Matrix Analysis and Applications

Chapter 1: Basic Concepts

Instructor: Kai Lu
(http://seit.sysu.edu.cn/teacher/1801)

School of Electronics and Information Technology
Sun Yat-sen University

November 8, 2020

Table of Contents

Textbooks and References

I. **Textbook** (application oriented)

[R1] James P. Reilly, *Matrix Computations for Signal Processing*, to be published (lecture notes available).

II. **References**

- **Undergraduate level**:
  [R2] Gilbert Strang, *Introduction to Linear Algebra*, 5th Ed. Wellesley-Cambridge, 2016.

  [R3] Sheldon Axler, *Linear Algebra Done Right*, 3rd Ed., Springer, 2015.
- **Graduate level**:
  [R4] Gene H. Golub and Charles F. van Loan, *Matrix Computations*, 4th Ed., John Hopkins University Press, 2013.

  [R5] Roger A. Horn and Charles R. Johnson, *Matrix Analysis*, 2nd Ed., Cambridge University Press, 2013.
- **Handbook of matrix computations**:

  [R6] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*, version November 15, 2012.

Table of Contents

Notation

$\mathbb{R}$	real space, or the set of real numbers
$\mathbb{C}$	complex space, or the set of complex numbers
$\mathbb{R}^n$, $\mathbb{C}^n$	n-dimensional real/complex space
$\mathbb{R}^{m\times n}$, $\mathbb{C}^{m\times n}$	set of all $m \times n$ real/complex valued matrices
$\boldsymbol{x}$	column vector
x_i	the i^{th} entry of $\boldsymbol{x}$
$\boldsymbol{A}$	matrix
a_{ij}	the $(i, j)^{\text{th}}$ entry of $\boldsymbol{A}$
$\boldsymbol{a}_i$	the i^{th} column of $\boldsymbol{A}$
$(\cdot)^T$	transpose
$(\cdot)^*$	conjugate
$(\cdot)^H$	Hermitian transpose, i.e., conjugate transpose
$(\cdot)^{-1}$	inverse
$(\cdot)^\dagger$	pseudo-inverse
$\mathrm{tr}(\cdot)$	trace
$\|\cdot\|_p$	vector ℓ_p-norm, $p \in \mathbb{R}$ yet, usually, $p \in \{0, 1, 2, \infty\}$

Notation (cont'd)

1. **Vector**: $\boldsymbol{x} \in \mathbb{R}^n$ means that $\boldsymbol{x}$ is a real-valued n-dimensional column vector, i.e.,

$$\boldsymbol{x} \triangleq \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \ x_i \in \mathbb{R} \text{ for all } i.$$

Similarly, $\boldsymbol{x} \in \mathbb{C}^n$ means that $\boldsymbol{x}$ is a complex-valued n-dimensional column vector.

Notation (cont'd)

1. **Vector**: $\boldsymbol{x} \in \mathbb{R}^n$ means that $\boldsymbol{x}$ is a real-valued n-dimensional column vector, i.e.,

$$\boldsymbol{x} \triangleq \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \, x_i \in \mathbb{R} \text{ for all } i.$$

   Similarly, $\boldsymbol{x} \in \mathbb{C}^n$ means that $\boldsymbol{x}$ is a complex-valued n-dimensional column vector.

2. **Transpose**: Let $\boldsymbol{x} \in \mathbb{R}^n$. The notation $\boldsymbol{x}^T$ means that

$$\boldsymbol{x}^T = \begin{bmatrix} x_1, & x_2, & \cdots, & x_n \end{bmatrix}.$$

Notation (cont'd)

3. **Hermitian transpose**: Let $\boldsymbol{x} \in \mathbb{C}^n$. The notation $\boldsymbol{x}^H$ means that

$$\boldsymbol{x}^H = \left[x_1^*, \quad x_2^*, \quad \cdots, \quad x_n^*\right],$$

where the superscript * denotes the complex conjugate.

Notation (cont'd)

3. **Hermitian transpose**: Let $\boldsymbol{x} \in \mathbb{C}^n$. The notation $\boldsymbol{x}^H$ means that

$$\boldsymbol{x}^H = \begin{bmatrix} x_1^*, & x_2^*, & \cdots, & x_n^* \end{bmatrix},$$

where the superscript * denotes the complex conjugate.

4. **Matrix**: $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ (resp. $\boldsymbol{A} \in \mathbb{C}^{m\times n}$) means that $\boldsymbol{A}$ is a real-valued (resp. complex-valued) $m \times n$ matrix, i.e.,

$$\boldsymbol{A} \triangleq \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \; a_{ij} \in \mathbb{R} \text{ (resp. } a_{ij} \in \mathbb{C}) \text{ for all } i.$$

Notation (cont'd)

5. Unless otherwise stated, we denote the i^{th} **column** of a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ as $\boldsymbol{a}_i \in \mathbb{R}^m$, i.e.,

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1, & \boldsymbol{a}_2, & \cdots, & \boldsymbol{a}_n \end{bmatrix}.$$

The same notation applies to $\boldsymbol{A} \in \mathbb{C}^{m\times n}$.

6. **Transpose**: Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. The notation $\boldsymbol{A}^T$ means that

$$\boldsymbol{A}^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \cdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{n\times m}.$$

Or, if $\boldsymbol{B} = \boldsymbol{A}^T$, we have $b_{ij} = a_{ji}$.
- Property: $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$.

Notation (cont'd)

7. **Hermitian transpose**: Let $\boldsymbol{A} \in \mathbb{C}^{m\times n}$. The notation $\boldsymbol{A}^H$ means that

$$\boldsymbol{A}^H = \begin{bmatrix} a_{11}^* & a_{21}^* & \cdots & a_{m1}^* \\ a_{12}^* & a_{22} & \cdots & a_{m2}^* \\ \vdots & \cdots & \ddots & \vdots \\ a_{1n}^* & a_{2n}^* & \cdots & a_{mn}^* \end{bmatrix} \in \mathbb{C}^{n\times m}.$$

Or, if $\boldsymbol{B} = \boldsymbol{A}^H$, we have $b_{ij} = a_{ji}^*$.
- Property: $(\boldsymbol{A}\boldsymbol{B})^H = \boldsymbol{B}^H\boldsymbol{A}^H$.

Notation (cont'd)

7. **Hermitian transpose**: Let $\boldsymbol{A} \in \mathbb{C}^{m \times n}$. The notation $\boldsymbol{A}^H$ means that

$$\boldsymbol{A}^H = \begin{bmatrix} a_{11}^* & a_{21}^* & \cdots & a_{m1}^* \\ a_{12}^* & a_{22} & \cdots & a_{m2}^* \\ \vdots & \dots & \ddots & \vdots \\ a_{1n}^* & a_{2n}^* & \cdots & a_{mn}^* \end{bmatrix} \in \mathbb{C}^{n \times m}.$$

   Or, if $\boldsymbol{B} = \boldsymbol{A}^H$, we have $b_{ij} = a_{ji}^*$.

   - Property: $(\boldsymbol{A}\boldsymbol{B})^H = \boldsymbol{B}^H \boldsymbol{A}^H$.

8. **Trace**: Given $\boldsymbol{A} \in \mathbb{R}^{m \times m}$, the trace of $\boldsymbol{A}$ is

$$\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{m} a_{ii}.$$

   - Properties:
     - $\operatorname{tr}(\boldsymbol{A}^T) = \operatorname{tr}(\boldsymbol{A})$
     - $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{B}\boldsymbol{A})$ for $\boldsymbol{A}, \boldsymbol{B}$ of appropriate size.

