Lecture 2: Preliminaries II. Linear Algebra

School of Mathematical Sciences, Xiamen University

1. Basics.

- We will entirely focus on matrices and vectors over the *reals*.
- $\mathbf{x} \in \mathbb{R}^n$: an n-dimensional real vector

  $\mathbf{0}$: the vector of all zeros

  $\mathbf{1}$: the vector of all ones
- $\mathbf{A} \in \mathbb{R}^{m \times n}$: an $m \times n$ matrix with the ith row $\mathbf{A}_{i,:}$ and the jth column $\mathbf{A}_{:,j}$

  $\mathbf{I}_n$: the $n \times n$ identity matrix with the ith column $\mathbf{e}_i$
- Standard properties of the matrix inverse:

$$\mathbf{A}^{-\top} = (\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1} \quad \text{and} \quad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}.$$

- **Orthogonal matrix**

  A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is orthogonal if $\mathbf{A}^\top = \mathbf{A}^{-1}$.

2. Norms.

- **Definition**: Any function, $\|\cdot\| : \mathbb{R}^{m\times n} \mapsto \mathbb{R}$ that satisfies the following properties is called a **norm**:
  (1) Non-negativity:
  $$\|\mathbf{A}\| \geq 0; \quad \|\mathbf{A}\| = 0 \quad \text{if and only if} \quad \mathbf{A} = \mathbf{0}.$$
  (2) Triangle inequality:
  $$\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|.$$
  (3) Scalar multiplication:
  $$\|\alpha\mathbf{A}\| = |\alpha|\|\mathbf{A}\|, \quad \text{for all} \quad \alpha \in \mathbb{R}.$$
- For any norm, we have
  $$\|-\mathbf{A}\| = \|\mathbf{A}\|, \quad |\|\mathbf{A}\| - \|\mathbf{B}\|| \leq \|\mathbf{A} - \mathbf{B}\|.$$
  The latter property is known as the reverse triangle inequality.

3. Vector norms.

- Given $\mathbf{x} \in \mathbb{R}^n$ and $p \geq 1$, we define the vector p-norm as:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

The most common vector p-norms are:

(1) One norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$.

(2) Euclidean (two) norm: $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} = \sqrt{\mathbf{x}^\top \mathbf{x}}$.

(3) Infinity (max) norm: $\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$.

- Cauchy–Schwartz inequality:

$$|\mathbf{x}^\top \mathbf{y}| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

- Hölder's inequality:

$$|\mathbf{x}^\top \mathbf{y}| \le \|\mathbf{x}\|_1 \|\mathbf{y}\|_\infty, \qquad |\mathbf{x}^\top \mathbf{y}| \le \|\mathbf{x}\|_\infty \|\mathbf{y}\|_1$$

- Pythagorean theorem.

  Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are orthogonal, i.e., $\mathbf{x}^\top \mathbf{y} = 0$, if and only if

$$\|\mathbf{x} \pm \mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2.$$

- Another interesting property of the Euclidean norm is that it does not change after pre(post)-multiplication by a matrix with orthonormal columns (rows).

  Given a vector $\mathbf{x} \in \mathbb{R}^n$ and a matrix $\mathbf{V} \in \mathbb{R}^{m \times n}$ with $m \ge n$ and $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_n$:

$$\|\mathbf{V}\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \quad \text{and} \quad \|\mathbf{x}^\top \mathbf{V}^\top\|_2 = \|\mathbf{x}^\top\|_2 = \|\mathbf{x}\|_2.$$

4. Matrix norms

- The Frobenius norm of $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{m\times n}$:

$$\mathbf{A} := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} = \sqrt{\operatorname{tr}(\mathbf{A}^\top \mathbf{A})} = \sqrt{\operatorname{tr}(\mathbf{A}\mathbf{A}^\top)}$$

- Induced matrix norms: Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and an integer $p \geq 1$ we define the matrix p-norm as:

$$\|\mathbf{A}\|_p = \max_{\mathbf{x}\neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \max_{\|\mathbf{x}\|_p=1} \|\mathbf{A}\mathbf{x}\|_p.$$

