

MSBD5004 Math Methods for Data Analysis

- Introduce Math Tools for machine learning

- Supervised Learning (Regression)

Given a training data set

$$\{(x_i, y_i)\}_{i=1,2,\dots,m}^{\text{input}} \quad \text{label}$$
$$y_i \in \mathbb{R}$$

Find a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ s.t.

$$f(x_i) \approx y_i, \quad i=1, \dots, m.$$

With f , given a new $x \in \mathbb{R}^n$, we use

$f(x)$ as the predicted output.

- Which class \mathcal{F} should f be in?

- We need to know functions on \mathbb{R}^n

- \mathcal{F} is known as the hypothesis space.

- Which function f is the best in \mathcal{F} for our task?

- Loss function:

if $f \in \mathcal{F}$, we assign a loss
 $L(f)$.

$$\text{So } L: \mathcal{F} \rightarrow \mathbb{R}$$

function of function
functional

-

$$\boxed{\min_{f \in \mathcal{F}} L(f)}$$

- Numerical optimization

Ch.2. Vector spaces (Linear spaces), Norms, Limits/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:

$$\text{Addition } + : V \times V \rightarrow V \quad (\text{i.e., } \begin{array}{l} \forall x, y \in V \\ x+y \in V \end{array})$$

$$\text{Scalar multiplication } \cdot : \mathbb{R} \times V \rightarrow V \quad (\text{i.e., } \begin{array}{l} \forall \alpha \in \mathbb{R}, x \in V \\ \alpha \cdot x \in V \end{array})$$

that satisfy the following:

$$\textcircled{1} \quad (x+y) + z = x + (y+z) \quad \forall x, y, z \in V$$

$$\textcircled{2} \quad x+y = y+x \quad \forall x, y \in V$$

$$\textcircled{3} \quad \exists \text{ an element, denoted by } 0, \text{ in } V \text{ s.t.} \\ x+0 = 0+x = x \quad \forall x \in V.$$

$$\textcircled{4} \quad \forall x \in V, \quad \exists \text{ an element, denoted by } -x, \text{ on } V \text{ s.t.} \\ x+(-x) = (-x)+x = 0.$$

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

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

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

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

Remark: - We can define vector spaces over \mathbb{C} (complex domain)

- We will assume vector space over \mathbb{R} unless specified.

- A vector space V is also called a linear space,

- linear combination:

Given a set of vectors $\{v_1, \dots, v_p\} \subseteq V$, 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 coefficients.

- linear dependence/independence:
A set of vectors $\{v_1, \dots, v_p\} \subseteq V$ are linearly independent
 $\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_p v_p = 0 \iff \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$
- linear subspace (vector subspace):
A non-empty subset $W \subseteq V$ is a subspace of V if
 $\forall v_1, v_2 \in W, \quad \alpha_1 v_1 + \alpha_2 v_2 \in W$
 $\alpha_1, \alpha_2 \in \mathbb{R}$
 $(W \text{ is closed under } "+" \text{ and } "\cdot")$
— $\{0\}$ is a subspace of V .
 \nwarrow zero subspace
- linear span:
Given a subset $S \subseteq V$ (S can finite/infinite countable/uncountable),
the linear span of S , denoted by $\text{span}\{S\}$, is
 $\text{span}\{S\} = \left\{ \sum_{i=1}^k \alpha_i v_i \mid \alpha_i \in \mathbb{R}, v_i \in S, k \in \mathbb{N} \right\}$
 \mathbb{N} — the set of natural numbers.
— $\text{span}\{S\}$ is a subspace of V .
- Basis and dimension.
— A subset $B \subset V$ is a basis of V if
 $\text{span}\{B\} = V$ and,
{ The elements in B are linearly independent.
— Every vector space has at least one basis.
— All bases of the same vector space has the same cardinality.
— The dimension of V is
 $\dim(V) = |B|$ for a basis B of V .

- Examples of vector spaces

Example 1: \mathbb{R} is a vector space with

+ : standard number addition

\cdot : ~~$\$ - \cdot - \cdot$~~ multiplication

- $\{1\}$ is basis of \mathbb{R}

- $\dim(\mathbb{R}) = 1$.

Ex. 2: \mathbb{R}^n is a vector space with

+ : $\forall x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

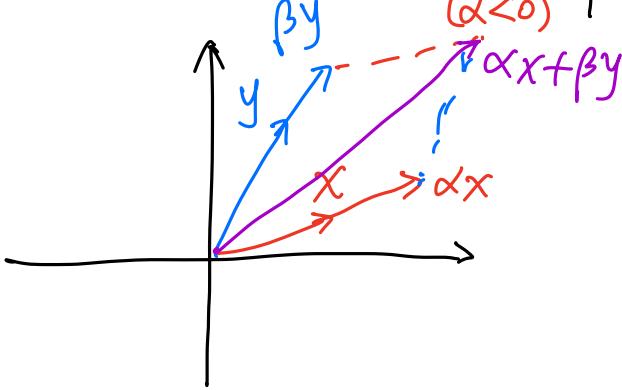
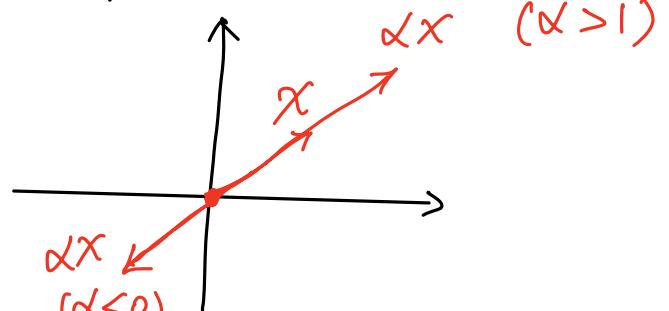
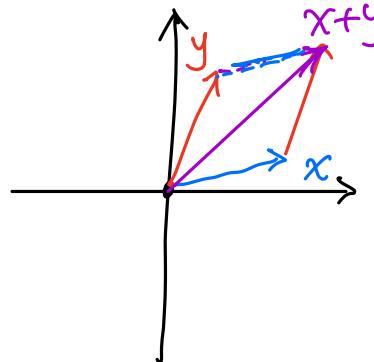
$$x + y = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix}$$

\cdot : $\forall \alpha \in \mathbb{R},$
 $x \in \mathbb{R}^n$

$$\alpha \cdot x = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{pmatrix}$$

- Zero vector: $0 = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$

- In \mathbb{R}^2 , we use $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ to represent "arrows" that points x_1 units rightward, x_2 units upward.



$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad \leftarrow i\text{-th entry}$$

- $\{e_1, e_2, \dots, e_n\}$ is a basis of \mathbb{R}^n

- $\dim(\mathbb{R}^n) = n$
- Many input data can be modeled by vectors on \mathbb{R}^n
 - Digital sound signal of length n .
 - n different numerical features of a single thing.

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

$$+ : \forall X \in \mathbb{R}^{m \times n}, Y \in \mathbb{R}^{m \times n}, X + Y = \begin{bmatrix} x_{ij} + y_{ij} \end{bmatrix}_{i=1, j=1}^{m, n}$$

$$\begin{bmatrix} x_{ij} \end{bmatrix}_{i=1, j=1}^{m, n} \quad \begin{bmatrix} y_{ij} \end{bmatrix}_{i=1, j=1}^{m, n}$$

$$\cdot : \forall \alpha \in \mathbb{R}, X \in \mathbb{R}^{m \times n}, \alpha \cdot X = \begin{bmatrix} \alpha x_{ij} \end{bmatrix}_{i=1, j=1}^{m, n}$$

is a vector space.

In this vectors

$$- \text{Zero: } 0 = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$$

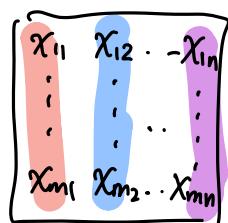
- $\{E_{ij} \mid i=1, \dots, m, j=1, \dots, n\}$, where $E_{ij} = \begin{bmatrix} & \cdots & & \\ & \downarrow & & \\ & & 1 & \\ & \cdots & & \\ & & & 0 \end{bmatrix}$
- is a basis of $\mathbb{R}^{m \times n}$

$$- \dim(\mathbb{R}^{m \times n}) = mn$$

$$- \mathbb{R}^{m \times n} \text{ is the same as } \mathbb{R}^{mn}$$

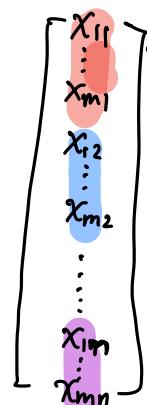
(by concatenating columns of a matrix to form a long vector
in $\mathbb{R}^{m \times n}$ in \mathbb{R}^{mn})

(vectorization)



$\mathbb{R}^{m \times n}$

vectorization



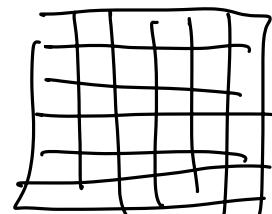
\mathbb{R}^{mn}

- \mathbb{R}^{mxn} can be used to represent
 - Black-white digital images of resolution mxn pixels
 - In recommender systems, ratings of movies by viewers

 $\xrightarrow{n \text{ viewers}}$



- Numerical tabular data of m rows and n columns



Ex. 4: All 3-way arrays of size $mxnxl$, \mathbb{R}^{mxnxl} , with

$$+ : X, Y \in \mathbb{R}^{mxnxl}, \quad X + Y = [x_{ijk} + y_{ijk}]_{i=1, j=1, k=1}^{m n l}$$

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

is a vector space

On this vector space

$$- \quad 0 = [0]_{i=1, j=1, k=1}^{m n l}$$

$$- \quad \{E_{ijk} \mid \begin{matrix} i=1, \dots, m \\ j=1, \dots, n \\ k=1, \dots, l \end{matrix}\}$$

is a basis of \mathbb{R}^{mxnxl}

$$E_{ijk} \in \mathbb{R}^{mxnxl}$$

with only (i, j, k) -entry 1.
and others 0.

$$- \quad \dim(\mathbb{R}^{mxnxl}) = mnl$$

$$- \quad \mathbb{R}^{mxnxl} \text{ is the same as } \mathbb{R}^{mnl}$$

$$- \quad \mathbb{R}^{mxnxl} \text{ can be used to represent}$$

- color image with mxn pixels

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

- Hyperspectral images of $m \times n$ pixels and l spectral channels.
- Black-white videos of resolution $m \times n$ and l frames
⋮
⋮
- Similarly, the set of 4-way, 5-way, ... arrays also form vector spaces.
- 3-way, 4-way, ..., d-way, ..., arrays are called tensors.

Example 5: Consider the set of all strings

Define addition: ' I ' + 'am' = 'I am' (Non-commutable)
and some scalar multiplication

This is not a vector space.

- How to map texts (words) to vectors is a fundamental task in text data analysis and natural language processing (Word embedding)

Example 6: Consider $C[a,b] = \{ f \mid f \text{ is a continuous function on } [a,b] \}$

(with

$$+ : \forall f, g \in C[a,b],$$

$$\text{define } f+g \text{ by } (f+g)(t) = f(t) + g(t)$$

$$\cdot : \begin{array}{c} \forall \alpha \in \mathbb{R} \\ f \in C[a,b] \end{array} \quad \begin{array}{c} \text{define } \alpha \cdot f \text{ by } \forall t \in [a,b] \\ (\alpha f)(t) = \alpha \cdot f(t) \end{array}$$

Then it forms a vector space.

- $C[a,b]$ is a function space, because elements of $C[a,b]$ are functions.
- The 0 vector in $C[a,b]$ is the 0 function that maps every number in $[a,b]$ to 0.

- $\dim(C[a,b]) = +\infty$
- $C[a,b]$ could be a hypothesis space

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

$i=1, \dots, m$

Ex. 7. The infinite sequence set

$$\ell_\infty = \left\{ \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} \mid \begin{array}{l} \exists \text{ a finite } C \in \mathbb{R} \text{ s.t. } \\ |a_i| \leq C \quad \forall i \end{array} \right\}$$

with:

$$+ : (a+fb)_i = a_i + b_i \quad \forall i. \quad \text{and} \quad \forall a, b \in \ell_\infty$$

$$\cdot : (\alpha a)_i = \alpha a_i \quad \forall i \quad \begin{matrix} \alpha \in \mathbb{R} \\ a \in \ell_\infty \end{matrix}$$

It forms a vector space.

$$0 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$e_i = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{pmatrix} \leftarrow i\text{-th entry.}$$

$\{e_i \mid i \in \mathbb{N}\}$ is a basis of ℓ_∞

$$\dim(\ell_\infty) = +\infty$$

§ 1.2 Metric in vector spaces

We need to define "closeness" of two vectors
"distance"

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

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

↑
shift invariant

Therefore, we only need to define

"distance of a vector" to 0

Let $x \in V$. Let $\|x\|$ be its length/magnitude
called norm, \oplus

"magnitude of the vector"
"length of the vector"

It should satisfy:

$$\textcircled{1} \quad \|x\| \geq 0 \quad \forall x \in V \quad \text{and} \quad \|x\|=0 \Leftrightarrow x=0$$

$$\textcircled{2} \quad \|\alpha x\| = |\alpha| \|x\| \quad \forall \alpha \in \mathbb{R} \quad \forall x \in V$$

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

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

$$\textcircled{1} \quad \|x\| \geq 0 \quad \forall x \in V \quad \text{and} \quad \|x\|=0 \Leftrightarrow x=0$$

$$\textcircled{2} \quad \|\alpha x\| = |\alpha| \|x\| \quad \forall \alpha \in \mathbb{R} \quad \forall x \in V$$

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

(Triangle inequality)

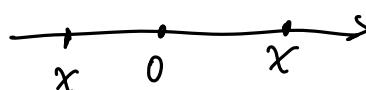
With a norm $\|\cdot\|$,

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

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

— Define $\|x\| = |x| \quad \forall x \in \mathbb{R}$

Then we can check it is a norm.



— Define $\|x\| = 2|x| \quad \forall x \in \mathbb{R}$

It is still a norm on \mathbb{R} .

- ① We can define infinitely many norms on the same vector space
- ② Norms are generalizations of the absolute value.

Ex. 2: \mathbb{R}^n is a vector space.

- Euclidean norm (2-norm)

$$\begin{aligned}\|x\|_2 &= \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} \\ &= \left(\sum_{i=1}^n |x_i|^2 \right)^{\frac{1}{2}}\end{aligned}$$

- Indeed, we can prove $\|\cdot\|_2$ is a norm on \mathbb{R}^n

- **1-norm:** $\|x\|_1 = \sum_{i=1}^n |x_i|$

Indeed, $\|\cdot\|_1$ is a norm on \mathbb{R}^n

- **p-norm ($p \geq 1$):** $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}$

$\|\cdot\|_p$ is a norm on \mathbb{R}^n

- **Infinity norm:** $\|x\|_\infty = \lim_{p \rightarrow +\infty} \|x\|_p$
 $\quad \quad \quad = \max_{i=1}^n |x_i|$
 Can prove \rightarrow

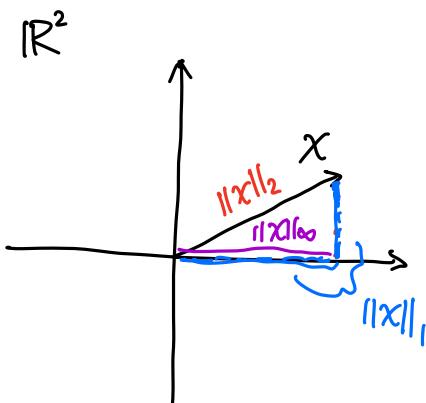
Let $i_0 = \arg \max_{i=1}^n |x_i|$

$$\begin{aligned}|x_{i_0}|^p &\leq \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}} \\ &\leq \underbrace{(n |x_{i_0}|^p)^{\frac{1}{p}}}_{n^{\frac{1}{p}} |x_{i_0}|}\end{aligned}$$

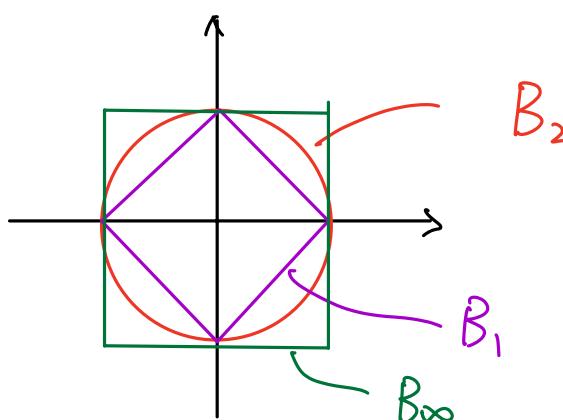
Let $p \rightarrow +\infty$

$$|x_{i_0}| \leq \lim_{p \rightarrow +\infty} \|x\|_p \leq |x_{i_0}|$$

$$\Rightarrow \lim_{p \rightarrow +\infty} \|x\|_p = |x_{i_0}| = \max_{i=1}^n |x_i|$$



Unit Balls : $B_p = \{x \in \mathbb{R}^n \mid \|x\|_p = 1\}$



- A common technique in machine learning to find vectors with

- different structures is to minimize p -norms with constraints
- e.g. — for sparse vectors, l_1 -norm minimization
(LASSO)
 - for vectors following Gaussian distribution, l_2 -norm minimization
 - for vector with ± 1 entries, l_∞ -norm minimization

Ex. 3. $\mathbb{R}^{m \times n}$ is a vector space.

