# Ch. 2. Vector spaces, metric and convergence

## § 2.1. Vector spaces

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

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

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

  that satisfying the following.

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

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

  ③ Zero vector: $\exists$ an element, denoted by $0$, in $V$, s.t.
  $$x+0 = 0+x = x \quad \forall x \in V.$$

  ④ Negative vector: $\forall x \in V$, $\exists$ an element, denoted by $-x$, in $V$, s.t.
  $$x+(-x) = (-x)+x = 0$$

  ⑤ $\forall x \in V, \quad 1x = x$.

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

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

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

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.

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

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

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 is a vector space, with standard matrix addition and standard scalar multiplication.

- This vector space is the same as $\mathbb{R}^{mn}$.
- An $m \times n$ matrix can be used to represent a black-white digital image of $m \times n$ pixels.

Example 4: All 3-arrays of size $m \times n \times l$ is a vector space, if "$+$" and "$\cdot$" is defined by

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

$$x + y = [x_{ijk} + y_{ijk}]_{i=1,}^{n}{}_{j=1,}^{m}{}_{k=1}^{l}$$

$\cdot$: $\forall x = [x_{ijk}]_{i=1}^{n}{}_{j=1}^{m}{}_{k=1}^{l}$ and $\alpha \in \mathbb{R}$,

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

- This vector space is the same as $\mathbb{R}^{mnl}$.
- An $n \times m \times 3$ 3-array can be used to model a color digital image, where $x_{ijk}$ means $(i,j)$-th pixel in channel $k$, and channel 1, 2, 3 means Red, Green, Blue channels of the image.
- An $n \times m \times l$ 3-array can be used to model a black-white video, where $x_{ijk}$ means the $(i,j)$-th pixel at $k$-th frame.

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 $C[a,b] = \{f \mid f \text{ is continous on } [a,b]\}$ is a vector space if we define "+" and "$\cdot$" by:

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

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

- $C[a,b]$ is referred to as a function space, since any vector in the vector space is a function.
- $C[a,b]$ could be the hypothesis space of a learner with one input and one output, i.e.,

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

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

Example 7: The infinite sequence

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

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

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

It forms a vector space.

- 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.)

$\|\ \|$ 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.,

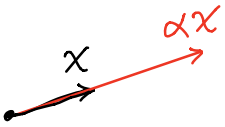
$$\|x\| \geq 0 \quad \forall\, x \in V.$$

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

$$\|x\| = 0 \iff x = 0.$$

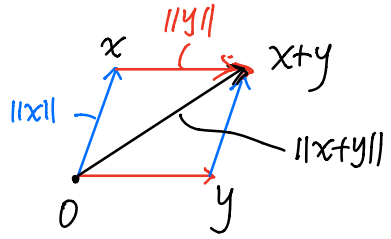
② The length of a multiple of a vector should be the multiple of the length of the vector. i.e.,

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




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 $\|\cdot\| : V \to \mathbb{R}$ such that:

① $\|x\| \geq 0, \forall x \in V$ and $\|x\| = 0 \Leftrightarrow x = 0$.

② $\|\alpha x\| = |\alpha| \, \|x\|, \quad \forall x \in V$ and $\alpha \in \mathbb{R}$.

③ $\|x+y\| \leq \|x\| + \|y\|, \quad \forall x, y \in V$.

**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 $\mathbb{R}$.

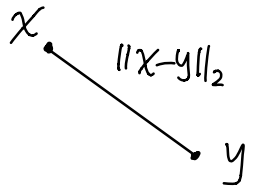
(Can you find other norms on $\mathbb{R}$?)

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

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

* 2-norm: (Euclidean norm)

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

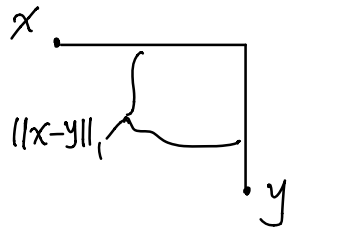



The induced distance

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

* 1-norm:

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



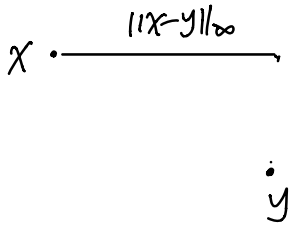

The induced distance

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

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

* $\infty$-norm:

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



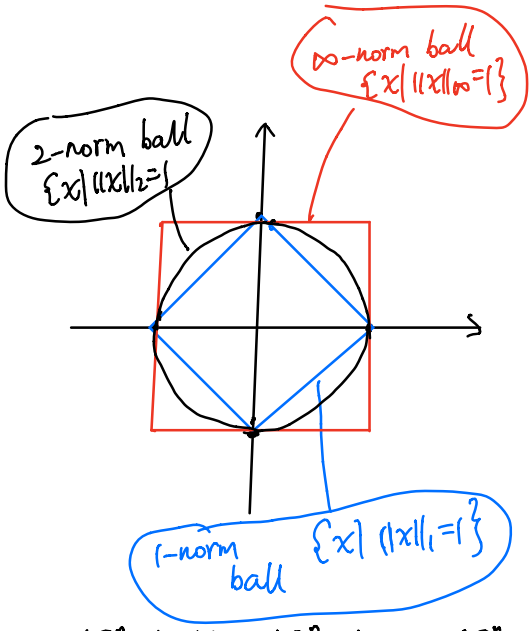

The induced distance is

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

* $p$-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 $x = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$

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

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

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)

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

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

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

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

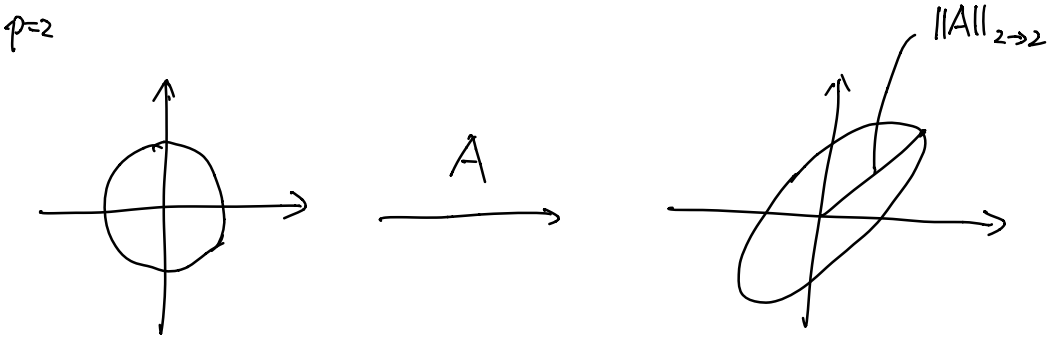
- $A \in \mathbb{R}^{m\times n}$ is also a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$

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

We can define matrix p-norm of a matrix by

$$\|A\|_{p\to p} = \sup_{\substack{x\in\mathbb{R}^n \\ x\neq 0}} \frac{\|Ax\|_p}{\|x\|_p} \quad \text{(Check it is a norm!)}$$

$$= \sup_{\substack{x\in\mathbb{R}^n \\ \|x\|_p=1}} \|Ax\|_p$$




- Find the vector 2-norm (Frobenius norm) and the matrix 2-norm of $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

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

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

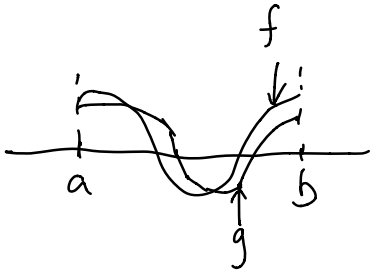
Example 4: $C[a,b]$ is a vector space over $\mathbb{R}$.

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: • $l_\infty$ is a vector space over $\mathbb{R}$

$\forall\ a \in l_\infty$, define $\|a\|_\infty = \sup_i |a_i|$

It is a norm on $l_\infty$, called $l_\infty$-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 $l_p = \{a \mid \|a\|_p < +\infty\} \subset l_\infty$

We can show that $l_p$ is a vector space

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

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

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

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

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

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.

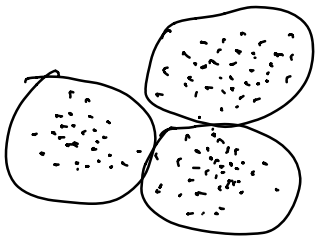
2. A common technique in machine learning to find the solution is to optimize some norm of the unknown. Different norms lead to very different solutions.