Notation (cont'd)

Two typical applications of **trace**:

- Some operations that are difficult to specify without resorting to summation notation can be specified using matrix products and the trace operator. For example,

$$\|\boldsymbol{A}\|_F \triangleq \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\operatorname{tr}\left(\boldsymbol{A}\boldsymbol{A}^T\right)}.$$

- Writing an expression in terms of the trace operator opens up opportunities to manipulate the expression using many useful identities. For example,

$$\operatorname{tr}(\boldsymbol{ABC}) = \operatorname{tr}(\boldsymbol{BCA}) = \operatorname{tr}(\boldsymbol{CAB}),$$

which is invariant to cyclic permutation.

Table of Contents

Vector Space

A **vector space** is a set $\mathbb{V}$ along with an addition on $\mathbb{V}$ and a scalar multiplication on $\mathbb{V}$ such that the following properties hold:

1) **commutativity**: $\boldsymbol{u}+\boldsymbol{v}=\boldsymbol{v}+\boldsymbol{u}$ for all $\boldsymbol{u},\boldsymbol{v}\in\mathbb{V}$;
2) **associativity**: $(\boldsymbol{u}+\boldsymbol{v})+\boldsymbol{w}=\boldsymbol{u}+(\boldsymbol{v}+\boldsymbol{w})$ and $(ab)\boldsymbol{v}=a(b\boldsymbol{v})$ for all $\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}\in\mathbb{V}$ and all $a,b\in\mathbb{R}$;
3) **additive identity**: there exists an element $\mathbf{0}\in\mathbb{V}$ such that $\boldsymbol{v}+\mathbf{0}=\boldsymbol{v}$ for all $\boldsymbol{v}\in\mathbb{V}$;
4) **additive inverse**: for every $\boldsymbol{v}\in\mathbb{V}$, there exists $\boldsymbol{w}\in\mathbb{V}$ such that $\boldsymbol{v}+\boldsymbol{w}=\mathbf{0}$;
5) **multiplicative identity**: $1\boldsymbol{v}=\boldsymbol{v}$ for all $\boldsymbol{v}\in\mathbb{V}$;
6) **distributive properties**: $a(\boldsymbol{u}+\boldsymbol{v})=a\boldsymbol{u}+a\boldsymbol{v}$ and $(a+b)\boldsymbol{v}=a\boldsymbol{v}+b\boldsymbol{v}$ for all $a,b\in\mathbb{R}$ and all $\boldsymbol{u},\boldsymbol{v}\in\mathbb{V}$.

Examples

Example 1

The simplest vector space contains only one point. In other words, $\{\mathbf{0}\}$ is a vector space (widely known as **trivial subspace**).

Examples

Example 1

The simplest vector space contains only one point. In other words, $\{\mathbf{0}\}$ is a vector space (widely known as **trivial subspace**).

Example 2

$\mathbb{R}^\infty$ is defined to be the set of all sequences of elements of $\mathbb{R}$:

$$\mathbb{R}^\infty \triangleq \{(x_1, x_2, \cdots) : x_j \in \mathbb{R} \text{ for } j = 1, 2, \cdots\}.$$

Addition and scalar multiplication on $\mathbb{R}^\infty$ are defined as expected:

$$(x_1, x_2, \cdots) + (y_1, y_2, \cdots) = (x_1 + y_1, x_2 + y_2, \cdots),$$

$$\lambda(x_1, x_2, \cdots) = (\lambda x_1, \lambda x_2, \cdots).$$

With these definitions, $\mathbb{R}^\infty$ becomes a vector space over $\mathbb{R}$.

Subspace

Definition: A subset $\mathbb{U}$ of $\mathbb{V}$ is called a **subspace** of $\mathbb{V}$ if $\mathbb{U}$ is also a vector space (using the same addition and scalar multiplication as on $\mathbb{V}$).

Subspace

Definition: A subset $\mathbb{U}$ of $\mathbb{V}$ is called a **subspace** of $\mathbb{V}$ if $\mathbb{U}$ is also a vector space (using the same addition and scalar multiplication as on $\mathbb{V}$).

Conditions for a subspace: A subset $\mathbb{U}$ of $\mathbb{V}$ is a subspace of $\mathbb{V}$ if and only if $\mathbb{U}$ satisfies the following three conditions:

1) **additive identity**: $\mathbf{0} \in \mathbb{U}$;
2) **closed under addition**: $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{U}$ implies $\boldsymbol{u} + \boldsymbol{v} \in \mathbb{U}$;
3) **closed under scalar multiplication**: $a \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{U}$ implies $a\boldsymbol{u} \in \mathbb{U}$.

Subspace

Definition: A subset $\mathbb{U}$ of $\mathbb{V}$ is called a **subspace** of $\mathbb{V}$ if $\mathbb{U}$ is also a vector space (using the same addition and scalar multiplication as on $\mathbb{V}$).

Conditions for a subspace: A subset $\mathbb{U}$ of $\mathbb{V}$ is a subspace of $\mathbb{V}$ if and only if $\mathbb{U}$ satisfies the following three conditions:

1) **additive identity**: $\mathbf{0} \in \mathbb{U}$;
2) **closed under addition**: $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{U}$ implies $\boldsymbol{u} + \boldsymbol{v} \in \mathbb{U}$;
3) **closed under scalar multiplication**: $a \in \mathbb{R}$ and $\boldsymbol{u} \in \mathbb{U}$ implies $a\boldsymbol{u} \in \mathbb{U}$.

Example 3 (straight lines in planar space)

Consider the vector space $\mathbb{R}^2$, and let $\mathbb{L} \triangleq \{(x, y)\}|y = \alpha x\}$ be a straight line through the origin. Clearly, $\mathbb{L}$ is a subset of $\mathbb{R}^2$. Moreover, $\mathbb{L}$ is a vector space of $\mathbb{R}^2$ because $\mathbb{L}$ satisfies the conditions above for vector space. This example shows that it is possible for one vector space to properly contain other vector spaces.

Example 4 (spanning space)

Given a collection of vectors $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\} \subseteq \mathbb{R}^m$,

$$\mathrm{span}\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\} \triangleq \left\{ \boldsymbol{y} \in \mathbb{R}^m \,\middle|\, \boldsymbol{y} = \sum_{i=1}^{n} \beta_i \boldsymbol{a}_i = \boldsymbol{A}\boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^n \right\}.$$

In other words, $\mathrm{span}\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$ is the set of all linear combinations of $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n$.

Example 4 (spanning space)

Given a collection of vectors $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\} \subseteq \mathbb{R}^m$,

$$\text{span}\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\} \triangleq \left\{ \boldsymbol{y} \in \mathbb{R}^m \,\middle|\, \boldsymbol{y} = \sum_{i=1}^{n} \beta_i \boldsymbol{a}_i = \boldsymbol{A}\boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^n \right\}.$$

In other words, $\text{span}\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$ is the set of all linear combinations of $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n$.

Example 5 (orthogonal complement subspace)

Given a subset $\mathbb{S} \subseteq \mathbb{R}^m$,

$$\mathbb{S}_\perp \triangleq \left\{ \boldsymbol{y} \in \mathbb{R}^m \,\middle|\, \boldsymbol{y}^T \boldsymbol{x} = 0 \text{ for all } \boldsymbol{x} \in \mathbb{S} \right\}. \tag{1}$$

Note that $\boldsymbol{y}^T \boldsymbol{x} = \sum_{i=1}^{m} x_i y_i$ is the inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$.

Question 6

Example 3 shows that straight lines through the origin in $\mathbb{R}^2$ *are subspaces, but what about straight lines not through the origin? Further, what about curved lines through the origin – can some of them be subspaces of* $\mathbb{R}^2$*?*

Question 6

Example 3 shows that straight lines through the origin in $\mathbb{R}^2$ are subspaces, but what about straight lines not through the origin? Further, what about curved lines through the origin – can some of them be subspaces of $\mathbb{R}^2$?