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

We can define vector p -norms for matrices

- $p=1$: $\|A\|_{1,\text{vec}} = \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|$
- $p=2$: $\|A\|_{2,\text{vec}} = \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 \right)^{\frac{1}{2}}$

(also known as Frobenius norm, denoted by $\|\cdot\|_F$)
(i.e., $\|A\|_F = \|A\|_{2,\text{vec}}$)

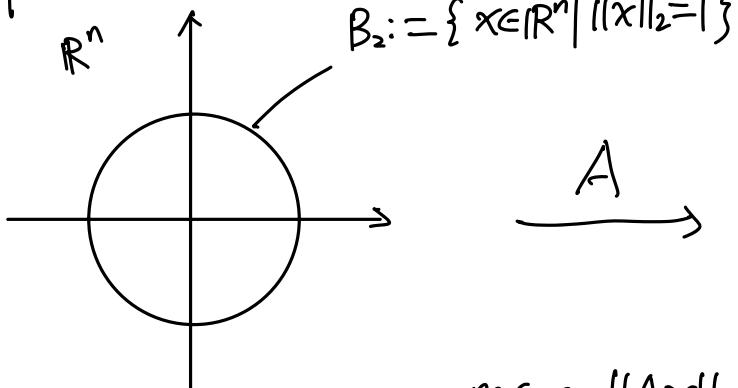
- $p=\infty$: $\|A\|_{\infty,\text{vec}} = \max_{i \in 1}^m \max_{j=1}^n |a_{ij}|$

- $\mathbb{R}^{m \times n}$ can be viewed as linear transformations $\mathbb{R}^n \rightarrow \mathbb{R}^m$

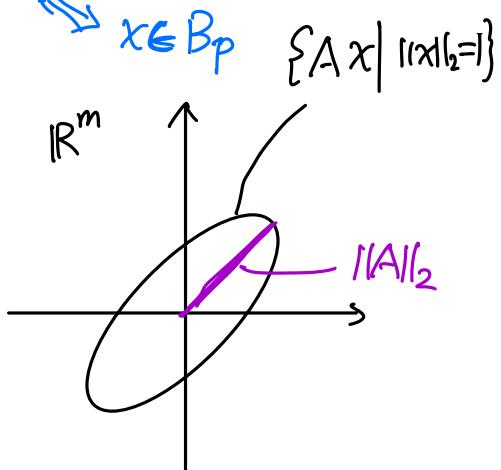
We can define the matrix p -norm

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} \stackrel{\substack{\text{Can prove} \\ \downarrow}}{=} \max_{\substack{x: \|x\|_p=1}} \|Ax\|_p$$

- $p=2$:



$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$$



$$\Leftrightarrow \|A\|_2^2 = \max_{\|x\|_2=1} \|Ax\|_2^2 = \max_{x^T x=1} x^T A^T A x$$

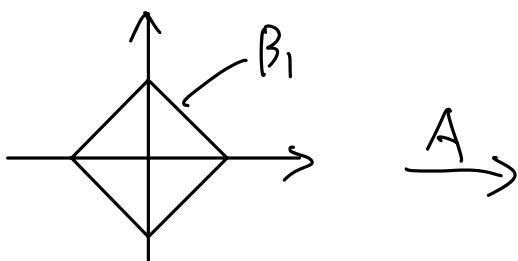
$$\boxed{\|x\|_2^2 = x^T x}$$

= the max eigenvalue of $A^T A$

$$\Leftrightarrow \|A\|_2 = (\text{max eig value of } A^T A)^{1/2} = \text{max singular value of } A$$

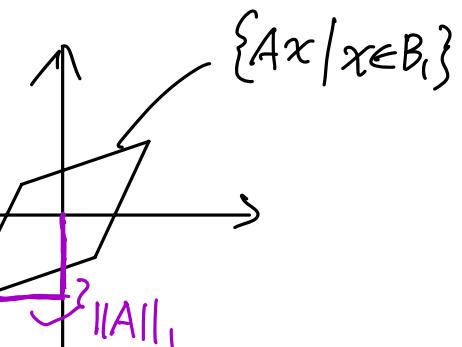
In numerical analysis, $\|\cdot\|_2$ is the default norm of a matrix, also denoted by $\|\cdot\|$

- $p=1$: $\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1$



can prove

$$\|A\|_1 \stackrel{?}{=} \max_{j=1}^n \|a_j\|_1$$



if $A = [a_1 \ a_2 \ \dots \ a_n]$

where $a_i \in \mathbb{R}^m$.

- $p=\infty$:

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty$$

can prove

$$\stackrel{?}{=} \max_{i=1}^m \|a^{(i)}\|_1$$

if $A = \begin{bmatrix} (a^{(1)})^T \\ \vdots \\ (a^{(m)})^T \end{bmatrix}$
where $a^{(i)} \in \mathbb{R}^n$.

- We can also define a mixed matrix norm

$$\|A\|_{p \rightarrow q} = \max_{\|x\|_p=1} \|Ax\|_q$$

$$\mathbb{R}^n \rightarrow \mathbb{R}^m$$

$$\|\cdot\|_p \quad \|\cdot\|_q$$

- We can define other norms, e.g.,

the nuclear norm $\|\cdot\|_*$

$$\|A\|_* = (\text{the sum of all singular values of } A)$$

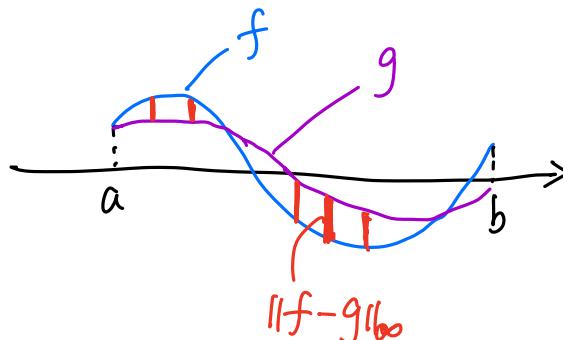
Ex. 4. $C[a,b] = \{f: [a,b] \rightarrow \mathbb{R} \mid f \text{ is continuous}\}$ is a vector space.

$\forall f \in C[a,b]$, define $\|f\|_\infty = \max_{t \in [a,b]} |f(t)|$

— We can check $\|\cdot\|_\infty$ is a norm on $C[a,b]$.

— Then $\forall f,g \in C[a,b]$, their distance is

$$\|f-g\|_\infty = \max_{t \in [a,b]} |f(t)-g(t)|$$



— Some examples of other norms on $C[a,b]$

— p-norm $\|f\|_p = \left(\int_a^b |f(t)|^p dt \right)^{1/p}$ ($p \geq 1$)

— $p=1$: $\|f\|_1 = \int_a^b |f(t)| dt$

— $p=2$: $\|f\|_2 = \left(\int_a^b |f(t)|^2 dt \right)^{1/2}$

Ex 5.: $\ell_\infty = \left\{ a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{pmatrix} \mid \exists C > 0 \text{ st. } |a_i| \leq C \ \forall i \right\}$

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

Then $\|\cdot\|_\infty$ is a norm on ℓ_∞ .

• define $\|a\|_p = \left(\sum_{i=1}^{+\infty} |a_i|^p \right)^{1/p}$ ($p \geq 1$)

but $\|\cdot\|_p$ is NOT a norm on ℓ_∞ .

Ex: $a = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{pmatrix} \in \ell_\infty \quad \|a\|_p = \left(\sum_{i=1}^{+\infty} 1 \right)^{1/p} = +\infty \notin \mathbb{R}$

• We consider the set $\ell_p = \left\{ a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{pmatrix} \mid \|a\|_p < +\infty \right\} \subseteq \ell_\infty$

we can prove that

$\|\cdot\|_p$ is a norm on ℓ_p .

"sup" can ~~not~~ be replaced by "max".

Ex: $a = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{3} \\ \frac{1}{4} \\ \vdots \end{pmatrix}$

$\max_i |a_i|$ doesn't exist because $a_i < a_{i+1} \ \forall i$

Remarks: 1. For the same vector space, we can define infinitely

many norms on it.

2. A common tech in ML is to optimize norms of the unknown vector.
Different norms leads very different solutions.

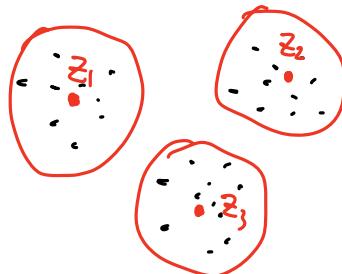
§ 2.3. Case Study: K-means for Clustering

Clustering:

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

$\left(\begin{array}{l} \mathbb{R}^n \text{ can be replaced} \\ \text{with other vector} \\ \text{spaces} \end{array} \right)$

Group them into different K groups.



Applications:

- Image Clustering
- Text data Clustering
- Recommender System

— Let c_i — be the group that x_i belongs to
 $i=1, 2, \dots, N$.

G_j — be the groups, $G_j = \{ i \mid c_i = j \}$

$j=1, 2, \dots, K$

and

z_j — be the representative vector in G_j

$j=1, 2, \dots, K$

(and $z_j \in \mathbb{R}^n$ and z_j is not necessarily from $\{x_1, x_2, \dots, x_N\}$)

— Define the quality of a grouping $\{G_1, \dots, G_K\}$

① Within G_j , all vectors should be close to z_j :

Define

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

We want J_j small.

② Consider all groups: J_j should be small for all j .

Define

$$J = J_1 + J_2 + \dots + J_K = \sum_{j=1}^K J_j$$

We want J small.

So we solve

$$\boxed{\min J}$$



$$\boxed{\min_{\substack{G_1, G_2, \dots, G_K \\ z_1, z_2, \dots, z_K}} \sum_{j=1}^K \sum_{i \in G_j} \|x_i - z_j\|_2^2}$$

Alternating minimizations.

Step 0: Initialize z_1, z_2, \dots, z_K

→ Step 1: Fix z_1, \dots, z_K , solve

$$\min_{G_1, G_2, \dots, G_K} \sum_{j=1}^K \sum_{i \in G_j} \|x_i - z_j\|_2^2 \quad (1)$$

Step 2: Fix G_1, \dots, G_K , Solve

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

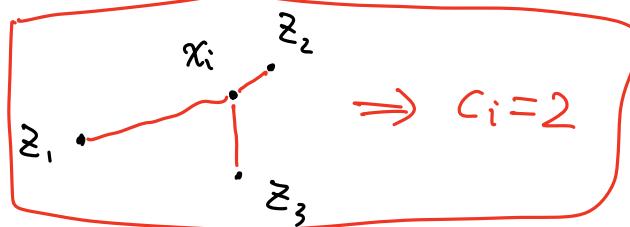
Repeat

To solve subproblem (1):

$$\sum_{j=1}^K \sum_{i \in G_j} \|x_i - z_j\|_2^2 = \sum_{i=1}^N \|x_i - z_{c_i}\|_2^2$$

Then (1) \Leftrightarrow $\min_{c_1, c_2, \dots, c_N} \sum_{i=1}^N \|x_i - z_{c_i}\|_2^2$

 $\Leftrightarrow \min_{c_1, c_2, \dots, c_N} \|x_1 - z_{c_1}\|_2^2 + \|x_2 - z_{c_2}\|_2^2 + \dots + \|x_N - z_{c_N}\|_2^2$
 $\Leftrightarrow \min_{c_i} \|x_i - z_{c_i}\|_2^2, \quad i=1, 2, \dots, N$
 $\Leftrightarrow \min_{c_i \in \{1, 2, \dots, k\}} \|x_i - z_{c_i}\|_2^2 \quad i=1, 2, \dots, N$
 $\Leftrightarrow c_i = \operatorname{argmin}_{c_i} \{\|x_i - z_1\|_2^2, \|x_i - z_2\|_2^2, \dots, \|x_i - z_k\|_2^2\}$



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

Then we use c_i to define G_j :

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

To solve subproblem (2):

$$\min_{z_1, \dots, z_k} \sum_{j=1}^k \sum_{i \in G_j} \|x_i - z_j\|_2^2 \quad (2)$$



$$\min_{z_1, z_2, \dots, z_k} \sum_{i \in G_1} \|x_i - z_1\|_2^2 + \sum_{i \in G_2} \|x_i - z_2\|_2^2 + \dots + \sum_{i \in G_k} \|x_i - z_k\|_2^2$$



$$\min_{z_j} \sum_{i \in G_j} \|x_i - z_j\|_2^2, \quad j=1, \dots, k$$

$$z_j = \frac{1}{|G_j|} \sum_{i \in G_j} x_i$$

$|G_j|$ — number of elements in G_j

— mean of vectors in G_j

for $j=1, 2, \dots, k$.

To see this, consider $n=1$

$$\min_{z_j \in \mathbb{R}} \sum_{i \in G_j} (x_i - z_j)^2$$

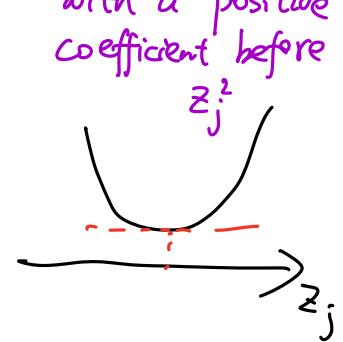
a quadratic function of z_j with a positive n

$$\sum_{i \in G_j} z_j(z_j - x_i) = 0$$

↑ take derivative and set it 0.

$$\sum_{i \in G_j} z_j = \sum_{i \in G_j} x_i$$

$$z_j = \frac{1}{|G_j|} \sum_{i \in G_j} x_i$$



z_j is the mean of vectors in G_j

Algorithm: (K-means algorithm)

Step 0: Initialize z_1, z_2, \dots, z_k

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

(i.e., we assign x_i to the nearest representative vector)

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

Step 2: $z_j = \frac{1}{|G_j|} \sum_{i \in G_j} x_i, \quad j=1, 2, \dots, k.$

(z_j is the mean of vectors in Group j)

Repeat

Replace 2-norm with 1-norm.

$$\min_{G_1, G_2, \dots, G_K} \sum_{j=1}^k \sum_{i \in G_j} \|x_i - z_j\|_1$$

$$z_1, z_2, \dots, z_k$$

Alternating minimizations.

Step 0: Initialize z_1, z_2, \dots, z_k

→ Step 1: Fix z_1, \dots, z_k , solve

$$\min_{G_1, G_2, \dots, G_K} \sum_{j=1}^k \sum_{i \in G_j} \|x_i - z_j\|_1 \quad (1)$$

Step 2: Fix G_1, \dots, G_K . Solve

$$\min_{z_1, \dots, z_K} \sum_{j=1}^K \sum_{i \in G_j} \|x_i - z_j\|_1, \quad (2)$$

Repeat

To solve subproblem (1):

$$\sum_{j=1}^K \sum_{i \in G_j} \|x_i - z_j\|_1 = \sum_{i=1}^N \|x_i - z_{c_i}\|_1$$

$$\begin{aligned} \text{Then } (1) &\iff \min_{c_1, c_2, \dots, c_N} \sum_{i=1}^N \|x_i - z_{c_i}\|_1 \\ &\iff \min_{c_1, c_2, \dots, c_N} \|x_1 - z_{c_1}\|_1 + \|x_2 - z_{c_2}\|_1 + \dots + \|x_N - z_{c_N}\|_1 \\ &\iff \min_{c_i} \|x_i - z_{c_i}\|_1, \quad i=1, 2, \dots, N \\ &\iff \min_{c_i \in \{1, 2, \dots, K\}} \|x_i - z_{c_i}\|_1, \quad i=1, 2, \dots, N \\ &\iff c_i = \operatorname{argmin}_{c_i} \{\|x_i - z_1\|_1, \|x_i - z_2\|_1, \dots, \|x_i - z_K\|_1\} \end{aligned}$$

$\Rightarrow c_i = 2$

Assign x_i to the nearest representative in 1-norm distance

To solve subproblem (2):

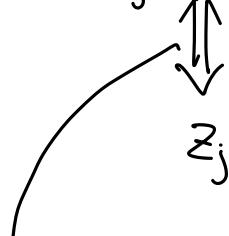
$$\min_{z_1, \dots, z_K} \sum_{j=1}^K \sum_{i \in G_j} \|x_i - z_j\|_1, \quad (2)$$



$$\min_{z_1, z_2, \dots, z_K} \sum_{i \in G_1} \|x_i - z_1\|_1 + \sum_{i \in G_2} \|x_i - z_2\|_1 + \dots + \sum_{i \in G_K} \|x_i - z_K\|_1$$



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

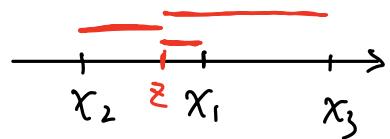


$$z_j = \operatorname{median} \{x_i \mid i \in G_j\}$$

median is taken entrywisely.

Consider $n=1$.

$$\min_{z_j \in \mathbb{R}} \sum_{i \in G_j} |x_i - z_j|$$



Indeed, the optimal

$$z_j = \text{median } \{x_i \mid i \in G_j\}$$

$$|x_1 - z| + |x_2 - z| + |x_3 - z|$$

$$= |x_3 - x_2| + |z - x_1| \text{ if } z \in (x_2, x_3)$$

the optimal

$$z = x_1 = \text{median } \{x_1, x_2, x_3\}$$

Algorithm: (K-medians algorithm)

Step 0: Initialize z_1, z_2, \dots, z_k

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

(i.e., we assign x_i to the nearest representative vector)

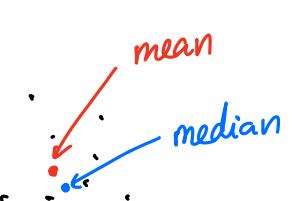
$$G_j = \{i \mid c_i = j\}, \quad j=1, \dots, k.$$

Step 2: $z_j = \text{median } \{x_i \mid i \in G_j\} \quad j=1, 2, \dots, k.$

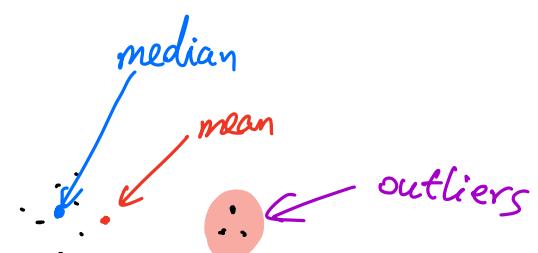
(z_j is the mean of vectors in Group j)

Repeat

Comparison of K-means and K-medians



Both mean and median
are good ~~for~~ for
a representative vec.



Median seems better
than the mean.

- Mean is sensitive to outliers
- Median is robust
- In ML algorithms,

l -norm distance is more robust to outliers than 2-norm.

Approximation of vectors on vector spaces.

— Iterative alg.

— Approximate of functions on vector spaces
(Calculus on vector spaces)

§ 2.3. Limit and convergence of vectors

Let V be a vector space with a norm $\|\cdot\|$ ($V, \|\cdot\|$)
(V is a normed vector space)

Let $\{x^{(k)}\}_{k \in \mathbb{N}} \subset V$.