  There exists a unit norm vector (unit norm in the p-norm) $\mathbf{x}$ such that $\|\mathbf{A}\|_p = \|\mathbf{A}\mathbf{x}\|_p$. The induced matrix p-norms follow the submultiplicativity laws:

$$\|\mathbf{A}\mathbf{x}\|_p \leq \|\mathbf{A}\|_p \|\mathbf{x}\|_p, \qquad \|\mathbf{A}\mathbf{B}\|_p \leq \|\mathbf{A}\|_p \|\mathbf{B}\|_p.$$

- The most common matrix p-norms are:
  (1) One norm: the maximum absolute column sum,

$$\|\mathbf{A}\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}| = \max_{1\le j\le n} \|\mathbf{A}\mathbf{e}_j\|_1.$$

  (2) Infinity norm: the maximum absolute row sum,

$$\|\mathbf{A}\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}| = \max_{1\le i\le m} \|\mathbf{A}^\top\mathbf{e}_i\|_1.$$

  (3) Two (or spectral) norm:

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \max_{\|\mathbf{x}\|_2=1} \sqrt{\mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x}} = \sqrt{\lambda_{\max}(\mathbf{A}^\top\mathbf{A})}.$$

- We have

$$\|\mathbf{A}^\top\|_1 = \|\mathbf{A}\|_\infty, \quad \|\mathbf{A}^\top\|_\infty = \|\mathbf{A}\|_1, \quad \|\mathbf{A}^\top\|_2 = \|\mathbf{A}\|_2,$$

- The matrix two-norm and Frobenius norm are not affected by pre-(or post-) multiplication with matrices whose columns (or rows) are orthonormal vectors:

$$\|\mathbf{U}\mathbf{A}\mathbf{V}^\top\|_2 = \|\mathbf{A}\|_2, \quad \|\mathbf{U}\mathbf{A}\mathbf{V}^\top\|_{\mathrm{F}} = \|\mathbf{A}\|_{\mathrm{F}},$$

  where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices ($\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$) of appropriate dimensions.
- The two and the Frobenius norm can be related by:

$$\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_{\mathrm{F}} \leq \sqrt{\mathrm{rank}(\mathbf{A})}\|\mathbf{A}\|_2 \leq \sqrt{\min\{m,n\}}\|\mathbf{A}\|_2.$$

- The Frobenius norm satisfies :

$$\|\mathbf{A}\mathbf{B}\|_{\mathrm{F}} \leq \|\mathbf{A}\|_2\|\mathbf{B}\|_{\mathrm{F}}, \quad \|\mathbf{A}\mathbf{B}\|_{\mathrm{F}} \leq \|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{B}\|_2.$$

- Matrix Pythagoras. Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$. If $\mathbf{A}^\top\mathbf{B} = \mathbf{0}$ then

$$\|\mathbf{A}+\mathbf{B}\|_{\mathrm{F}}^2 = \|\mathbf{A}\|_{\mathrm{F}}^2 + \|\mathbf{B}\|_{\mathrm{F}}^2.$$

5. Singular value decomposition (SVD)

- Definition: Let m and n be arbitrary positive integers ($m \geq n$ or $m < n$). Given $\mathbf{A} \in \mathbb{R}^{m \times n}$, not necessarily of full rank, a *singular value decomposition* (SVD) of $\mathbf{A}$ is a factorization

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

where $\mathbf{U} \in \mathbb{R}^{m \times m}$ is orthogonal ($\mathbf{U}^{-1} = \mathbf{U}^\top$), $\mathbf{V} \in \mathbb{R}^{n \times n}$ is orthogonal ($\mathbf{V}^{-1} = \mathbf{V}^\top$), and $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is diagonal. In addition, it is assumed that the diagonal entries σ_i of $\mathbf{\Sigma}$ are nonnegative and in nonincreasing order; that is

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0,$$

where $p = \min\{m, n\}$.

- $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$ are called the *singular values* of $\mathbf{A}$.