Question 7

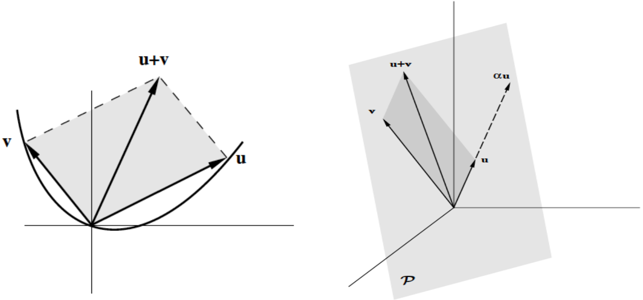
What about planes through the origin in $\mathbb{R}^3$?

Question 6

Example 3 shows that straight lines through the origin in $\mathbb{R}^2$ are subspaces, but what about straight lines not through the origin? Further, what about curved lines through the origin – can some of them be subspaces of $\mathbb{R}^2$?

Question 7

What about planes through the origin in $\mathbb{R}^3$?



(a) Curves in $\mathbb{R}^2$. (b) Planes in $\mathbb{R}^3$.

Figure 1: Subspaces or not?

Commonly Seen Subspaces

1. **Range space**: Given $\boldsymbol{A} \in \mathbb{R}^{m\times n}$,

$$\mathcal{R}(\boldsymbol{A}) \triangleq \{\boldsymbol{y} \in \mathbb{R}^m \,|\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \text{ for } \boldsymbol{x} \in \mathbb{R}^n\},$$

which is essentially the same as span: $\mathcal{R}(\boldsymbol{A}) = \text{span}\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$.

2. **Nullspace**: Given $\boldsymbol{A} \in \mathbb{R}^{m\times n}$,

$$\mathcal{N}(\boldsymbol{A}) \triangleq \{\boldsymbol{x} \in \mathbb{R}^n \,|\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}.$$

In light of Eq. (1), it is straightforward that $\mathcal{R}(\boldsymbol{A})_\perp = \mathcal{N}(\boldsymbol{A}^T)$.
(Note: the proof is left as an exercise.)

Commonly Seen Subspaces

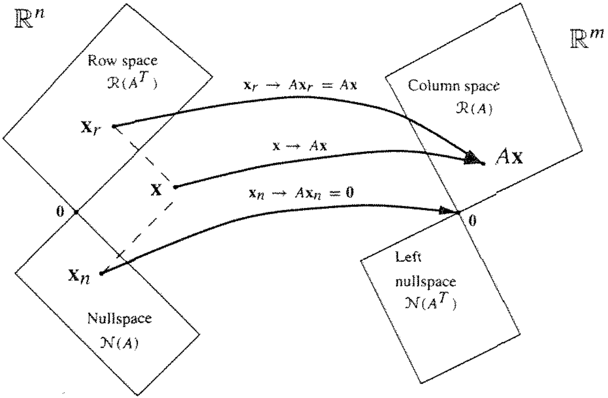


Figure 2: The four fundamental subspaces of a matrix operator.

More Definitions

1. **Linear independence**: A collection of vectors $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\} \subseteq \mathbb{R}^m$, is said to be **linear independence** if

$$\sum_{i=1}^{n} \beta_i \boldsymbol{a}_i = \mathbf{0} \text{ for some } \boldsymbol{\beta} \in \mathbb{R}^n \Longrightarrow \boldsymbol{\beta} = \mathbf{0}.$$

2. A **maximal linearly independent set** is a collection of vectors that contains the maximum number of linearly independent vectors spanning the subspace.

More Definitions

1. **Linear independence**: A collection of vectors $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\} \subseteq \mathbb{R}^m$, is said to be **linear independence** if

$$\sum_{i=1}^{n} \beta_i \boldsymbol{a}_i = \mathbf{0} \text{ for some } \boldsymbol{\beta} \in \mathbb{R}^n \Longrightarrow \boldsymbol{\beta} = \mathbf{0}.$$

2. A **maximal linearly independent set** is a collection of vectors that contains the maximum number of linearly independent vectors spanning the subspace.
3. A **basis** for a subspace is *any* maximal linearly independent set within the subspace.
4. The **dimension** of a subspace $\mathbb{S}$, denoted by $\dim \mathbb{S}$, is the maximum number of linear independent vectors that spans $\mathbb{S}$.
   - Properties:
     - Given $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ where $m \geq n$, $\dim \mathcal{R}(\boldsymbol{A}) \leq n$.
     - Given $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\dim \mathcal{R}(\boldsymbol{A}) + \dim \mathcal{N}(\boldsymbol{A}) = n$.

Some Basic Concepts

Rank: The rank of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, denoted by $\operatorname{rank}(\boldsymbol{A})$, is the maximum number of linearly *independent* columns of $\boldsymbol{A}$.

1. By definition, we have $\dim \mathcal{R}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A})$ and, thus, range space is also known as column space (cf. Fig. 2).

2. $\operatorname{rank}(\boldsymbol{A})$ is the maximum number of linearly independent rows of $\boldsymbol{A}$.
   - Property: $\operatorname{rank}(\boldsymbol{A}^T) = \operatorname{rank}(\boldsymbol{A})$.

Some Basic Concepts (cont'd)

3. $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is said to have
   - **full rank** if $\mathrm{rank}(\boldsymbol{A})$, i.e., either the collection of all columns of $\boldsymbol{A}$ is linearly independent, or the collection of all rows of $\boldsymbol{A}$ is linearly independent.
   - **rank deficient** if $\mathrm{rank}(\boldsymbol{A}) < \min\{m, n\}$.
4. Additionally, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is said to have
   - **full column rank** if the collection of all columns of $\boldsymbol{A}$ is linearly independent, or equivalently, if $\mathrm{rank}(\boldsymbol{A}) = n$.
   - **full row rank** if the collection of all rows of $\boldsymbol{A}$ is linearly independent, or equivalently, if $\mathrm{rank}(\boldsymbol{A}) = m$.

Some Basic Concepts (cont'd)

Spark: The spark of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, denoted by $\mathrm{spark}(\boldsymbol{A})$, is the minimum number of linearly *dependent* columns of $\boldsymbol{A}$.

Properties: Given $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with $m \geq n$, we have

(1) $\mathrm{spark}(\boldsymbol{A}) \in \{1, 2, \cdots, n\} \cup \{+\infty\}$,

(2) $\mathrm{spark}(\boldsymbol{A}) = 1$ if and only if $\boldsymbol{A}$ has a zero column,

(3) $\mathrm{spark}(\boldsymbol{A}) = +\infty \Longleftrightarrow \mathrm{rank}(\boldsymbol{A}) = n$,

(4) if $\mathrm{spark}(\boldsymbol{A}) \neq +\infty$, then $\mathrm{spark}(\boldsymbol{A}) \leq \mathrm{rank}(\boldsymbol{A}) + 1$.

---

[1] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization", *Proc. Natl. Acad. Sci., vol. 100, no. 5, pp. 2197–2202, 2003.*

Some Basic Concepts (cont'd)

Spark: The spark of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, denoted by $\mathrm{spark}(\boldsymbol{A})$, is the minimum number of linearly *dependent* columns of $\boldsymbol{A}$.

Properties: Given $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with $m \geq n$, we have

(1) $\mathrm{spark}(\boldsymbol{A}) \in \{1, 2, \cdots, n\} \cup \{+\infty\}$,

(2) $\mathrm{spark}(\boldsymbol{A}) = 1$ if and only if $\boldsymbol{A}$ has a zero column,

(3) $\mathrm{spark}(\boldsymbol{A}) = +\infty \Longleftrightarrow \mathrm{rank}(\boldsymbol{A}) = n$,

(4) if $\mathrm{spark}(\boldsymbol{A}) \neq +\infty$, then $\mathrm{spark}(\boldsymbol{A}) \leq \mathrm{rank}(\boldsymbol{A}) + 1$.

