Example ($f : \mathbb{R} \to \mathbb{R}$ and $a(x) = lx + b$ with $l,\, b \in \mathbb{R}$)**

$$\xrightarrow{\text{by (4)}} \lim_{x\to x_0} \frac{|f(x) - f(x_0) - l(x - x_0)|}{|x - x_0|} = 0 \overset{\text{why?}}{\iff} \lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0} = l.$$

$\therefore$ The scalar $l$ is called the derivative of $f$ at $x_0$, denoted by $f'(x_0)$.

# Differentiability

**Example ($\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{a}(\boldsymbol{x}) = \boldsymbol{l}^\top \boldsymbol{x} + b$ with $\boldsymbol{l} \in \mathbb{R}^n$, $b \in \mathbb{R}$)**

$$\overset{(4)}{\Longrightarrow} \lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \frac{|\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{x}_0)-\boldsymbol{l}(\boldsymbol{x}-\boldsymbol{x}_0)|}{\|\boldsymbol{x}-\boldsymbol{x}_0\|} = 0.$$

The vector $\boldsymbol{l}^\top$ is called the derivative of $\boldsymbol{f}$ at $\boldsymbol{x}_0$, denoted by $\boldsymbol{Df}(\boldsymbol{x}_0)$

$$\boldsymbol{Df} = \begin{bmatrix} \frac{\partial \boldsymbol{f}}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}}{\partial x_n} \end{bmatrix} = (\nabla \boldsymbol{f})^\top.$$

**Example ($\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{a}(\boldsymbol{x}) = \boldsymbol{l}\boldsymbol{x} + \boldsymbol{b}$ with $\boldsymbol{l} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$)**

$$\overset{(4)}{\Longrightarrow} \lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \frac{\|\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{x}_0)-\boldsymbol{l}(\boldsymbol{x}-\boldsymbol{x}_0)\|}{\|\boldsymbol{x}-\boldsymbol{x}_0\|} = 0.$$

The matrix $\boldsymbol{l} \in \mathbb{R}^{m\times n}$ is called the derivative of $\boldsymbol{f}$ at $\boldsymbol{x}_0$, denoted by $\boldsymbol{Df}(\boldsymbol{x}_0)$.

$$\boldsymbol{Df} = \begin{bmatrix} \frac{\partial \boldsymbol{f}_1}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}_1}{\partial x_n} \\ \frac{\partial \boldsymbol{f}_2}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}_2}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial \boldsymbol{f}_m}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Df}_1 \\ \vdots \\ \boldsymbol{Df}_m \end{bmatrix}$$

(also Jacobian of $\boldsymbol{f}$)

## Definition (Hessian matrix of $\boldsymbol{f}:\mathbb{R}^n\to\mathbb{R}$)

For a function $\boldsymbol{f}:\mathbb{R}^n\to\mathbb{R}$, if $\nabla\boldsymbol{f}$ is differentiate, then $\boldsymbol{f}$ is twice differentiate, and the derivative of $\nabla\boldsymbol{f}$ is defined as

$$\boldsymbol{D}^2\boldsymbol{f}=\begin{bmatrix}\frac{\partial^2\boldsymbol{f}}{\partial x_1^2} & \frac{\partial^2\boldsymbol{f}}{\partial x_1x_2} & \cdots & \frac{\partial^2\boldsymbol{f}}{\partial x_1x_n}\\ \frac{\partial^2\boldsymbol{f}}{\partial x_2x_1} & \frac{\partial^2\boldsymbol{f}}{\partial x_2^2} & \cdots & \frac{\partial^2\boldsymbol{f}}{\partial x_2x_n}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial^2\boldsymbol{f}}{\partial x_nx_1} & \frac{\partial^2\boldsymbol{f}}{\partial x_nx_2} & \cdots & \frac{\partial^2\boldsymbol{f}}{\partial x_n^2}\end{bmatrix}$$

## Definition (continuously differentiate function)

A function $\boldsymbol{f}:\Omega\subset\mathbb{R}^n\to\mathbb{R}^m$ is said to be continuously differentiate if it is differentiate componentwisely, and $\boldsymbol{D}\boldsymbol{f}:\Omega\to\mathbb{R}^{m\times n}$ is continuous.

$\mathcal{C}^p$: all components of $\boldsymbol{f}$ have continuous partial derivatives until order $p$.