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

### Clustering

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

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., singals, 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 patition the documents into $k$ groups, which typically can be intepreted 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:

- Representation:

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

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

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

  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,

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

  should be small.

  Altogether, we solve the following

  $$\min_{\substack{G_1, \cdots, G_k \\ z_1, \cdots, z_k}} J \iff \min_{\substack{G_1, \cdots G_k \\ z_1, \cdots z_k}} \sum_{j=1}^{k} J_j \iff \min_{\substack{G_1, \cdots, G_k \\ z_1, \cdots, z_k}} \sum_{j=1}^{k} \left( \sum_{i \in G_j} \| x_i - z_j \|_2^2 \right)$$

- Optimization.

  We may use an alternating minimization to solve the minimization.

  Step 1: Fix the representatives $z_1, \cdots, z_k$, find the best partitions $G_1, \cdots, G_k$, i.e., solve

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

Step 2: Fix the groups $G_1, \cdots, G_K$, find the best representatives $z_1, \cdots, z_K$, i.e., solve

$$\min_{z_1,\cdots,z_K} \sum_{j=1}^{K} \left( \sum_{i \in G_j} \|x_i - z_j\|_2^2 \right) \quad ---- \quad (2)$$

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 $G_1, \cdots, G_K$ is equivalent to finding $C_1, C_2, \cdots, C_N$. So (1) becomes

$$\min_{C_1, C_2, \cdots, C_N} \Big( \underbrace{\|x_1 - z_{C_1}\|_2^2}_{\text{depends on } C_1 \text{ only}} + \underbrace{\|x_2 - z_{C_2}\|_2^2}_{\text{depends on } C_2 \text{ only}} + \cdots + \underbrace{\|x_N - z_{C_N}\|^2}_{\text{depends on } C_N \text{ only}} \Big)$$

$$\Updownarrow$$

$$\min_{C_i} \|x_i - z_{C_i}\|_2^2 \qquad i = 1, 2, \cdots, N.$$

Since $C_i \in \{1, 2, \cdots, k\}$, to get $C_i$, we only need to compare

$$\|x_i - z_1\|_2^2, \ \|x_i - z_2\|_2^2, \ \cdots, \ \|x_i - z_k\|_2^2$$

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

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

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

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

Obviously, it is equivalent to minimize each term independently,

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

$$\min_{z_j}\left(\sum_{i\in G_j}\|x_i-z_j\|_2^2\right), \qquad j=1,2,\cdots,K.$$

Note that

$$\sum_{i\in G_j}\|x_i-z_j\|_2^2=\sum_{i\in G_j}\sum_{l=1}^{n}(x_{il}-z_{jl})^2$$
$$=\sum_{l=1}^{n}\left(\sum_{i\in G_j}(x_{il}-z_{jl})^2\right)$$

($x_{il}$, $z_{jl}$ are $l$-th component of $x_i$, $z_j$ respectively)

Each term is this summation are independent again.

Thus, $\min_{z_j}\left(\sum_{i\in G_j}\|x_i-z_j\|_2^2\right) \Longleftrightarrow \boxed{\min_{z_{jl}}\sum_{i\in G_j}(x_{il}-z_{jl})^2}$, $l=1,2,\cdots,n$.

↑ one variable minimization.

Taking derivative w.r.t. $z_{jl}$ and setting it to 0, we obtain that the solution $z_{jl}$ satisfies

$$2\sum_{i\in G_j}(z_{jl}-x_{il})=0$$
$$\Rightarrow \quad z_{jl}=\left(\sum_{i\in G_j}x_{il}\right)/|G_j| \qquad (|G_j| \text{ is the number of elements in } G_j)$$
$$l=1,2,\cdots,n.$$

In vector form,

$$\begin{pmatrix} z_{j1}\\ z_{j2}\\ \vdots\\ z_{jn}\end{pmatrix}=\frac{1}{|G_j|}\cdot\sum_{i\in G_j}\begin{pmatrix} x_{i1}\\ x_{i2}\\ \vdots\\ x_{in}\end{pmatrix} \Longleftrightarrow \quad z_j=\frac{1}{|G_j|}\left(\sum_{i\in G_j}x_i\right)$$

In other words,

$\boxed{z_j \text{ is the mean of all vectors in } G_j.}$

Altogether, we get the following clustering algorithm.