Remark 8

The notion of spark of a matrix was introduced by Donoho and Elad.[1] *It is strictly related to Compressed Sensing. The word "spark" comes from a verbal fusion of "sparse" and "rank".*

[1] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization", *Proc. Natl. Acad. Sci., vol. 100, no. 5, pp. 2197–2202, 2003.*

Some Basic Concepts (cont'd)

Example 9

Given $\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 1 & 2 & 0 & 2 \\ 1 & 2 & 0 & 3 \\ 1 & 0 & -3 & 4 \end{bmatrix}$, the spark of this matrix equals 3 because:

- There is no set of 1 column of $\boldsymbol{A}$ that are linearly dependent.
- There is no set of 2 columns of $\boldsymbol{A}$ that are linearly dependent.
- But there is a set of 3 column of $\boldsymbol{A}$ that are linearly dependent.

Specifically, the first three columns of $\boldsymbol{A}$ are linearly dependent because

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 2 \\ 2 \\ 2 \\ 0 \end{pmatrix} + \frac{1}{3}\begin{pmatrix} 0 \\ 0 \\ 0 \\ -3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{2}$$

Some Matrix Concepts (cont'd)

1. **Singularity/Nonsingularity**: A square matrix $\boldsymbol{A} \in \mathbb{R}^{m\times m}$, is called **nonsingular** if

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \Longleftrightarrow \boldsymbol{x} = \boldsymbol{0},$$

and **singular** otherwise.

   - A nonsingular matrix $\boldsymbol{A}$ can simply be seen as a square matrix whose columns $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m$ are linearly independent.

2. **Inverse**: A square matrix $\boldsymbol{A} \in \mathbb{R}^{m\times m}$ is **invertible** if there exists a matrix $\boldsymbol{A}^{-1}$, called the inverse of $\boldsymbol{A}$, such that $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$.
- Properties:
   - If $\boldsymbol{A}$ is nonsingular, it is also invertible (the converse is also true).
   - $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$.
   - $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$, where $\boldsymbol{A}$, $\boldsymbol{B} \in \mathbb{R}^{m\times m}$.
   - $(\boldsymbol{A}^{-1})^T = (\boldsymbol{A}^T)^{-1}$.

Table of Contents

Determinant

Consider a square matrix $\boldsymbol{A} \in \mathbb{R}^{m \times m}$.

1. If $\boldsymbol{A} = a \in \mathbb{R}$, **determinant** is $\det \boldsymbol{A} = a$.

2. If $\boldsymbol{A} \in \mathbb{R}^{m \times m}$, $m > 1$, **determinant** is defined inductively:
   - Define $\boldsymbol{A}_{ij}$ as the submatrix obtained from $\boldsymbol{A}$ by deleting the i^{th} row and j^{th} column of $\boldsymbol{A}$.
   - $\det(\boldsymbol{A}_{ij})$ is called the **minor** associated with a_{ij} of $\boldsymbol{A}$, and

$$c_{ij} = (-1)^{i+j} \det(\boldsymbol{A}_{ij}) \tag{3}$$

     is called the **cofactor** of a_{ij}.
   - **cofactor expansion**:

$$\det(\boldsymbol{A}) = \sum_{j=1}^{m} a_{ij} c_{ij}, \text{ for any } i = 1, \cdots, m$$

$$\det(\boldsymbol{A}) = \sum_{i=1}^{m} a_{ij} c_{ij}, \text{ for any } j = 1, \cdots, m$$

Properties of Determinant

1) $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$, $\forall \boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times m}$
2) $\det(\boldsymbol{A}) = \det(\boldsymbol{A}^T)$
3) $\det(\alpha\boldsymbol{A}) = \alpha^m \det(\boldsymbol{A})$, $\forall \boldsymbol{A} \in \mathbb{R}^{m\times m}$
4) $\det(\boldsymbol{A}) = 0 \Longleftrightarrow \boldsymbol{A}$ is singular (recalling that $\det(\boldsymbol{A}) = \prod_i \lambda_i$)
5) For a nonsingular $\boldsymbol{A}$, $\det(\boldsymbol{A}^{-1}) = 1/\det(\boldsymbol{A})$
6) If $\boldsymbol{B} \in \mathbb{R}^{m\times m}$ is nonsingular, then $\det(\boldsymbol{B}^{-1}\boldsymbol{AB}) = \det(\boldsymbol{A})$
7) $\boldsymbol{A}^{-1} = \tilde{\boldsymbol{A}}/\det(A)$, where the **adjoint** $\tilde{\boldsymbol{A}}$ of $\boldsymbol{A}$ is defined as

$$\tilde{\boldsymbol{A}} \triangleq \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{bmatrix}^T, \; {c_{ij}}'s \text{ being the cofactors of } \boldsymbol{A},$$

as defined in Eq. (3).

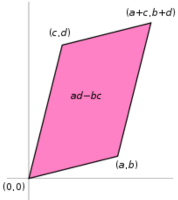
Physical Meaning of Determinant

1) When $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ has linearly independent columns, the volume of the n-dimensional parallelepiped generated by the columns of $\boldsymbol{A}$ is given by

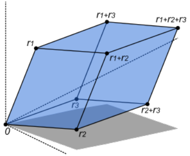
$$V_n = \left[\det\left(\boldsymbol{A}^T\boldsymbol{A}\right)\right]^{1/2}. \tag{4}$$

In particular, if $\boldsymbol{A}$ is square (i.e., $m = n$), then

$$V_n = |\det(\boldsymbol{A})|. \tag{5}$$



(a) Parallelogram



(b) Parallelepiped

2) $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} \Longrightarrow \det(\boldsymbol{Y}) = \det(\boldsymbol{A})\det(\boldsymbol{X}) \Longrightarrow \det(\boldsymbol{A}) = \frac{\det(\boldsymbol{Y})}{\det(\boldsymbol{X})}$

Table of Contents

Inner Product and Angle

The motivation for defining inner product came initially from the norms of vectors on $\mathbb{R}^2$ and $\mathbb{R}^3$.

1. **Inner product** of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \triangleq \sum_{i=1}^{n} y_i x_i = \boldsymbol{y}^T \boldsymbol{x}$.
    - $\boldsymbol{x}, \boldsymbol{y}$ are called **orthogonal** if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.
    - $\boldsymbol{x}, \boldsymbol{y}$ are called **parallel** if $\boldsymbol{x} = \alpha \boldsymbol{y}$ for some $\alpha \in \mathbb{R}$. In this case we have $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \pm \|\boldsymbol{x}\|_2 \, \|\boldsymbol{y}\|_2$.

Inner Product and Angle

The motivation for defining inner product came initially from the norms of vectors on $\mathbb{R}^2$ and $\mathbb{R}^3$.

1. **Inner product** of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \triangleq \sum_{i=1}^{n} y_i x_i = \boldsymbol{y}^T \boldsymbol{x}$.
   - $\boldsymbol{x}, \boldsymbol{y}$ are called **orthogonal** if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.
   - $\boldsymbol{x}, \boldsymbol{y}$ are called **parallel** if $\boldsymbol{x} = \alpha \boldsymbol{y}$ for some $\alpha \in \mathbb{R}$. In this case we have $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \pm \|\boldsymbol{x}\|_2 \, \|\boldsymbol{y}\|_2$.
2. **Angle** between $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$: $\theta \triangleq \cos^{-1}\left(\frac{\boldsymbol{y}^T \boldsymbol{x}}{\|\boldsymbol{x}\|_2 \, \|\boldsymbol{y}\|_2} \right)$.
   - $\boldsymbol{x}$ and $\boldsymbol{y}$ are called **orthogonal** if $\theta = \pm\pi/2$.
   - $\boldsymbol{x}$ and $\boldsymbol{y}$ are called **parallel** if $\theta = 0$ or $\theta = \pm\pi$.