## Theorem (Schwarz's theorem: symmetric of Hessian)

*If $\boldsymbol{f}$ is twice continuously differentiate at $\boldsymbol{x}\Longrightarrow\boldsymbol{D}^2\boldsymbol{f}(\boldsymbol{x})$ is symmetric.*

**Example**

$\boldsymbol{f}(\boldsymbol{x}) = 5x_1 + 8x_2 + x_1x_2 - x_1^2 - 2x_2^2$, $\boldsymbol{Df}(\boldsymbol{x}) =?$ $\boldsymbol{D^2f}(\boldsymbol{x}) =?$

Ans: $\boldsymbol{Df}(\boldsymbol{x}) = [5 + x_2 - 2x_1, 8 + x_1 - 4x_2]$, $\boldsymbol{D^2f}(\boldsymbol{x}) = \begin{bmatrix} -2 & 1 \\ 1 & -4 \end{bmatrix}$.

**Example**

$f(\boldsymbol{x}) = 2x_1^4 + 3x_1^2x_2 + 2x_1x_2^3 + 4x_2^2$, $\boldsymbol{Df}(\boldsymbol{x}) =?$ $\boldsymbol{D^2f}(\boldsymbol{x}) =?$

Ans: $\boldsymbol{Df}(\boldsymbol{x}) = \left[\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}\right] = \left[8x_1^3 + 6x_1x_2 + 2x_2^3, 3x_1^2 + 6x_1x_2^2 + 8x_2\right]$.

$$\boldsymbol{D^2f}(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 24x_1^2 + 6x_2 & 6x_1 + 6x_2^2 \\ 6x_1 + 6x_2^2 & 12x_1x_2 + 8 \end{bmatrix}.$$

For a point $\boldsymbol{x}_0 = [-2, 3]^\top$, $\boldsymbol{Df}(\boldsymbol{x}_0) =?$ $\boldsymbol{D^2f}(\boldsymbol{x}_0) =?$

then $\boldsymbol{Df}(\boldsymbol{x}_0) = \left[-46, -72\right]$ and $\boldsymbol{D^2f}(\boldsymbol{x}_0) = \begin{bmatrix} 114 & 42 \\ 42 & -64 \end{bmatrix}$.

## compute the gradient and Hessian

(1) $f(\boldsymbol{x}) = f(x_1, x_2) = \ln(e^{x_1} + e^{x_2})$
(2) $f(\boldsymbol{x}) = f(x_1, x_2) = \frac{x_1^2}{x_2}$
(3) $f(\boldsymbol{x}) = f(x_1, x_2) = -\sqrt{x_1 x_2}$
(4) $f(\boldsymbol{x}) = f(x_1, x_2) = x_1^2 - 2x_1 + 3x_1x_2^2 + 4x_2^3$. $x_0 = (1,1)^T$, $p = (-2,1)^T$.
$f(x_0) =?$ $f(x_0 + p) =?$ $\nabla f(x_0) =?$ $\nabla^2 f(x_0) =?$

Check the structure of $\boldsymbol{Df}(\boldsymbol{x})$ and $\boldsymbol{D^2f}(\boldsymbol{x})$.

- $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$, $\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} 2(x_2^3 - x_1^2) \\ 3(x_2^3 - x_1^2) + 2(x_3^3 - x_2^2) \\ 3(x_3^3 - x_2^2) + 2(x_4^3 - x_3^2) \\ \cdots \\ 3(x_n^3 - x_{n-1}^2) \end{bmatrix}$, $\boldsymbol{Df}(\boldsymbol{x}) = \begin{bmatrix} * & * & & \\ * & * & * & \\ & * & * & * \\ & & * & * \end{bmatrix}$.
- $f(\boldsymbol{x}) = x_1 \sum_{i=1}^n i^2 x_i^2$, $\boldsymbol{D}^2 f(\boldsymbol{x}) = \begin{bmatrix} * & * & * & * & * \\ * & * & & & \\ * & & * & & \\ * & & & * & \\ * & & & & * \end{bmatrix}$.

# Differentiation Rules

| function $\boldsymbol{f}$ | derivative $\boldsymbol{Df}$ |
|---|---|
| $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{a}$, $\boldsymbol{a} \in \mathbb{R}^n$ | $0$ |
| $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x}$, $\boldsymbol{a} \in \mathbb{R}^n$ | $\boldsymbol{a}^\top$ |
| $\boldsymbol{f}(\boldsymbol{x}) = \\|\boldsymbol{x}\\|_2^2 = \boldsymbol{x}^\top \boldsymbol{x}$ | $2\boldsymbol{x}^\top$ |
| $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ | $\boldsymbol{x}^\top(\boldsymbol{A} + \boldsymbol{A}^\top)$ |
| $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$ | $I_n$ |
| $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ | $\boldsymbol{A}$ |
| $\boldsymbol{f}(x) = x\boldsymbol{a}$, $x \in \mathbb{R}$, $\boldsymbol{a} \in \mathbb{R}^n$ | $\boldsymbol{a}$ |
| $\boldsymbol{f}(\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{X})$, $\boldsymbol{A}, \boldsymbol{X} \in \mathbb{R}^{m\times n}$ | $\boldsymbol{A}$ |

