

## Ch.2. Vector spaces, metric and convergence

### § 2.1. Vector spaces (Linear spaces)

- Definition: A vector space over  $\mathbb{R}$  (the real domain) is a set  $V$  together with two functions:

vector addition:  $+ : V \times V \rightarrow V$  (i.e.,  $x+y$ , where  $x, y \in V$ )

scalar multiplication:  $\cdot : \mathbb{R} \times V \rightarrow V$  (i.e.,  $\alpha \cdot x$ , where  $\alpha \in \mathbb{R}, x \in V$ )  
or simply  $\alpha x$

that satisfying the following.

$$\textcircled{1} \text{ Associativity of addition: } x + (y + z) = (x + y) + z \quad \forall x, y, z \in V$$

$$\textcircled{2} \text{ Commutativity of addition: } x + y = y + x \quad \forall x, y \in V.$$

$$\textcircled{3} \text{ Zero vector: } \exists \text{ an element, denoted by } 0, \text{ in } V, \text{ s.t.}$$

$$x + 0 = 0 + x = x \quad \forall x \in V.$$

$$\textcircled{4} \text{ Negative vector: } \forall x \in V, \exists \text{ an element, denoted by } -x, \text{ in } V, \text{ s.t.}$$

$$x + (-x) = (-x) + x = 0$$

$$\textcircled{5} \quad \forall x \in V, \quad 1x = x.$$

$$\textcircled{6} \quad \forall x \in V \text{ and } \alpha, \beta \in \mathbb{R}, \quad \alpha(\beta x) = (\alpha\beta)x.$$

$$\textcircled{7} \quad \forall x \in V \text{ and } \alpha, \beta \in \mathbb{R}, \quad (\alpha + \beta)x = \alpha x + \beta x$$

$$\textcircled{8} \quad \forall x, y \in V \text{ and } \alpha \in \mathbb{R}, \quad \alpha(x + y) = \alpha x + \alpha y \quad \checkmark$$

Remark: • We can define vector space over  $\mathbb{C}$  (the complex domain) similarly.  
 • We will assume vector space over  $\mathbb{R}$  for default. Vector space over  $\mathbb{C}$  is used very rarely.

- A vector space  $V$  is also called a **linear space**, where many linear operations/concepts are available.

— **linear combination:** Given a set of vectors  $\{v_1, \dots, v_p\} \subset V$ ,

a linear combination of  $\{v_1, \dots, v_p\}$  is a vector  $w \in V$  in the form of  $w = \alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_p v_p$ , where  $\alpha_1, \dots, \alpha_p \in \mathbb{R}$  are called coefficients of the linear combination.

- **linear independence:** A set of vectors  $\{v_1, \dots, v_p\} \subset V$  are linearly independent if

$$\alpha_1 v_1 + \dots + \alpha_p v_p = 0 \iff \alpha_1 = \dots = \alpha_p = 0$$

- **linear subspace** (vector subspace): A non-empty subset  $W \subseteq V$  is a subspace of  $V$  if

$$\forall v_1, v_2 \in W \text{ and } \alpha_1, \alpha_2 \in \mathbb{R}, \quad \alpha_1 v_1 + \alpha_2 v_2 \in W$$

(i.e.,  $W$  is closed under vector addition and scalar multiplication)

- $\{0\}$  is a subspace of  $V$ , called zero subspace.

- **linear span:** Given a subset  $S \subseteq V$  ( $S$  can be finite or infinite, countable or uncountable), the set of all linear combinations of finite subsets of  $S$  is the linear span of  $S$ , denoted by  $\text{span}\{S\}$ .

$$(\text{i.e., } \text{span}\{S\} = \left\{ \sum_{i=1}^k \alpha_i v_i \mid v_i \in S, \alpha_i \in \mathbb{R}, k \in \mathbb{N} \right\})$$

(Here  $\mathbb{N}$  stands for the set of all natural numbers)

- $\text{span}\{S\}$  is a subspace of  $V$ .

- **Basis and dimension**

- A subset  $B \subset V$  is called a basis of  $V$  if

$$\{\text{span}\{B\} = V \text{ and }$$

The elements of  $B$  are linearly independent.

- Every vector space has at least a basis,
- All bases of a vector space have the same cardinality.

- The dimension of  $V$  is

$$\dim(V) = |\mathcal{B}| \text{ for a basis } \mathcal{B} \text{ of } V.$$

- Examples of vector spaces

Example 1:  $\mathbb{R}$  is a vector space, with "+" the standard addition of real numbers and " $\cdot$ " the standard multiplication of real numbers.

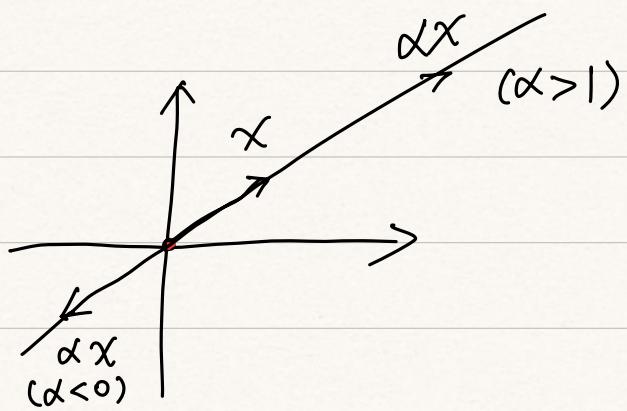
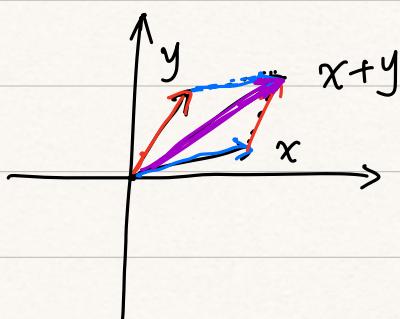
Example 2:  $\mathbb{R}^n$  is a vector space, with "+" and " $\cdot$ " defined by:

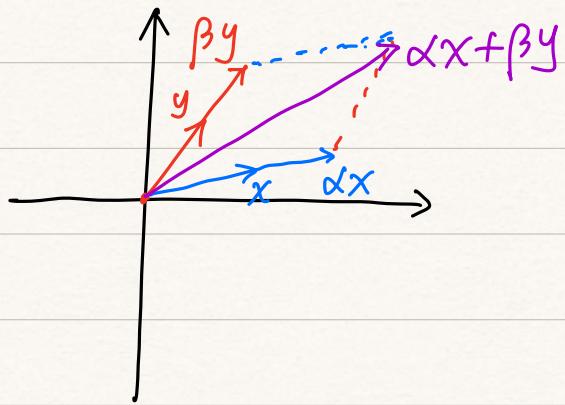
addition:  $\forall \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1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

scalar multiplication:  $\forall \alpha \in \mathbb{R} \text{ and } \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n, \alpha \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{bmatrix}$ .

This vector space is studied in linear algebra

- Zero vector:  $0 = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$
- $-x = \begin{pmatrix} -x_1 \\ -x_2 \\ \vdots \\ -x_n \end{pmatrix}$  if  $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$
- In  $\mathbb{R}^2$ , we use  $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$  to denote the arrow that points rightward and upward by  $x_1$  and  $x_2$  units respectively.





