Lecture 1: Mathematical Preliminaries

Shi Pu

School of Data Science (SDS)
The Chinese University of Hong Kong, Shenzhen

Outline

Outline

- $\mathbb{R}^n$ - the set of n-dimensional column vectors with real components endowed with the component-wise addition operator:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix},$$

and the scalar-vector product

$$\lambda \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \lambda x_1 \\ \lambda x_2 \\ \vdots \\ \lambda x_n \end{pmatrix}.$$

- $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ - standard/canonical basis.
- $\mathbf{e}$ and $\mathbf{0}$ - all ones and all zeros column vectors..

- Nonnegative orthant:

$$\mathbb{R}^n_+ = \{(x_1, x_2, \ldots, x_n)^\top : x_1, x_2, \ldots, x_n \geq 0\}.$$

- Positive orthant:

$$\mathbb{R}^n_{++} = \{(x_1, x_2, \ldots, x_n)^\top : x_1, x_2, \ldots, x_n > 0\}.$$

- If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the closed line segment between $\mathbf{x}$ and $\mathbf{y}$ is given by

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

- The open line segment $(\mathbf{x}, \mathbf{y})$ is similarly defined as

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

  for $\mathbf{x} \neq \mathbf{y}$ and $(\mathbf{x}, \mathbf{x}) = \emptyset$
- Unit-simplex:

$$\Delta_n = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0, \mathbf{e}^\top \mathbf{x} = 1\}.$$

- The set of all real valued matrices is denoted by $\mathbb{R}^{m\times n}$.
- $\mathbf{I}_n$ - $n \times n$ identity matrix.
- $\mathbf{0}_{m\times n}$ - $m \times n$ zeros matrix.

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) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
2. (additivity) $\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$ for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$.
3. (homogeneity) $\langle \lambda \mathbf{x}, \mathbf{y} \rangle = \lambda \langle \mathbf{x}, \mathbf{y} \rangle$ for any $\lambda \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
4. (positive definiteness) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ for any $\mathbf{x} \in \mathbb{R}^n$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

Inner Products

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) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
2. (additivity) $\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$ for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$.
3. (homogeneity) $\langle \lambda \mathbf{x}, \mathbf{y} \rangle = \lambda \langle \mathbf{x}, \mathbf{y} \rangle$ for any $\lambda \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
4. (positive definiteness) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ for any $\mathbf{x} \in \mathbb{R}^n$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

- The "dot product"

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

- The "weighted dot product"

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{w}} = \sum_{i=1}^{n} w_i x_i y_i, \text{ where } \mathbf{w} \in \mathbb{R}^n_{++}.$$

Definition

An norm $\|\cdot\|$ on $\mathbb{R}^n$ is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ satisfying

1. (nonnegativity) $\|\mathbf{x}\| \geq 0$ for any $\mathbf{x} \in \mathbb{R}^n$ and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
2. (positive homogeneity) $\|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$.
3. (triangle inequality) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

- One natural way to generate a norm on $\mathbb{R}^n$ is to take any inner product $\langle\cdot,\cdot\rangle$ defined on $\mathbb{R}^n$, and define the associated norm

$$\|\mathbf{x}\| \equiv \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}, \text{ for all } \mathbf{x} \in \mathbb{R}^n,$$

- The norm associated with the dot-product is the so-called Euclidean norm or l_2-norm:

$$\|\mathbf{x}\|_2 \equiv \sqrt{\sum_{i=1}^{n} x_i^2} \text{ for all } \mathbf{x} \in \mathbb{R}^n.$$

- The l_p-norm ($p \geq 1$) is defined by $\|x\|_p \equiv \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$.
- The l_∞-norm is

$$\|\mathbf{x}\|_\infty \equiv \max_{i=1,2,\ldots,n} |x_i|.$$

- It can be shown that

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

Example: $l_{1/2}$ is not a norm. why?

Lemma

For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$*:*

$$|\mathbf{x}^\top \mathbf{y}| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|.$$

Proof.

For any $\lambda \in \mathbb{R}$:

$$\|\mathbf{x} + \lambda \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + 2\lambda \langle \mathbf{x}, \mathbf{y} \rangle + \lambda^2 \|\mathbf{y}\|^2$$

Therefore (why?),

$$4\langle \mathbf{x}, \mathbf{y} \rangle^2 - 4\|\mathbf{x}\|^2 \|\mathbf{y}\|^2 \leq 0,$$

establishing the desired result. $\square$

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

1. (nonnegativity) $\|\mathbf{A}\| \geq 0$ for any $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = \mathbf{0}$.
2. (positive homogeneity) $\|\lambda\mathbf{A}\| = |\lambda|\|\mathbf{A}\|$ for any $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\lambda \in \mathbb{R}$.
3. (triangle inequality) $\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$ for any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$.
4. (submultiplicativity) $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$ for any compatible $\mathbf{A}, \mathbf{B}$.

- Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, the induced matrix norm $\|\mathbf{A}\|_{a,b}$ (called (a,b)-norm) is defined by
$$\|\mathbf{A}\|_{a,b} = \max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_b : \|\mathbf{x}\|_a \leq 1\}.$$
Conclusion:
$$\|\mathbf{A}\mathbf{x}\|_b \leq \|\mathbf{A}\|_{a,b}\|\mathbf{x}\|_a$$
- An induced norm is a norm (satisfies nonnegativity, positive homogeneity and triangle inequality, and submultiplicativity).
- 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 Contd

- Spectral norm: If $\|\cdot\|_a = \|\cdot\|_b = \|\cdot\|_2$, the induced $(2,2)$-norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the maximum singular value of $\mathbf{A}$

$$\|\mathbf{A}\|_2 = \|\mathbf{A}\|_{2,2} = \sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})} \equiv \sigma_{\max}(\mathbf{A}).$$

- l_1-norm: when $\|\cdot\|_a = \|\cdot\|_b = \|\cdot\|_1$, the induced $(1,1)$-matrix norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is given by

$$\|\mathbf{A}\|_1 = \max_{j=1,2,\ldots,n} \sum_{i=1}^{m} |A_{i,j}|.$$

- l_∞-norm: when $\|\cdot\|_a = \|\cdot\|_b = \|\cdot\|_\infty$, the induced (∞,∞)-matrix norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is given by

$$\|\mathbf{A}\|_\infty = \max_{i=1,2,\ldots,m} \sum_{j=1}^{n} |A_{i,j}|.$$

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2}, \ \mathbf{A} \in \mathbb{R}^{m\times n}$$

The Frobenius norm is not an induced norm.
Why is it a norm?

Outline

- Let $\mathbf{A} \in \mathbb{R}^{n \times n}$. Then a nonzero vector $\mathbf{v} \in \mathbb{R}^n$ is called an eigenvector of $\mathbf{A}$ if there exists a $\lambda \in \mathbb{C}$ for which

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}.$$

  The scalar λ is the eigenvalue corresponding to the eigenvector $\mathbf{v}$.
- In general, real-valued matrices can have complex eigenvalues, but when the matrix is *symmetric* the eigenvalues are necessarily *real*.
- The eigenvalues of a symmetric $n \times n$ matrix $\mathbf{A}$ are denoted by

$$\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \ldots \geq \lambda_n(\mathbf{A}).$$

- The maximum eigenvalue is also denoted by $\lambda_{\max}(\mathbf{A})(= \lambda_1(\mathbf{A}))$, and the minimum eigenvalue is also denoted by $\lambda_{\min}(\mathbf{A})(= \lambda_n(\mathbf{A}))$.

Theorem

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

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

- The columns of the matrix **U** constitute an orthogonal basis comprising eigenvectors of **A** and the diagonal elements of **D** are the corresponding eigenvalues.
- A direct result is that $\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^n \lambda_i(\mathbf{A})$ and $\det(\mathbf{A}) = \prod_{i=1}^n \lambda_i(\mathbf{A})$.

Outline

- The open ball with center $\mathbf{c} \in \mathbb{R}^n$ and radius r:

$$B(\mathbf{c}, r) = \{\mathbf{x} : \|\mathbf{x} - \mathbf{c}\| < r\}.$$

- The closed ball with center $\mathbf{c} \in \mathbb{R}^n$ and radius r:

$$B[\mathbf{c}, r] = \{\mathbf{x} : \|\mathbf{x} - \mathbf{c}\| \leq r\}.$$

Definition

Given a set $U \subseteq \mathbb{R}^n$, a point $\mathbf{c} \in U$ is called an interior point of U if there exists $r > 0$ for which $B(\mathbf{c}, r) \subseteq U$.

The set of all interior points of a given set U is called the interior of the set and is denoted by $\mathrm{int}(U)$:

$$\mathrm{int}(U) = \{\mathbf{x} \in U : B(\mathbf{x}, r) \subseteq U \text{ for some } r > 0.\}$$

Examples:

$$\mathrm{int}(\mathbb{R}^n_+) = \mathbb{R}^n_{++}, \quad \mathrm{int}(B[\mathbf{c}, r]) = B(\mathbf{c}, r), \quad \mathrm{int}([\mathbf{x}, \mathbf{y}]) = ?$$

Open and Closed Sets I

- An open set is a set that contains only interior points, meaning that $U = \mathrm{int}(U)$.
- Examples of open sets are open balls (hence the name...) and the positive orthant $\mathbb{R}^n_{++}$.

Result: a union of any number of open sets is an open set and the intersection of a finite number of open sets is open.

