MA 5037: Optimization Methods and Applications
Chapter 1: Mathematical Preliminaries

Suh-Yuh Yang (楊肅煜)

Department of Mathematics, National Central University
Jhongli District, Taoyuan City 32001, Taiwan

syyang@math.ncu.edu.tw
http://www.math.ncu.edu.tw/∼syyang/

Vector space $\mathbb{R}^n$

- **Vector space $\mathbb{R}^n$:** the set of n-dimensional column vectors with real components endowed with the following component-wise addition operator and the scalar-vector product,

$$\boldsymbol{x}+\boldsymbol{y} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} := \begin{bmatrix} x_1+y_1 \\ x_2+y_2 \\ \vdots \\ x_n+y_n \end{bmatrix}, \quad \forall\, \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^n,$$

$$\lambda\boldsymbol{x} = \lambda \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} := \begin{bmatrix} \lambda x_1 \\ \lambda x_2 \\ \vdots \\ \lambda x_n \end{bmatrix}, \quad \forall\, \boldsymbol{x} \in \mathbb{R}^n, \lambda \in \mathbb{R}.$$

- **Standard/canonical basis of $\mathbb{R}^n$:** $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n\}$, where $\boldsymbol{e}_i := (0,\cdots,0,\underbrace{1}_{ith},0,\cdots,0)^\top$ for $i=1,2,\cdots,n$.
- $\boldsymbol{e} := (1,1,\cdots,1)^\top$ and $\mathbf{0} := (0,0,\cdots,0)^\top$ are all ones and all zeros column vectors, respectively.

Important subsets of $\mathbb{R}^n$

- **Nonnegative orthant**

$$\mathbb{R}^n_+ := \{(x_1, x_2, \cdots, x_n)^\top : x_i \geq 0 \; \forall \, i\}.$$

- **Positive orthant**

$$\mathbb{R}^n_{++} := \{(x_1, x_2, \cdots, x_n)^\top : x_i > 0 \; \forall \, i\}.$$

- **Closed line segment:** let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$[\boldsymbol{x}, \boldsymbol{y}] := \{(1-\alpha)\boldsymbol{x} + \alpha\boldsymbol{y} : \alpha \in [0, 1]\}.$$

  Note: $[\boldsymbol{x}, \boldsymbol{x}] = \{\boldsymbol{x}\}$.

- **Open line segment:** let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$(\boldsymbol{x}, \boldsymbol{y}) := \{(1-\alpha)\boldsymbol{x} + \alpha\boldsymbol{y} : \alpha \in (0, 1)\}.$$

  Note: $(\boldsymbol{x}, \boldsymbol{x}) = \varnothing$.

- **Unit-simplex**

$$\Delta_n := \{\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^\top \in \mathbb{R}^n : x_1, x_2, \cdots, x_n \geq 0, \boldsymbol{e}^\top \boldsymbol{x} = 1\}.$$

Vector space $\mathbb{R}^{m\times n}$

- The set of all real-valued matrices of order $m \times n$ is denoted by $\mathbb{R}^{m\times n}$.
- The $n \times n$ identity matrix is denoted by $\boldsymbol{I}_n$.
- The $m \times n$ zero matrix is denoted by $\mathbf{0}_{m\times n}$.
- We will frequently omit the subscripts of these matrices when the dimensions will be clear from the context.

Inner product on $\mathbb{R}^n$

- **Definition:** An inner product on $\mathbb{R}^n$ is a map $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with the following properties:
  (1) *symmetry:* $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle, \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.
  (2) *additivity:* $\langle \boldsymbol{x}, \boldsymbol{y} + \boldsymbol{z} \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \langle \boldsymbol{x}, \boldsymbol{z} \rangle, \forall\, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$.
  (3) *homogeneity:* $\langle \lambda \boldsymbol{x}, \boldsymbol{y} \rangle = \lambda \langle \boldsymbol{x}, \boldsymbol{y} \rangle, \forall\, \lambda \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.
  (4) *positive definiteness:* $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0, \forall\, \boldsymbol{x} \in \mathbb{R}^n$,
  and $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ if and only if $\boldsymbol{x} = \mathbf{0}$.
- **Example 1:** (dot product) The *standard inner product* is defined by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i, \quad \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

- **Example 2:** (weighted dot product) Let $\boldsymbol{w} \in \mathbb{R}^n_{++}$. Then the following weighted dot product is also an inner product:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{w}} := \sum_{i=1}^{n} w_i x_i y_i, \quad \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Vector norms