Remark 10

The word **orthogonal** *comes from the Greek word* **orthogonios**, *which means right-angled.*

Vector Norms

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called a **vector norm** if

1) $f(\boldsymbol{x}) \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^n$; (positivity)
2) $f(\boldsymbol{x}) = 0$ if and only if $\boldsymbol{x} = \mathbf{0}$; (separate points)
3) $f(\boldsymbol{x} + \boldsymbol{y}) \leq f(\boldsymbol{x}) + f(\boldsymbol{y})$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$; (triangle inequality)
4) $f(\alpha \boldsymbol{x}) = |\alpha| f(\boldsymbol{x})$ for any $\alpha \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^n$. (absolute homogeneity)

- Vector norm is used to measure the length of a vector.
- We usually use the notation $\|\cdot\|$ to denote a norm.

Examples

ℓ_p**-norm:**

$$\|\boldsymbol{x}\|_p \triangleq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \ p \geq 1 \tag{6}$$

1. ℓ_1**-norm (Manhattan norm)**:

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i| \tag{7}$$

2. ℓ_2**-norm (Euclidean norm)**:

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} = (\boldsymbol{x}^T\boldsymbol{x})^{1/2} \tag{8}$$

3. ℓ_∞**-norm**:

$$\|\boldsymbol{x}\|_\infty = \max_{i=1,\cdots,n} |x_i| \tag{9}$$

Note: Strictly speaking, $\|\cdot\|_p$ for $0 < p < 1$ is not a norm because it violates the triangle inequality.

Important Inequalities for Inner Product and Vector Norms

1. **Cauchy-Schwartz inequality**:

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| = \|\boldsymbol{x}^T \boldsymbol{y}\|_1 \leq \|\boldsymbol{x}\|_2 \, \|\boldsymbol{y}\|_2,$$

where the equality holds if and only if $\boldsymbol{x} = \alpha \boldsymbol{y}$ for some $\alpha \in \mathbb{R}$.

2. **Hölder inequality**:

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\|_p \, \|\boldsymbol{y}\|_q,$$

where $1/p + 1/q = 1$, $\forall p \geq 1$.

Some Notable Special Cases of Hölder Inequality

1. **Counting measure**

$$\sum_{k=1}^{n} |x_k\, y_k| \leq \left(\sum_{k=1}^{n} |x_k|^p\right)^{\frac{1}{p}} \left(\sum_{k=1}^{n} |y_k|^q\right)^{\frac{1}{q}},$$

where $\forall (x_1, \cdots, x_n), (y_1, \cdots, y_n) \in \mathbb{R}^n$ or $\mathbb{C}^n$.

2. **Lebesgue measure**

$$\int_S |f(x)\, g(x)| \leq \left(\int_{\mathcal{S}} |f(x)|^p\right)^{\frac{1}{p}} \left(\int_{\mathcal{S}} |g(x)|^q\right)^{\frac{1}{q}},$$

where $\mathcal{S}$ is a measurable subset of $\mathbb{R}^n$ with the Lebesgue measure, and f and g are measurable real- or complex-valued functions on $\mathcal{S}$.

Some Notable Special Cases of Hölder Inequality (cont'd)

3. **Probability measure**

$$\mathbb{E}\left[|X\,Y|\right] \leq \left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}} \left(\mathbb{E}\left[|Y|^q\right]\right)^{\frac{1}{q}},$$

where X and Y are RVs on Ω w.r.t. the probability space $(\Omega, \mathcal{F}, \mathcal{P})$.

4. **Product measure**

$$\begin{aligned} &\int_{\mathcal{S}_1}\int_{\mathcal{S}_2} |f(x,y)\,g(x,y)|\,\mu_2(\mathrm{d}y)\mu_1(\mathrm{d}x) \\ \leq\ & \left(\int_{\mathcal{S}_1}\int_{\mathcal{S}_2} |f(x,y)|^p\,\mu_2(\mathrm{d}y)\mu_1(\mathrm{d}x)\right)^{\frac{1}{p}} \\ & \times \left(\int_{\mathcal{S}_1}\int_{\mathcal{S}_2} |g(x,y)|^q\,\mu_2(\mathrm{d}y)\mu_1(\mathrm{d}x)\right)^{\frac{1}{q}}, \end{aligned}$$

where two σ-finite measure spaces $(\mathcal{S}_1, \Sigma_1, \mu_1)$ and $(\mathcal{S}_2, \Sigma_2, \mu_2)$ define the product measure space by $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$, $\Sigma = \Sigma_1 \times \Sigma_2$, and $\mu = \mu_1 \times \mu_2$.

The Relation Between Various Norms

1. Basic relation:

$$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_1 \le \sqrt{n}\,\|\boldsymbol{x}\|_2 \le n\,\|\boldsymbol{x}\|_\infty$$

2. **Minkowski inequality** $(1 \le p \le \infty)$:

$$\|\boldsymbol{x}+\boldsymbol{y}\|_p \le \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p,$$

with equality for $1 < p < \infty$ if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are positively linearly dependent, i.e., $\boldsymbol{x} = \lambda \boldsymbol{y}$ for some $\lambda \ge 0$ or $\boldsymbol{y} = \mathbf{0}$.

The Relation Between Various Norms

1. Basic relation:

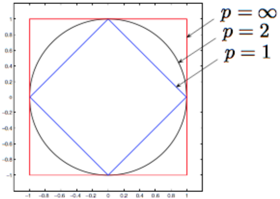
$$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_1 \le \sqrt{n}\,\|\boldsymbol{x}\|_2 \le n\,\|\boldsymbol{x}\|_\infty$$

2. **Minkowski inequality** $(1 \le p \le \infty)$:

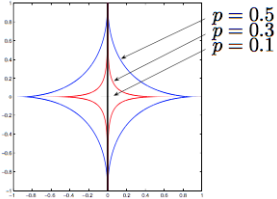
$$\|\boldsymbol{x}+\boldsymbol{y}\|_p \le \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p,$$

with equality for $1 < p < \infty$ if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are positively linearly dependent, i.e., $\boldsymbol{x} = \lambda\boldsymbol{y}$ for some $\lambda \ge 0$ or $\boldsymbol{y} = \boldsymbol{0}$.

3. Regions of $\|\boldsymbol{x}\|_p = 1$ are convex sets for all $p \ge 1$:



(a) Region of $\|x\|_p = 1$, $p \ge 1$.



(b) Region of $\|x\|_p = 1$, $p \le 1$.

More about Vector Norms (ℓ_0-norm)

In principle, according to (6), ℓ_0-norm of $\boldsymbol{x}$ can be defined as

$$\|\boldsymbol{x}\|_0 = \sqrt[0]{\sum_i x_i^0}\,. \tag{10}$$

Strictly speaking, ℓ_0-norm is not actually a norm. It is a cardinality function which has its definition in the form of ℓ_p-norm, though many people call it a norm. It is a bit tricky to work with because there is a presence of 0^{th}-power and 0^{th}-root in it.

More about Vector Norms (ℓ_0-norm)

In principle, according to (6), ℓ_0-norm of $\boldsymbol{x}$ can be defined as

$$\|\boldsymbol{x}\|_0 = \sqrt[0]{\sum_i x_i^0}\,. \tag{10}$$

Strictly speaking, ℓ_0-norm is not actually a norm. It is a cardinality function which has its definition in the form of ℓ_p-norm, though many people call it a norm. It is a bit tricky to work with because there is a presence of 0^{th}-power and 0^{th}-root in it.