- $\{e_1, e_2, \dots, e_n\}$  is a basis of  $\mathbb{R}^n$ . Here  $e_i = \begin{pmatrix} 0 \\ \vdots \\ i \\ \vdots \\ 0 \end{pmatrix}$  ← i-th entry

So,  $\dim(\mathbb{R}^n) = n$ .  
Many input data can be modeled by vectors in  $\mathbb{R}^n$ .

- Digital sound signals of length  $n$ .
- Time series of length  $n$ .
- $n$  different attributes/features of a single thing or object.

Example 3 : All real  $m \times n$  matrices, denoted by  $\mathbb{R}^{m \times n}$ , with

$+$  : standard matrix addition

$$\forall X = [x_{ij}]_{i=1, j=1}^{m, n}, \quad Y = [y_{ij}]_{i=1, j=1}^{m, n}$$

$$X + Y = [x_{ij} + y_{ij}]_{i=1, j=1}^{m, n}$$

- : Standard scalar multiplication

$$\forall \alpha \in \mathbb{R}, \quad X \in \mathbb{R}^{m \times n},$$

$$\alpha \cdot X = [\alpha x_{ij}]_{i=1, j=1}^{m, n}$$

is a vector space.

In this vector space  $\mathbb{R}^{m \times n}$ :

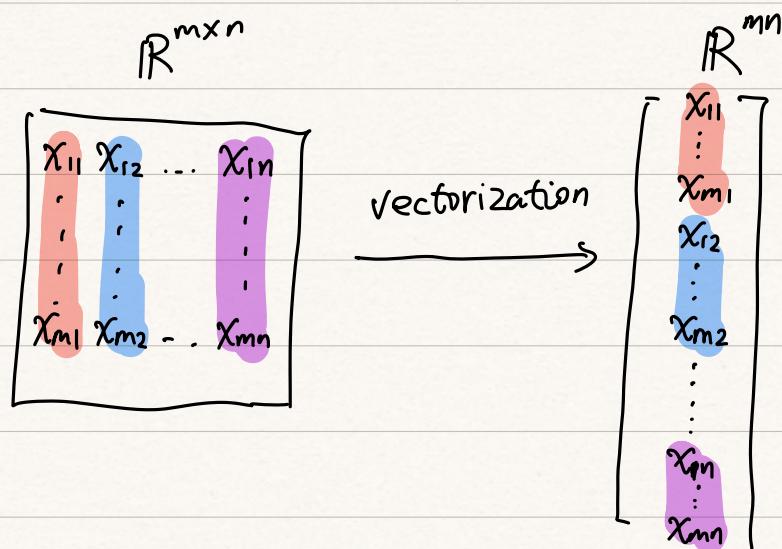
- Zero vector :  $0 =$  the matrix of all zeros.
- $\{E_{ij} \mid i=1, \dots, n, j=1, \dots, m\}$  is a basis of  $\mathbb{R}^{m \times n}$ ,

where  $E_{ij} = \begin{bmatrix} & \\ & 1 \\ & \end{bmatrix}$

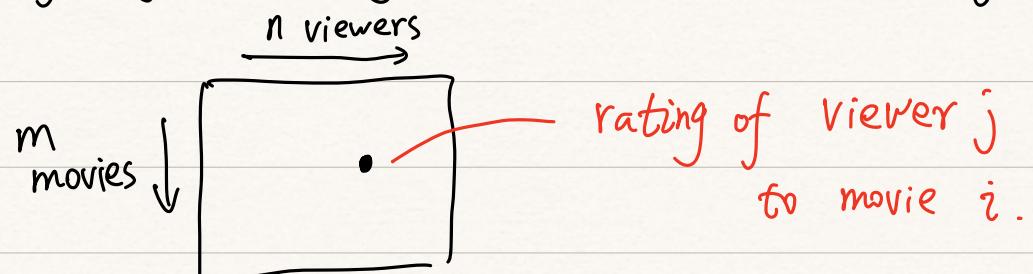
$(i,j)$ -entry

all other entries are 0.

- So,  $\dim(\mathbb{R}^{m \times n}) = mn$ .
- $\mathbb{R}^{m \times n}$  is the same as  $\mathbb{R}^{mn}$  by concatenating columns of a matrix in  $\mathbb{R}^{m \times n}$  to a long vector in  $\mathbb{R}^{mn}$  (called vectorization)



- $\mathbb{R}^{m \times n}$  can be used to represent:
  - Black-white digital images of resolution  $m \times n$ .
  - Ratings of movies by viewers in a recommender system.



- Numerical tabular data of  $m$  rows and  $n$  columns.

Example 4: All 3-arrays of size  $m \times n \times l$ , denoted by  $\mathbb{R}^{m \times n \times l}$ , is a vector space if "+" and ":" is defined by

+:  $\forall x = [x_{ijk}]_{i=1}^n \begin{matrix} m \\ j=1 \\ k=1 \end{matrix}$  and  $y = [y_{ijk}]_{i=1}^n \begin{matrix} m \\ j=1 \\ k=1 \end{matrix}$ ,

$$x+y = [x_{ijk}+y_{ijk}]_{i=1}^n \begin{matrix} m \\ j=1 \\ k=1 \end{matrix}$$

• "  $\forall x = [x_{ijk}]_{i=1}^n \begin{matrix} m \\ j=1 \\ k=1 \end{matrix}$  and  $\alpha \in \mathbb{R}$ ,

$$\alpha x = [\alpha x_{ijk}]_{i=1}^n \begin{matrix} m \\ j=1 \\ k=1 \end{matrix}$$

•  $\dim(\mathbb{R}^{m \times n \times l}) = mn l$ .

•  $\mathbb{R}^{m \times n \times l}$  is the same as  $\mathbb{R}^{mn l}$  by vectorization.

•  $\mathbb{R}^{m \times n \times l}$  can be used to represent:

— Color images with  $m \times n$  pixels

(The 3rd dimension are used for channels, so  $l=3$ )

— Hyper spectral images of  $m \times n$  pixels and  $l$  spectral channels.

— Black-white videos of  $m \times n$  pixels and  $l$  frames

• Similarly, the set of 4-arrays, 5-arrays, ..., are all vector spaces with similar additions and scalar multiplications.

• 3-arrays, 4-arrays, ..., are called **tensors**.

Example 5: Consider the set of all strings.

Define the addition by, e.g.,

$$'I' + 'am' = 'I am'$$

and some scalar multiplication.

Then it doesn't form a vector space.

• Therefore, we cannot use vector space to model text data in this naive way.

How to "vectorize" the texts is a fundamental research topic in text data analysis and natural language processing.

Example 6: The function space  $C[a,b] = \{f \mid f \text{ is continuous on } [a,b]\}$   
 is a vector space if we define "+" and ":" by:

$$+ : \forall f,g \in C[a,b], \quad (f+g)(t) = f(t)+g(t), \quad \forall t \in [a,b].$$

$$\cdot : \forall f \in C[a,b], \alpha \in \mathbb{R}, \quad (\alpha f)(t) = \alpha f(t).$$

- $C[a,b]$  is referred to as a function space, since any vector in the vector space is a function.
- The zero vector in  $C[a,b]$  is the zero function, which maps every  $t \in [a,b]$  to 0.
- $\dim(C[a,b]) = +\infty$
- $C[a,b]$  could be the hypothesis space of a learner with one input and one output, i.e.,

$$x_i \rightarrow \boxed{?} \rightarrow y_i, \quad \text{with } x_i \in [a, b] \text{ and } y_i \in \mathbb{R}.$$

Leave a  $f \in C[a,b]$  s.t.  $f(x_i) \approx y_i$  for all  $i$ .