- **Definition:** A norm $\|\cdot\|$ on $\mathbb{R}^n$ is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ satisfying the following:
  (1) *nonnegativity:* $\|\boldsymbol{x}\| \geq 0, \forall\, \boldsymbol{x} \in \mathbb{R}^n$, and $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$.
  (2) *positive homogeneity:* $\|\lambda \boldsymbol{x}\| = |\lambda|\|\boldsymbol{x}\|, \forall\, \lambda \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^n$.
  (3) *triangle inequality:* $\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|, \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.
- **The associated norm with an inner product:** One natural way to generate a norm on $\mathbb{R}^n$ is to take any inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^n$ and define the associated norm

$$\|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}, \quad \forall\, \boldsymbol{x} \in \mathbb{R}^n.$$

  If the inner product is the dot product (i.e., the standard inner product), then the associated norm is the so-called *Euclidean norm or ℓ_2-norm:*

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}, \quad \forall\, \boldsymbol{x} \in \mathbb{R}^n.$$

  By default, the underlying norm on $\mathbb{R}^n$ is $\|\cdot\|_2$ and the subscript 2 will be frequently omitted.

ℓ_p-norms: $p \geq 1$

- The ℓ_p-norm, $p \geq 1$, is defined by

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \ \forall\, \boldsymbol{x} \in \mathbb{R}^n.$$

Note: explain why $\|\cdot\|_{\frac{1}{2}}$ is not a norm!

- The ℓ_∞-norm is defined by

$$\|\boldsymbol{x}\|_\infty = \max_{i=1,2,\cdots,n} |x_i|, \ \forall\, \boldsymbol{x} \in \mathbb{R}^n,$$

and unsurprisingly, it can be shown that

$$\|\boldsymbol{x}\|_\infty = \lim_{p \to \infty} \|\boldsymbol{x}\|_p.$$

- **The Cauchy-Schwarz inequality:** *For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have*

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| (= |\boldsymbol{x}^\top \boldsymbol{y}|) \leq \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2.$$

Equality is satisfied if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly dependent.

Matrix norms

- **Definition:** A norm $\|\cdot\|$ on $\mathbb{R}^{m\times n}$ is a function $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ satisfying the following:
  (1) *nonnegativity:* $\|\boldsymbol{A}\| \geq 0, \forall\, \boldsymbol{A} \in \mathbb{R}^{m\times n}$, and $\|\boldsymbol{A}\| = 0$ if and only if $\boldsymbol{A} = \boldsymbol{0}$.
  (2) *positive homogeneity:* $\|\lambda \boldsymbol{A}\| = |\lambda|\|\boldsymbol{A}\|, \forall\, \lambda \in \mathbb{R}$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.
  (3) *triangle inequality:* $\|\boldsymbol{A}+\boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|, \forall\, \boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$.
- **Induced norms:** Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, the induced matrix norm $\|\boldsymbol{A}\|_{a,b}$ is defined by

$$\|\boldsymbol{A}\|_{a,b} := \max\{\|\boldsymbol{A}\boldsymbol{x}\|_b : \boldsymbol{x} \in \mathbb{R}^n \text{ and } \|\boldsymbol{x}\|_a \leq 1\}.$$

  An induced norm is a norm.
- It can be shown that for any $\boldsymbol{x} \in \mathbb{R}^n$, we have

$$\|\boldsymbol{A}\boldsymbol{x}\|_b \leq \|\boldsymbol{A}\|_{a,b}\|\boldsymbol{x}\|_a.$$

- We refer to the matrix-norm $\|\cdot\|_{a,b}$ as the (a,b)-norm. When $a = b$, we will simply refer to it as an a-norm.

Matrix norms (cont'd)