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

Output: $c_1, c_2, \cdots, c_N$ and $z_j$, $j=1,\cdots,K$.

Initialization: Initialize $z_1, z_2, \cdots, z_K$ by choosing $K$ vectors

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

Step 1: Given $z_1, z_2, \cdots, z_K$, compute

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

and define

$$G_j = \{ i \mid C_i = j \}, \quad j = 1, 2, \cdots, K.$$

step 2: Given $G_1, G_2, \cdots, G_K$, compute

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

Go back to step 1.

This algorithm is know 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 1-norm. We solve

$$\min_{\substack{G_1,\cdots,G_K \\ z_1,\cdots,z_K}} \sum_{j=1}^{K} \left( \sum_{i \in G_j} \|x_i - z_j\|_1 \right)$$

The numerical solver is

Step 1: Fix $z_1, \cdots, z_K$, solve

$$\min_{G_1,\cdots,G_K} \sum_{j=1}^{K} \left( \sum_{i \in G_j} \|x_i - z_j\|_1 \right).$$

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

$$C_i = \arg\min_{j \in \{1,2,\cdots,K\}} \|x_i - z_j\|_1, \quad i = 1, 2, \cdots, N.$$

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

Step 2: Fix $G_1, G_2, \cdots, G_K$, solve

$$\min_{z_1,\cdots,z_K} \sum_{j=1}^{K} \left( \sum_{i \in G_j} \|x_i - z_j\|_1 \right.$$

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

$$\min_{z_j} \sum_{i \in G_j} \|x_i - z_j\|_1 \qquad j = 1, 2, \cdots, K.$$

It is well known (Galileo) that the solution is

$$z_j = \underset{i \in G_j}{\text{median}} (x_i).$$

where $\underset{i \in G_j}{\text{median}} (x_i)$ 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 $\{x^{(k)}\}_{k\in\mathbb{N}}$ be a sequence in $V$, (i.e., $x^{(k)}\in V \ \forall k = 1,2,3,\cdots$).

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

$$\lim_{k\to\infty} \|x^{(k)}-x\| = 0$$

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

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

Then $$\|x^{(k)}-x\|_2 = \|x^{(k)}\|_2 = \left(\sum_{i=1}^{n}\left(\frac{i}{k}\right)^2\right)^{1/2}$$

$$= \frac{1}{k}\left(1^2+2^2+\cdots+n^2\right)^{1/2}$$

$$\lim_{k\to\infty}\|x^{(k)}-x\|_2 = \left(\lim_{k\to\infty}\frac{1}{k}\right)\underbrace{\left(1^2+2^2+\cdots+n^2\right)^{1/2}}$$

↘ constant in $k$.

$$= 0$$

Therefore $x^{(k)} \to 0$ as $k\to\infty$

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

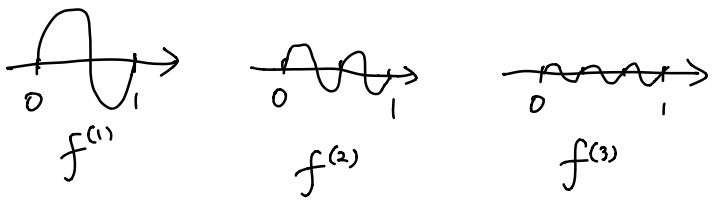
Let $f^{(k)}(t) = \sin(2\pi k t)/k^2$

Let $0$ be the $0$ function.

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

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

Therefore, $f^{(k)} \to 0$




Example 3: Consider infinite sequences

$$a^{(k)} = \begin{pmatrix} 1/k \\ 1/k \\ \vdots \\ 1/k \\ 0 \\ \vdots \\ 0 \\ \vdots \end{pmatrix} \left.\vphantom{\begin{pmatrix} 1/k \\ 1/k \\ \vdots \\ 1/k \\ 0 \\ \vdots \\ 0 \\ \vdots \end{pmatrix}}\right\} k\text{-terms} \quad \in \ell_1, \ell_2, \ell_\infty \qquad a = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ \vdots \end{pmatrix} = 0 \in \ell_1, \ell_2, \ell_\infty$$

- In $\ell_2$ with $\|\cdot\|_2$ norm

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

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

Thus, $a^{(k)} \to a$ in $\ell_2$.