**Theorem (chain rule)**

*Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, $\boldsymbol{g} : \mathbb{R}^m \to \mathbb{R}^p$, and $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))$. Then $\boldsymbol{Dh} = ?$*

Ans: It follows by definition that $\boldsymbol{Df} \in \mathbb{R}^{m\times n}$, $\boldsymbol{Dg} \in \mathbb{R}^{p\times m}$.

$\because$ $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^p \Longrightarrow \boldsymbol{Dh} \in \mathbb{R}^{p\times n}$. By matrix multiplications,

$\boldsymbol{Dh}(\boldsymbol{x}) = \boldsymbol{Dg}(\boldsymbol{f}(\boldsymbol{x}))\boldsymbol{Df}(\boldsymbol{x})$.

# Differentiation Rules

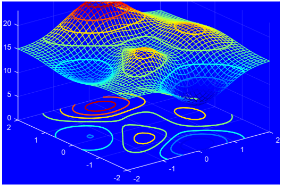
## Examples (hint on blackboard)

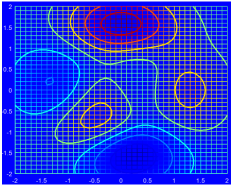
1. $\boldsymbol{f}(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$; 2. $\varphi(t) = \boldsymbol{f}(\boldsymbol{x}_0 + t\boldsymbol{d})$.
3. $\boldsymbol{f}(\boldsymbol{x}) = (\boldsymbol{a}^\top \boldsymbol{x})(\boldsymbol{b}^\top \boldsymbol{x})$; 4. $\boldsymbol{f}(\boldsymbol{x}) = \mathrm{e}^{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}$.

## Definition (level set)

The level set of a function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}$ at level $c$ is the set of points $S = \{\boldsymbol{x} \mid \boldsymbol{f}(\boldsymbol{x}) = c\}$.

★ For $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}$, $S$ is a curve; For $\boldsymbol{f} : \mathbb{R}^3 \to \mathbb{R}$, $S$ are surfaces.

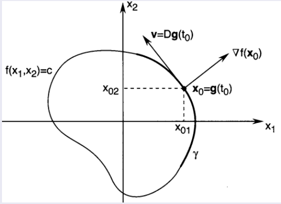


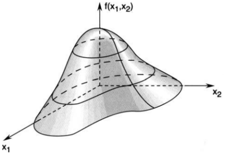

Properties of level set:
1. $S_1 \cap S_2 = \emptyset$;
2. distributions imply slope;
3. $S \perp \nabla \boldsymbol{f}$;
4. locally ellipsoid.

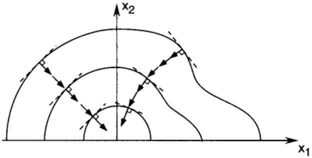
## Theorem

*let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{x}_0 \in \mathbb{R}^n$. Then, $\nabla \boldsymbol{f}(\boldsymbol{x}_0)$ is normal to the tangent vector to an arbitrary smooth curve passing through $\boldsymbol{x}_0$ on the level set $S = \{\boldsymbol{x} \mid \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_0)\}$. [key point of proof: chain rule]*




- ★ $\nabla \boldsymbol{f}(\boldsymbol{x}_0)$ is the direction of maximum rate of increase of $\boldsymbol{f}$ $\boldsymbol{x}_0$.
- ★ The direction of maximum rate of increase of a differentiable function at a point is normal to the level set of the function through that point.




level curves and is called a path of steepest ascent.

# Level Sets and Gradients

## Definition (graph)

The graph of $\boldsymbol{f}:\mathbb{R}^n \to \mathbb{R}$ is the set $G(\boldsymbol{f}) = \{[\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})] \mid \boldsymbol{x} \in \mathbb{R}^n\} \subset \mathbb{R}^{n+1}$.

## Definition (nonvertical tangent hyperplane)

if $\boldsymbol{f}:\mathbb{R}^n \to \mathbb{R}$ is differentiable at $\boldsymbol{x}_0$, its tangent hyperplane is defined by $\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{x}_0) + \boldsymbol{Df}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) = \boldsymbol{f}(\boldsymbol{x}_0) + (\boldsymbol{x} - \boldsymbol{x}_0)^\top \nabla \boldsymbol{f}(\boldsymbol{x}_0)$.