Example 7: The infinite sequence

$$l_{\infty} = \left\{ \begin{pmatrix} a_1 \\ a_2 \\ \vdots \end{pmatrix} \mid \exists \text{ a finite } C \text{ s.t. } |a_i| \leq C \quad \forall i \right\}$$

with: "+" :  $(a+b)_i = a_i + b_i$  for all  $a, b \in l_{\infty}$  for all  $i$

"·" :  $(\alpha a)_i = \alpha a_i$  for all  $a \in l_{\infty}$ ,  $\alpha \in \mathbb{R}$  for all  $i$

It forms a vector space.

- The 0 vector is the sequence of all zeros.
- $\dim(l_{\infty}) = +\infty$
- This vector space can be used to model, e.g., time series with a very long time and/or a very fine time resolution.



## § 1.2. Metric in vector spaces

In order to do calculus on vector spaces, we need to define 'distance, closeness between vectors.'

Let  $V$  be a vector space. Let  $x, y \in V$ . Then,

$$\text{distance}(x, y) = \text{distance}(x-y, y-y) = \text{distance}(x-y, 0)$$

(distance should be shift invariant.)

$\parallel$   
length of  $x-y$ .

Therefore, to define a distance, we only need to define a length for each vector in  $V$ .

Let  $x \in V$ . Let  $\|x\|$  be its length, called **norm**, which should satisfy

① a length should be non negative, i.e.

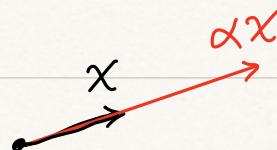
$$\|x\| \geq 0 \quad \forall x \in V.$$

Moreover, only the zero vector has a zero length, i.e.,

$$\|x\|=0 \Leftrightarrow x=0.$$

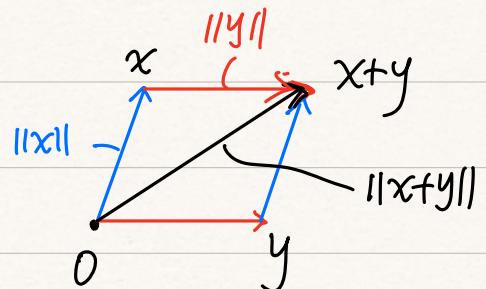
② The length of a multiple of a vector should be the multiple of the length of the vector, i.e.,

$$\forall \alpha \in \mathbb{R}, \quad \|\alpha x\| = |\alpha| \|x\|$$



③ Triangular inequality: the length of the direct path is the smallest

$$\|x+y\| \leq \|x\| + \|y\|$$



Definition: Let  $V$  be a vector space over  $\mathbb{R}$ . A norm on  $V$  is a function  $\|\cdot\| : V \rightarrow \mathbb{R}$  such that:

- ①  $\|x\| \geq 0, \forall x \in V$  and  $\|x\|=0 \iff x=0$ .
- ②  $\|\alpha x\| = |\alpha| \|x\|, \forall x \in V$  and  $\alpha \in \mathbb{R}$ .
- ③  $\|x+y\| \leq \|x\| + \|y\|, \forall x, y \in V$ .

Example 1:  $\mathbb{R}$  is a vector space over  $\mathbb{R}$ .

Let  $\|x\| = |x| \quad \forall x \in \mathbb{R}$ . Then it is a norm on  $\mathbb{R}$ .

(Can you find other norms on  $\mathbb{R}$ ?)

Example 2:  $\mathbb{R}^n$  is a vector space over  $\mathbb{R}$ .

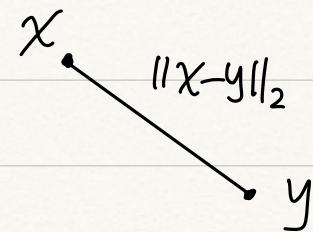
There are many norms on  $\mathbb{R}^n$ .

\* 2-norm: (Euclidean norm)

$$\|x\|_2 = \left( \sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}}$$

The induced distance

$\|x-y\|_2 = \left( \sum_{i=1}^n (x_i - y_i)^2 \right)^{\frac{1}{2}}$  is the Euclidean distance

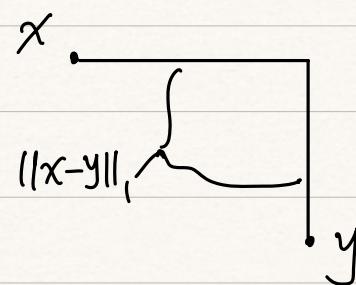


\* 1-norm:

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

The induced distance

$$\|x-y\|_1 = \sum_{i=1}^n |x_i - y_i|$$



is known as Manhattan distance (You can walk only horizontally and vertically)

\*  $\infty$ -norm:

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

$$x \xrightarrow{\|x-y\|_\infty} y$$

The induced distance is

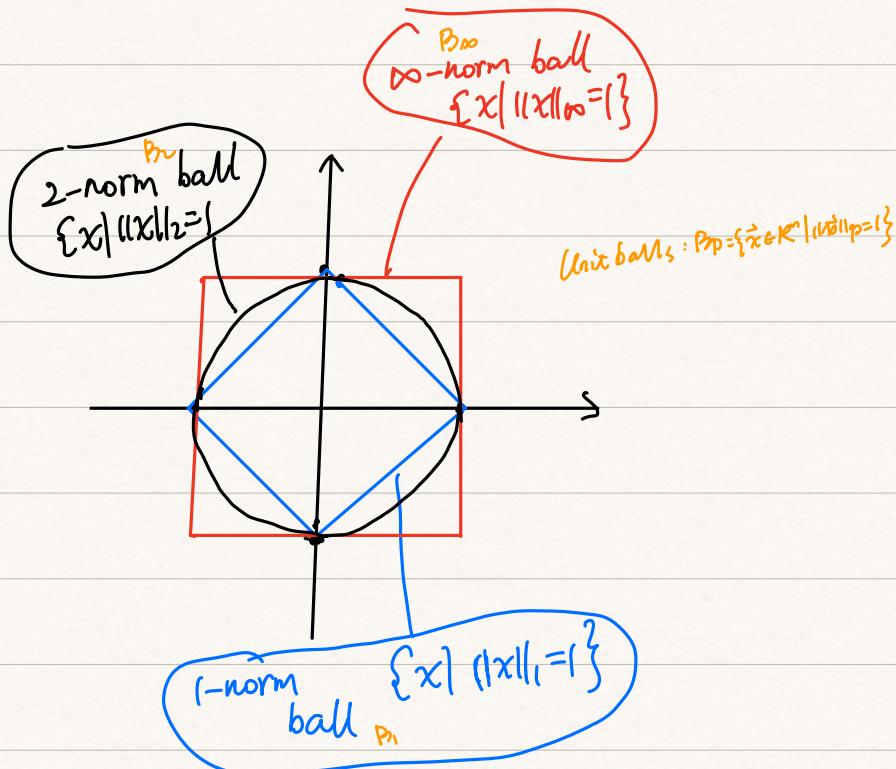
$$\|x-y\|_\infty = \max_{1 \leq i \leq n} |x_i - y_i|$$



## \* p-norm ( $p \geq 1$ )

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

- Comparison of unit balls.