- A set $U \subseteq \mathbb{R}^n$ is closed if it contains all the limits of convergent sequences of vectors in U, that is, if $\{\mathbf{x}_i\}_{i=1}^{\infty} \subseteq U$ satisfies $\mathbf{x}_i \to \mathbf{x}^*$ as $i \to \infty$, then $\mathbf{x}^* \in U$.
- A known result states that U is closed iff its complement U^c is open.
- Examples of closed sets are the closed ball $B[\mathbf{c}, r]$, closed lines segments, the nonnegative orthant $\mathbb{R}^n_+$ and the unit simplex Δ_n.

What about $\mathbb{R}^n$? $\emptyset$?

Definition

Given a set $U \subseteq \mathbb{R}^n$, a boundary point of U is a vector $\mathbf{x} \in \mathbb{R}^n$ satisfying the following: any neighborhood of $\mathbf{x}$ contains at least one point in U and at least one point in its complement U^c.

- The set of all boundary points of a set U is denoted by $\mathrm{bd}(U)$.

Examples:

$$\begin{aligned}
\mathrm{bd}(B(\mathbf{c}, r)) &= \\
\mathrm{bd}(B[\mathbf{c}, r]) &= \\
\mathrm{bd}(\mathbb{R}^n_{++}) &= \\
\mathrm{bd}(\mathbb{R}^n_{+}) &= \\
\mathrm{bd}(\mathbb{R}^n) &= \\
\mathrm{bd}(\Delta_n) &=
\end{aligned}$$

- The closure of a set $U \subseteq \mathbb{R}^n$ is denoted by $\mathrm{cl}(U)$ and is defined to be the smallest closed set containing U:

$$\mathrm{cl}(U) = \bigcap \{\, T : U \subseteq T, T \text{ is closed } \}$$

- Another equivalent definition of $\mathrm{cl}(U)$ is:

$$\mathrm{cl}(U) = U \cup \mathrm{bd}(U).$$

Examples:

$$\begin{aligned} \mathrm{cl}(\mathbb{R}^n_{++}) &= \\ \mathrm{cl}(B(\mathbf{c}, r)) &= \\ (\mathbf{x} \neq \mathbf{y}), \mathrm{cl}((\mathbf{x}, \mathbf{y})) &= \end{aligned}$$

- A set $U \subseteq \mathbb{R}^n$ is called bounded if there exists $M > 0$ for which $U \subseteq B(\mathbf{0}, M)$.
- A set $U \subseteq \mathbb{R}^n$ is called compact if it is closed and bounded.
- Examples of compact sets: closed balls, unit simplex, closed line segments.

Outline

Definition

Let f be a function defined on a set $S \subseteq \mathbb{R}^n$. Let $\mathbf{x} \in \mathrm{int}(S)$ and let $\mathbf{d} \in \mathbb{R}^n$. If the limit

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

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

- For any $i = 1, 2, \ldots, n$, if the limit

  $$\lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x})}{t}$$

  exists, then its value is called the i-th partial derivative and is denoted by $\frac{\partial f}{\partial x_i}(\mathbf{x})$.
- If all the partial derivatives of a function f exist at a point $\mathbf{x} \in \mathbb{R}^n$, then the gradient of f at $\mathbf{x}$ is $\nabla f(\mathbf{x}) = (\frac{\partial f}{\partial x_1}(\mathbf{x}), \frac{\partial f}{\partial x_2}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{x}))^\top$.

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. In that case,

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

Proposition

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

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

Another way to write the above result is as follows:

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

where $o(\cdot) : \mathbb{R}^n_+ \to \mathbb{R}$ is a one-dimensional function satisfying $\frac{o(t)}{t} \to 0$ as $t \to 0^+$.

- The partial derivatives ∂f are themselves real-valued functions that can be partially differentiated. The (i, j)-partial derivatives of f at $x \in U$ (if exists) is defined by

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

- 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. In that case, for any $i \neq j$ and any $\mathbf{x} \in U$:

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

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

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

- For twice continuously differentiable functions, the Hessian is a symmetric matrix.

Linear Approximation Theorem

Theorem

Let $f : U \to \mathbb{R}$ be defined on an open set $U \subseteq \mathbb{R}^n$. Suppose that f is twice continuously differentiable over U. Let $\mathbf{x} \in U$ and $r > 0$ satisfy $B(\mathbf{x}, r) \subseteq U$. Then for any $\mathbf{y} \in B(\mathbf{x}, r)$ there exists $\xi \in [\mathbf{x}, \mathbf{y}]$ such that:

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

Theorem

Let $f : U \to \mathbb{R}$ be defined on an open set $U \subseteq \mathbb{R}^n$. Suppose that f is twice continuously differentiable over U. Let $\mathbf{x} \in U$ and $r > 0$ satisfy $B(\mathbf{x}, r) \subseteq U$. Then for any $\mathbf{y} \in B(\mathbf{x}, r)$:

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