- **spectral norm or ℓ_2-norm:** If $\|\cdot\|_a = \|\cdot\|_b = \|\cdot\|_2$, the induced $(2,2)$-norm of a matrix $\boldsymbol{A} = (A_{ij}) \in \mathbb{R}^{m\times n}$ is the maximum singular value of $\boldsymbol{A}$,
$$\|\boldsymbol{A}\|_2 = \|\boldsymbol{A}\|_{2,2} := \sqrt{\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})} =: \sigma_{\max}(\boldsymbol{A}).$$
This norm is called the spectral norm or ℓ_2-norm. Note that the eigenvalues λ_i $(i = 1, 2, \cdots, n)$ of $\boldsymbol{A}^\top \boldsymbol{A}$ are real and nonnegative.
- **ℓ_1-norm:** If $\|\cdot\|_a = \|\cdot\|_b = \|\cdot\|_1$, the induced $(1,1)$-norm of $\boldsymbol{A} = (A_{ij}) \in \mathbb{R}^{m\times n}$ is given by
$$\|\boldsymbol{A}\|_1 = \|\boldsymbol{A}\|_{1,1} := \max_{j=1,2,\cdots,n} \sum_{i=1}^{m} |A_{ij}|.$$
- **ℓ_∞-norm:** If $\|\cdot\|_a = \|\cdot\|_b = \|\cdot\|_\infty$, the induced (∞,∞)-norm of $\boldsymbol{A} = (A_{ij}) \in \mathbb{R}^{m\times n}$ is given by
$$\|\boldsymbol{A}\|_\infty = \|\boldsymbol{A}\|_{\infty,\infty} := \max_{i=1,2,\cdots,m} \sum_{j=1}^{n} |A_{ij}|.$$
- **Frobenius norm:** A non-induced norm is defined by
$$\|\boldsymbol{A}\|_F := \left(\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2\right)^{1/2}, \quad \forall\, \boldsymbol{A} = (A_{ij}) \in \mathbb{R}^{m\times n}.$$

Eigenvalues and eigenvectors

- **Definition:** *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Then a nonzero vector $\boldsymbol{v} \in \mathbb{C}^n$ is called an eigenvector of $\boldsymbol{A}$ if there exists a $\lambda \in \mathbb{C}$ (the complex field) for which $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$. The scalar λ is called the eigenvalue corresponding to the eigenvector $\boldsymbol{v}$.*

  Note: $\exists\, \mathbf{0} \neq \boldsymbol{v} \in \mathbb{C}^n$ s.t. $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v} \;\Rightarrow \boldsymbol{A}\boldsymbol{v} - \lambda\boldsymbol{I}\boldsymbol{v} = (\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{v} = \mathbf{0}$.
  $\Rightarrow\; \det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$.

- $f_{\boldsymbol{A}}(\lambda) := \det(\boldsymbol{A} - \lambda\boldsymbol{I})$ is called the characteristic polynomial of $\boldsymbol{A}$.

  $$f_{\boldsymbol{A}}(\lambda) = (-1)^n\lambda^n + (-1)^{n-1}\underbrace{(a_{11} + \cdots + a_{nn})}_{:=\text{trace}(\boldsymbol{A})}\lambda^{n-1} + \cdots + \det(\boldsymbol{A}).$$

- In general, real-valued matrices can have complex eigenvalues, but it is well known that *all the eigenvalues of symmetric matrices are real.* The eigenvalues of a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ are denoted by

  $$\underbrace{\lambda_1(\boldsymbol{A})}_{:=\lambda_{\max}(\boldsymbol{A})} \geq \lambda_2(\boldsymbol{A}) \geq \cdots \geq \lambda_{n-1}(\boldsymbol{A}) \geq \underbrace{\lambda_n(\boldsymbol{A})}_{:=\lambda_{\min}(\boldsymbol{A})}$$

The spectral decomposition (factorization) theorem

The spectral decomposition theorem: *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Then there exists an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{n\times n}$, $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}$, and a diagonal matrix $\boldsymbol{D} = \mathrm{diag}(d_1, d_2, \cdots, d_n)$ for which*

$$\boldsymbol{U}^\top \boldsymbol{A}\boldsymbol{U} = \boldsymbol{D}.$$

- The columns of the matrix $\boldsymbol{U}$ in the factorization constitute an orthonormal basis comprised of eigenvectors of $\boldsymbol{A}$ and the diagonal elements of $\boldsymbol{D}$ are the corresponding eigenvalues.
- A direct result is that the trace and the determinant of $\boldsymbol{A}$ can be expressed via its eigenvalues:

$$\mathrm{trace}\, \boldsymbol{A} = \sum_{i=1}^{n} \lambda_i(\boldsymbol{A}) \quad \text{and} \quad \det \boldsymbol{A} = \prod_{i=1}^{n} \lambda_i(\boldsymbol{A}).$$

Hint: $f_{\boldsymbol{D}}(\lambda) = \det(\boldsymbol{D} - \lambda \boldsymbol{I}) = \det(\boldsymbol{U}^\top (\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{U})$
$= \det(\boldsymbol{U}^\top) \det(\boldsymbol{A} - \lambda \boldsymbol{I}) \det(\boldsymbol{U}) = \det(\boldsymbol{A} - \lambda \boldsymbol{I}) = f_{\boldsymbol{A}}(\lambda).$