## Definition (Taylor series for $f:(a,b) \to \mathbb{R}$)

If $f$ is $m$ times continuously differentiable, then, for any $x$, $x_0 \in (a,b)$, $f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + R_m$, where $R_m = \frac{h^m(1-\theta)^{m-1}}{(m-1)!} f^{(m)}(\xi)$ with $\xi \in (x - x_0, x + x_0)$ is called residual/remainder.

★ The residual can be variant form. It usually use the order symbols $O$ and $o$.

# Taylor Series

## Definition (order symbols $O$ and $o$)

Let $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}$ be defined in some neighborhood $N(0)$. $\boldsymbol{g}(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \neq 0$.
Let $\boldsymbol{f} : \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$ and $0 \in \Omega$.

- $\boldsymbol{f}(\boldsymbol{x}) = O(\boldsymbol{g}(\boldsymbol{x})) \iff \exists K > 0,\ \delta > 0$ such that if $\boldsymbol{x} \in \Omega$, then
$$\frac{\|\boldsymbol{f}(\boldsymbol{x})\|}{|\boldsymbol{g}(\boldsymbol{x})|} \leq K, \quad \forall \|\boldsymbol{x}\| \leq \delta.$$
- $\boldsymbol{f}(\boldsymbol{x}) = o(\boldsymbol{g}(\boldsymbol{x})) \iff \lim_{\boldsymbol{x} \to \boldsymbol{x}_0} \frac{\|\boldsymbol{f}(\boldsymbol{x})\|}{|\boldsymbol{g}(\boldsymbol{x})|} = 0.$

## Definition (Taylor series for $\boldsymbol{f} : \Omega \subset \mathbb{R}^n \to \mathbb{R}$)

If $\boldsymbol{f}$ is $m$ times continuously differentiable, then, for any $\boldsymbol{x}$, $\boldsymbol{x}_0 \in \Omega$,
$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_0) + \boldsymbol{D}\boldsymbol{f}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^\top \boldsymbol{D}^{\boldsymbol{2}}\boldsymbol{f}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + R_m$,
where $R_m = o(\|\boldsymbol{x} - \boldsymbol{x}_0\|^2)$ or $R_m = O(\|\boldsymbol{x} - \boldsymbol{x}_0\|^3)$.

# Taylor Series

Remark: Approximation $\boldsymbol{f}:\Omega\subset\mathbb{R}^n\to\mathbb{R}$ by Taylor series

1. Linear approximation: by denoting $\boldsymbol{g}(\boldsymbol{x})=\nabla\boldsymbol{f}(\boldsymbol{x})$,
   $\boldsymbol{l}(\boldsymbol{x})=\boldsymbol{f}(\boldsymbol{x}_0)+\boldsymbol{Df}(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0)=\boldsymbol{f}(\boldsymbol{x}_0)+[\boldsymbol{g}(\boldsymbol{x}_0)]^\top(\boldsymbol{x}-\boldsymbol{x}_0)$.
2. Quadratic approximation: by denoting $\boldsymbol{F}(\boldsymbol{x})=\boldsymbol{D^2f}(\boldsymbol{x})$,
   $$\begin{aligned}\boldsymbol{q}(\boldsymbol{x})=&\quad \boldsymbol{f}(\boldsymbol{x}_0)+\boldsymbol{Df}(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0)+\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^\top\boldsymbol{D^2f}(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0)\\ =&\quad \boldsymbol{f}(\boldsymbol{x}_0)+[\boldsymbol{g}(\boldsymbol{x}_0)]^\top(\boldsymbol{x}-\boldsymbol{x}_0)+\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^\top\boldsymbol{F}(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0).\end{aligned}$$

## Theorem (mean value theorem)

*Let $\boldsymbol{f}:\Omega\subset\mathbb{R}^n\to\mathbb{R}$ be differentiable on open set $\Omega$, then for any $\boldsymbol{x},\boldsymbol{y}\in\Omega$, there exists a matrix $M$ such that $\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{f}(\boldsymbol{y})=M(\boldsymbol{x}-\boldsymbol{y})$.*

Special case: If $\Omega\subset\mathbb{R}$ is open interval and $f:\Omega\to\mathbb{R}$ is differentiable, then there exists $M\in\mathbb{R}$ such that $f(a)-f(b)=M(a-b),\ \forall a,b\in\Omega$.