- Note that  $(\mathbb{R}^n, \|\cdot\|_1)$ ,  $(\mathbb{R}^n, \|\cdot\|_2)$ ,  $(\mathbb{R}^n, \|\cdot\|_\infty)$ , ... are all different normed spaces. So, for a given vector space, we can obtain various normed space by choosing different norms.

- Calculate the norms of  $x = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$

$$\|x\|_2 = (3^2 + 4^2)^{1/2} = 5 \quad \|x\|_1 = |3| + |4| = 7$$

$$\|x\|_\infty = \max \{3, 4\} = 4.$$

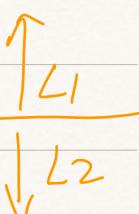
Example 3:  $\mathbb{R}^{m \times n}$  is a vector space

- Since  $\mathbb{R}^{m \times n}$  can be viewed as  $\mathbb{R}^{mn}$ , we can define

vector p-norms for matrices

$$\|A\|_{p,\text{vector}} = \left( \sum_{j=1}^n \sum_{i=1}^m |a_{ij}|^p \right)^{1/p}$$

- $p=1$ : sum of absolute values of entries of  $A$ .
- $p=2$ : sum of squares and then take square root  
(also known as Frobenius norm)



$$\|A\|_F = \left( \sum_{j=1}^n \sum_{i=1}^m |a_{ij}|^2 \right)^{\frac{1}{2}}$$

$\rightarrow p=\infty$ : max entry of  $A$  in magnitude

Q Calculate the Frobenius norm of  $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \in \mathbb{R}^{2 \times 2}$

$$\|A\|_F = \left( \sum_{j=1}^2 \sum_{i=1}^2 |a_{ij}|^2 \right)^{\frac{1}{2}} = (1^2 + 2^2 + 3^2 + 4^2)^{\frac{1}{2}} = \sqrt{30}$$

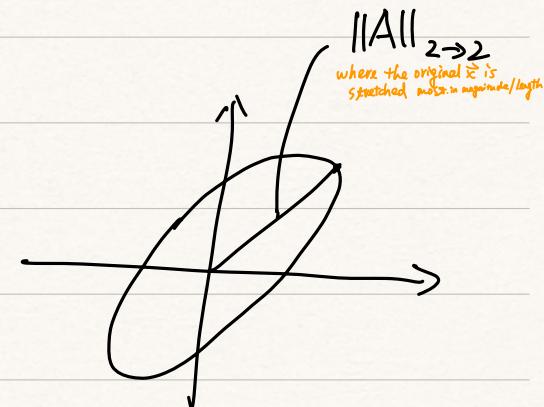
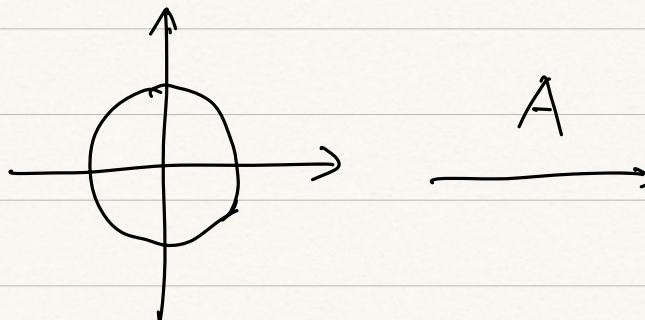
$A \in \mathbb{R}^{m \times n}$  is also a linear transformation from  $\mathbb{R}^n$  to  $\mathbb{R}^m$

Define  $p$ -norms on  $\mathbb{R}^n$  and  $\mathbb{R}^m$  respectively.

We can define matrix  $p$ -norm of a matrix by

$$\begin{aligned} \|A\|_{p \rightarrow p} &= \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Ax\|_p}{\|x\|_p} \quad (\text{check it is a norm!}) \\ &= \sup_{\substack{x \in \mathbb{R}^n \\ \|x\|_p=1}} \|Ax\|_p \end{aligned}$$

$p=2$



Find the vector 2-norm (Frobenius norm) and the matrix

2-norm of  $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

(Vector 2-norm)

Frobenius norm :  $\|A\|_F = (1^2 + 0^2 + 0^2 + 1^2)^{\frac{1}{2}} = \sqrt{2}$

matrix 2-norm :  $\|A\|_2 \rightarrow 2 = \sup_{\substack{x \in \mathbb{R}^2 \\ x \neq 0}} \frac{\|Ax\|_2}{\|x\|_2} = \sup_{\substack{x \in \mathbb{R}^2 \\ x \neq 0}} \frac{\|x\|_2}{\|x\|_2} = 1$

Example 4:  $C[a, b]$  is a vector space over  $\mathbb{R}$ .

(连通形式)

The 0 vector in  $C[a, b]$  is the function that takes value 0 on  $[a, b]$ .

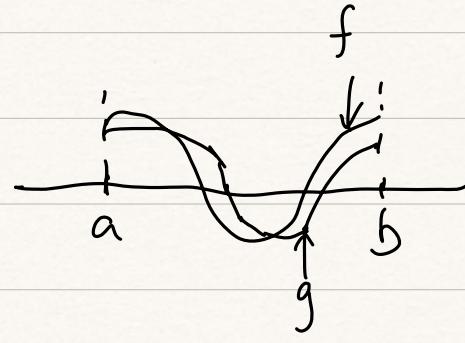
To measure how large a function  $f$  is,

we can use the following norms

$$\|f\|_{\infty} = \sup_{x \in [a, b]} |f(x)|$$

Then distance of two functions  $f, g \in C[a, b]$

is  $\|f-g\|_{\infty} = \sup_{x \in [a, b]} |f(x)-g(x)|.$



Some other norms of  $C[a, b]$  can be

$$\|f\|_1 = \int_a^b |f(x)| dx$$

$$\|f\|_2 = \left( \int_a^b |f(x)|^2 dx \right)^{1/2}$$

$$\|f\|_p = \left( \int_a^b |f(x)|^p dx \right)^{1/p}$$

Example 5:  $\ell_{\infty}$  is a vector space over  $\mathbb{R}$

If  $a \in \ell_{\infty}$ , define  $\|a\|_{\infty} = \sup_i |a_i|$

It is a norm on  $\ell_{\infty}$ , called  $\ell_{\infty}$ -norm.

Similarly, for any infinite sequence  $a = (a_i)_{i \in \mathbb{Z}}$

we define  $\|a\|_p = \left( \sum_{i=1}^{\infty} |a_i|^p \right)^{1/p}$ , where  $p \geq 1$ .

Consider the space  $\ell_p = \{a \mid \|a\|_p < +\infty\} \subset \ell_{\infty}$

We can show that  $\ell_p$  is a vector space

and  $\|\cdot\|_p$  is a norm on  $\ell_p$ , call  $\ell_p$ -norm.