In reality, most mathematicians and engineers use this definition of ℓ_0-norm instead:

$$\|\boldsymbol{x}\|_0 \triangleq \#(i \,|\, x_i \neq 0), \tag{11}$$

which means the total number of non-zero elements in $\boldsymbol{x}$. Notice that this ℓ_0-norm defined by (11) is widely used to measure the sparsity of a large matrix (*to be further discussed in Chap. 10*).

Applications of Vector Norms

1. ℓ_0-norm

   In the field of **Compressive Sensing**, in general we aim at finding the sparsest solution of the under-determined linear system. The sparsest solution means the solution which has fewest non-zero entries, i.e., the smallest ℓ_0-norm, which can be formulated as:

$$\mathcal{P}_1: \quad \min \|\boldsymbol{x}\|_0 \tag{12}$$

$$\text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{13}$$

Applications of Vector Norms

1. ℓ_0-norm

   In the field of **Compressive Sensing**, in general we aim at finding the sparsest solution of the under-determined linear system. The sparsest solution means the solution which has fewest non-zero entries, i.e., the smallest ℓ_0-norm, which can be formulated as:

$$\mathcal{P}_1 : \quad \min \|\boldsymbol{x}\|_0 \tag{12}$$

$$\text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{13}$$

   However, solving $\mathcal{P}_1$ is not an easy task. Because the lack of ℓ_0-norm's mathematical representation, ℓ_0-minimization is regarded by computer scientist as an NP-hard problem, simply says that it's too complex and almost impossible to solve. In many case, ℓ_0-minimization problem is relaxed to be higher-order norm problem such as ℓ_1-minimization and ℓ_2-minimization.

Applications of Vector Norms (cont'd)

2. ℓ_1-norm
In the field of computer vision, the **Sum of Absolute Difference** (SAD) between two vectors $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$ is computed as

$$\mathrm{SAD}(\boldsymbol{x}_1, \boldsymbol{x}_2) \triangleq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_1 = \sum_{i=1}^{n} |\boldsymbol{x}_{1_i} - \boldsymbol{x}_{2_i}|. \tag{14}$$

In more general case of signal difference measurement, the **Mean-Absolute Error** (MAE) between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2 \in \mathbb{R}^n$ is defined as

$$\mathrm{MAE}(\boldsymbol{x}_1, \boldsymbol{x}_2) \triangleq \frac{1}{n} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_1 = \frac{1}{n} \sum_{i=1}^{n} |\boldsymbol{x}_{1_i} - \boldsymbol{x}_{2_i}|. \tag{15}$$

Applications of Vector Norms (cont'd)

3. ℓ_2-norm (a.k.a. Euclidean norm)
   ℓ_2-norm is widely used as a standard quantity for measuring the difference of two vectors, i.e., $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 = \sqrt{\sum_{i=1}^{n} (\boldsymbol{x}_{1_i} - \boldsymbol{x}_{2_i})^2}$, or in its squared form, known as a **Sum of Squared Difference** (SSD) in the field of computer vision:

$$\mathrm{SSD}(\boldsymbol{x}_1, \boldsymbol{x}_2) \triangleq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 = \sum_{i=1}^{n} (\boldsymbol{x}_{1_i} - \boldsymbol{x}_{2_i})^2. \tag{16}$$

Applications of Vector Norms (cont'd)

3. ℓ_2-norm (a.k.a. Euclidean norm)
   ℓ_2-norm is widely used as a standard quantity for measuring the difference of two vectors, i.e., $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 = \sqrt{\sum_{i=1}^{n} (\boldsymbol{x}_{1_i} - \boldsymbol{x}_{2_i})^2}$, or in its squared form, known as a **Sum of Squared Difference** (SSD) in the field of computer vision:

$$\mathrm{SSD}(\boldsymbol{x}_1, \boldsymbol{x}_2) \triangleq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 = \sum_{i=1}^{n} (\boldsymbol{x}_{1_i} - \boldsymbol{x}_{2_i})^2. \tag{16}$$

   In the field of signal processing, the **Mean-Squared Error** (MSE) measurement is widely used to compute a similarity, a quality, or a correlation between two signals, defined as

$$\mathrm{MSE}(\boldsymbol{x}_1, \boldsymbol{x}_2) \triangleq \frac{1}{n} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_{1_i} - \boldsymbol{x}_{2_i})^2. \tag{17}$$

Applications of Vector Norms (cont'd)

As previously discussed in ℓ_0-optimization section, because of many issues from both a computational view and a mathematical view, many ℓ_0-optimization problems relax themselves to become ℓ_1- and ℓ_2-optimization instead. Because of this, we will now discuss about the optimization of ℓ_2 (least squares problem):

$$\mathcal{P}_2 : \quad \min \|\boldsymbol{x}\|_2 \tag{18}$$

$$\text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{19}$$

which can be readily solved by using the Lagrangian multiplier method, and the final solution is the Moore-Penrose pseudoinverse.

Applications of Vector Norms (cont'd)

4. Revising ℓ_1-optimization
Even though the solution of Least Squares problem is easy to compute, it's not necessary be the best solution. Because of the smooth nature of ℓ_2-norm itself, it is hard to find a single, best solution for the problem. In contrary, the ℓ_1-optimization can provide much better result than this solution,

$$\mathcal{P}_3: \quad \min \|\boldsymbol{x}\|_1 \tag{20}$$

$$\text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{21}$$

Since the nature of ℓ_1-norm is not smooth as in the ℓ_2-norm case, the solution of this problem is much better and more unique than the ℓ_2-optimization. However, even though the problem of ℓ_1-minimization has almost the same form as the ℓ_2-minimization, it's much harder to solve. Because this problem does not have a smooth function, the trick we used to solve ℓ_2-problem is no longer valid. The only way left to find its solution is to search for it directly, which means that we have to compute every single possible solution to find the best one from the pool of "infinitely many" possible solutions.

Applications of Vector Norms (cont'd)

4. Revising ℓ_1-optimization
Even though the solution of Least Squares problem is easy to compute, it's not necessary be the best solution. Because of the smooth nature of ℓ_2-norm itself, it is hard to find a single, best solution for the problem. In contrary, the ℓ_1-optimization can provide much better result than this solution,

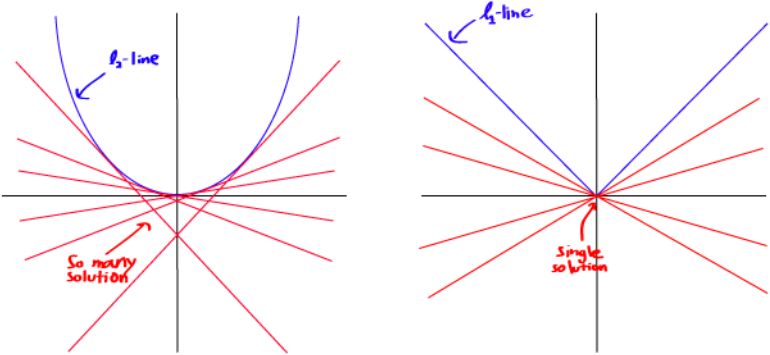
$$\mathcal{P}_3: \quad \min \|\boldsymbol{x}\|_1 \tag{20}$$
$$\text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{21}$$

Since the nature of ℓ_1-norm is not smooth as in the ℓ_2-norm case, the solution of this problem is much better and more unique than the ℓ_2-optimization. However, even though the problem of ℓ_1-minimization has almost the same form as the ℓ_2-minimization, it's much harder to solve. Because this problem does not have a smooth function, the trick we used to solve ℓ_2-problem is no longer valid. The only way left to find its solution is to search for it directly, which means that we have to compute every single possible solution to find the best one from the pool of "infinitely many" possible solutions.