Rayleigh quotient

- **Definition:** For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, the Rayleigh quotient is defined by
$$R_{\boldsymbol{A}}(\boldsymbol{x}) := \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\|\boldsymbol{x}\|^2}, \quad \forall\, \boldsymbol{x} \neq \boldsymbol{0}.$$
- **Lower and upper bounds on the Rayleigh quotient:**
Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be symmetric. Then
$$\lambda_{\min}(\boldsymbol{A}) \leq R_{\boldsymbol{A}}(\boldsymbol{x}) \leq \lambda_{\max}(\boldsymbol{A}), \quad \forall\, \boldsymbol{x} \neq \boldsymbol{0}.$$

Proof.

(i) By the spectral decomposition theorem, $\exists$ an orthogonal $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ such that $\boldsymbol{U}^\top \boldsymbol{A}\boldsymbol{U} = \boldsymbol{D}$, $\boldsymbol{D} = \mathrm{diag}(d_1, d_2, \cdots, d_n)$, and $\lambda_{\max}(\boldsymbol{A}) = d_1 \geq d_2 \geq \cdots \geq d_n = \lambda_{\min}(\boldsymbol{A})$.

(ii) Making the change of variables $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{y}$, we have
$$\max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\|\boldsymbol{x}\|^2} = \max_{\boldsymbol{y}\neq\boldsymbol{0}} \frac{\boldsymbol{y}^\top \boldsymbol{U}^\top \boldsymbol{A}\boldsymbol{U}\boldsymbol{y}}{\|\boldsymbol{U}\boldsymbol{y}\|^2} = \max_{\boldsymbol{y}\neq\boldsymbol{0}} \frac{\boldsymbol{y}^\top \boldsymbol{D}\boldsymbol{y}}{\boldsymbol{y}^\top \underbrace{\boldsymbol{U}^\top \boldsymbol{U}}\boldsymbol{y}} = \max_{\boldsymbol{y}\neq\boldsymbol{0}} \frac{\sum_i d_i y_i^2}{\sum_i y_i^2}.$$
Since $d_i \leq d_1 \;\forall\, i$, $R_{\boldsymbol{A}}(\boldsymbol{x}) \leq \dfrac{d_1(\sum_i y_i^2)}{\sum_i y_i^2} = d_1 = \lambda_{\max}(\boldsymbol{A})$. □

The minimal and maximal eigenvalues

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *be symmetric. Then*

- $\min\limits_{\boldsymbol{x}\neq\boldsymbol{0}} R_{\boldsymbol{A}}(\boldsymbol{x}) = \lambda_{\min}(\boldsymbol{A})$*, and the eigenvectors of* $\boldsymbol{A}$ *corresponding to the minimal eigenvalue are minimizers.*

  Proof: Let $\boldsymbol{v}$ be an eigenvector corresponding to the minimal eigenvalue of $\boldsymbol{A}$. Then
  $$R_{\boldsymbol{A}}(\boldsymbol{v}) = \frac{\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v}}{\|\boldsymbol{v}\|^2} = \frac{\lambda_{\min}(\boldsymbol{A})\|\boldsymbol{v}\|^2}{\|\boldsymbol{v}\|^2} = \lambda_{\min}(\boldsymbol{A}),$$
  which combined with the lower bound on the Rayleigh quotient lead to the desired result. □

- $\max\limits_{\boldsymbol{x}\neq\boldsymbol{0}} R_{\boldsymbol{A}}(\boldsymbol{x}) = \lambda_{\max}(\boldsymbol{A})$*, and the eigenvectors of* $\boldsymbol{A}$ *corresponding to the maximal eigenvalue are maximizers.*

  Proof: Let $\boldsymbol{w}$ be an eigenvector corresponding to the maximal eigenvalue of $\boldsymbol{A}$. Then
  $$R_{\boldsymbol{A}}(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{A}\boldsymbol{w}}{\|\boldsymbol{w}\|^2} = \frac{\lambda_{\max}(\boldsymbol{A})\|\boldsymbol{w}\|^2}{\|\boldsymbol{w}\|^2} = \lambda_{\max}(\boldsymbol{A}),$$
  which combined with the upper bound on the Rayleigh quotient lead to the desired result. □