Let $x \in V$

Then define: $\{x^{(k)}\}_{k \in \mathbb{N}}$ converges to x , denoted by $x^{(k)} \rightarrow x$, if

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

i.e.,

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

Ex 1: $(\mathbb{R}^n, \|\cdot\|_2)$

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

Then $\|x^{(k)} - x\|_2 = \|x^{(k)}\|_2 = \left(\sum_{i=1}^n \left(\frac{i}{k}\right)^2 \right)^{\frac{1}{2}} = \frac{1}{k} \cdot \left(\sum_{i=1}^n i^2 \right)^{\frac{1}{2}}$ constant of k $\rightarrow 0$ as $k \rightarrow \infty$

i.e., $\lim_{k \rightarrow \infty} \|x^{(k)} - x\|_2 = 0$

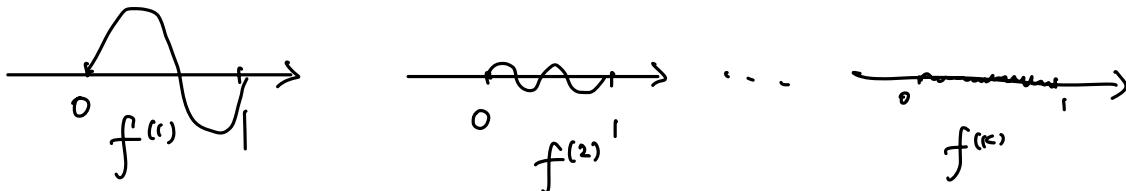
$$x^{(k)} \rightarrow x$$

Ex 2: Consider $C[0,1] = \{f \mid f \text{ is a continuous function on } [0,1]\}$
with $\|\cdot\|_\infty$ -norm $(\|f\|_\infty = \max_{t \in [0,1]} |f(t)|)$

Consider $f^{(k)}(t) = \frac{\sin(2\pi kt)}{k^2}$. Let 0 be the 0 function.

$$\lim_{k \rightarrow \infty} \|f^{(k)} - 0\|_\infty = \lim_{k \rightarrow \infty} \|f^{(k)}\|_\infty = \lim_{k \rightarrow \infty} \left\| \frac{\sin(2\pi kt)}{k^2} \right\|_\infty = \lim_{k \rightarrow \infty} \frac{1}{k^2} = 0$$

$$\text{So, } f^{(k)} \rightarrow 0$$



Ex 3. Consider infinite sequences:

$$a^{(k)} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{k-terms} \quad \in l_1, l_2, l_\infty$$

$$\forall k: \|a^{(k)}\|_1 = \sum_{i=1}^k \frac{1}{k} = 1 < +\infty$$

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

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

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

- In l_2 (with $\|\cdot\|_2$ -norm) $l_2 = \{a \mid \|a\|_2 < +\infty, a \in l_\infty\}$

$$\lim_{k \rightarrow \infty} \|a^{(k)} - a\|_2 \leq \lim_{k \rightarrow \infty} \|a^{(k)}\|_2 = \lim_{k \rightarrow \infty} \sqrt{\frac{1}{k}} = 0$$

So, $a^{(k)} \rightarrow a$ in $\|\cdot\|_2$ -norm

- In l_∞ (with $\|\cdot\|_\infty$ -norm) $l_\infty = \{a \mid \exists C \text{ s.t. } |a_i| \leq C \text{ if } i\}$

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

So, $a^{(k)} \rightarrow a$ in $\|\cdot\|_\infty$ -norm

- In l_1 (with $\|\cdot\|_1$ -norm) $l_1 = \{a \mid \|a\|_1 < +\infty, a \in l_\infty\}$

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

So, $a^{(k)} \not\rightarrow a$ in $\|\cdot\|_1$ -norm

The convergence/limit of vectors depends on the norm of the normed vector space

Ex. 4. Consider $V = \{a \mid \|a\|_1 < +\infty, a \in \ell^\infty\}$

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

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

$$\forall k: \|a^{(k)}\|_\infty = 1 < +\infty$$

$$\|a^{(k)}\|_1 = \sum_{i=1}^k y_i < +\infty$$

$$\Rightarrow a^{(k)} \in V$$

$$a = \begin{pmatrix} 1 \\ y_2 \\ \vdots \\ y_k \\ \vdots \end{pmatrix}$$

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

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



The limit of vectors in V may not be in the same vector space V .

If this happens, the normed vector space is called incomplete

Completeness of normed vector spaces

- Completeness matters

- Iterative alg in ML.

$$x^{(k)} \in V$$

If not complete, then

$$\lim_{k \rightarrow +\infty} x^{(k)} \notin V$$

- Calculus of functions on vector spaces.

$f(\lim_{k \rightarrow +\infty} x^{(k)}) = f(x)$ — not defined.

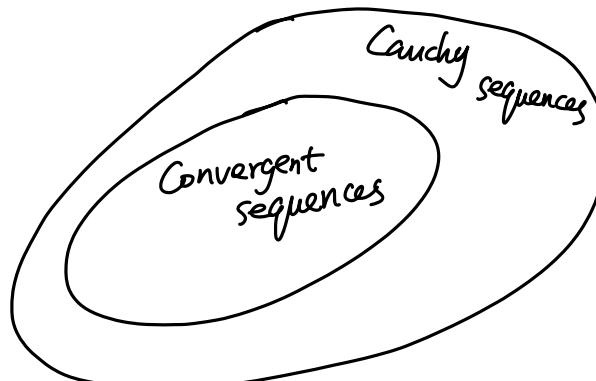
Completeness:

Cauchy sequence:

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

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

Fact: ① If $x^{(k)} \rightarrow x \in V$ in $(V, \|\cdot\|)$, then $\{x^{(k)}\}_{k \in \mathbb{N}}$ must be a Cauchy sequence.



② The reverse may not be true. (See Ex. 4)

A normed vector space $(V, \|\cdot\|)$ is complete if the limit of all Cauchy sequences is in V

$$\{\text{Convergent sequences}\} = \{\text{Cauchy sequences}\}$$



$(V, \|\cdot\|)$ is complete

We can complete any incomplete vector spaces, by including all limits of its Cauchy sequences

$$(V, \|\cdot\|) \rightarrow (\bar{V}, \|\cdot\|)$$

We call complete normed vector space a **Banach Space**

Examples of Banach spaces (complete normed vector spaces)

- \mathbb{R}^n with any norm.
- $\mathbb{R}^{m \times n}$ with any norm.
- Tensor spaces $\mathbb{R}^{m \times n \times l}$ with any norm.
- $C[a, b]$ with $\| \cdot \|_\infty$
- l_p with p -norm, $p \geq 1$ or $p = \infty$.

Examples of incomplete normed vector spaces.

- $V = \{a | \|a\|_1 < \infty, a \in l_\infty\}$ with $\| \cdot \|_\infty$ -norm is incomplete.
Its completion is $(l_\infty, \| \cdot \|_\infty)$
- $C[a, b]$ with p -norm ($p \geq 1$, p finite)
is incomplete
$$\|f\|_p = \left(\int_a^b |f(t)|^p dt \right)^{\frac{1}{p}}$$

Its completion is $L^p(a, b)$

Applications:

- Supervised learning:

Given (x_i, y_i) , $i=1, \dots, m$ $x_i \in [a, b]$ $y_i \in \mathbb{R}$

Find $f \in C[a, b]$ s.t. $f(x_i) \approx y_i$ $i=1, \dots, m$

An iter alg. generate

$f^{(k)} \in C[a, b]$, $k=1, 2, \dots$

- $\| \cdot \|_1$ -norm: We may find f s.t.

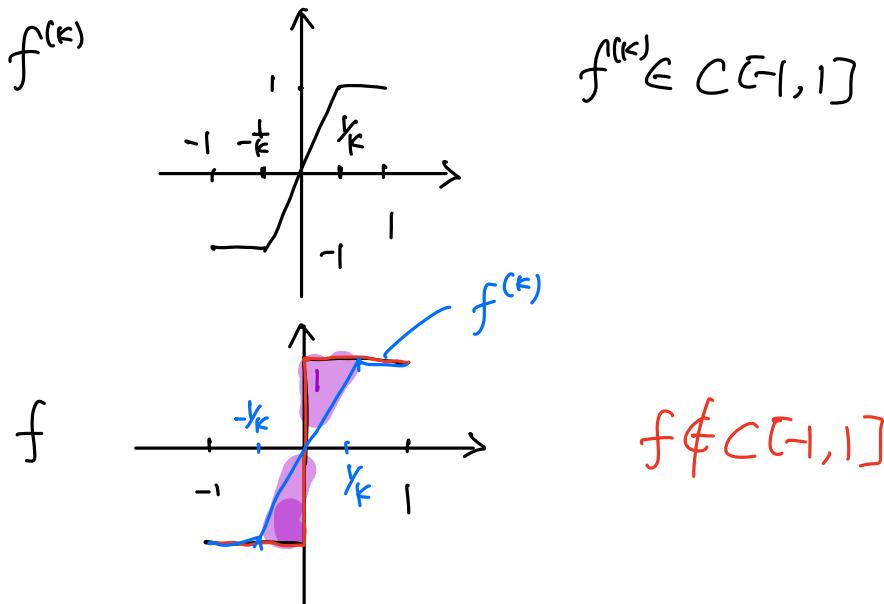
$\lim_{k \rightarrow \infty} \|f^{(k)} - f\|_1 = 0$ for some f .

$(C[a,b], \|\cdot\|_1)$ is incomplete.

\Rightarrow it might happen $f \notin C[a,b]$.

$\Rightarrow f$ is discontinuous.

Explicit Example:



$$\|f^{(k)} - f\|_1 = \int_{-1}^1 |f^{(k)}(t) - f(t)| dt = \text{area of two purple triangles}$$

$\rightarrow 0$ as $k \rightarrow \infty$

$$\Rightarrow \lim_{k \rightarrow \infty} \|f^{(k)} - f\|_1 = 0$$

• $\|\cdot\|_\infty$ -norm: We may find f s.t

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

$(C[a,b], \|\cdot\|_\infty)$ is complete $\Rightarrow f \in C[a,b]$

— Iterative algorithm.

Generate a sequence of vectors $\{\chi^{(k)}\}_{k \in \mathbb{N}} \subset V$ with $\|\cdot\|_1$.

Check the convergence?

— $\lim_{k \rightarrow \infty} \|\chi^{(k)} - \chi\| = 0$ is not practical, because we don't have χ .

— Use Cauchy sequence.

We check

$$\|x^{(k)} - x^{(l)}\| \leq \varepsilon$$

for large
K, l.

practically, we just check

$$\|x^{(k)} - x^{(k+1)}\| \stackrel{?}{\leq} \varepsilon$$

§ 2.5. Finite dimensional vector spaces

- In most of the cases, we are dealing with finite dim vector spaces, e.g. \mathbb{R}^n , $\mathbb{R}^{m \times n}$, $\mathbb{R}^{m \times n \times l}$
- Properties of finite dim vector spaces.
 - ① Any finite dim vector space with any norm is complete.
 - ② For a finite dim vector space V , all norms are equivalent.

for any two norms $\|\cdot\|_A$ and $\|\cdot\|_B$ on a finite dim vector space V ,

$$\exists C_1, C_2 > 0 \text{ s.t. } C_1 \|a\|_A \leq \|a\|_B \leq C_2 \|a\|_A, \forall a \in V$$

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: " \Rightarrow "

Since $x^{(k)} \rightarrow x$ in $\|\cdot\|_A$, $\lim_{k \rightarrow \infty} \|x^{(k)} - x\|_A = 0$

Due to the norm equivalence,

$$c_1 \|x^{(k)} - x\|_B \leq \|x^{(k)} - x\|_A \leq c_2 \|x^{(k)} - x\|_B$$

$$0 = C_1 \cdot \lim_{k \rightarrow \infty} \|x^{(k)} - x\|_A \leq \lim_{k \rightarrow \infty} \|x^{(k)} - x\|_B \leq C_2 \cdot \lim_{k \rightarrow \infty} \|x^{(k)} - x\|_A = 0$$

$$\Rightarrow \lim_{k \rightarrow \infty} \|x^{(k)} - x\|_B = 0$$

i.e. $x^{(k)} \rightarrow x$ in $\|\cdot\|_B$.

" \Leftarrow " Similar. ⊗

Examples of norm equivalence

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

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

$$\|a\|_2 \leq \|a\|_1 \leq \sqrt{n} \|a\|_2 \quad \forall a \in \mathbb{R}^n$$

$$\|a\|_2^2 = \sum_{i=1}^n |a_i|^2$$

$$\|a\|_1^2 = (\sum_i |a_i|)^2 = \sum_{i,j} |a_i||a_j| = \sum_{i=1}^n |a_i|^2 + \sum_{i \neq j} |a_i||a_j| \quad \left. \begin{array}{l} \\ \end{array} \right\} \|a\|_2^2 \leq \|a\|_1^2$$

$$\begin{aligned} \|a\|_1^2 &= \sum_{i=1}^n |a_i|^2 = \sum_{i=1}^n \sum_{j=1}^n |a_i||a_j| \leq \sum_{i=1}^n \sum_{j=1}^n \left(\frac{1}{2}(|a_i|^2 + |a_j|^2) \right) \\ &= n \left(\sum_{i=1}^n |a_i|^2 \right) = n \cdot \|a\|_2^2 \end{aligned}$$

- $\|\cdot\|_2$ and $\|\cdot\|_\infty$ are equ.

$$\|a\|_\infty \leq \|a\|_2 \leq \sqrt{n} \|a\|_\infty \quad \forall a \in \mathbb{R}^n$$

- $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are equ.

$$\|a\|_\infty \leq \|a\|_1 \leq n \|a\|_\infty \quad \forall a \in \mathbb{R}^n$$

- The convergence speed depends on norms.

Ex: $\mathbb{R}^2, x^{(k)} = \frac{1}{k} \begin{bmatrix} \cos\left(\frac{k+1}{2}\pi\right) \\ \sin\left(\frac{k+1}{2}\pi\right) \end{bmatrix} \subset \mathbb{R}^2$

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

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

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

To achieve an ε -precision

$$2\text{-norm: } \|x^{(k)} - 0\|_2 = \frac{1}{k} \leq \varepsilon \Rightarrow k \geq \frac{1}{\varepsilon}$$

$$1\text{-norm: } \|x^{(k)} - 0\|_1 = \frac{\sqrt{2}}{k} \leq \varepsilon \Rightarrow k \geq \frac{\sqrt{2}}{\varepsilon}$$

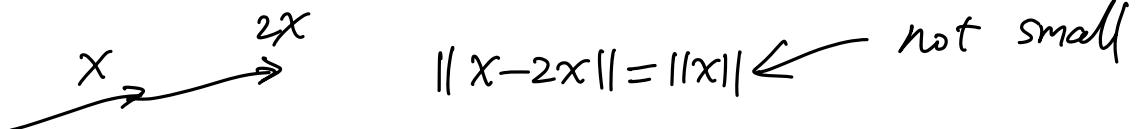
③ The computation can only be done in finite dim vector spaces.

Ch.3. Inner product, Hilbert Spaces

In many applications, norm/distance is not enough.

How two vectors are correlated / aligned?

-

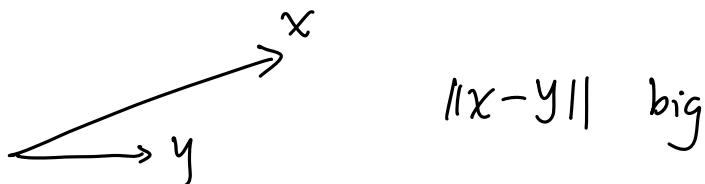


-

x, y pictures of the same scene

at noon
at night

$$x \approx 2y$$



We need correlation of two vectors

§. 3.1 Inner product.

Let V be a vector space over \mathbb{R}

A function $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$ is called an inner product over \mathbb{R} if:

① $\forall x \in V, \langle x, x \rangle \geq 0$

and $\langle x, x \rangle = 0 \iff x = 0$

② $\langle \alpha x_1 + \beta x_2, y \rangle = \alpha \langle x_1, y \rangle + \beta \langle x_2, y \rangle \quad \forall x_1, x_2, y \in V$
 $\alpha, \beta \in \mathbb{R}$

$$③ \quad \forall x, y \in V, \quad \langle x, y \rangle = \langle y, x \rangle$$

- ② & ③ $\Rightarrow \langle x, \alpha y_1 + \beta y_2 \rangle = \alpha \langle x, y_1 \rangle + \beta \langle x, y_2 \rangle$

and

$$\langle \alpha_1 x_1 + \alpha_2 x_2, \beta_1 y_1 + \beta_2 y_2 \rangle$$

$$= \alpha_1 \beta_1 \langle x_1, y_1 \rangle + \alpha_2 \beta_1 \langle x_2, y_1 \rangle + \alpha_1 \beta_2 \langle x_1, y_2 \rangle + \alpha_2 \beta_2 \langle x_2, y_2 \rangle$$

- We can also define inner prod. over \mathbb{C} .

Ex 1: \mathbb{R}^n . we can define

$$\forall x, y \in \mathbb{R}^n \quad \langle x, y \rangle = \sum_{i=1}^n x_i y_i = x^T y = y^T x$$

(Euclidean Inner Product)

Ex 2: \mathbb{R}^n . We can define a weighted inner product.

$$\langle x, y \rangle_A = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i y_j = x^T A y \quad \text{where } A \in \mathbb{R}^{n \times n} \text{ is}$$

Let's check $\langle \cdot, \cdot \rangle_A$ is an inner prod.

a symmetric positive definite (SPD) matrix,

$$① \quad \langle x, x \rangle_A = x^T A x \geq 0 \quad \forall x \in \mathbb{R}^n$$

$$\langle x, x \rangle_A = 0 \Leftrightarrow x^T A x = 0 \Leftrightarrow x = 0$$

i.e. $\begin{cases} A = A^T \text{ and} \\ x^T A x > 0 \quad \forall x \neq 0 \end{cases}$

$$② \quad \langle \alpha x_1 + \beta x_2, y \rangle_A = (\alpha x_1 + \beta x_2)^T A y$$

$$= \alpha x_1^T A y + \beta x_2^T A y = \alpha \langle x_1, y \rangle_A + \beta \langle x_2, y \rangle_A$$

$$③ \quad \langle x, y \rangle_A = x^T A y = (x^T A y)^T = y^T A^T x = y^T A x = \langle y, x \rangle_A \quad \blacksquare$$

If we choose $A = I$ (I is SPD)

$$\text{then } \langle x, y \rangle_I = x^T I y = x^T y = \langle x, y \rangle$$

For the same vector space, we can infinitely many inner products.

Ex. 3. $\mathbb{R}^{m \times n}$. We define $\langle \cdot, \cdot \rangle$

$$\forall A, B \in \mathbb{R}^{m \times n}, \quad \langle A, B \rangle = \sum_{i=1}^m \sum_{j=1}^n a_{ij} b_{ij} \quad \left\{ \begin{array}{l} = \text{trace}(A^T B) \\ = \text{trace}(B^T A) \\ = \text{trace}(AB^T) \\ = \text{trace}(BA^T) \end{array} \right.$$

recall $\text{trace}(C) = \sum_{i=1}^n C_{ii}$
 $\forall C \in \mathbb{R}^{n \times n}$

Ex.4. For two infinite sequences, $a, b \in \ell^2 = \left\{ \begin{pmatrix} c_1 \\ c_2 \\ \vdots \end{pmatrix} : \sum_{i=1}^{+\infty} c_i^2 < +\infty \right\}$

$$\langle a, b \rangle = \sum_{i=1}^{+\infty} a_i b_i$$

Ex.5. In $C[a,b] = \{ f \mid f: [a,b] \rightarrow \mathbb{R} \text{ is a continuous function}\}$

$$\langle f, g \rangle = \int_a^b f(t)g(t) dt$$

§ 3.2. Properties of inner products

Cauchy-Schwarz inequality (C-S)

If $\langle \cdot, \cdot \rangle$ is an inner product in a vector space V ,

$$\text{then } \forall x, y \in V, \quad |\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$$

The " $=$ " holds if and only if $x = \alpha y$ or $y = \alpha x$ for some $\alpha \in \mathbb{R}$.

proof. We first prove the inequality.