Remark 11 (Computation of ℓ_1-optimization)

Since there is no easy way to find the solution for this problem mathematically, the usefulness of ℓ_1-optimization is very limited for decades. Until recently, the advancement of computer with high computational power allows us to "sweep" through all the solutions. By using many helpful algorithms, namely the Convex Optimization algorithm such as linear programming, or non-linear programming, etc. it's now possible to find the best solution to this question. Many applications that rely on ℓ_1-optimization, including the Compressive Sensing, are now possible. There are many toolboxes for ℓ_1-optimization available nowadays. These toolboxes usually use different approaches and/or algorithms to solve the same question. The example of these toolboxes are ℓ_1-magic, SparseLab, ISAL1.

Applications of Vector Norms (cont'd)



(c) ℓ_2-optimization

(d) ℓ_1-optimization

Figure 3: Comparison between ℓ_2- and ℓ_1-optimization

Table of Contents

Matrix Product Representations

Definition: Let $\boldsymbol{A} \in \mathbb{R}^{m\times k}$, $\boldsymbol{B} \in \mathbb{R}^{k\times n}$, the matrix product is defined as

$$\boldsymbol{C} \triangleq \boldsymbol{A}\boldsymbol{B}, \text{ with } c_{ij} = \sum_{l=1}^{k} a_{il} b_{lj} \text{ and } \boldsymbol{C} \in \mathbb{R}^{m\times n}. \tag{22}$$

Matrix Product Representations

Definition: Let $\boldsymbol{A} \in \mathbb{R}^{m \times k}$, $\boldsymbol{B} \in \mathbb{R}^{k \times n}$, the matrix product is defined as

$$\boldsymbol{C} \triangleq \boldsymbol{A}\boldsymbol{B}, \text{ with } c_{ij} = \sum_{l=1}^{k} a_{il} b_{lj} \text{ and } \boldsymbol{C} \in \mathbb{R}^{m \times n}. \tag{22}$$

Question 12

Why should the product of matrices be defined in the form of (22)*?*

Matrix Product Representations

Definition: Let $\boldsymbol{A} \in \mathbb{R}^{m\times k}$, $\boldsymbol{B} \in \mathbb{R}^{k\times n}$, the matrix product is defined as

$$\boldsymbol{C} \triangleq \boldsymbol{A}\boldsymbol{B}, \text{ with } c_{ij} = \sum_{l=1}^{k} a_{il} b_{lj} \text{ and } \boldsymbol{C} \in \mathbb{R}^{m\times n}. \tag{22}$$

Question 12

Why should the product of matrices be defined in the form of (22)?

Answer: In essence, the product of two matrices represents the composition of the two associated linear functions. More specifically, let

$$f(\boldsymbol{x}) = f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} ax_1 + bx_2 \\ cx_1 + dx_2 \end{pmatrix} \text{ and } g(\boldsymbol{x}) = g\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} Ax_1 + Bx_2 \\ Cx_1 + Dx_2 \end{pmatrix}. \tag{23}$$

Then, let's compose f and g to create another linear function $h \triangleq f(g(\boldsymbol{x}))$, yielding

$$h(\boldsymbol{x}) \triangleq f(g(\boldsymbol{x})) = f\begin{pmatrix} Ax_1 + Bx_2 \\ Cx_1 + Dx_2 \end{pmatrix} = \begin{pmatrix} (aA + bC)x_1 + (aB + bD)x_2 \\ (cA + dC)x_1 + (cB + dD)x_2 \end{pmatrix}.$$

Matrix Product Representations (cont'd)

On the other hand, if we use matrices of coefficients to represent these linear functions, i.e., $f(\boldsymbol{x})$, $g(\boldsymbol{x})$, and $h(\boldsymbol{x})$, we have

$$\boldsymbol{F} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \boldsymbol{G} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \text{ and } \boldsymbol{H} = \begin{pmatrix} aA+bC & aB+bD \\ cA+dC & cB+dD \end{pmatrix}, \tag{25}$$

respectively. After making this association, it is natural to call $\boldsymbol{H}$ the composition (or product) of $\boldsymbol{F}$ and $\boldsymbol{G}$, and to write

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} aA+bC & aB+bD \\ cA+dC & cB+dD \end{pmatrix}. \tag{26}$$

Note: This idea was contributed by Arthur Cayley* around 1855.[2]

---

[2] **Arthur Cayley** (1821-1895, British mathematician). It was not until the work of Arthur Cayley that the matrix was singled out as a separate entity, distinct from the notion of a determinant, and algebraic operations between matrices were defined. In an 1855 paper, Cayley first introduced his basic ideas that were presented mainly to simplify notation. In 1857, Cayley expanded on his original ideas and wrote *A Memoir on the Theory of Matrices*. This laid the foundations for the modern theory and is generally credited for being the birth of the subjects of matrix analysis and linear algebra. (BTW, Cayley published three math papers before completing his undergraduate degree in 1842. Irish mathematician William Rowan Hamilton (1805-1865) was made a professor in 1827 when he was 22 years old and still an undergraduate!)

Matrix Product Representations (cont'd)

Application: Feedforward neural networks of chain structure:

$$f(\boldsymbol{x}) = f^{(n)}\left(f^{(n-1)}\left(f^{(\cdots)}\left(f^{(1)}(\boldsymbol{x})\right)\right)\right)$$

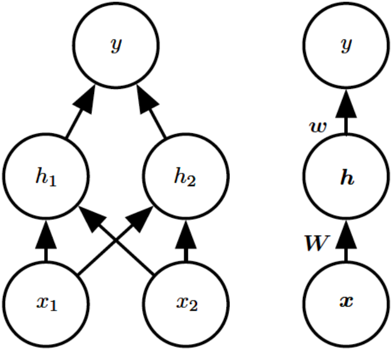


Figure 4: An illustrative feedforward network of two layers.

Matrix Product Representations (cont'd)

Define

$$\boldsymbol{h} = f^{(1)}\left(\boldsymbol{x}; \boldsymbol{W}\right), \tag{27}$$

$$y = f^{(2)}\left(\boldsymbol{h}; \boldsymbol{w}\right). \tag{28}$$

Then, the output of the network is

$$\begin{aligned} y &= f\left(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{w}\right) && (29)\\ &= f^{(2)}\left(f^{(1)}(\boldsymbol{x})\right) && (30)\\ &= \boldsymbol{w}^T\left(\boldsymbol{W}^T\boldsymbol{x}\right) && (31)\\ &= \left(\boldsymbol{W}\boldsymbol{w}\right)^T\boldsymbol{x} && (32)\\ &= \tilde{\boldsymbol{W}}^T\boldsymbol{x}, && (33) \end{aligned}$$

where $\tilde{\boldsymbol{W}} \triangleq \boldsymbol{W}\boldsymbol{w}$ implies that the composition of two functions is equivalent to the product of their coefficient matrices.

Remark 13 (linear maps and matrix multiplication)

Please refer to Chapter 3 of [R1].[3]

[3] [R1] Sheldon Axler, *Linear Algebra Done Right*, 3rd Ed., Springer, 2015.

Matrix Product Representations (cont'd)

1. **Inner-product representation**: Let

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \\ \vdots \\ \boldsymbol{a}_m^T \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_1, & \boldsymbol{b}_2, & \cdots, & \boldsymbol{b}_n \end{bmatrix}$$

(note: abuse of notations with $\boldsymbol{A}$ in the above equation). We have

$$c_{ij} = \boldsymbol{a}_i^T \boldsymbol{b}_j.$$

2. **Column representation**: Recall

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_1, & \boldsymbol{b}_2, & \cdots, & \boldsymbol{b}_n \end{bmatrix}, \boldsymbol{C} = \begin{bmatrix} \boldsymbol{c}_1, & \boldsymbol{c}_2, & \cdots, & \boldsymbol{c}_n \end{bmatrix}.$$

We have

$$\boldsymbol{c}_i = \boldsymbol{A}\boldsymbol{b}_i, \; i = 1, 2, \cdots, n.$$

Matrix Product Representations (cont'd)

3. **Outer-product representation**: Let

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1, & \boldsymbol{a}_2, & \cdots, & \boldsymbol{a}_k \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_1^T \\ \boldsymbol{b}_2^T \\ \vdots \\ \boldsymbol{b}_k^T \end{bmatrix}$$

We have

$$\boldsymbol{C} = \sum_{i=1}^{k} \boldsymbol{a}_i \boldsymbol{b}_i^T.$$

- If $\boldsymbol{A}$ has full column rank and $\boldsymbol{B}$ has full row rank, then $\operatorname{rank}(\boldsymbol{C}) = k$.