Example:

$$a = \begin{pmatrix} 1 \\ 1/2 \\ 1/3 \\ \vdots \end{pmatrix}$$

Then  $\|a\|_{\infty} = \sup_i |a_i| = 1$

$$\|a\|_2 = \left( \sum_{i=1}^{\infty} \left(\frac{1}{i}\right)^2 \right)^{1/2} = \left( \frac{\pi^2}{6} \right)^{1/2} = \frac{\pi}{\sqrt{6}}$$

$\ell_p$  (

$$\|a\|_1 = \left( \sum_{i=1}^{\infty} \frac{1}{i} \right) = +\infty$$

Therefore,  $a \in l_\infty$ ,  $a \in l_2$ , but  $a \notin l_1$ .

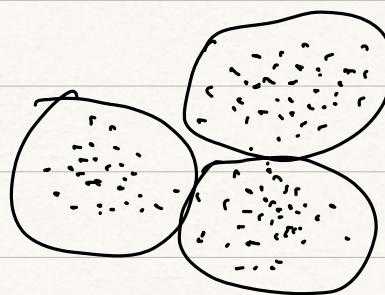
- Remarks:
1. For the same vector space, we can define infinitely many different norms.
  2. A common technique in machine learning to find the solution is to optimize some norm of the unknown. Different norms lead to very different solutions.

## § 1.2 Case study: Clustering, k-means, k-medians

### Clustering

Suppose we are given  $N$  vectors  $x_1, x_2, \dots, x_N \in \mathbb{R}^n$

The goal of clustering is to group or partition the vectors into  $K$  groups or clusters, with the vectors in each group close to each other.



- We use  $\mathbb{R}^n$  because it is simple yet able to model a variety of data sets (e.g., signals, images, videos, attributes of things)
- Actually, the methods can be extended to any normed spaces.
- Applications:

— Topic discovery. Suppose the  $N$  vectors are word histograms

with  $N$  documents respectively, i.e., the  $j$ -th component in  $x_i$  is the counts of the  $j$ -th word in document  $i$ .

A clustering algorithm partitions the documents into  $K$  groups, which typically can be interpreted as groups of documents with the same topics, genre, or author.

— Patient Clustering. If  $\{x_i\}_{i=1}^N$  are feature vectors associated with  $N$  patients admitted to a hospital, a clustering algorithm clusters the patient into  $K$  groups of similar patients.

— Recommendation system. A group of  $N$  people respond to ratings of  $n$  movies. A clustering algorithm can be used to cluster the people into  $K$  groups, each with similar taste.

Then we can recommend new movies liked by someone to people in the same group as him/her.

— Many other applications.

Mathematical formulation: (群=类)

- Representation:

Let  $C_i \in \{1, 2, \dots, k\}$  be the group that  $x_i$  belongs to.  $i = 1, 2, \dots, N$ .

Then, group  $j$ , denoted by  $G_j$ , is  $G_j = \{i \mid C_i = j\}$ .  $j = 1, 2, \dots, k$ .

We assign each group a representative vector, denoted by  $z_1, z_2, \dots, z_k$ .

The representative vectors are not necessarily one of given vectors.

- Evaluation:

First of all, within one specific group  $j$   $G_j$ , all vectors should be close to the representative vector  $z_j$ . More precisely, let

$$J_j = \sum_{i \in G_j} \|x_i - z_j\|_2^2$$

Then,  $J_j$  should be small.

Secondly, consider all groups, since each  $J_j$  is small,

$$J = J_1 + J_2 + \dots + J_k$$

should be small.

Altogether, we solve the following

$$\min_{G_1, \dots, G_k} J \iff \min_{G_1, \dots, G_k} \sum_{j=1}^k J_j \iff \min_{G_1, \dots, G_k} \sum_{j=1}^k \left( \sum_{i \in G_j} \|x_i - z_j\|_2^2 \right)$$

- Optimization.

We may use an alternating minimization to solve the minimization.

Step 1: Fix the representatives  $z_1, \dots, z_k$ , find the best partitions

$G_1, \dots, G_k$ , i.e., solve

$$\min_{G_1, \dots, G_k} \sum_{j=1}^k \left( \sum_{i \in G_j} \|x_i - z_j\|_2^2 \right) \quad \dots \quad (1)$$

Step 2: Fix the groups  $G_1, \dots, G_K$ , find the best representatives

$\bar{z}_1, \dots, \bar{z}_K$ , i.e., solve

$$\min_{\bar{z}_1, \dots, \bar{z}_K} \sum_{j=1}^K \left( \sum_{i \in G_j} \|x_i - \bar{z}_j\|_2^2 \right) \quad \dots \quad (2)$$

The two steps are repeated until convergence.

Let's find the solutions of the sub-problems (1) and (2) respectively.

For (1):

finding the partition  $G_1, \dots, G_K$  is equivalent to  
 找出最佳的聚类  
 找出最佳的簇  
 找出最佳的类别

finding  $c_1, c_2, \dots, c_N$ . So (1) becomes

$$\min_{c_1, c_2, \dots, c_N} \left( \underbrace{\|x_1 - z_{c_1}\|_2^2}_{\text{depends on } c_1 \text{ only}} + \underbrace{\|x_2 - z_{c_2}\|_2^2}_{\text{depends on } c_2 \text{ only}} + \dots + \underbrace{\|x_N - z_{c_N}\|_2^2}_{\text{depends on } c_N \text{ only}} \right)$$



$$\min_{c_i} \|x_i - z_{c_i}\|_2^2 \quad i=1, 2, \dots, N.$$

Since  $c_i \in \{1, 2, \dots, K\}$ , to get  $c_i$ , we only need to compare

$$\|x_i - z_1\|_2^2, \|x_i - z_2\|_2^2, \dots, \|x_i - z_K\|_2^2$$

and choose the minimum from it. i.e.,

$$c_i = \arg \min_{j \in \{1, \dots, K\}} \|x_i - z_j\|_2^2, \quad i=1, 2, \dots, N.$$

In other words,

$x_i$  is assigned to the group whose representative vector is the closest to  $x_i$ .

For (2): It is rewritten as

$$\min_{z_1, \dots, z_K} \left( \underbrace{\sum_{i \in G_1} \|x_i - z_1\|_2^2}_{\text{depends on } z_1 \text{ only}} + \underbrace{\sum_{i \in G_2} \|x_i - z_2\|_2^2}_{\text{depends on } z_2 \text{ only}} + \dots + \underbrace{\sum_{i \in G_K} \|x_i - z_K\|_2^2}_{\text{depends on } z_K \text{ only}} \right)$$

Obviously, it is equivalent to minimize each term independently,

i.e., solve  $K$  independent problems.

$$\min_{\bar{z}_j} \left( \sum_{i \in G_j} \|x_i - \bar{z}_j\|_2^2 \right), \quad j=1, 2, \dots, K.$$

Note that

$$\begin{aligned} \sum_{i \in G_j} \|x_i - \bar{z}_j\|_2^2 &= \sum_{i \in G_j} \sum_{l=1}^n (x_{il} - \bar{z}_{jl})^2 \\ &= \sum_{l=1}^n \left( \sum_{i \in G_j} (x_{il} - \bar{z}_{jl})^2 \right) \end{aligned}$$

$x_{il}$  are  $l$ -th component  
of  $\frac{x_i}{z_j}$  respectively

[Each term] in this summation are independent again. 单独求解

Thus,  $\min_{\bar{z}_j} \left( \sum_{i \in G_j} \|x_i - \bar{z}_j\|_2^2 \right) \Leftrightarrow$

$$\min_{\bar{z}_{jl}} \sum_{i \in G_j} (x_{il} - \bar{z}_{jl})^2, \quad l=1, 2, \dots, n.$$

One variable minimization.