- If $y = 0$, then

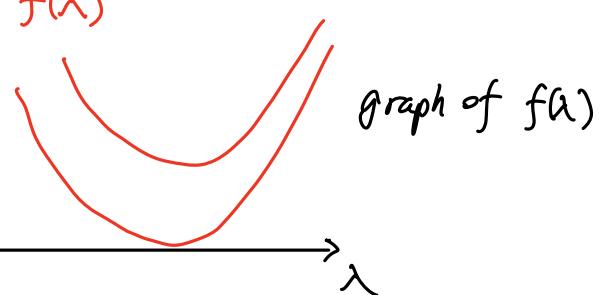
$$\begin{aligned} |\langle x, y \rangle|^2 &= |\langle x, 0 \rangle|^2 = 0 \\ \langle x, x \rangle \langle y, y \rangle &= \langle x, x \rangle \langle 0, 0 \rangle = 0 \end{aligned} \quad \Rightarrow \quad |\langle x, y \rangle|^2 = 0 \leq 0 = \langle x, x \rangle \langle y, y \rangle$$

- If $y \neq 0$, then

Let $\lambda \in \mathbb{R}$ be arbitrary. Consider

$$\begin{aligned} 0 \leq \langle x + \lambda y, x + \lambda y \rangle &= \langle x, x \rangle + \lambda \langle x, y \rangle + \lambda \langle y, x \rangle + \lambda^2 \langle y, y \rangle \\ &= \underbrace{\lambda^2 \langle y, y \rangle}_{f(\lambda)} + \lambda \cdot 2 \langle x, y \rangle + \underbrace{\langle x, x \rangle}_{f(x)} \end{aligned}$$

Since $y \neq 0$, $\langle y, y \rangle > 0$
 $f(\lambda) \geq 0$



\Rightarrow there exists at most one real solution of $f(\lambda) = 0$

$$\Rightarrow \Delta = (2 \langle x, y \rangle)^2 - 4 \cdot \langle y, y \rangle \cdot \langle x, x \rangle \leq 0$$

$$\Rightarrow |\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$$

Secondly, we prove: " $=$ " $\Leftrightarrow x = \alpha y$ or $y = \alpha x$ for some $\alpha \in \mathbb{R}$.

$$\Leftarrow: \text{ If } x = \alpha y : \Rightarrow |\langle x, y \rangle|^2 = |\langle \alpha y, y \rangle|^2 = |\alpha| |\langle y, y \rangle|^2 = \alpha^2 \langle y, y \rangle^2$$

$$\langle x, x \rangle \langle y, y \rangle = \langle \alpha y, \alpha y \rangle \langle y, y \rangle = \alpha^2 \langle y, y \rangle^2$$

$$\Rightarrow " = "$$

If $y = \alpha x$, similar.

\Rightarrow : If " $=$ ": two cases: $\begin{cases} y=0 \\ y \neq 0 \end{cases}$

- If $y=0$, then $y=0 \cdot x$ \square

- If $y \neq 0$, then: Because " $=$ ", $\Delta=0$.

Thus, $f(\lambda)$ has exactly one real root.

i.e., $\exists \beta \in \mathbb{R}$ s.t. $f(\beta) = 0$

$$\Rightarrow \langle x + \beta y, x + \beta y \rangle = 0$$

$$\Rightarrow x + \beta y = 0$$

$$\Rightarrow x = (-\beta) \cdot y$$



With C-S, we can:

- We can define a norm thru the inner product.

Let V be a vector space with $\langle \cdot, \cdot \rangle$

Define

$$\|x\| = (\langle x, x \rangle)^{\frac{1}{2}} \quad \forall x \in V$$

Then $\|\cdot\|$ is a norm on V .

proof. ① $\|x\| = (\langle x, x \rangle)^{\frac{1}{2}} \geq 0 \quad \forall x \in V$

$$\|x\|=0 \Leftrightarrow \langle x, x \rangle^{\frac{1}{2}}=0 \Leftrightarrow \langle x, x \rangle=0 \Leftrightarrow x=0$$

$$\begin{aligned} ② \| \alpha x \| &= \langle \alpha x, \alpha x \rangle^{\frac{1}{2}} = (\alpha^2 \langle x, x \rangle)^{\frac{1}{2}} = |\alpha| \cdot \langle x, x \rangle^{\frac{1}{2}} \\ &= |\alpha| \cdot \|x\| \end{aligned}$$

$$③ \forall x, y \in V \quad \text{WTS: } \|x+y\| \leq \|x\| + \|y\|$$

$$\begin{aligned} \|x+y\|^2 &= \langle x+y, x+y \rangle \\ &= \langle x, x \rangle + 2 \langle x, y \rangle + \langle y, y \rangle \\ &= \|x\|^2 + 2 \langle x, y \rangle + \|y\|^2 \end{aligned}$$

the norm induced by the inner product

the default norm on innerprod vector

$$\begin{aligned}
 & \stackrel{\text{C-S}}{\leq} \|x\|^2 + 2|\langle x, y \rangle| + \|y\|^2 \\
 & \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 \\
 & = (\|x\| + \|y\|)^2
 \end{aligned}$$

$$\Rightarrow \|x+y\| \leq \|x\| + \|y\| \quad \boxed{\checkmark}$$

With the norm on $(V, \langle \cdot, \cdot \rangle)$, Cauchy-Schwartz can be simplified to.

$$|\langle x, y \rangle| \leq \|x\| \|y\| \quad \forall x, y \in V$$

Here $\|\cdot\|$ is the norm induced by $\langle \cdot, \cdot \rangle$

Ex 1: $(\mathbb{R}^n, \langle \cdot, \cdot \rangle) \quad \langle x, y \rangle = x^T y$

The induced norm is

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

Ex 2: \mathbb{R}^n with $\langle \cdot, \cdot \rangle_A \quad \langle x, y \rangle_A = x^T A y,$

where $A \in \mathbb{R}^{n \times n}$ is SPD.

The induced norm is

$$\|x\|_A = (\langle x, x \rangle_A)^{\frac{1}{2}} = (x^T A x)^{\frac{1}{2}} = \left(\sum_{i,j} a_{ij} x_i x_j \right)^{\frac{1}{2}}$$

Ex. 3. the p -norm in \mathbb{R}^n : $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}} \quad (p \geq 1)$

Unless $p=2$, p -norm is NOT induced by any inner prod.

Ex. 4. $\mathbb{R}^{m \times n}$ with $\langle \cdot, \cdot \rangle \quad \langle A, B \rangle = \text{trace}(A^T B)$

The induced norm is

$$\|A\| = \langle A, A \rangle^{\frac{1}{2}} = (\text{trace}(A^T A))^{\frac{1}{2}} = \left(\sum_{i,j} a_{ij}^2 \right)^{\frac{1}{2}} = \|A\|_F$$

- Angles between vectors ($\forall x, y \in V, \langle x, y \rangle \neq 0$)

By C-S, $\forall x, y \in V$

$$|\langle x, y \rangle| \leq \|x\| \|y\|$$

$$-\|x\| \|y\| \leq \langle x, y \rangle \leq \|x\| \|y\|$$

Assume $x, y \neq 0$

inner prod

norm

Then

$$\boxed{-1 \leq \frac{\langle x, y \rangle}{\|x\| \|y\|} \leq 1}$$

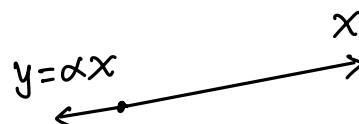
- If $\frac{\langle x, y \rangle}{\|x\| \|y\|} = 1$, then $\langle x, y \rangle = \|x\| \|y\|$
"=" holds in C-S, $\Rightarrow x = \alpha y$ or $y = \alpha x$ with $\alpha > 0$.



So, $\langle x, y \rangle = 0$ $\left| \langle x, y \rangle \right|$

- If $\frac{\langle x, y \rangle}{\|x\| \|y\|} = -1$ then $-\langle x, y \rangle = \|x\| \|y\|$

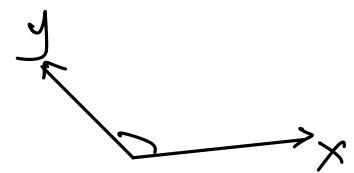
"=" holds in C-S \Rightarrow $x = \alpha y$ or $y = \alpha x$ with $\alpha < 0$



So, $\langle x, y \rangle = \pi$

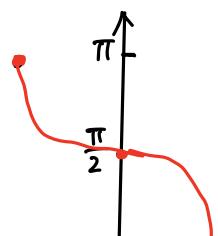
- If $-1 < \frac{\langle x, y \rangle}{\|x\| \|y\|} < 1$, then

$$\langle x, y \rangle \in (0, \pi)$$

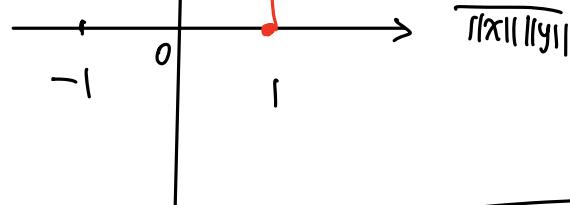


So, the angles should satisfy

$$\langle x, y \rangle = \begin{cases} 0 & \frac{\langle x, y \rangle}{\|x\| \|y\|} = 1 \\ \pi & \frac{\langle x, y \rangle}{\|x\| \|y\|} = -1 \end{cases}$$



$$\langle x, y \rangle$$



We define

$$\langle x, y \rangle = \arccos \frac{\langle x, y \rangle}{\|x\| \|y\|}$$



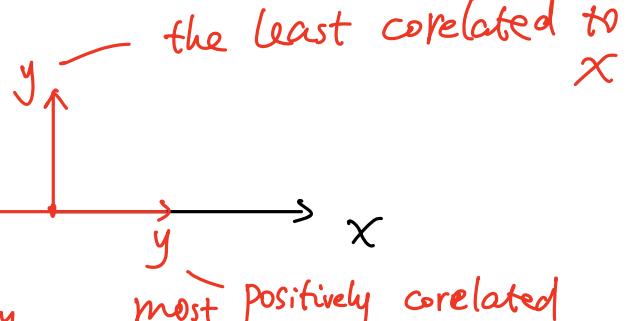
This definition is consistent with angles of vectors in \mathbb{R}^2 and \mathbb{R}^3 with Euclidean inner prod.

Orthogonality:

- $\langle x, y \rangle = 0$, then

$$\left\{ \begin{array}{l} \langle x, y \rangle = \frac{\pi}{2} \\ x, y \text{ are the least correlated.} \end{array} \right.$$

mostly negatively correlated

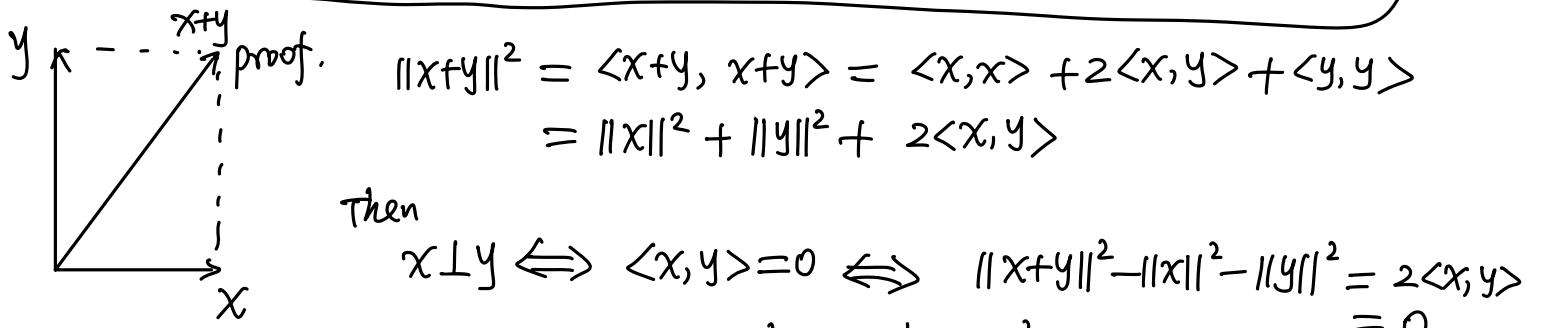


We call x, y are orthogonal, denoted $x \perp y$

- Pythagoras' thm

Let x, y be two vectors on V with $\langle \cdot, \cdot \rangle$ and the induced norm $\|\cdot\|$. Then

$$x \perp y \iff \|x+y\|^2 = \|x\|^2 + \|y\|^2$$



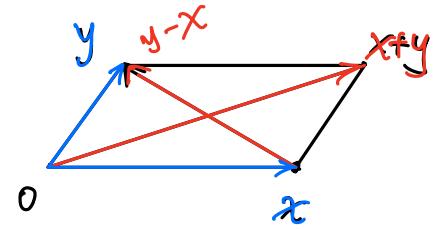
- Parallelogram Law:

Let V be a vector space with $\langle \cdot, \cdot \rangle$ and the induced norm $\|\cdot\|$. Then

$$2(\|x\|^2 + \|y\|^2) = \|x+y\|^2 + \|x-y\|^2 \quad \forall x, y \in V$$

proof.

$$\begin{aligned} & \|x+y\|^2 + \|x-y\|^2 \\ &= \langle x+y, x+y \rangle + \langle x-y, x-y \rangle \\ &= \langle x, x \rangle + \langle y, y \rangle + \cancel{2\langle x, y \rangle} \\ &\quad + \cancel{\langle x, x \rangle + \langle y, y \rangle - 2\langle x, y \rangle} \\ &= 2\langle x, x \rangle + 2\langle y, y \rangle \\ &= 2(\|x\|^2 + \|y\|^2) \end{aligned}$$



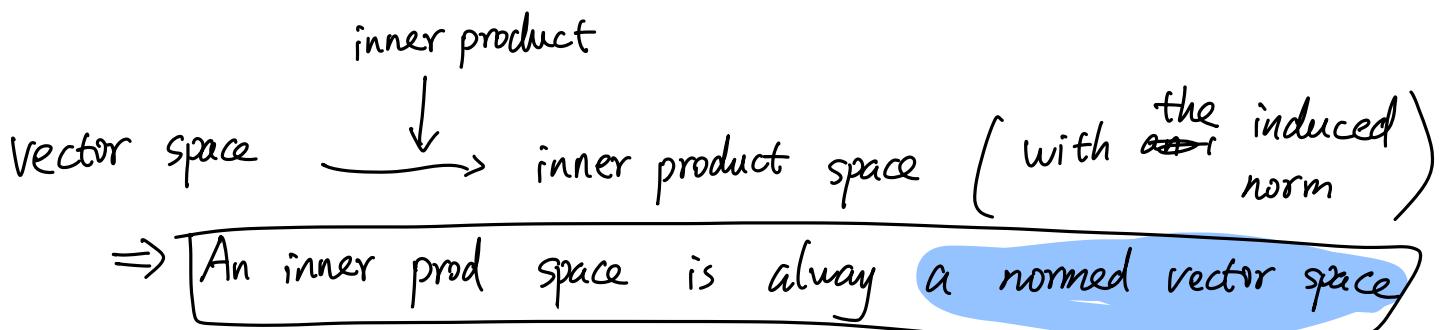
Actually, let V be a vector space with a norm $\|\cdot\|$.

If

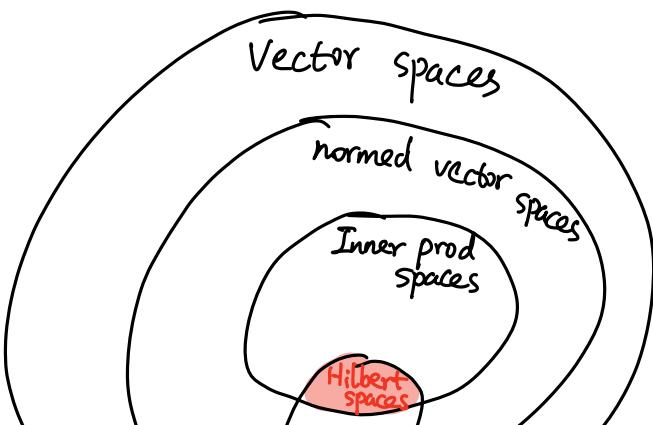
$$2(\|x\|^2 + \|y\|^2) = \|x+y\|^2 + \|x-y\|^2 \quad \forall x, y \in V$$

then $\exists \langle \cdot, \cdot \rangle$ on V such that

$$\|x\| = (\langle x, x \rangle)^{\frac{1}{2}} \quad \forall x \in V.$$



Complete Inner Product spaces are called
Hilbert Spaces



Banach spaces

Example 1: \mathbb{R}^n with Euclidean inner product

$$\langle x, y \rangle = x^T y$$

is a Hilbert space.

Ex. 2. \mathbb{R}^n with any inner product is a Hilbert space.

Ex. 3. $\mathbb{R}^{m \times n}$ with any inner product is a Hilbert space.

Ex. 4. Tensor space $\mathbb{R}^{m \times n \times l}$ with any inner product is a Hilbert space.

Ex. 5. $l_2 = \left\{ \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \mid \sum_{i=1}^{+\infty} a_i^2 < +\infty \right\}$ with inner prod

$$\langle a, b \rangle = \sum_{i=1}^{+\infty} a_i b_i$$

is a Hilbert space.

Ex. 6. $C[a, b] = \{f \mid f: [a, b] \rightarrow \mathbb{R} \text{ is continuous}\}$

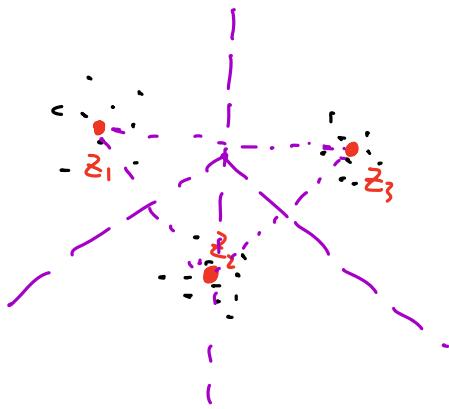
with inner product $\langle f, g \rangle = \int_a^b f(t)g(t) dt$

is NOT a Hilbert space.

The completion of $C[a, b]$ with $\langle \cdot, \cdot \rangle$ and its induced norm is $L^2(a, b)$

§ 3.3 Case Study: Kernel K-means / Kernel trick

— K-means works well for clusters whose boundaries are linear

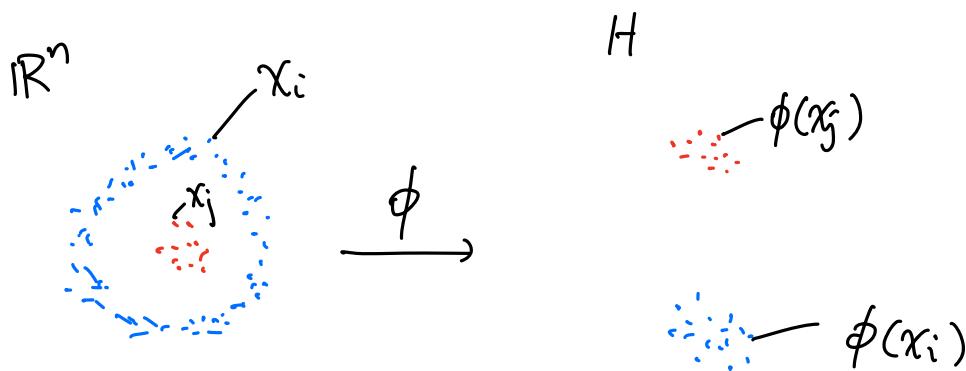


— K-means will not work for



— To cluster the above datasets

- ① Transform the data points to some transformed domain (called the feature space)
- ② Apply K-means in the feature space.



$$\phi: \mathbb{R}^n \rightarrow H$$

An explicit example of ϕ :

$$\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^3$$

$$\phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{cases} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix} & \text{if } x \text{ is red} \\ \begin{pmatrix} x_1 \\ x_2 \\ 0 \end{pmatrix} & \text{if } x \text{ is blue} \end{cases}$$

Given data points $x_1, x_2, \dots, x_N \in \mathbb{R}^n$,

we want to group them into K groups.

Let $\phi: \mathbb{R}^n \rightarrow H$
 \uparrow
feature transformation
feature space
some Hilbert space.