Matrix Product Representations (cont'd)

3. **Outer-product representation**: Let

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1, & \boldsymbol{a}_2, & \cdots, & \boldsymbol{a}_k \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_1^T \\ \boldsymbol{b}_2^T \\ \vdots \\ \boldsymbol{b}_k^T \end{bmatrix}$$

We have

$$\boldsymbol{C} = \sum_{i=1}^{k} \boldsymbol{a}_i \boldsymbol{b}_i^T.$$

- If $\boldsymbol{A}$ has full column rank and $\boldsymbol{B}$ has full row rank, then $\mathrm{rank}(\boldsymbol{C}) = k$.

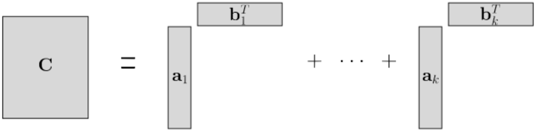


Figure 5: Decomposition of the product of two matrices.

Illustration of the Outer-Product Representation

4. **Strassen Matrix Multiplication**: Consider a 2×2 block matrix with each block being square:

$$\begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix} \tag{34}$$

If we use the ordinary algorithm, $C_{ij} = A_{i1}B_{1j} + A_{i2}B_{2j}$. There are 8 multiplications and 4 adds. In 1967, Strassen has shown how to compute $\boldsymbol{C}$ with just 7 multiplications and 18 adds. For more details, please refer to Section 1.3.11 of [R2].[4]

[4] [R2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th Ed., The John Hopkins University Press, 2013.

Illustration of the Outer-Product Representation

4. **Strassen Matrix Multiplication**: Consider a 2×2 block matrix with each block being square:

$$\begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix} \tag{34}$$

If we use the ordinary algorithm, $C_{ij} = A_{i1}B_{1j} + A_{i2}B_{2j}$. There are 8 multiplications and 4 adds. In 1967, Strassen has shown how to compute $\boldsymbol{C}$ with just 7 multiplications and 18 adds. For more details, please refer to Section 1.3.11 of [R2].[4]

Remark 14 (Advanced topics on matrix multiplication)

For the comparison of the four aforementioned matrix multiplications and more advanced topics on matrix multiplication such as efficiency, fast matrix-vector products, and parallel matrix multiplication, please refer to Chapter 1 of [R2].

[4] [R2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th Ed., The John Hopkins University Press, 2013.

Table of Contents

Extension: Complex-Valued Vectors and Matrices

1. The aforementioned subspace and matrix concepts for real vector/matrix space $\mathbb{R}^n/\mathbb{R}^{m\times n}$ also apply to the complex space $\mathbb{C}^n/\mathbb{C}^{m\times n}$, e.g.,
   - A subset $\mathbb{S} \subseteq \mathbb{C}^m$ is called a subspace if, for any $\alpha, \beta \in \mathbb{C}$,
   $$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S} \Longrightarrow \alpha\boldsymbol{x} + \beta\boldsymbol{y} \in \mathbb{S}.$$
   - $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\} \subseteq \mathbb{C}^m$, is said to be linearly independent if $\sum_{i=1}^n \alpha_i \boldsymbol{a}_i = \boldsymbol{0}$ for some $\boldsymbol{\alpha} \in \mathbb{C}^n$, implies $\boldsymbol{\alpha} = \boldsymbol{0}$.
   - $\mathbb{R}(\boldsymbol{A}) \triangleq \{\boldsymbol{y} \in \mathbb{C}^m \,|\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \in \mathbb{C}^n\}$.
   - Complex inner product of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$:
   $$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^n x_i y_i^* = \boldsymbol{y}^H \boldsymbol{x} = \langle \boldsymbol{y}, \boldsymbol{x} \rangle^*. \tag{35}$$
2. Inner product can also be defined for matrices: Given $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{C}^{m\times n}$,
$$\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \sum_{i=1}^m \sum_{j=1}^n y_{ij}^* x_{ij} = \mathrm{tr}(\boldsymbol{X}\boldsymbol{Y}^H).$$

More About Complex Inner Product

A **complex inner product** is a function $\mathcal{F} : \boldsymbol{x}, \boldsymbol{y} \to \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ that assigns to every ordered pair of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in vector space $\mathcal{V}$ a scalar in complex space $\mathbb{C}$, denoted $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$, such that $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{V}$ and $c \in \mathbb{C}$, satisfies

- $\langle \boldsymbol{x} + \boldsymbol{z}, \boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \langle \boldsymbol{z}, \boldsymbol{y} \rangle$,
- $\langle c\boldsymbol{x}, \boldsymbol{y} \rangle = c\langle \boldsymbol{x}, \boldsymbol{y} \rangle$,
- $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle^*$,
- $\langle \boldsymbol{x}, \boldsymbol{x} \rangle > 0$ for all $\boldsymbol{x} \neq \mathbf{0}$ and $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ if and only if $\boldsymbol{x} = \mathbf{0}$.

A complex vector space with a complex inner product defined above is called a **complex inner product space** or **unitary space**. By virtue of this definition, (35) can also be rewritten as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{n} x_i^* y_i = \boldsymbol{x}^H \boldsymbol{y} = \langle \boldsymbol{y}, \boldsymbol{x} \rangle^*. \tag{36}$$

[5] [R3] P. J. Schreier and L. L. Scharf, *Statistical Signal Processing of Complex-Valued Data*, Cambridge University Press, 2010.

More About Complex Inner Product

A **complex inner product** is a function $\mathcal{F} : \boldsymbol{x}, \boldsymbol{y} \to \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ that assigns to every ordered pair of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in vector space $\mathcal{V}$ a scalar in complex space $\mathbb{C}$, denoted $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$, such that $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{V}$ and $c \in \mathbb{C}$, satisfies

- $\langle \boldsymbol{x} + \boldsymbol{z}, \boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \langle \boldsymbol{z}, \boldsymbol{y} \rangle$,
- $\langle c\boldsymbol{x}, \boldsymbol{y} \rangle = c\langle \boldsymbol{x}, \boldsymbol{y} \rangle$,
- $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle^*$,
- $\langle \boldsymbol{x}, \boldsymbol{x} \rangle > 0$ for all $\boldsymbol{x} \neq \mathbf{0}$ and $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ if and only if $\boldsymbol{x} = \mathbf{0}$.

A complex vector space with a complex inner product defined above is called a **complex inner product space** or **unitary space**. By virtue of this definition, (35) can also be rewritten as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{n} x_i^* y_i = \boldsymbol{x}^H \boldsymbol{y} = \langle \boldsymbol{y}, \boldsymbol{x} \rangle^*. \tag{36}$$

Remark: For more information on the theory of complex-valued matrix, see [R3].[5]

[5] [R3] P. J. Schreier and L. L. Scharf, *Statistical Signal Processing of Complex-Valued Data*, Cambridge University Press, 2010.

Thank you
for your attention!

Minghua Xia

E-mail: xiamingh@mail.sysu.edu.cn

Web: http://seit.sysu.edu.cn/teacher/86