@21:47

$\checkmark$  Taking derivative w.r.t.  $\bar{z}_{jl}$  and setting it to 0, we obtain that

the solution  $\bar{z}_{jl}$  satisfies

$$2 \sum_{i \in G_j} (\bar{z}_{jl} - x_{il}) = 0 \quad \Leftrightarrow$$

$$\Rightarrow \bar{z}_{jl} = \left( \sum_{i \in G_j} x_{il} \right) / |G_j| \quad \left( |G_j| \text{ is the number of elements in } G_j. \right)$$

$$l=1, 2, \dots, n.$$

$$\sum_{i \in G_j} \bar{z}_{jl} = \sum_{i \in G_j} x_{il} / |G_j| \cdot \bar{z}_{jl}$$

In vector form,

$$\begin{pmatrix} \bar{z}_{j1} \\ \bar{z}_{j2} \\ \vdots \\ \bar{z}_{jn} \end{pmatrix} = \frac{1}{|G_j|} \cdot \sum_{i \in G_j} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{pmatrix} \Leftrightarrow \bar{z}_j = \frac{1}{|G_j|} \left( \sum_{i \in G_j} x_i \right)$$

In other words,

$\bar{z}_j$  is the mean of all vectors in  $G_j$ .

Altogether, we get the following clustering algorithm.

Input:  $x_1, x_2, \dots, x_N \in \mathbb{R}^n$ .

Output:  $C_1, C_2, \dots, C_K$  and  $\bar{z}_j, j=1, \dots, K$ .

Initialization: Initialize  $\bar{z}_1, \bar{z}_2, \dots, \bar{z}_K$  by choosing  $K$  vectors

from  $x_1, x_2, \dots, x_N$  randomly.

Step 1: Given  $z_1, z_2, \dots, z_k$ , compute

$$c_i = \arg \min_{j \in \{1, 2, \dots, k\}} \|x_i - z_j\|_2^2, \quad i=1, 2, \dots, N.$$

and define

$$G_j = \{i \mid c_i = j\}, \quad j=1, 2, \dots, k.$$

Step 2: Given  $G_1, G_2, \dots, G_k$ , compute

$$z_j = \frac{1}{|G_j|} \left( \sum_{i \in G_j} x_i \right)$$

Go back to step 1.

This algorithm is known as "K-means" algorithm, because it computes K means of vectors at step 2.

### K-medians Algorithm

In K-means, the Euclidean norm is used. We can replace it by  $\ell$ -norm. We solve

$$\min_{\substack{G_1, \dots, G_k \\ z_1, \dots, z_k}} \sum_{j=1}^k \left( \sum_{i \in G_j} \|x_i - z_j\|_1 \right)$$

The numerical solver is

Step 1: Fix  $z_1, \dots, z_k$ , solve

$$\min_{\substack{G_1, \dots, G_k}} \sum_{j=1}^k \left( \sum_{i \in G_j} \|x_i - z_j\|_1 \right).$$

Similar to the discussion in K-means, the solution is

$$c_i = \arg \min_{j \in \{1, 2, \dots, k\}} \|x_i - z_j\|_1, \quad i=1, 2, \dots, N.$$

and  $G_j = \{i \mid c_i = j\}$ .

Step 2: Fix  $G_1, G_2, \dots, G_k$ , solve

$$\min_{\substack{z_1, \dots, z_k}} \sum_{j=1}^k \left( \sum_{i \in G_j} \|x_i - z_j\|_1 \right)$$

Similar to the discussion in k-means, it is decomposed into K sub problems

$$\min_{z_j} \sum_{i \in G_j} \|x_i - z_j\|_1, \quad j=1, 2, \dots, k.$$

Why median?  $\checkmark$  @ 10:24

It is well known (Galileo) that the solution is

$$z_j = \text{median}_{i \in G_j}(x_i),$$

where  $\text{median}_{i \in G_j}(x_i)$  takes component-wise median.

This algorithm is called "k-median" algorithm.

## §2.3 Limit and convergence in vector spaces

Convergence in normed vector spaces

To define calculus, we need first define convergent sequence.

Let  $V$  be a normed vector space.

Let  $\{\chi^{(k)}\}_{k \in \mathbb{N}}$  be a sequence in  $V$ , (i.e.,  $\chi^{(k)} \in V \quad \forall k = 1, 2, 3, \dots$ ).

Let  $\chi \in V$ . We say  $\{\chi^{(k)}\}_{k \in \mathbb{N}}$  converges to  $\chi$ , denoted by  $\chi^{(k)} \rightarrow \chi$ , if

$$\lim_{k \rightarrow \infty} \|\chi^{(k)} - \chi\| = 0$$

Example 1: Consider  $\mathbb{R}^n$  with  $\|\cdot\|_2$

$$\text{Let } \chi^{(k)} = \begin{pmatrix} 1/k \\ 2/k \\ \vdots \\ n/k \end{pmatrix} \in \mathbb{R}^n \quad \text{and} \quad \chi = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = 0$$

$$\text{Then } \|\chi^{(k)} - \chi\|_2 = \|\chi^{(k)}\|_2 = \left( \sum_{i=1}^n \left(\frac{i}{k}\right)^2 \right)^{\frac{1}{2}} \\ = \frac{1}{k} \left( 1^2 + 2^2 + \dots + n^2 \right)^{\frac{1}{2}}$$

$$\lim_{k \rightarrow \infty} \|\chi^{(k)} - \chi\|_2 = \left( \lim_{k \rightarrow \infty} \frac{1}{k} \underbrace{(1^2 + 2^2 + \dots + n^2)^{\frac{1}{2}}}_{\text{constant}} \right) \xrightarrow{k \rightarrow \infty} \text{constant in } K. \\ = 0$$

Therefore  $\chi^{(k)} \rightarrow 0$  as  $k \rightarrow \infty$

Example 2: Consider  $C[0, 1]$  with  $\|\cdot\|_\infty$  a vector space

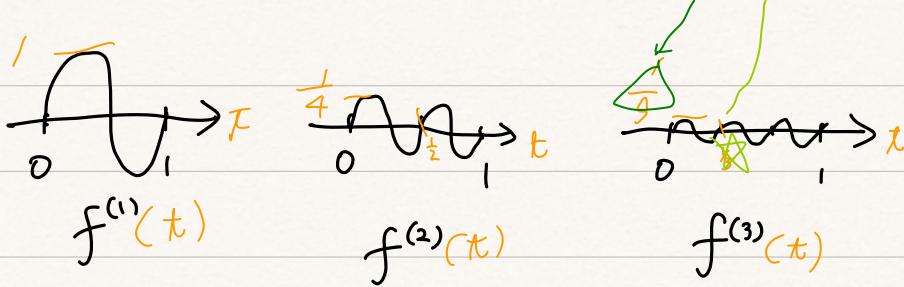
$$\text{Let } f^{(k)}(t) = \sin(2\pi kt)/k^2$$