The key of success is: How to choose ϕ and H ?

Kernel K-means: We define ϕ and H implicitly.

Let's first check K-means in H .

- We apply K-means for $\phi(x_1), \phi(x_2), \dots, \phi(x_N)$

Step 0: Initialize z_1, z_2, \dots, z_K in H

$$\rightarrow \text{Step 1: } c_i = \arg \min_{j \in \{1, 2, \dots, K\}} \left\{ \|\phi(x_i) - z_j\|_H^2 \right\}, i=1, \dots, N \quad (1)$$

(i.e., we assign x_i to the nearest representative vector)

$$G_j = \{i \mid c_i = j\}, j=1, \dots, K.$$

$$\text{Step 2: } z_j = \frac{1}{|G_j|} \sum_{i \in G_j} \phi(x_i), \quad j=1, 2, \dots, K.$$

(z_j is the mean of vectors in Group j)

Repeat

Eliminate z_1, z_2, \dots, z_K by plug them into (1)

Step 0: Initialize G_1, \dots, G_K

$$\rightarrow \text{Step 1: } c_i = \arg \min_{j \in \{1, \dots, K\}} \left\{ \left\| \phi(x_i) - \frac{1}{|G_j|} \sum_{l \in G_j} \phi(x_l) \right\|_H^2 \right\}, i=1, \dots, N \quad (2)$$

$$G_j = \{i \mid c_i = j\}, j=1, \dots, K$$

Repeat

We re-arrange the function in (2):

$$\left\| \phi(x_i) - \frac{1}{|G_j|} \sum_{l \in G_j} \phi(x_l) \right\|_H^2$$

$$= \left\langle \phi(x_i) - \frac{1}{|G_j|} \sum_{l \in G_j} \phi(x_l), \phi(x_i) - \frac{1}{|G_j|} \sum_{l \in G_j} \phi(x_l) \right\rangle_H$$

$$= \langle \phi(x_i), \phi(x_i) \rangle_H + \frac{1}{|G_j|^2} \left\langle \sum_{l \in G_j} \phi(x_l), \sum_{l \in G_j} \phi(x_l) \right\rangle_H$$

$$- \frac{2}{|G_j|} \langle \phi(x_i), \sum_{l \in G_j} \phi(x_l) \rangle_H$$

$$= \langle \phi(x_i), \phi(x_i) \rangle_H + \frac{1}{|G_j|^2} \sum_{l_1 \in G_j} \sum_{l_2 \in G_j} \langle \phi(x_{l_1}), \phi(x_{l_2}) \rangle_H - \frac{2}{|G_j|} \sum_{l \in G_j} \langle \phi(x_i), \phi(x_l) \rangle_H$$

All terms involving ϕ, H are in the form

$$\langle \phi(x), \phi(y) \rangle_H : (\mathbb{R}^n, \mathbb{R}^n) \rightarrow \mathbb{R}$$

Then, instead of define ϕ, H explicitly, we just define

$$K : (\mathbb{R}^n, \mathbb{R}^n) \rightarrow \mathbb{R}$$

Such that $K(x, y) = \langle \phi(x), \phi(y) \rangle_H$ for some $\phi : \mathbb{R}^n \rightarrow H$
(Kernel function) $\langle \cdot, \cdot \rangle_H$ on H .

The resulting algorithm is:

Choose a kernel function $K : (\mathbb{R}^n, \mathbb{R}^n) \rightarrow \mathbb{R}$

Step 0: Initialize G_1, \dots, G_K

→ Step 1:

$$c_i = \operatorname{argmin}_{j \in \{1, \dots, K\}} \left\{ K(x_i, x_i) + \frac{1}{|G_j|^2} \sum_{l_1 \in G_j} \sum_{l_2 \in G_j} K(x_{l_1}, x_{l_2}) - \frac{2}{|G_j|} \sum_{l \in G_j} K(x_i, x_l) \right\}, \quad i = 1, \dots, N$$

$$G_j = \{i \mid c_i = j\}, \quad j = 1, \dots, K$$

Repeat

Kernel K-means

How to choose the kernel K ?

- Necessary conditions for K :

$$\textcircled{1} \quad K(x, y) = \langle \phi(x), \phi(y) \rangle_H = \langle \phi(y), \phi(x) \rangle_H = K(y, x)$$

$$\boxed{K(x, y) = K(y, x) \quad \forall x, y \in \mathbb{R}^n} \quad (K \text{ is symmetric})$$

② Let y_1, y_2, \dots, y_m be m vectors on \mathbb{R}^n

Then, for any $c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} \in \mathbb{R}^m$

$$0 \leq \left\langle \sum_{i=1}^m c_i \phi(y_i), \sum_{i=1}^m c_i \phi(y_i) \right\rangle_H = \sum_{i=1}^m \sum_{j=1}^m c_i c_j \langle \phi(y_i), \phi(y_j) \rangle_H$$

$$= \sum_{i=1}^m \sum_{j=1}^m c_i c_j K(y_i, y_j) \quad \forall c \in \mathbb{R}^m$$

Define a Kernel matrix

$$K = [K(y_i, y_j)]_{i=1, j=1}^m \in \mathbb{R}^{m \times m}$$

$$= \sum_{i=1}^m \sum_{j=1}^m K_{ij} c_i c_j = c^T K c$$

(i.e., $\forall c \in \mathbb{R}^m, c^T K c \geq 0$)

Also, ① $\Rightarrow K_{ij} = K_{ji}$, i.e., $K^T = K \quad \forall i, j$

\Rightarrow the kernel matrix K is Symmetric positive semi-definite (SPSD)

① ② together: A necessary condition for a good kernel is:

For any m , for any $y_1, \dots, y_m \in \mathbb{R}^n$, the kernel matrix

$$K = [K(y_i, y_j)]_{i=1, j=1}^m \in \mathbb{R}^{m \times m}$$

is SPSD

The kernel function is SPSD

- Sufficient condition

Mercer's theorem tells us: If a function $K: (\mathbb{R}^n, \mathbb{R}^n) \rightarrow \mathbb{R}$ is SPSD and continuous, then there exists a Hilbert

space H with inner prod $\langle \cdot, \cdot \rangle_H$ and a transformation $\phi: \mathbb{R}^n \rightarrow H$ such that

$$K(x, y) = \langle \phi(x), \phi(y) \rangle_H \quad \forall x, y \in \mathbb{R}^n$$

Remark: • $K(\cdot, \cdot): (\mathbb{R}^n, \mathbb{R}^n) \rightarrow \mathbb{R}$ with K SPSD and continuous

The we have:

$$\left\{ \begin{array}{l} K(x, y) = k(y, x) \\ \forall y_1, \dots, y_m \in \mathbb{R}^n \end{array} \right.$$

$K := [k(x_i, x_j)]_{i,j=1}^m$ is SPSD
non-linear if one vector is fixed

• $\langle \cdot, \cdot \rangle$ in $\mathbb{R}^n: (\mathbb{R}^n, \mathbb{R}^n) \rightarrow \mathbb{R}$

$$\left\{ \begin{array}{l} \langle x, y \rangle = \langle y, x \rangle \\ \forall y_1, \dots, y_m \in \mathbb{R}^n \end{array} \right.$$

$[\langle y_i, y_j \rangle]_{i,j=1}^m$ is SPSD

② linear if one vector is fixed.

Kernel function can be viewed as a "nonlinear" inner product.

Two views of k -means \longrightarrow Kernel k -means

1. Linear inner product

$$\langle \cdot, \cdot \rangle$$

Kernel k -means

non-linear inner product

$$K(\cdot, \cdot)$$

2. No feature transformation

k -means in \mathbb{R}^n

$x_i \mapsto \phi(x_i)$ in H

k -means in H

Some popular kernel functions

① $K(x, y) = \langle x, y \rangle_A$

$$\phi(x) = A^{\frac{1}{2}}x$$

$$H = \mathbb{R}^n$$

② $K(x, y) = (x^T y + 1)^\alpha$, where α is a positive integer.

(polynomial kernel)

we can find ϕ and H explicitly.

e.g. in \mathbb{R}^2 and $\alpha=2$

$$\begin{aligned}
 K(x, y) &= (x^T y + 1)^2 = (x_1 y_1 + x_2 y_2 + 1)^2 \\
 &= x_1^2 y_1^2 + x_2^2 y_2^2 + 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2 \\
 &= \left\langle \begin{pmatrix} x_1^2 \\ x_2^2 \\ 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ \sqrt{2}x_1 x_2 \end{pmatrix}, \begin{pmatrix} y_1^2 \\ y_2^2 \\ 1 \\ \sqrt{2}y_1 \\ \sqrt{2}y_2 \\ \sqrt{2}y_1 y_2 \end{pmatrix} \right\rangle \equiv \langle \phi(x), \phi(y) \rangle_{\mathbb{R}^6}
 \end{aligned}$$

So, $\phi(x) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ \sqrt{2}x_1 x_2 \end{pmatrix}$ and $H = \mathbb{R}^6$

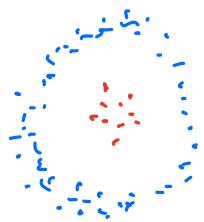
③ $K(x, y) = \mathcal{T}(x^T y)$, where $\mathcal{T}: \mathbb{R} \rightarrow \mathbb{R}$ is some function.

The kernel is used in 1-layer neural network.

④ $K(x, y) = e^{-\frac{\|x-y\|_2^2}{\sigma^2}}$, where $\sigma > 0$ is a parameter
(Gaussian Kernel)

The corresponding H is a Reproducing Kernel Hilbert Space.
(RKHS)

Why Kernel k-means work?



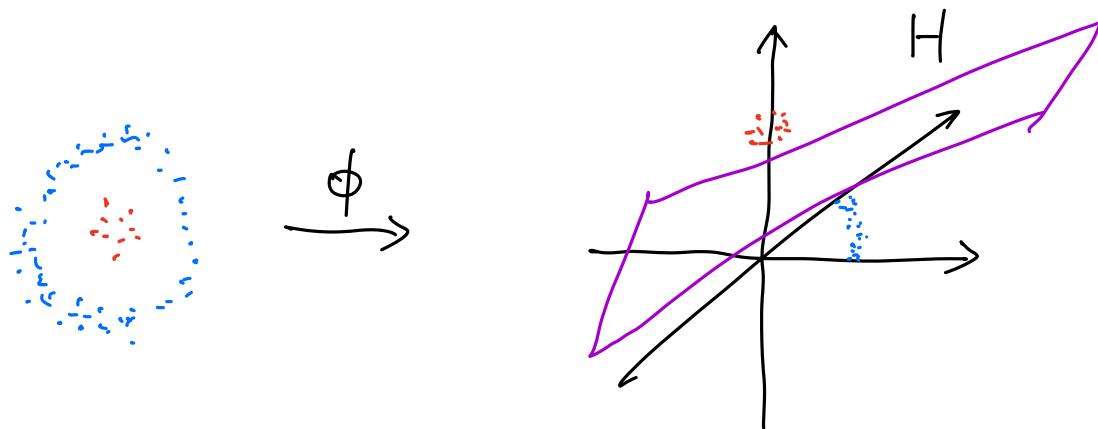
with $K(x, y) = e^{-\frac{\|x-y\|_2^2}{\sigma^2}}$

- $K(x_i, x_i) = e^{-\frac{\|x_i-x_i\|_2^2}{\sigma^2}} = e^0 = 1 \quad \forall i$
 - Since $\|\phi(x_i)\|_H^2 = \langle \phi(x_i), \phi(x_i) \rangle_H = K(x_i, x_i) = 1 \quad \forall i$
i.e., all $\phi(x_i)$, $i=1, \dots, N$ are on the unit sphere on H .
- $K(x_i, x_j) \begin{cases} \approx 0 & \text{if } \|x_i - x_j\|_2 \text{ is large} \\ \approx 1 & \text{if } \|x_i - x_j\|_2 \text{ is small.} \end{cases}$

— Since $\langle \phi(x_i), \phi(x_j) \rangle_H = K(x_i, x_j)$

$\phi(x_i) \perp \phi(x_j)$ if $\|x_i - x_j\|_2$ is large

$\phi(x_i) \approx \phi(x_j)$ if $\|x_i - x_j\|_2$ is small.



3.4 Case Study : Metric Learning

Given a set of vectors $x_1, x_2, \dots, x_N \in \mathbb{R}^n$

and certain pairs of them are similar / dissimilar

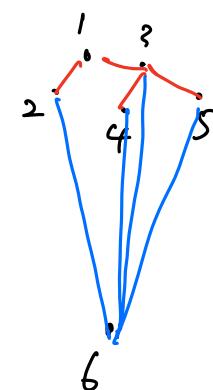
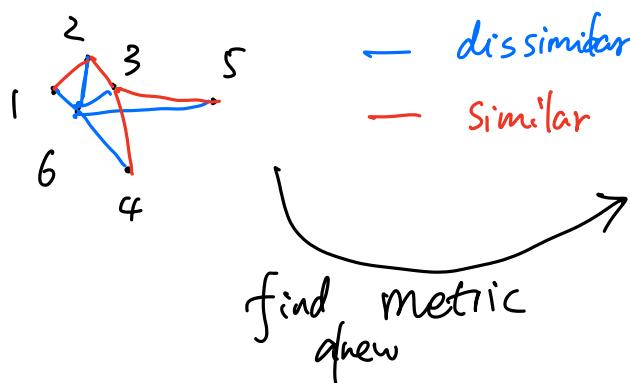
$S : (x_i, x_j) \in S$ if x_i and x_j are similar

$D : (x_i, x_j) \in D$ if x_i, x_j are dis-similar

We want to find a metric such that

"Similar pairs are close,

Dis-Similar pairs are far away from each other.



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

- p-norm: $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}} \quad (p \geq 1)$

- This set of norm functions is too small
(only 1. parameter p to tune)

- norms induced by inner products

Given an SPD matrix $A \in \mathbb{R}^{n \times n}$, $\langle x, y \rangle_A = x^T A y \quad \forall x, y \in \mathbb{R}^n$
and $\|x\|_A = (x^T A x)^{\frac{1}{2}} \quad \forall x \in \mathbb{R}^n$

- This set of norms is large enough
(there are $\frac{n(n+1)}{2}$ parameters for SPD A)

Then find a metric \iff find an SPD matrix $A \in \mathbb{R}^{n \times n}$
(The distance of x, y is
 $\|x-y\|_A = (x-y)^T A (x-y)$)

However, the set of all SPD matrices is NOT closed
(i.e., a sequence of SPD matrices can converge
to a non-SPD matrix)

- We work on its closure

$$\overline{\{ \text{All SPD matrices} \}} = \{ \text{All SPSD matrices} \}$$
$$\{ A \mid \begin{array}{l} A = A^T \\ x^T A x > 0 \quad \forall x \neq 0 \end{array} \} \quad \begin{matrix} \downarrow \\ \text{symmetric positive} \\ \text{semi-definite} \end{matrix}$$
$$\{ A \mid \begin{array}{l} A = A^T \\ x^T A x \geq 0 \quad \forall x \end{array} \}$$

- Instead of SPD matrix, we find an SPSD matrix

$$A \in \mathbb{R}^{n \times n} \quad \text{and} \quad \text{define} \quad \|x-y\|_A = ((x-y)^T A (x-y))^{\frac{1}{2}}$$

- $\|x\|_A$ is not a norm, because
 $\|x\|_A = 0 \Leftrightarrow x^T A x = 0 \Leftrightarrow x = 0$
- $\|x\|_A$ is still good enough, because
 - ① $\|x\|_A \geq 0 \quad \forall x \in \mathbb{R}^n. \quad \|x\|_A = 0 \Leftrightarrow x = 0$
 - ② $\|\alpha x\|_A = |\alpha| \|x\|_A \quad \forall \alpha \in \mathbb{R}, x \in \mathbb{R}^n$
 - ③ $\|x+y\|_A \leq \|x\|_A + \|y\|_A \quad \forall x, y \in \mathbb{R}^n$
- $\|\cdot\|_A$ is called a pseudo-norm

which SPSD A?

- For $(x_i, x_j) \in S$, their distance small, i.e.

$$\sum_{(x_i, x_j) \in S} \|x_i - x_j\|_A^2 \text{ is small.}$$

- For $(x_i, x_j) \in D$, their dist is large, i.e.

$$\|x_i - x_j\|_A^2 \text{ large for } (x_i, x_j) \in D$$

Then we solve

$$\left\{ \begin{array}{l} \min_{A \in \mathbb{R}^{n \times n}} \sum_{(x_i, x_j) \in S} \|x_i - x_j\|_A^2 \\ \text{s.t.} \quad A \text{ is SPSD} \\ \quad \|x_i - x_j\|_A^2 \geq 1 \quad \forall (x_i, x_j) \in D \end{array} \right.$$

This optimization is a Semi-Definite Programming (SDP).
 There are many softwares available to solve SDPs.

Ch. 4. Linear functions and Differentiation

§ 4.1. Linear functions.

Let $f: V \rightarrow \mathbb{R}$ be a function on a vector space V .

Then f is linear if:

$$f(\alpha x + \beta y) = \alpha \cdot f(x) + \beta \cdot f(y) \quad \forall x, y \in V \\ \alpha, \beta \in \mathbb{R}.$$

Ex 1: The mean of entries of vectors on \mathbb{R}^n

$$\forall x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n \quad f(x) = \frac{x_1 + x_2 + \dots + x_n}{n} \text{ is linear}$$

Ex. 2. The max entry of a vector on \mathbb{R}^n

$$\forall x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n, \quad f(x) = \max\{x_i \mid i=1, \dots, n\} \text{ is not linear}$$

For example: on \mathbb{R}^2 , $f(0) = 1$, $f(1) = 1$

$$\text{But } f((0) + (1)) = f(1) = 1$$

if

$$f(0) + f(1) = 1 + 1 = 2$$

Ex 3: $f: \mathbb{R}^n \rightarrow \mathbb{R}$ defined by $f(x) = a^T x$, where $a \in \mathbb{R}^n$ is fixed.,
is linear.

Ex. 4: $F: C[-1, 1] \rightarrow \mathbb{R}$ $\quad C[-1, 1] = \{f \mid f: [-1, 1] \rightarrow \mathbb{R} \text{ is continuous}\}$

$$F(f) = f(0) \quad \forall f \in C[-1, 1]$$

is linear,

$$\text{because } F(\alpha f + \beta g) = (\alpha f + \beta g)(0) = \alpha f(0) + \beta g(0) \\ = \alpha F(f) + \beta F(g)$$

Ex 5: $F: C[a,b] \rightarrow \mathbb{R}$ defined by

$F(f) = \int_a^b f(t) dt$ is linear, because

$$\begin{aligned} F(\alpha f + \beta g) &= \int_a^b (\alpha f + \beta g)(t) dt = \alpha \int_a^b f(t) dt + \beta \int_a^b g(t) dt \\ &= \alpha F(f) + \beta F(g) \end{aligned}$$

Ex. 6. Let $(V, \langle \cdot, \cdot \rangle)$ be an inner prod space.