- Rank SVD or compact SVD or condensed SVD:

$$\mathbf{A} = \begin{bmatrix} \mathbf{U}_r & \mathbf{U}_{\mathrm{c}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_r^\top \\ \mathbf{V}_{\mathrm{c}}^\top \end{bmatrix} = \mathbf{U}_r \boldsymbol{\Sigma}_r \mathbf{V}_r^\top = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

  where $r = \mathrm{rank}(\mathbf{A})$,

$$\mathbf{U}_r = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r \end{bmatrix}, \quad \mathbf{U}_{\mathrm{c}} = \begin{bmatrix} \mathbf{u}_{r+1} & \mathbf{u}_{r+2} & \cdots & \mathbf{u}_m \end{bmatrix},$$

$$\mathbf{V}_r = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_r \end{bmatrix}, \quad \mathbf{V}_{\mathrm{c}} = \begin{bmatrix} \mathbf{v}_{r+1} & \mathbf{v}_{r+2} & \cdots & \mathbf{v}_n \end{bmatrix},$$

  and

$$\boldsymbol{\Sigma}_r = \mathrm{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_r\}.$$

- $\{\sigma_i^2, \mathbf{u}_i\}$ are eigenvalue-eigenvector pairs of $\mathbf{A}\mathbf{A}^\top$, and $\{\sigma_i^2, \mathbf{v}_i\}$ are eigenvalue-eigenvector pairs of $\mathbf{A}^\top\mathbf{A}$:

$$\mathbf{A}\mathbf{A}^\top \mathbf{u}_i = \sigma_i^2 \mathbf{u}_i, \quad \mathbf{A}^\top \mathbf{A} \mathbf{v}_i = \sigma_i^2 \mathbf{v}_i, \quad i = 1, 2, \ldots, p$$

- $\mathbf{u}_i$ is called *left singular vector*, and $\mathbf{v}_i$ is called *right singular vector*: $\mathbf{u}_i^\top \mathbf{A} = \sigma_i \mathbf{v}_i^\top, \quad \mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i, \quad i = 1, 2, \ldots, p$

5.1. Matrix properties via SVD

- Two-norm and Frobenius norm

$$\|\mathbf{A}\|_2 = \sigma_1, \quad \|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2}$$

- range($\mathbf{A}$): *column space* of $\mathbf{A}$, spanned by the columns of $\mathbf{A}$

$$\begin{aligned} \mathrm{range}(\mathbf{A}) : &= \{\mathbf{y} \in \mathbb{R}^m \mid \exists \mathbf{x} \in \mathbb{R}^n \quad s.t. \quad \mathbf{y} = \mathbf{A}\mathbf{x}\} \\ &= \mathrm{span}\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_r\} \end{aligned}$$

- null($\mathbf{A}$): *kernel* or *null space* of $\mathbf{A}$

$$\mathrm{null}(\mathbf{A}) : = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{0}\} = \mathrm{span}\{\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \cdots, \mathbf{v}_n\}$$

- Range and null space of $\mathbf{A}^\top$:

$$\mathrm{range}(\mathbf{A}^\top) = \mathrm{span}\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r\} = \mathrm{null}(\mathbf{A})^\perp$$

$$\mathrm{null}(\mathbf{A}^\top) = \mathrm{span}\{\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \cdots, \mathbf{u}_m\} = \mathrm{range}(\mathbf{A})^\perp$$

5.2. Low-rank approximation

Theorem 1 (Eckart-Young-Mirski)

For any integer k with $1 \leq k < r = \operatorname{rank}(\mathbf{A})$, define

$$\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

Then

$$\|\mathbf{A} - \mathbf{A}_k\|_2 = \min_{\substack{\mathbf{B} \in \mathbb{R}^{m \times n}, \\ \operatorname{rank}(\mathbf{B}) \leq k}} \|\mathbf{A} - \mathbf{B}\|_2 = \sigma_{k+1},$$

and

$$\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}} = \min_{\substack{\mathbf{B} \in \mathbb{R}^{m \times n}, \\ \operatorname{rank}(\mathbf{B}) \leq k}} \|\mathbf{A} - \mathbf{B}\|_{\mathrm{F}} = \sqrt{\sigma_{k+1}^2 + \cdots + \sigma_r^2}.$$