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

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

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

Thus, $a^{(k)} \to a$ in $\ell_\infty$.

- In $\ell_1$ with $\|\cdot\|_1$ norm

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

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

Thus, $a^{(k)} \not\to a$ in $\ell_1$

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

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

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

Let $a^{(k)} = \begin{pmatrix} 1 \\ 1/2 \\ \vdots \\ 1/k \\ 0 \\ \vdots \end{pmatrix}$ $a = \begin{pmatrix} 1 \\ 1/2 \\ 1/3 \\ \vdots \\ 1/k \\ \vdots \end{pmatrix}$

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

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

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

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

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

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

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

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

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
- $l_p$ with $p \geq 1$ and finite, $l_\infty$.

Example of incomplete normed vector spaces.

- Example 4 is the last section:

  $V = \{a \mid \begin{matrix} a \text{ is a infinite sequence} \\ \|a\|_\infty < +\infty \end{matrix}\}$ with norm $\|\cdot\|_\infty$ is incomplete

  The completion of this space is $l_\infty$.

- $C[a,b]$ 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

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

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

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

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

for iter $= 1, 2, 3, \cdots$.

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

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

end.

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

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

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

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

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

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

  Assume an algorithem generates a sequence of functions

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

  If $\|f^{(iter)} - \tilde{f}\|_2 \to 0$ 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 $\{x_s^{(iter)}\}$ is convergent ?

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

We use "Cauchy sequence" to check the convergence.

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

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

Example: In K-means algorithm,

for iter = 1, 2, ···

· — —

· — —.

if $\|z_s^{(iter+1)} - z_s^{(iter)}\|_2 \leq \varepsilon$ for all $s = 1, 2, \cdots, K$, the 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., $\mathbb{R}^n$, $\mathbb{R}^{m\times n}$, $\mathbb{R}^{m\times n\times k}$), 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: $C_1\|a\|_A \le \|a\|_B \le C_2\|a\|_A \quad \forall a \in V$.

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

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

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

① $\|x^{(k)}-x\|_A \to 0 \Rightarrow 0 \le \|x^{(k)}-x\|_B \le C_2\|x^{(k)}-x\|_A \to 0$

$\Rightarrow \|x^{(k)}-x\|_B \to 0.$

② $\|x^{(k)}-x\|_B \to 0 \Rightarrow 0 \le \|x^{(k)}-x\|_A \le \frac{1}{C_1}\|x^{(k)}-x\|_B \to 0$

$\Rightarrow \|x^{(k)}-x\|_A \to 0$ ⊠

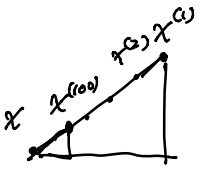
Note that the constants $C_1, C_2$ depend on $V$.

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

- $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent because
$$\|a\|_2 \le \|a\|_1 \le \sqrt{n}\,\|a\|_2$$
- $\|\cdot\|_2$ and $\|\cdot\|_\infty$ are equivalent
$$\|a\|_\infty \le \|a\|_2 \le \sqrt{n}\,\|a\|_\infty$$
- $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are equivalent
$$\|a\|_\infty \le \|a\|_1 \le n\,\|a\|_\infty$$

- The convergence speed depends on norms.

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




$x^{(k)} \to 0$ under any norm.

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

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

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: $l_1 = \{a \mid a \text{ is an infinite sequence}, \|a\|_1 < +\infty\}$ — Infinite dimensional

$\|\cdot\|_1$ and $\|\cdot\|_\infty$ are NOT equivalent in $l_1$, because

$a^{(k)} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \\ \vdots \end{pmatrix}$ $\}k$ terms $\quad a^{(k)} \in l_1$

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

$\|a^{(k)}\|_1 = k$

$\Rightarrow \lim_{k\to+\infty} \frac{\|a^{(k)}\|_1}{\|a^{(k)}\|_\infty} = +\infty$

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

$$\|a\|_1 \le C\|a\|_\infty \quad \forall a \in l_1$$

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