Define $f(x) = \langle a, x \rangle$, where $a \in V$ is fixed,
then f is linear, because

$$\begin{aligned} f(\alpha x + \beta y) &= \langle a, \alpha x + \beta y \rangle = \alpha \langle a, x \rangle + \beta \langle a, y \rangle \\ &= \alpha f(x) + \beta f(y). \end{aligned}$$

Ex. 7: A norm function is NOT linear, because

— $\|-x\| = \|x\|$

— If $\|\cdot\|$ is linear, then

$$\begin{aligned} \|-x\| &= \|(-1) \cdot x + 0 \cdot x\| = (-1) \cdot \|x\| + 0 \cdot \|x\| \\ &= -\|x\|. \text{ Contradiction.} \end{aligned}$$

Properties of linear functions.

- Homogeneity: $f(\alpha x) = \alpha \cdot f(x) \quad \forall \alpha \in \mathbb{R}, x \in V$

(Because f is linear, $\Rightarrow f(\alpha x + 0 \cdot y) = \alpha f(x) + 0 \cdot f(y) = \alpha \cdot f(x)$)

Choose $\alpha = 0$. $f(0) = 0$

- Additivity: $f(x+y) = f(x) + f(y) \quad \forall x, y \in V$.

Linearity \iff Homogeneity + Additivity

- $f(\alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_k x_k) \quad \forall \alpha_1, \dots, \alpha_k \in \mathbb{R}$
 $= \alpha_1 f(x_1) + f(\alpha_2 x_2 + \dots + \alpha_k x_k) \quad x_1, \dots, x_k \in V$
 \vdots
 $= \alpha_1 f(x_1) + \alpha_2 f(x_2) + \dots + \alpha_k f(x_k)$

Linear functions on Hilbert Spaces.

- Let H be a Hilbert space with inner prod $\langle \cdot, \cdot \rangle$ and the induced norm $\|\cdot\|$
- $\forall a \in H, f(x) = \langle a, x \rangle$ is a linear function on H .

- Its reverse: \forall linear function $f: H \rightarrow \mathbb{R}$,
- $\exists a \in H$ s.t. $f(x) = \langle a, x \rangle \quad \forall x \in H$
?

The answer is affirmative.

- For simplicity, consider $H = \mathbb{R}^n$

Thm: For any linear function $f: \mathbb{R}^n \rightarrow \mathbb{R}$

\exists a unique $a \in \mathbb{R}^n$ s.t. $f(x) = \langle a, x \rangle \quad \forall x \in \mathbb{R}^n$

proof. $\forall x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$, we have

$$x = x_1 e_1 + x_2 e_2 + \dots + x_n e_n$$

Since $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is linear,

$$f(x) = f(x_1 e_1 + x_2 e_2 + \dots + x_n e_n)$$

$$= x_1 \cdot f(e_1) + x_2 \cdot f(e_2) + \dots + x_n \cdot f(e_n)$$

$$= \left\langle \begin{pmatrix} f(e_1) \\ f(e_2) \\ \vdots \\ f(e_n) \end{pmatrix}, \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \right\rangle \stackrel{\text{(def)}}{=} \langle a, x \rangle$$

$$e_i = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{pmatrix} \leftarrow \begin{matrix} & \\ & \text{i-th} \\ & \text{entry} \end{matrix}$$

Now we show $a \in \mathbb{R}^n$ is unique.

Suppose we have $a, b \in \mathbb{R}^n$ s.t. $f(x) = \langle a, x \rangle = \langle b, x \rangle \quad \forall x \in \mathbb{R}^n$,

$$f(e_i) = \langle a, e_i \rangle = \langle b, e_i \rangle \quad \begin{matrix} \parallel \\ a_i \end{matrix} \quad \begin{matrix} \parallel \\ b_i \end{matrix} \quad \forall i = 1, \dots, n$$

$$\Rightarrow a_i = b_i \quad i = 1, \dots, n \Rightarrow a = b \quad \blacksquare$$

It can be extended to general Hilbert spaces.

Riesz Representation Theorem:

Let H be a Hilbert space. Let $f: H \rightarrow \mathbb{R}$. Then

f is linear and bounded $\iff f(x) = \langle a, x \rangle$ for some unique $a \in H$.

Ex. 1: The mean function on \mathbb{R}^n is linear,

$$f(x) = \frac{x_1 + x_2 + \dots + x_n}{n} = \left\langle \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \right\rangle$$

Ex. 2: Let H be a Hilbert space.

$\|\cdot\|$ is non-linear

So, $\nexists a \in H$ s.t. $\|x\| = \langle a, x \rangle \quad \forall x \in H$.

Ex. 3. $\mathbb{R}^{n \times n}$ with inner prod $\langle A, B \rangle = \sum_{i,j} a_{ij} b_{ij} = \text{trace}(A^T B)$
 $\wedge A, B \in \mathbb{R}^{n \times n}$

The trace function

$\text{trace}(A) = \sum_i a_{ii}$ is linear

So, $\text{trace}(A) = \text{trace}(I^T A) = \langle I, A \rangle$

Remark: • f is bounded $\iff \exists C > 0$ s.t

$$|f(x)| \leq C \cdot \|x\| \quad \forall x \in H.$$

• If H is a finite dimensional Hilbert space, then

linear \iff linear and bounded.

• In infinite dim Hilbert spaces

\exists linear but unbounded functions.

— Explicit example is impossible.

— We can still construct some linear and unbounded

function on incomplete inner prod space.

$$\text{Ex : } C'[-1,1] = \{f \mid f, f' \in C[-1,1]\} \quad (\text{This space is incomplete})$$

with $\langle f, g \rangle = \int_{-1}^1 f(t)g(t) dt$

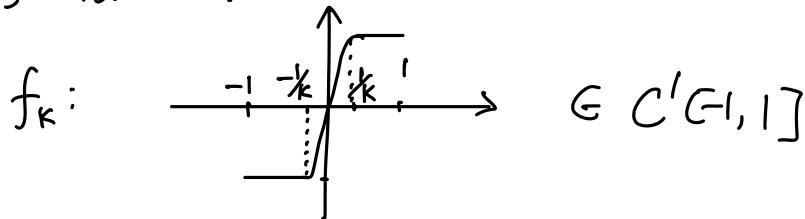
Define $F : C'[-1,1] \rightarrow \mathbb{R}$ by

$$F(f) = f'(0)$$

- F is linear because

$$\begin{aligned} F(\alpha f + \beta g) &= (\alpha f + \beta g)'(0) = \alpha f'(0) + \beta g'(0) \\ &= \alpha F(f) + \beta F(g) \end{aligned}$$

- F is unbounded.



$$F(f_k) = f'_k(0) = k \rightarrow +\infty \text{ as } k \rightarrow +\infty$$

$$\text{but } \|f_k\| = \left(\int_{-1}^1 f_k^2(t) dt \right)^{\frac{1}{2}} \leq \left(\int_{-1}^1 1 dt \right)^{\frac{1}{2}} = \sqrt{2}$$

$$\text{So, } \frac{F(f_k)}{\|f_k\|} \geq \frac{k}{\sqrt{2}} \rightarrow +\infty \text{ as } k \rightarrow +\infty$$

Ex. 4: $L^2(-1,1) = \{f \mid \left(\int_{-1}^1 f^2(t) dt \right)^{\frac{1}{2}} < +\infty\}$ is a Hilbert space.

$$\langle f, g \rangle = \int_{-1}^1 f(t)g(t) dt$$

Consider $G : L^2(-1,1) \rightarrow \mathbb{R}$ defined by

$$G(f) = \int_{-1}^1 f(t) dt$$

- G is well defined. i.e., $\forall f \in L^2(-1,1)$, $G(f)$ finite

$$\begin{aligned} |G(f)| &= \left| \int_{-1}^1 f(t) dt \right| = \left| \int_{-1}^1 f(t) \cdot 1 dt \right| = |\langle f, 1 \rangle| \\ &\stackrel{CS}{\leq} \|f\| \cdot \|1\| = \sqrt{2} \cdot \|f\| < +\infty \end{aligned}$$

- G is linear: $G(\alpha f + \beta g) = \int_{-1}^1 (\alpha f(t) + \beta g(t)) dt$

$$= \alpha G(f) + \beta G(g)$$

- G is bounded:

$$|G(f)| \leq \sqrt{2} \cdot \|f\|$$

Riesz $\Rightarrow \exists g \in L^2(-1,1)$ s.t. $G(f) = \langle g, f \rangle$

Indeed, $g : \boxed{g(t) = 1 \quad \forall t \in [-1,1]}$,

$$\begin{aligned} \text{because } G(f) &= \int_{-1}^1 f(t) dt = \int_{-1}^1 (f(t) \cdot 1) dt \\ &= \langle 1, f \rangle \end{aligned}$$

Hyperplane

Let $H, \langle \cdot, \cdot \rangle$ be a Hilbert space

Let $a \in H$

• Consider

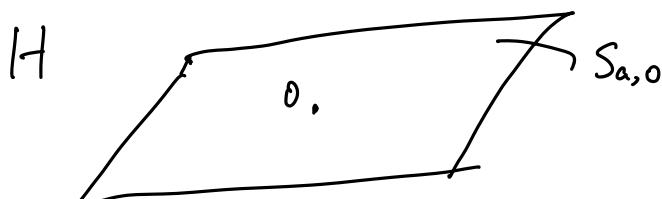
$$S_{a,0} = \{x \in H \mid \langle a, x \rangle = 0\} \subseteq H$$

- $\forall \alpha, \beta \in \mathbb{R}, x, y \in S_{a,0}$

$$\langle a, \alpha x + \beta y \rangle = \alpha \langle a, x \rangle + \beta \langle a, y \rangle = \alpha \cdot 0 + \beta \cdot 0 = 0$$

$$\Rightarrow \alpha x + \beta y \in S_{a,0}$$

$\Rightarrow \boxed{S_{a,0} \text{ is a subspace of } H}$



$$S_{a,b} = \{x \in H \mid \langle a, x \rangle = b\} \subseteq H$$

for some $a \in H$
 $b \in \mathbb{R}$

Let $x_0 \in S_{a,b}$ ($S_0, \langle a, x_0 \rangle = b$)

① $\forall x \in S_{a,b}$

$$\langle a, x - x_0 \rangle = \langle a, x \rangle - \langle a, x_0 \rangle = b - b = 0$$

$$\Rightarrow x - x_0 \in S_{a,0}$$

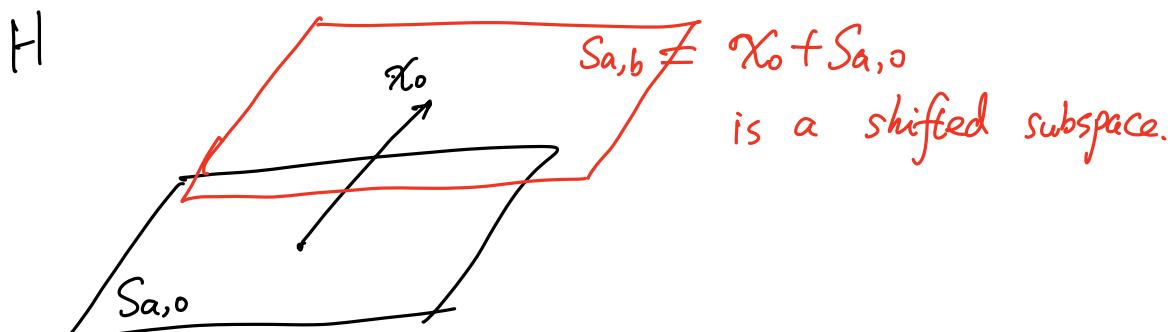
$$\Rightarrow x \in x_0 + S_{a,0} \Rightarrow \boxed{S_{a,b} \subseteq x_0 + S_{a,0}}$$

② $\forall x \in S_{a,0}$

$$\langle a, x_0 + x \rangle = \langle a, x_0 \rangle + \langle a, x \rangle = b + 0 = b$$

$$\Rightarrow x_0 + x \in S_{a,b} \Rightarrow x_0 + S_{a,0} \subseteq S_{a,b}$$

$$\Rightarrow S_{a,b} = x_0 + S_{a,0}$$



$S_{a,b}$ is a plane on H

Also, its co-dimension is 1

(It is defined in terms of
1 equation only)

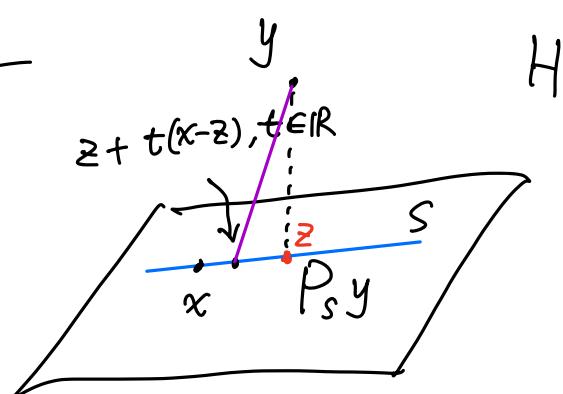
We call $S_{a,b}$
Hyperplane.

Projection onto Hyperplanes

- $S = \{x \in H \mid \langle a, x \rangle = b\}$ in $H, \langle \cdot, \cdot \rangle$

- Given $y \in H$,
Find a vector on S , denoted by $P_S y$, which is
the closest to y ,

$$P_S y = \arg \min_{x \in S} \|y - x\|$$



- Let's find the closed form of $P_S y$, in terms of a, b, y .

Thm: z is a solution of $\min_{x \in S} \|y - x\|$

$$\begin{cases} z \in S \\ \Leftrightarrow \langle y - z, x - z \rangle = 0 \quad \forall x \in S \end{cases}$$

proof. " \Downarrow ": If z is a solution of $\min_{x \in S} \|y-x\|$, then:

- Obviously, $z \in S$

- $\forall x \in S$, consider the line passing thru x and z

Any ^{vector} ~~point~~ on the line is $\boxed{z+t(x-z), t \in \mathbb{R}}$

$$\text{Since } \langle a, z+t(x-z) \rangle = \langle a, z \rangle + t \langle a, x-z \rangle$$

$$= \langle a, z \rangle + t \langle a, x \rangle - t \langle a, z \rangle$$

$$= b + t \cdot b - t \cdot b = b$$

$$\Rightarrow z+t(x-z) \in S \quad \forall t \in \mathbb{R}.$$

Since z is the $\min_{x \in S} \|y-x\|$,

$$\cancel{\|z-y\|^2} \leq \|(z+t(x-z))-y\|^2 = \|(z-y)+t(x-z)\|^2$$

$$= \langle (z-y)+t(x-z), (z-y)+t(x-z) \rangle$$

$$= \cancel{\|z-y\|^2} + t^2 \|x-z\|^2 + 2t \langle z-y, x-z \rangle$$

$$\Rightarrow 2t \langle z-y, x-z \rangle \geq -t^2 \|x-z\|^2$$

- If $t > 0$, then

$$\langle z-y, x-z \rangle \geq -\frac{t}{2} \|x-z\|^2$$

Let $t \rightarrow 0_+$

$$\boxed{\langle z-y, x-z \rangle \geq 0}$$

- If $t < 0$, then

$$\langle z-y, x-z \rangle \leq -\frac{t}{2} \|x-z\|^2$$

Let $t \rightarrow 0_-$

$$\boxed{\langle z-y, x-z \rangle \leq 0}$$

Together, $\langle z-y, x-z \rangle = 0$

" \Updownarrow ": If $z \in S$, and $\langle z-y, x-z \rangle = 0 \quad \forall x \in S$,
then: $\forall x \in S$,

$$\|x-y\|^2 = \|(x-z)+(z-y)\|^2$$

$$\begin{aligned}
 &= \|x-z\|^2 + \|z-y\|^2 + 2\langle x-z, z-y \rangle \\
 &= \|x-z\|^2 + \|z-y\|^2 \\
 &\geq \|z-y\|^2 \\
 \Rightarrow z = \arg \min_{x \in S} \|x-y\|. &\quad \blacksquare
 \end{aligned}$$

Thm: The solution of $\min_{x \in S} \|y-x\|$ exists and is unique,

which is given by $y - \left(\frac{\langle a, y \rangle - b}{\|a\|^2} \right) a$

proof. Let $z = y - \left(\frac{\langle a, y \rangle - b}{\|a\|^2} \right) a$. Then

$$\begin{aligned}
 \textcircled{1} \quad \langle a, z \rangle &= \langle a, y \rangle - \left\langle a, \frac{\langle a, y \rangle - b}{\|a\|^2} a \right\rangle \\
 &= \langle a, y \rangle - \frac{\langle a, y \rangle - b}{\|a\|^2} \cancel{\langle a, a \rangle} \\
 &= b \\
 \Rightarrow z \in S
 \end{aligned}$$

$$\begin{aligned}
 \textcircled{2} \quad \forall x \in S, \quad \langle z-y, x-z \rangle &= \left\langle -\frac{\langle a, y \rangle - b}{\|a\|^2} a, x-z \right\rangle \\
 &= -\frac{\langle a, y \rangle - b}{\|a\|^2} \langle a, x-z \rangle \\
 &= -\frac{\cancel{\langle a, y \rangle} \cancel{-b}}{\|a\|^2} (\langle a, x \rangle - \langle a, z \rangle) \\
 &= -\frac{\cancel{\langle a, y \rangle} \cancel{-b}}{\|a\|^2} (b-b) = 0
 \end{aligned}$$

$\Rightarrow z$ is a solution of $\min_{x \in S} \|x-y\|$

For the uniqueness, Suppose we have 2 solutions z_1, z_2 .

Then, $z_1 \in S, z_2 \in S$

and

z_1 is a solution $\Rightarrow z_1 \in S, \langle z_1-y, z_2-z_1 \rangle = 0$

z_2 is a solution $\Rightarrow z_2 \in S, \langle z_2-y, z_1-z_2 \rangle = 0$

Add the two identities:

$$\begin{aligned}
 \langle z_1-y, z_2-z_1 \rangle + \langle z_2-y, z_1-z_2 \rangle &= 0 \\
 \Leftrightarrow \langle z_1-y+y-z_2, z_2-z_1 \rangle &= 0
 \end{aligned}$$

$$\begin{aligned} &\Leftrightarrow \langle z_1 - z_2, z_2 - z_1 \rangle = 0 \\ &\Leftrightarrow -\|z_1 - z_2\|^2 = 0 \Rightarrow z_1 = z_2 \quad \text{☒} \end{aligned}$$

In summary,

Let H be a Hilbert space, and

$$S = \{x \in H \mid \langle a, x \rangle = b\}, \text{ where } a \in H, b \in \mathbb{R}$$

Let $y \in H$. Then the projection of y onto S is unique and

$$P_S y := \arg \min_{x \in S} \|x - y\|$$

and given by

$$P_S y = y - \left(\frac{\langle a, y \rangle - b}{\|a\|^2} \right) a$$

Furthermore,

$$\langle y - P_S y, P_S y - x \rangle = 0 \quad \forall x \in S$$

$$\Leftrightarrow y - P_S y \perp S_{a,0}$$

Affine functions

A linear function plus a constant is called an affine function.

i.e. an affine function f can be written as

$$f(x) = g(x) + b,$$

where $g: H \rightarrow \mathbb{R}$ is linear.