5.3. Moore–Penrose pseudoinverse

- Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ have an SVD (rank form) $\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^\top$. The *Moore–Penrose pseudoinverse* of $\mathbf{A}$, denoted by $\mathbf{A}^\dagger$:

$$\mathbf{A}^\dagger := \mathbf{V}_r \mathbf{\Sigma}_r^{-1} \mathbf{U}_r^\top = \sum_{i=1}^{r} \frac{1}{\sigma_i} \mathbf{v}_i \mathbf{u}_i^\top.$$

- The matrix $\mathbf{A}^\dagger$ is the *unique* matrix satisfying the four equations

$$\mathbf{AXA} = \mathbf{A}, \quad \mathbf{XAX} = \mathbf{X}, \quad (\mathbf{AX})^\top = \mathbf{AX}, \quad (\mathbf{XA})^\top = \mathbf{XA}.$$

- If $\mathbf{A}$ has full column rank, then $\mathbf{A}^\dagger = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top$.
  If $\mathbf{A}$ has full row rank, then $\mathbf{A}^\dagger = \mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top)^{-1}$.

6. QR factorization

- Definition: Let m and n be arbitrary positive integers ($m \geq n$ or $m < n$). Given $\mathbf{A} \in \mathbb{R}^{m \times n}$, not necessarily of full rank, a *full* QR *factorization* of $\mathbf{A}$ is a factorization

$$\mathbf{A} = \mathbf{QR}$$

where $\mathbf{Q} \in \mathbb{R}^{m \times m}$ is orthogonal, and $\mathbf{R} \in \mathbb{R}^{m \times n}$ is upper triangular. For $m \geq n$, a *reduced* QR *factorization* of $\mathbf{A}$ is a factorization

$$\mathbf{A} = \mathbf{Q}_n \mathbf{R}_n$$

where $\mathbf{Q}_n \in \mathbb{R}^{m \times n}$ has orthonormal columns, and

$$\mathbf{R}_n = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \ddots & \vdots \\ & & \ddots & \vdots \\ & & & r_{nn} \end{bmatrix}.$$

7. The least squares problem (LSP)

- LSP: Given $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m$; find $\mathbf{x}_{\text{ls}} \in \mathbb{R}^n$ such that

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}_{\text{ls}}\|_2 = \min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2.$$

  The *least squares solution*, $\mathbf{x}_{\text{ls}}$, maybe *not* unique. Why?

- Note that the 2-norm corresponds to Euclidean distance.

  LSP means we seek a vector $\mathbf{x}_{\text{ls}} \in \mathbb{R}^n$ such that the vector $\mathbf{A}\mathbf{x}_{\text{ls}}$ is the closest point in range($\mathbf{A}$) to $\mathbf{b}$.

  The *residual*, $\mathbf{r}_{\text{ls}} = \mathbf{b} - \mathbf{A}\mathbf{x}_{\text{ls}}$, is unique. Why?

- Define

$$f(\mathbf{x}) := \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 = \mathbf{b}^\top\mathbf{b} - \mathbf{x}^\top\mathbf{A}^\top\mathbf{b} - \mathbf{b}^\top\mathbf{A}\mathbf{x} + \mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x}.$$

  Then the gradient of $f(\mathbf{x})$ is

$$\nabla f(\mathbf{x}) = 2\mathbf{A}^\top\mathbf{A}\mathbf{x} - 2\mathbf{A}^\top\mathbf{b}.$$

- A vector $\mathbf{x}$ is a least squares solution if and only if $\mathbf{x}$ satisfies

$$\mathbf{A}^\top \mathbf{A}\mathbf{x} = \mathbf{A}^\top \mathbf{b},$$

  which is called the *normal equations*.

- The least squares solution $\mathbf{x}$ is unique if and only if $\mathbf{A}^\top \mathbf{A}$ has full rank.

- Moore–Penrose pseudoinverse solution $\mathbf{A}^\dagger \mathbf{b}$:

  Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ have rank $r < n$ and $\mathbf{b} \in \mathbb{R}^m$. Then the vector $\mathbf{A}^\dagger \mathbf{b}$ is the unique least squares solution with minimal 2-norm.