Basic topological concepts

- **Open ball:** The open ball with center $\mathbf{c} \in \mathbb{R}^n$ and radius $r > 0$ is defined by
$$B(\mathbf{c}, r) := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{c}\| < r\}.$$
The open ball $B(\mathbf{c}, r)$ is also referred to as a neighborhood of $\mathbf{c}$.
- **Close ball:** The close ball with center $\mathbf{c} \in \mathbb{R}^n$ and radius $r > 0$ is defined by
$$B[\mathbf{c}, r] := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{c}\| \leq r\}.$$
- **Interior point:** Given a set $U \subseteq \mathbb{R}^n$, a point $\mathbf{c} \in U$ is an interior point of U if there exists $r > 0$ for which $B(\mathbf{c}, r) \subseteq U$.
- **Interior:** The set of all interior points of a given set U is called the interior of the set and is denoted by $int(U)$, i.e.,
$$int(U) := \{\mathbf{x} \in U : B(\mathbf{x}, r) \subseteq U \text{ for some } r > 0\}.$$

Examples:

$int(\mathbb{R}^n_+) = \mathbb{R}^n_{++}$.

$int(B[\mathbf{c}, r]) = B(\mathbf{c}, r)$, for $\mathbf{c} \in \mathbb{R}^n$ and $r \in \mathbb{R}_{++}$.

Open set, closed set, and boundary point

- **Open set:** $U \subseteq \mathbb{R}^n$ is an open set if and only if for every $\boldsymbol{x} \in U$ there exists $r > 0$ such that $B(\boldsymbol{x}, r) \subseteq U$.

  Examples: $\mathbb{R}^n$, open balls, positive orthant $\mathbb{R}^n_{++}$ are open sets.

  Note: (1) *A union of any number of open sets is an open set.*

  (2) *The intersection of a finite number of open sets is open.*

- **Closed set:** A set $U \subseteq \mathbb{R}^n$ is said to be closed if for every sequence of points $\{\boldsymbol{x}_k\} \subseteq U$ satisfying $\boldsymbol{x}_k \to \boldsymbol{x}^*$ as $k \to \infty$, it holds that $\boldsymbol{x}^* \in U$.

  Examples: closed ball $B[\boldsymbol{c}, r]$, closed lines segments, nonnegative orthant $\mathbb{R}^n_+$, unit simplex Δ_n, $\mathbb{R}^n$ are closed sets.

- **Proposition:** *Let f be a continuous function defined over a closed set $S \subseteq \mathbb{R}^n$. Then for any $\alpha \in \mathbb{R}$ the following sets are closed:*

$$
\begin{aligned}
Lev(f, \alpha) &:= \{\boldsymbol{x} \in S : f(\boldsymbol{x}) \leq \alpha\}, \\
Con(f, \alpha) &:= \{\boldsymbol{x} \in S : f(\boldsymbol{x}) = \alpha\}.
\end{aligned}
$$

Boundedness and compactness

- **Boundary point:** Given a set $U \subseteq \mathbb{R}^n$, a boundary point of U is a point $\boldsymbol{x} \in \mathbb{R}^n$ satisfying the following: any neighborhood of $\boldsymbol{x}$ contains at least one point in U and at least one point in U^c, i.e., $\forall\, r > 0$, $B(\boldsymbol{x}, r) \cap U \neq \varnothing$ and $B(\boldsymbol{x}, r) \cap U^c \neq \varnothing$.
- **Boundary of U:** The set of all boundary points of a set U is called the boundary of U and is denoted by $bd(U)$.
  Examples: $bd(B(\boldsymbol{c}, r)) = bd(B[\boldsymbol{c}, r]) = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{c}\| = r\}$.
- **Closure of U:** The closure of a set $U \subseteq \mathbb{R}^n$ is defined to be the smallest closed set containing U and denoted by $cl(U)$, i.e.,

$$cl(U) := \cap\{T : U \subseteq T,\ T \textit{ is closed}\}.$$

  Notes: (1) *The closure set is indeed a closed set as an intersection of closed sets.* (2) $cl(U) = U \cup bd(U)$.
- **Boundedness:** A set $U \subseteq \mathbb{R}^n$ is called *bounded* if $\exists\, M > 0$ s.t. $U \subseteq B(\mathbf{0}, M)$.
- **Compactness:** A set $U \subseteq \mathbb{R}^n$ is called *compact* if it is closed and bounded.