Properties:

① If $f: H \rightarrow \mathbb{R}$ is affine, then

$$f(\alpha x + \beta y) = g(\alpha x + \beta y) + b$$

$$= \alpha g(x) + \beta g(y) + b$$

$$= \alpha g(x) + \beta g(y) + \alpha b + \beta b$$

$$= \alpha(g(x) + b) + \beta(g(y) + b)$$

g is linear

if $\alpha + \beta = 1$

$$= \alpha f(x) + \beta f(y)$$

If $\alpha + \beta = 1$, then $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$

② If H is a Hilbert space. and if $f: H \rightarrow \mathbb{R}$ is bounded then

f is affine $\Leftrightarrow f(x) = \langle a, x \rangle + b$. for some $a \in H$
 $b \in \mathbb{R}$

§ 4.2. Case Studies: Linear regression and linear classifier

§ 4.2.1. Linear regression

- Given a set of data

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N),$$

where $x_i \in \mathbb{R}^n$ is the input vector

$y_i \in \mathbb{R}$ the label

Prediction: Given a new $x \in \mathbb{R}^n$, what is the $y \in \mathbb{R}$?

- We find a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ s.t.

$$f(x_i) \approx y_i, \quad i=1, 2, \dots, N.$$

Then we use $f(x)$ as the predicted y .

- The set of all functions $\mathbb{R}^n \rightarrow \mathbb{R}$ is too large.
 - We don't have enough data to determine f uniquely.
 - There are too many strange function that are not good for prediction
- Then, we find f in a subset Φ of all functions.
- Which Φ ?
 - How to choose Φ is fundamental.
 - Φ can not be too big.
 - Φ can not be too small. — weak approximation power.

- Linear regression

$$\Phi = \{ \text{affine functions } \mathbb{R}^n \rightarrow \mathbb{R} \}$$

- Since \mathbb{R}^n is a finite dim Hilbert space,

$$f \in \Phi \iff f(x) = \langle a, x \rangle + b \quad \text{for some } a \in \mathbb{R}^n, b \in \mathbb{R}$$

- So, we only need to find $a \in \mathbb{R}^n, b \in \mathbb{R}$ s.t

$$\langle a, x_i \rangle + b \approx y_i, \quad i=1, 2, \dots, N$$

- Least Squares: $x_i^T a$

$$\min_{\substack{a \in \mathbb{R}^n \\ b \in \mathbb{R}}} \sum_{i=1}^N ((\cancel{\langle a, x_i \rangle} + b) - y_i)^2 \quad (\text{LS})$$

- Write

$$X = \begin{bmatrix} x_1^T & | \\ x_2^T & | \\ \vdots & \vdots \\ x_N^T & | \end{bmatrix} \in \mathbb{R}^{N \times (n+1)}, \quad \beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}$$

Then

$$X\beta = \begin{bmatrix} x_1^T a + b \\ x_2^T a + b \\ \vdots \\ x_N^T a + b \end{bmatrix}$$

$$X\beta - y = \begin{bmatrix} x_1^T a + b - y_1 \\ x_2^T a + b - y_2 \\ \vdots \\ x_N^T a + b - y_N \end{bmatrix}$$

So,

$$(\text{LS}) \iff \min_{\beta \in \mathbb{R}^{n+1}} \|X\beta - y\|_2^2$$

- We have $n+1$ unknowns

... N linear equations

So $N \geq n+1$ is a necessary condition for a unique solution

- If $N \geq n+1$, then generally (LS) has a unique solution.
- If $N < n+1$, then (LS) cannot have a unique solution.

Example: Image recognition

x_i images $n \sim 10M$

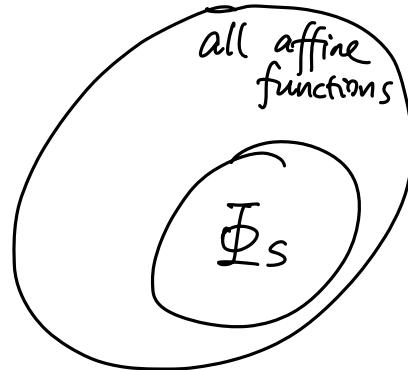
y_i label

N number of available ~~available~~ ^{images} $N \sim 100k$

In this case $N \ll n$

The $\Phi = \{\text{affine functions}\}$ is too big.

- Regularization: We use a subset of affine functions.



$$\Phi_s = \{ f \mid f(x) = \langle a, x \rangle + b \text{ and } \beta = \begin{bmatrix} a \\ b \end{bmatrix} \in S \subset \mathbb{R}^{n+1} \}$$

- Ridge regression.

We choose

$$S = \{ \beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1} \mid \|a\|_2 \leq C \} \text{ for some } C \geq 0$$

Then we solve

$$\min_{\beta \in S} \|X\beta - y\|_2^2 \iff$$

$$\min_{\substack{\|a\|_2 \leq C \\ b \in \mathbb{R}}} \|X(a)_b - y\|_2^2$$

regularization parameter.

by optimization theory

(Ridge regression)

$$\min_{a \in \mathbb{R}^n} \|X(a)_b - y\|_2^2 + \lambda \|a\|_2^2,$$

regularization term

$b \in \mathbb{R}$ data-fitting regularization parameter
 where $\lambda > 0$ is a constant
 depending on C and others.

(larger λ , small β s, looser fitting to the data.)

(smaller λ , larger β s, tighter fitting to the data.)

- LASSO regression

$$S = \left\{ \beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1} \mid \|a\|_1 \leq C \right\} \text{ for some } C > 0.$$

We solve $\min_{\beta \in S} \|X\beta - y\|_2^2$

\Downarrow

$$\min_{\substack{\|a\|_1 \leq C \\ b \in \mathbb{R}}} \|X \begin{pmatrix} a \\ b \end{pmatrix} - y\|_2^2$$

\Updownarrow by Lagrangian theory

$$\min_{\substack{a \in \mathbb{R}^n \\ b \in \mathbb{R}}} \|X \begin{pmatrix} a \\ b \end{pmatrix} - y\|_2^2 + \lambda \|a\|_1$$

(LASSO regression)

data-fitting regularization
regularization parameter

- LASSO regression will give us a sparse vector $a \in \mathbb{R}^n$.

$$x = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix} \quad \begin{aligned} & \langle a, x \rangle && \text{if } a \text{ is sparse} \\ &= \sum_{i=1}^n a_i \xi_i = \sum_{i \in I} a_i \xi_i && I = \{i \mid a_i \neq 0\} \\ & && |I| \ll n. \end{aligned}$$

The regression involves only

$$\xi_i, i \in I$$

Linear model has some limitations.

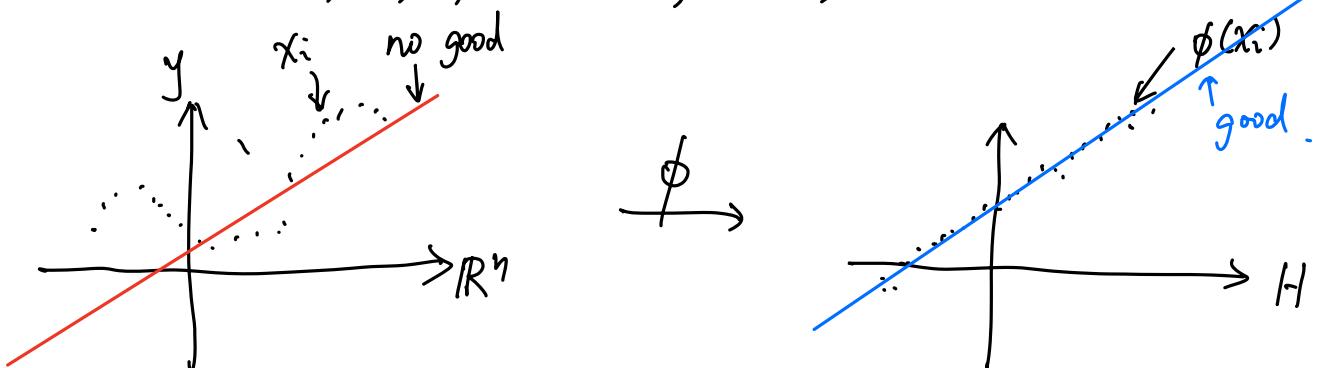
§ 4.2.2 Kernel Ridge regression

To improve the linear regression, we use kernel method.

① Transform: $\phi: \mathbb{R}^n \rightarrow H$ Hilbert space
 features space.

② Apply linear regression to

$(\phi(x_1), y_1), (\phi(x_2), y_2), \dots, (\phi(x_N), y_N)$ in H .



How to find ϕ , H , and f ?

— Linear and bounded functions on H is enough (No need affine)

Because: Let $f: H \rightarrow \mathbb{R}$ be affine and bounded

$$\begin{aligned} \text{Then } f(\phi(x)) &= g(\phi(x)) + b \\ &= \langle \phi(x), a \rangle_H + b \quad \text{for some } a \in H \\ &\quad b \in \mathbb{R} \end{aligned}$$

we can define a new $\tilde{\phi}$ and \tilde{H} by:

$$\tilde{\phi}(x) = \begin{pmatrix} \phi(x) \\ 1 \end{pmatrix} \quad \text{and} \quad \forall \begin{pmatrix} a \\ b \end{pmatrix}, \begin{pmatrix} c \\ d \end{pmatrix} \in \tilde{H}$$

$$\tilde{H} = (H, \mathbb{R}) \quad \langle \begin{pmatrix} a \\ b \end{pmatrix}, \begin{pmatrix} c \\ d \end{pmatrix} \rangle_{\tilde{H}} = \langle a, c \rangle_H + bd$$

So, $\tilde{\phi}: \mathbb{R}^n \rightarrow \tilde{H}$

Then

$$\begin{aligned} f(\phi(x)) &= \langle \phi(x), a \rangle_H + b \\ &\stackrel{\substack{\uparrow \\ \text{affine and bounded function}}}{=} \langle \begin{pmatrix} \phi(x) \\ 1 \end{pmatrix}, \begin{pmatrix} a \\ b \end{pmatrix} \rangle_{\tilde{H}} \\ &= \langle \tilde{\phi}(x), \tilde{a} \rangle_{\tilde{H}}, \text{ where } \tilde{a} \in \tilde{H}. \end{aligned}$$

on H

a linear and bounded
function on \tilde{H}

— The task is:

Find a linear and bounded function $f: H \rightarrow \mathbb{R}$ s.t.

$$f(\phi(x_i)) \approx y_i, \quad i=1, \dots, N$$

↑

$$\langle a, \phi(x_i) \rangle_H, \text{ where } a \in H.$$

— We solve

Find $a \in H$. s.t. $\langle a, \phi(x_i) \rangle_H \approx y_i, \quad i=1, \dots, N.$

Using least squares, we solve

$$\min_{a \in H} \sum_{i=1}^N (\langle a, \phi(x_i) \rangle_H - y_i)^2$$

impossible
to solve it

- H is infinitely dimensional
- We have only N data

— We need regularizations. $\|a\|_H^2 = \langle a, a \rangle_H$ as regularization.

Kernel Ridge Regression

$$\min_{a \in H} \sum_{i=1}^N (\langle a, \phi(x_i) \rangle_H - y_i)^2 + \lambda \|a\|_H^2, \quad (KR)$$

where $\lambda > 0$ is a regularization parameter.

- Still infinite dim problem
- Still need explicit ϕ, H

Representer Thm: The solution of (KR) must be in the form of

$$a = \sum_{i=1}^N c_i \phi(x_i),$$

$$\text{where } C = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N$$

Proof. For any $a \in H$, we first prove that a can be decomposed as

$$a = a_s + \sum_{i=1}^N c_i \phi(x_i),$$

where $C = \begin{bmatrix} C_1 \\ \vdots \\ C_N \end{bmatrix} \in \mathbb{R}^N$ and $a_s \in H$ satisfying $\langle a_s, \phi(x_i) \rangle_H = 0 \quad \forall i=1, \dots, N$.

proof. For simplicity, we prove only the case $N=1$.

$$S = \{v \in H \mid \langle v, \phi(x_1) \rangle_H = 0\}$$

Then S is a hyperplane and a subspace.

So, a can be decomposed as

$$a = P_S a + (a - P_S a)$$

- For $P_S a$: Since $P_S a$ is the projection

$$\left. \begin{array}{l} a - P_S a \perp S \\ \text{and} \\ P_S a \in S \end{array} \right\} \Rightarrow \langle a - P_S a, P_S a \rangle_H = 0$$

- For $a - P_S a$: By the explicit formula of projection,

$$a - P_S a = \left(\frac{\langle \phi(x_1), a \rangle_H}{\|\phi(x_1)\|_H^2} \right) \phi(x_1) \equiv C_1 \phi(x_1)$$

Define $P_S a = a_s$. Then

$$a = a_s + C_1 \phi(x_1),$$

and

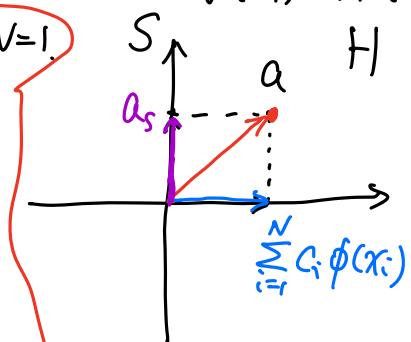
$$\langle a - P_S a, P_S a \rangle_H = \langle C_1 \phi(x_1), a_s \rangle_H = 0$$

$$\Rightarrow \langle \phi(x_1), a_s \rangle_H = 0 \quad \blacksquare$$

Then, the function in (KR) can be rewritten as.

$$\begin{aligned} & \sum_{i=1}^N (\langle a, \phi(x_i) \rangle_H - y_i)^2 + \lambda \|a\|_H^2 \\ &= \sum_{i=1}^N \left(\langle a_s + \sum_{j=1}^N C_j \phi(x_j), \phi(x_i) \rangle_H - y_i \right)^2 + \lambda \|a_s + \sum_{j=1}^N C_j \phi(x_j)\|_H^2 \\ &= \sum_{i=1}^N \left(\cancel{\langle a_s, \phi(x_i) \rangle} + \sum_{j=1}^N C_j \langle \phi(x_j), \phi(x_i) \rangle_H - y_i \right)^2 \\ &\quad + \lambda \left(\|a_s\|_H^2 + 2 \sum_{j=1}^N C_j \cancel{\langle a_s, \phi(x_j) \rangle_H} + \sum_{j=1}^N \sum_{i=1}^N C_i C_j \langle \phi(x_i), \phi(x_j) \rangle_H \right) \\ &\stackrel{i=1, \dots, N}{=} \sum_{i=1}^N \left(\sum_{j=1}^N C_j \langle \phi(x_j), \phi(x_i) \rangle_H - y_i \right)^2 + \lambda \sum_{j=1}^N \sum_{i=1}^N C_i C_j \langle \phi(x_i), \phi(x_j) \rangle_H \end{aligned}$$

Define $K = \left[\langle \phi(x_i), \phi(x_j) \rangle_H \right]_{i,j=1}^N + \lambda \|a_s\|_H^2$



$$\begin{aligned}
&\stackrel{\text{def}}{=} \sum_{i=1}^N \left(\sum_{j=1}^N K_{ji} c_j - y_i \right)^2 + \lambda \sum_{i=1}^N \sum_{j=1}^N c_i c_j K_{ij} + \lambda \|a_s\|_H^2 \\
&= \sum_{i=1}^N \left((K^T c)_i - y_i \right)^2 + \lambda c^T K c + \lambda \|a_s\|_H^2 \\
&= \|K^T c - y\|_2^2 + \lambda c^T K c + \lambda \cdot \|a_s\|_H^2
\end{aligned}$$

function of $c \in \mathbb{R}^N$, denoted by $F(c)$
function of $a_s \in H$ ($\langle a_s, \phi(x_i) \rangle_H = 0$), denoted by $G(a_s)$

Then,

$$(KR) \iff \begin{cases} \min_{\substack{c \in \mathbb{R}^N \\ a_s \in H}} & F(c) + \lambda G(a_s) \\ \text{s.t. } & \langle a_s, \phi(x_i) \rangle_H = 0, \quad i=1, \dots, N. \end{cases}$$

$$\iff \min_{c \in \mathbb{R}^N} F(c) \quad \textcircled{1}$$

and

$$\begin{cases} \min_{a_s \in H} & \lambda G(a_s) \\ \text{s.t. } & \langle a_s, \phi(x_i) \rangle_H = 0, \quad i=1, \dots, N. \end{cases} \quad \textcircled{2}$$

To solve \textcircled{2}: Because $\lambda > 0$,

$$\textcircled{2} \iff \begin{cases} \min_{a_s \in H} & \|a_s\|_H^2 \\ \text{s.t. } & \langle a_s, \phi(x_i) \rangle_H = 0, \quad i=1, \dots, N. \end{cases}$$

$$\iff a_s^* = 0$$

Let a^* be a solution of (KR). Then,

$$\begin{aligned}
a^* &= a_s^* + \sum_{i=1}^N c_i^* \phi(x_i) \\
&= \sum_{i=1}^N c_i^* \phi(x_i),
\end{aligned}$$

where $C^* = \begin{bmatrix} C_1^* \\ \vdots \\ C_N^* \end{bmatrix}$ is the solution of ①

As a by-product of the proof, we can apply the kernel trick
we only need to define the matrix K , not $\phi, H, \langle \cdot, \cdot \rangle_H$

why

$$\left[\langle \phi(x_i), \phi(x_j) \rangle_H \right]_{i,j=1}^N \in \mathbb{R}^{N \times N}$$

$\| \text{def} \|$
 $\langle x_i, x_j \rangle$

Full Kernel Ridge regression Alg:

① Choose a kernel function

$$K(x, y) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R},$$

$$\text{e.g., } K(x, y) = e^{-\frac{\|x-y\|_2^2}{\sigma^2}} \quad (\text{Gaussian kernel})$$

$$K(x, y) = (x^T y + 1)^\alpha \quad (\text{polynomial kernel})$$

② Calculate the kernel matrix

$$K = \left[K(x_i, x_j) \right]_{i,j=1}^N \in \mathbb{R}^{N \times N}$$

③ Solve C^* from

$$C^* = \arg \min_{C \in \mathbb{R}^N} \left(\|K^T C - y\|_2^2 + \lambda C^T K C \right)$$

④ Then the regression function is

$$\begin{aligned} f(x) &= \langle C^*, \phi(x) \rangle_H = \left\langle \sum_{i=1}^N C_i^* \phi(x_i), \phi(x) \right\rangle_H \\ &= \sum_{i=1}^N C_i^* \langle \phi(x_i), \phi(x) \rangle_H \\ &= \sum_{i=1}^N C_i^* K(x_i, x) \end{aligned}$$

§4.2.3. Linear Classification

Classification: Given

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N),$$

where $x_1, x_2, \dots, x_N \in \mathbb{R}^n$

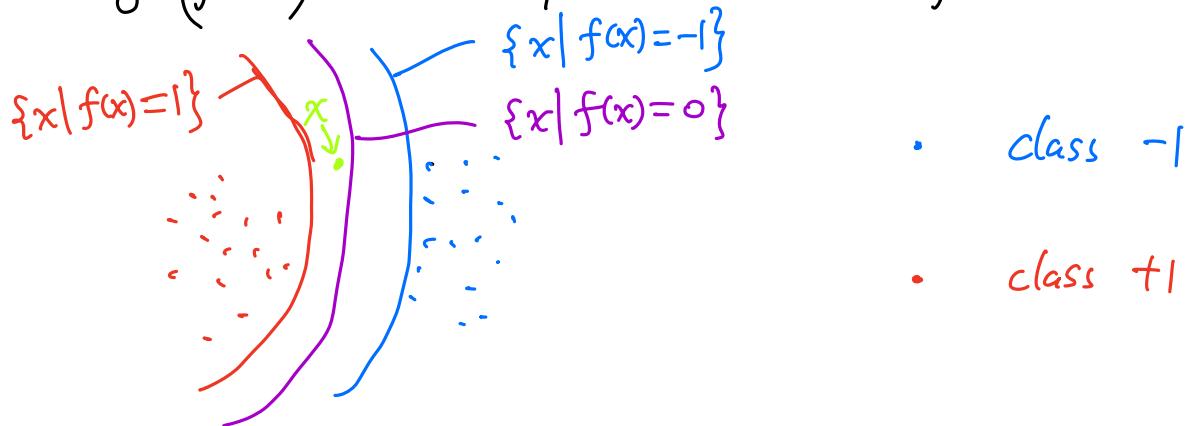
$$y_1, y_2, \dots, y_N \in \{-1, 1\}$$

Predict: given a new $x \in \mathbb{R}^n$, which class should x be in?