Let  $0$  be the  $0$  function.

$$\text{Then } \|f^{(k)} - 0\|_\infty = \max_{t \in [0, 1]} \left| \frac{\sin(2\pi kt)}{k^2} \right| = \frac{1}{k^2}$$

$$\text{So, } \lim_{k \rightarrow \infty} \|f^{(k)} - 0\|_\infty = \lim_{k \rightarrow \infty} \frac{1}{k^2} = 0$$

Therefore,  $f^{(k)} \rightarrow 0$  as  $k \rightarrow \infty$



Example 3: Consider infinite sequences

$$a^{(k)} = \begin{pmatrix} k \\ k \\ \vdots \\ k \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{k-terms}$$

$\in l_1, l_2, l_\infty$

$$a = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = 0 \in l_1, l_2, l_\infty$$

- In  $l_2$  with  $\|\cdot\|_2$  norm

$$\|a^{(k)} - a\|_2 = \|a^{(k)}\|_2 = \left( \sum_{i=1}^k \left(\frac{1}{k}\right)^2 \right)^{\frac{1}{2}} = \left( k \cdot \frac{1}{k^2} \right)^{\frac{1}{2}} = \frac{1}{\sqrt{k}}$$

$$\lim_{k \rightarrow \infty} \|a^{(k)} - a\|_2 = 0$$

Thus,  $a^{(k)} \rightarrow a$  in  $l_2$ .

- In  $l_\infty$  with  $\|\cdot\|_\infty$  norm

$$\|a^{(k)} - a\|_\infty = \|a^{(k)}\|_\infty = \frac{1}{k}$$

$$\lim_{k \rightarrow \infty} \|a^{(k)} - a\|_\infty = 0$$

Thus,  $a^{(k)} \rightarrow a$  in  $l_\infty$ .

- In  $l_1$  with  $\|\cdot\|_1$  norm

$$\|a^{(k)} - a\|_1 = \|a^{(k)}\|_1 = \sum_{i=1}^k \frac{1}{k} = 1$$

$$\lim_{k \rightarrow \infty} \|a^{(k)} - a\|_1 = 1 \neq 0$$

Thus,  $a^{(k)} \not\rightarrow a$  in  $l_1$ .

This example shows that: The convergence / limit depends on norms.

Example 4: Consider  $V = \{a \mid a \text{ is an infinite sequence, } \|a\|_1 < +\infty\}$

(元素之和有限)

with  $\|\cdot\|_\infty$  norm

Let

$$a^{(k)} = \begin{pmatrix} 1 \\ y_2 \\ \vdots \\ y_k \\ 0 \\ \vdots \end{pmatrix} \quad a = \begin{pmatrix} 1 \\ y_2 \\ y_3 \\ \vdots \\ y_k \\ \vdots \end{pmatrix}$$

$$\text{Then } \lim_{k \rightarrow \infty} \|a^{(k)} - a\|_\infty = \lim_{k \rightarrow \infty} \left\| \begin{pmatrix} 0 \\ \vdots \\ y_{(k+1)} \\ y_{(k+2)} \\ \vdots \end{pmatrix} \right\|_\infty = \lim_{k \rightarrow \infty} \frac{1}{k+1} = 0$$

i.e., the limit of  $a^{(k)}$  should be  $a$  in  $\ell^\infty$ -norm

However,  $a^{(k)} \in V$  because  $\|a^{(k)}\|_1 < +\infty$

but  $a \notin V$ , because  $\|a\|_1 = \sum_{i=1}^{\infty} \frac{1}{i} = +\infty$

 This example shows that: The limit may not be in a normed vector space.  
in Chinese

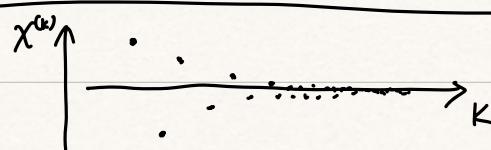
- If this happen, we call the normed vector space incomplete
- We can always complete a normed vector space by including all limit of "convergent" sequences

### Completeness of normed vector spaces

Given a sequence  $\{x^{(k)}\}_{k \in \mathbb{N}}$  in a vector space  $V$  with norm  $\|\cdot\|$ , how to determine that the sequence should be convergent?

 Cauchy sequence:  $\{x^{(k)}\}_{k \in \mathbb{N}}$  is called a cauchy sequence if

$\forall \varepsilon > 0, \exists K \text{ st. } \forall k, l \geq K \quad \|x^{(k)} - x^{(l)}\| < \varepsilon$



Lemma: If  $x^{(k)} \rightarrow x \in V$ , then  $\{x^{(k)}\}$  is a cauchy sequence.

Proof.  $x^{(k)} \rightarrow x$  implies:  $\forall \frac{\varepsilon}{2} > 0, \exists K$  s.t.  $\forall k \geq K \quad \|x^{(k)} - x\| \leq \frac{\varepsilon}{2}$ .

Therefore,  $\|x^{(k)} - x^{(l)}\| \leq \|x^{(k)} - x\| + \|x^{(l)} - x\| \leq \varepsilon$ .  $\forall k, l \geq K$ .  $\blacksquare$

The reverse of the lemma is not true. Consider example 4 above, where

$$\|a^{(k)} - a^{(l)}\|_{\infty} = \left\| \begin{bmatrix} 0 \\ \vdots \\ \frac{1}{k} \\ \vdots \\ \frac{1}{l} \\ 0 \end{bmatrix} \right\|_{\infty} = \frac{1}{k} < \varepsilon \quad \text{if } k, l \geq \frac{1}{\varepsilon} + 1 \quad (\text{assume } l \geq k)$$

But the limit of  $a^{(k)}$  is not in  $V$ .

A vector space  $V$  with norm  $\|\cdot\|$  is complete if all cauchy sequence in  $V$  is convergent.

A complete normed vector space is called a Banach space

We can always complete a normed vector space by including all limit of Cauchy sequences

Example of complete normed vector spaces (Banach spaces)

- $\mathbb{R}^n$  with any norm.
- $\mathbb{R}^{n \times m}$  with any norm.
- Tensor space  $\mathbb{R}^{m \times n \times l}$  with any norm
- $C[a, b]$  with  $\|\cdot\|_{\infty}$  norm
- $\ell_p$  with  $p \geq 1$  and finite,  $\ell_{\infty}$ .

Example of incomplete normed vector spaces.

- Example 4 is the last section:

$V = \{a \mid a \text{ is an infinite sequence}, \|a\|_1 < +\infty\}$  with norm  $\|\cdot\|_{\infty}$  is incomplete

The completion of this space is  $\ell_{\infty}$ .

- $C[a, b]$  with  $p$ -norm,  $p \neq +\infty, p \geq 1$ , are incomplete.

The completion of this space is done in "Real Analysis".

## Applications:

Many algorithms in data analysis are **Iterative algorithms**, which generates  $S$  sequences of vectors

$$\{x_s^{(\text{iter})}\}_{\text{iter} \in \mathbb{N}} \subset V \quad s=1, 2, \dots, S,$$

where  $V$  is endowed with a norm  $\|\cdot\|$ .

Example: in K-means algorithm, we generate a sequence of representative vectors  $\{z_1^{(\text{iter})}, \dots, z_K^{(\text{iter})}\}_{K \in \mathbb{N}}$ .