Differentiability

- **Directional derivative:** Let f be a real-valued function defined on a set $S \subseteq \mathbb{R}^n$. Let $\boldsymbol{x} \in int(S)$ and let $\mathbf{0} \neq \boldsymbol{d} \in \mathbb{R}^n$. If

$$\lim_{t \to 0} \frac{f(\boldsymbol{x} + t\boldsymbol{d}) - f(\boldsymbol{x})}{t}$$

exists, then it is called the directional derivative of f at $\boldsymbol{x}$ along the direction $\boldsymbol{d}$ and is denoted by $f'(\boldsymbol{x}; \boldsymbol{d})$.

Note that here we do not assume that $\boldsymbol{d}$ is a unit vector $\|\boldsymbol{d}\| = 1$.

- **Partial derivatives:** For $i = 1, 2, \cdots, n$, the directional derivative of f at $\boldsymbol{x}$ along the direction $\boldsymbol{e}_i$ is called the ith partial derivative and is denoted by $\frac{\partial f}{\partial x_i}(\boldsymbol{x})$, i.e.,

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}) = \lim_{t \to 0} \frac{f(\boldsymbol{x} + t\boldsymbol{e}_i) - f(\boldsymbol{x})}{t}.$$

- The gradient of f at $\boldsymbol{x}$ is defined as

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x}), \frac{\partial f}{\partial x_2}(\boldsymbol{x}), \cdots, \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \right)^\top.$$

Continuous differentiability

- **Definition:** A function f defined on an open set $U \subseteq \mathbb{R}^n$ is called continuously differentiable over U if all the partial derivatives exist and are continuous on U.
- **Definition:** A function f is said to be continuously differentiable over a set C if there exists an open set U containing C on which the function is also defined and continuously differentiable.
- Let f be continuously differentiable over U. Then we have

$$f'(\boldsymbol{x};\boldsymbol{d}) = \nabla f(\boldsymbol{x})^\top \boldsymbol{d}, \quad \forall\, \boldsymbol{x} \in U,\ \boldsymbol{d} \in \mathbb{R}^n.$$

- **Proposition:** *Let $f : U \to \mathbb{R}$ be defined on an open set $U \subseteq \mathbb{R}^n$. Assume that f is continuously differentiable over U. Then*

$$\lim_{\boldsymbol{d}\to\boldsymbol{0}} \frac{f(\boldsymbol{x}+\boldsymbol{d}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top \boldsymbol{d}}{\|\boldsymbol{d}\|} = 0, \quad \forall\, \boldsymbol{x} \in U,$$

or equivalently,

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y}-\boldsymbol{x}) + o(\|\boldsymbol{y}-\boldsymbol{x}\|),$$

where $o(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ satisfies $\frac{o(t)}{t} \to 0$ as $t \to 0^+$.

Twice continuous differentiability

- **Definition:** A function f defined on an open set $U \subseteq \mathbb{R}^n$ is called *twice continuously differentiable* over U if all the second order partial derivatives exist and are continuous over U.
- **Proposition:** *Let $f : U \to \mathbb{R}$ be defined on an open set $U \subseteq \mathbb{R}^n$. If f is twice continuously differentiable, then for any $i \neq j$ and any $\boldsymbol{x} \in U$,*

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

- The Hessian of f at a point $\boldsymbol{x} \in U$ is the $n \times n$ matrix

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

If f is twice continuously differentiable over U, the Hessian matrix is symmetric.

Linear and quadratic approximation theorems

There are two main approximation results which are consequences of Taylor's approximation theorem:

1. **Linear approximation theorem:** *Let $f : U \to \mathbb{R}$ be a twice continuously differentiable function over an open set $U \subseteq \mathbb{R}^n$, and let $\boldsymbol{x} \in U$, $r > 0$ satisfy $B(\boldsymbol{x}, r) \subseteq U$. Then for any $\boldsymbol{y} \in B(\boldsymbol{x}, r)$, there exists $\boldsymbol{\xi} \in (\boldsymbol{x}, \boldsymbol{y})$ such that*

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^\top \nabla^2 f(\boldsymbol{\xi})(\boldsymbol{y} - \boldsymbol{x}).$$

2. **Quadratic approximation theorem:** *Let $f : U \to \mathbb{R}$ be a twice continuously differentiable function over an open set $U \subseteq \mathbb{R}^n$, and let $\boldsymbol{x} \in U$, $r > 0$ satisfy $B(\boldsymbol{x}, r) \subseteq U$. Then for any $\boldsymbol{y} \in B(\boldsymbol{x}, r)$,*

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^\top \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + o(\|\boldsymbol{y} - \boldsymbol{x}\|^2).$$