Initialize  $z_1^{(0)}, \dots, z_K^{(0)}$  randomly.

for  $\text{iter} = 1, 2, 3, \dots$

given  $z_1^{(\text{iter})}, \dots, z_K^{(\text{iter})}$ , generate  $G_1^{(\text{iter})}, \dots, G_K^{(\text{iter})}$

given  $G_1^{(\text{iter})}, G_2^{(\text{iter})}, \dots, G_K^{(\text{iter})}$ , generate  $z_1^{(\text{iter}+1)}, \dots, z_K^{(\text{iter}+1)}$ .

end.

- $(V, \|\cdot\|)$  must be complete.

Otherwise, if the limit of  $x_s^{(\text{iter})}$  is not in  $V$ , then the the limit of  $x_s^{(\text{iter})}$  may not be a good solution.

(since  $V$  is the hypothesis space, or the set of good solutions )

Example: In a supervised learning, given  $(x_i, y_i), i=1, \dots, m, x_i \in [a, b], y_i \in \mathbb{R}$ .

We want to find a function  $f$  s.t.  $f(x_i) \approx y_i, i=1, \dots, m$ .

We aim at finding  $f \in C[a, b]$ , i.e., a continuous function.

Assume an algorithm generates a sequence of functions

$$\{f^{(\text{iter})}\}_{\text{iter} \in \mathbb{N}} \subset C[a, b].$$

If  $\|f^{(\text{iter})} - \tilde{f}\|_2 \rightarrow 0$  for some  $\tilde{f}$ , this  $\tilde{f}$  might not be in  $C[a, b]$  (i.e., might not be continuous) since  $(C[a, b], \|\cdot\|_2)$  is incomplete.

- How to check the sequences  $\{x_s^{(\text{iter})}\}$  is convergent ?

We can not use the definition of convergence since the limit is unknown.

We use "Cauchy sequence" to check the convergence.

We check whether or not  $\|\chi^{(k)} - \chi^{(l)}\|$  is small enough.

In practice, this condition is approximated by  $\|\chi^{(\text{iter})} - \chi^{(\text{iter}+1)}\|$  is small enough.

Example: In k-means algorithm,

for iter = 1, 2, ... .

- - -

- - - .

if  $\|z_s^{(\text{iter}+1)} - z_s^{(\text{iter})}\|_2 \leq \varepsilon$  for all  $s=1,2,\dots,K$ , then stop

end for

## § 2.4. Finite dimensional vector spaces

★ In most of the cases, we will be dealing with finite dimensional vector spaces (e.g.,  $\mathbb{R}^n$ ,  $\mathbb{R}^{m \times n}$ ,  $\mathbb{R}^{m \times n \times l}$ ), which have very nice properties:

① For a finite dimensional vector space  $V$ , all norms are equivalent in the sense that:

For any two norms  $\|\cdot\|_A$  and  $\|\cdot\|_B$ , there exists

$C_1, C_2 > 0$  such that:  $C_1 \|a\|_A \leq \|a\|_B \leq C_2 \|a\|_A \quad \forall a \in V$ .

- Consequently, the limit of the same sequence under any norm is the same.

$$x^{(k)} \rightarrow x \text{ in } \|\cdot\|_A \iff x^{(k)} \rightarrow x \text{ in } \|\cdot\|_B$$

proof. It is equivalent to prove:  $\|x^{(k)} - x\|_A \rightarrow 0 \iff \|x^{(k)} - x\|_B \rightarrow 0$

$$\begin{aligned} \textcircled{1} \quad \|x^{(k)} - x\|_A \rightarrow 0 &\Rightarrow 0 \leq \|x^{(k)} - x\|_B \leq C_2 \|x^{(k)} - x\|_A \rightarrow 0 \\ &\Rightarrow \|x^{(k)} - x\|_B \rightarrow 0. \end{aligned}$$

$$\begin{aligned} \textcircled{2} \quad \|x^{(k)} - x\|_B \rightarrow 0 &\Rightarrow 0 \leq \|x^{(k)} - x\|_A \leq \frac{1}{C_1} \|x^{(k)} - x\|_B \rightarrow 0 \\ &\Rightarrow \|x^{(k)} - x\|_A \rightarrow 0 \end{aligned}$$
⊗

Note that the constants  $C_1, C_2$  depend on  $V$ .

Example: Consider  $\mathbb{R}^n$ . and  $\|\cdot\|_1$ ,  $\|\cdot\|_2$ ,  $\|\cdot\|_\infty$

- $\|\cdot\|_1$  and  $\|\cdot\|_2$  are equivalent because

$$\|a\|_2 \leq \|a\|_1 \leq \sqrt{n} \|a\|_2$$

- $\|\cdot\|_2$  and  $\|\cdot\|_\infty$  are equivalent

$$\|a\|_\infty \leq \|a\|_2 \leq \sqrt{n} \|a\|_\infty$$

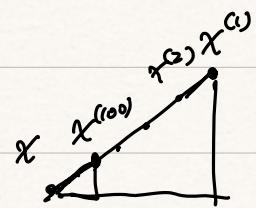
- $\|\cdot\|_1$  and  $\|\cdot\|_\infty$  are equivalent

$$\|a\|_\infty \leq \|a\|_1 \leq n \|a\|_\infty$$

- The convergence speed depends on norms.

Example:  $\chi^{(k)} = \frac{1}{k} \begin{bmatrix} \cos \frac{(k+\frac{1}{2})\pi}{2} \\ \sin \frac{(k+\frac{1}{2})\pi}{2} \end{bmatrix}$

$\chi^{(k)} \rightarrow 0$  under any norm.



However,  $\|\chi^{(k)} - 0\|_2 = \frac{1}{k}$

$$\|\chi^{(k)} - 0\|_1 = \frac{\sqrt{2}}{k}$$

To achieve  $\varepsilon$ -precision:

2-norm:  $\frac{1}{K_2} \sim \varepsilon \Rightarrow K_2 \sim \frac{1}{\varepsilon}$

1-norm:  $\frac{\sqrt{2}}{K_1} \sim \varepsilon \Rightarrow K_1 \sim \frac{\sqrt{2}}{\varepsilon}$

$K_1$  is about  $\sqrt{2}$  times of  $K_2$ .

Example:  $\ell_1 = \{a \mid a \text{ is an infinite sequence, } \|a\|_1 < +\infty\}$  — Infinite dimensional

$\|\cdot\|_1$  and  $\|\cdot\|_\infty$  are NOT equivalent in  $\ell_1$ , because

$$a^{(k)} = \left( \begin{array}{c} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \\ \vdots \end{array} \right) \text{ K terms} \quad a^{(k)} \in \ell_1$$

$$\begin{aligned} \|a^{(k)}\|_\infty &= 1 \\ \|a^{(k)}\|_1 &= k \end{aligned} \quad \Rightarrow \lim_{K \rightarrow +\infty} \frac{\|a^{(k)}\|_1}{\|a^{(k)}\|_\infty} = +\infty$$

So,  $\nexists$  a finite number  $C$  s.t.

$$\|a\|_1 \leq C \|a\|_\infty \quad \forall a \in \ell_1$$

② Any finite dimensional normed vector space is complete  
(i.e., they are Banach spaces)