Find a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ s.t.

$$\begin{cases} f(x_i) \geq 1 & \text{if } y_i = 1 \\ f(x_i) \leq -1 & \text{if } y_i = -1 \end{cases}$$

Then, $\operatorname{sgn}(f(x))$ as the predicted label of x .



Which function class should f be in?

— linear classifiers: Use affine functions.

— Any affine function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ must be in

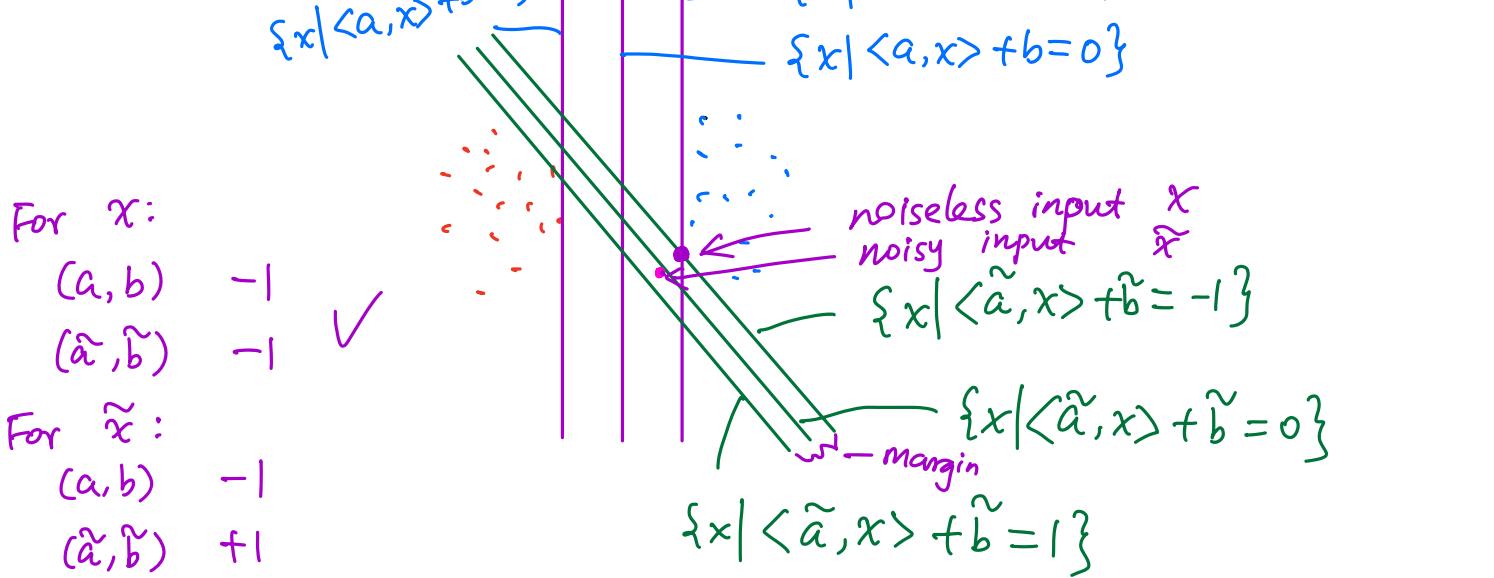
$$f(x) = \langle a, x \rangle + b, \quad \text{where } a \in \mathbb{R}^n, b \in \mathbb{R}$$

— So, we find $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ st

$$\begin{cases} \langle a, x_i \rangle + b \geq 1 & \text{if } y_i = 1 \\ \langle a, x_i \rangle + b \leq -1 & \text{if } y_i = -1 \end{cases} \quad i=1, 2, \dots, N$$

The solution is not unique.

$$\{x | \langle a, x \rangle + b = 1\} \quad \text{margin} \quad \{x | \langle a, x \rangle + b = -1\}$$



For x :

$$\begin{array}{ll} (a, b) & -1 \\ (\tilde{a}, \tilde{b}) & -1 \quad \checkmark \end{array}$$

For \tilde{x} :

$$\begin{array}{ll} (a, b) & -1 \\ (\tilde{a}, \tilde{b}) & +1 \end{array}$$

— which solution is better?

— larger width between separation hyperplanes,
larger tolerance to noise for mis-classification
the better classifier.

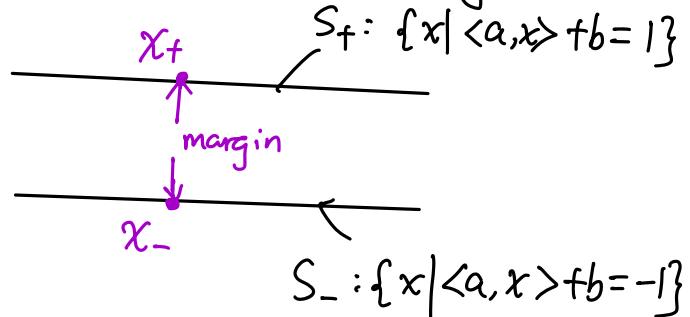
— larger margin, better classifier.

— Support vector machine (SVM): maximize
the margin

$$\text{margin} = \|x_f - x_- \|_2$$

$$\text{Notice that: } x_- = P_{S_-} x_f$$

$$x_f = P_{S_f} x_-$$



By projection formula: $x_f = P_{S_f} x_- = x_- - \frac{\langle a, x_- \rangle - (1-b)}{\|a\|_2^2} a$

$$\text{So, } \|x_f - x_- \|_2 = \left\| \frac{\langle a, x_- \rangle - (1-b)}{\|a\|_2^2} a \right\|_2 = \frac{|\langle a, x_- \rangle - (1-b)|}{\|a\|_2^2} \|a\|_2$$

$$\begin{aligned} x_- \in S_- \Rightarrow \langle a, x_- \rangle + b = -1 \\ \Rightarrow \langle a, x_- \rangle - (1-b) = -2 \end{aligned}$$

$$\text{So, } \|x_f - x_- \|_2 = \frac{|-2|}{\|a\|_2} \|a\|_2 = \frac{2}{\|a\|_2}$$

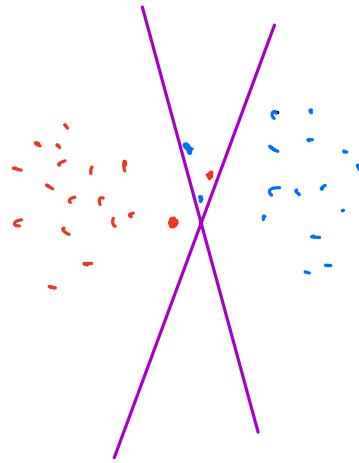
So, in SVM:

$$\boxed{\begin{array}{l} \max_{\substack{a \in \mathbb{R}^n \\ b \in \mathbb{R}}} \frac{2}{\|a\|_2} \\ \text{s.t. } \begin{cases} \langle a, x_i \rangle + b \geq 1 & \text{if } y_i = 1 \\ \langle a, x_i \rangle + b \leq -1 & \text{if } y_i = -1 \end{cases} \quad i=1, \dots, N \end{array}}$$

↔

$$\boxed{\begin{array}{l} \min_{\substack{a \in \mathbb{R}^n \\ b \in \mathbb{R}}} \|a\|_2^2 \\ \text{s.t. } y_i(\langle a, x_i \rangle + b) \geq 1, \quad i=1, \dots, N \end{array}} \quad (\text{SVM-1})$$

- (SVM-1) is not robust to noise.



- Noise makes the two classes non-separable by hyperplanes.
- No solutions (a, b) to $y_i(\langle a, x_i \rangle + b) \geq 1 \quad \forall i$

- Reformulate (SVM-1):

Define a generalized function: $h : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$

$$h(t) = \begin{cases} 0 & \text{if } t \geq 0 \\ +\infty & \text{if } t < 0 \end{cases}$$

Then

$$(\text{SVM-1}) \iff \min_{\substack{a \in \mathbb{R}^n \\ b \in \mathbb{R}}} \sum_{i=1}^N h(y_i(\langle a, x_i \rangle + b) - 1) + \lambda \|a\|_2^2,$$

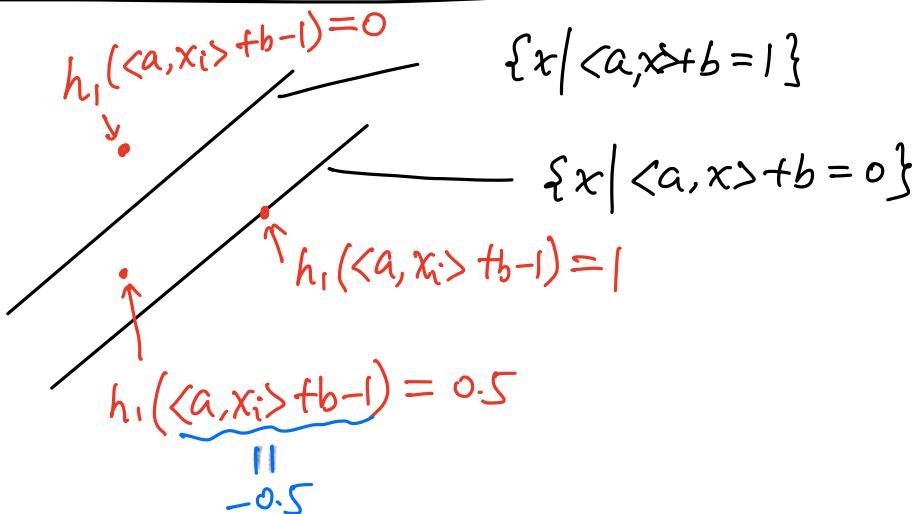
data-fitting *regularization*
where $\lambda > 0$

To avoid $+\infty$ function value for any (a, b) , we approximate h by some function that takes only finite values.

$$h_1(t) = \begin{cases} 0 & \text{if } t \geq 0 \\ |t| & \text{if } t < 0 \end{cases} = \max(0, -t)$$

Soft-Margin SVM

$$\min_{\alpha \in \mathbb{R}^N, b \in \mathbb{R}} \sum_{i=1}^N h_1(y_i(\langle \alpha, x_i \rangle + b) - 1) + \lambda \|\alpha\|_2^2 \quad (\text{SVM-2})$$



— (SVM-2) is non-smooth — optimization algorithm may not be available or converges slow.

— We further approximate h_1 by a smooth function.

$$h_2(t) = \ln(e^0 + e^{-t}) = \ln(1 + e^{-t})$$

— h_2 is smooth

— h_2 is a good approx. to h_1 , because

$$\begin{cases} \frac{h_2(t)}{e^{-t}} \rightarrow 1, & t \rightarrow -\infty \\ h_2(t) \rightarrow 0, & t \rightarrow +\infty \end{cases}$$

— h_2 is the logistic loss function

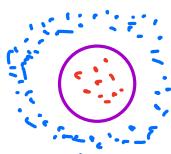
— h_2 has some statistical meaning.

So we solve

$$\min_{\substack{a \in \mathbb{R}^n \\ b \in \mathbb{R}}} \sum_{i=1}^N h_2(y_i(\langle a, x_i \rangle + b) - 1) + \lambda \|a\|_2^2$$

logistic regression

All SVMs are linear. They will not work if



We choose non-linear $f: \mathbb{R}^n \rightarrow \mathbb{R}$

Kernel SVM:

$$f: \quad \textcircled{1} \quad \phi: \mathbb{R}^n \rightarrow H \quad x_i \mapsto \phi(x_i)$$

\textcircled{2} SVM on H :

$$f(x_i) = \langle a, \phi(x_i) \rangle_H, \text{ for some } a \in H$$

and we solve

$$\min_{a \in H} \sum_{i=1}^N \tilde{h}(y_i \cdot \langle a, \phi(x_i) \rangle - 1) + \lambda \|a\|_H^2,$$

where $\lambda > 0$ is a regularization parameter.

(K-SVM)

$a \leftrightarrow (a_s, c)$

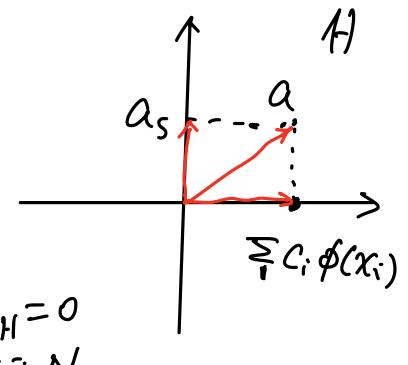
Representer Thm

Since $H \subset H$, we can decompose it as

$$a = a_s + \sum_{i=1}^N c_i \phi(x_i),$$

where $c = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N$ and a_s satisfies

$$\langle a_s, \phi(x_i) \rangle_H = 0 \quad i=1, \dots, N.$$



Then, the objective function in (K-SVM) is

$$\begin{aligned}
 & \sum_{i=1}^N \tilde{h}\left(y_i \cdot \langle a, \phi(x_i) \rangle_H - 1\right) + \lambda \|a\|_H^2 \\
 &= \sum_{i=1}^N \tilde{h}\left(y_i \cdot \left\langle a_s + \sum_{j=1}^N c_j \phi(x_j), \phi(x_i) \right\rangle_H - 1\right) + \lambda \left\| a_s + \sum_{j=1}^N c_j \phi(x_j) \right\|_H^2 \\
 &= \sum_{i=1}^N \tilde{h}\left(y_i \cdot \left(\sum_{j=1}^N c_j \langle \phi(x_j), \phi(x_i) \rangle_H \right) - 1\right) \\
 &\quad + \lambda \left(\|a_s\|_H^2 + \sum_{j_1=1}^N \sum_{j_2=1}^N c_{j_1} c_{j_2} \langle \phi(x_{j_1}), \phi(x_{j_2}) \rangle_H \right)
 \end{aligned}$$

Introduce a matrix $K = [\langle \phi(x_i), \phi(x_j) \rangle_H]_{i,j=1}^N \in \mathbb{R}^{N \times N}$

$$= \sum_{i=1}^N \tilde{h}\left(y_i (K^T c)_i - 1\right) + \lambda c^T K c + \lambda \|a_s\|_H^2$$

Then $a = a_s + \sum_{j=1}^N c_j \phi(x_j)$

$$\begin{aligned}
 (\text{K-SVM}) \iff & \min_{\substack{a_s \in H \\ c \in \mathbb{R}^N}} \sum_{i=1}^N \tilde{h}\left(y_i (K^T c)_i - 1\right) + \lambda c^T K c + \lambda \|a_s\|_H^2 \\
 & \text{s.t. } \langle a_s, \phi(x_j) \rangle_H = 0, \quad j=1, \dots, N
 \end{aligned}$$

$$\iff \min_{a_s \in H} \|a_s\|_H^2 \quad \text{s.t. } \langle a_s, \phi(x_j) \rangle_H = 0, \quad j=1, \dots, N$$

and

$$\min_{c \in \mathbb{R}^N} \sum_{i=1}^N \tilde{h}\left(y_i (K^T c)_i - 1\right) + \lambda c^T K c$$

$$\iff a_s^* = 0$$

and

$$c^* = \arg \min_{c \in \mathbb{R}^N} \sum_{i=1}^N \tilde{h}\left(y_i (K^T c)_i - 1\right) + \lambda c^T K c$$

So, the optimal a^* for (K-SVM)

$$a^* = \sum_{j=1}^N c_j^* \phi(x_j)$$

Full algorithm Kernel SVM:

① Define a kernel function $K: (\mathbb{R}^n, \mathbb{R}^n) \rightarrow \mathbb{R}$

② Form the kernel matrix $K = [K(x_i, x_j)]_{i,j=1}^N \in \mathbb{R}^{N \times N}$

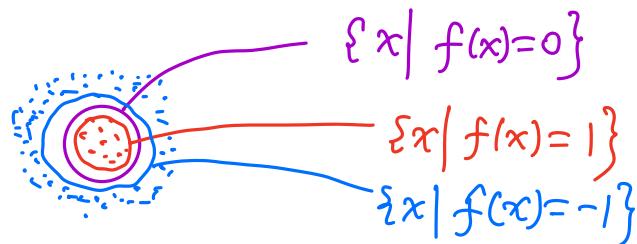
③ Solve

$$C^* = \arg \min_{C \in \mathbb{R}^N} \sum_{i=1}^N \tilde{h}\left(y_i((K^T C)_i - 1) + \lambda C^T K C\right)$$

④ The optimal $a^* = \sum_{j=1}^N c_j^* \phi(x_j)$, and the classification function is:

$$\begin{aligned} f(x) &= \langle a^*, \phi(x) \rangle_H \\ &= \left\langle \sum_{j=1}^N c_j^* \phi(x_j), \phi(x) \right\rangle_H = \sum_{j=1}^N c_j^* \langle \phi(x_j), \phi(x) \rangle_H \\ &= \sum_{j=1}^N c_j^* K(x_j, x) \end{aligned}$$

and $\text{sgn}(f(x))$ is the class that x is in.



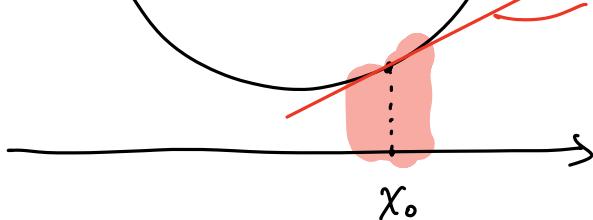
Linear functions have some limitations.

We still need non-linear functions

- functions by the kernel trick
- Neural network functions
- Optimization problem in linear models

§ 4.3. Linear approximation and differentiation

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$



Differentiation is
a local affine approximation

Recall for a function $f: \mathbb{R} \rightarrow \mathbb{R}$, its derivative at x_0 is

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

which is the same as

$$0 = \lim_{x \rightarrow x_0} \frac{|f(x) - (f(x_0) + f'(x_0)(x - x_0))|}{|x - x_0|}$$

The function $f(x_0) + f'(x_0)(x - x_0)$ satisfies:

- ① It is affine.
- ② It passes thru $(x_0, f(x_0))$

③ Its error to approximate f satisfies $\lim_{x \rightarrow x_0} \frac{\text{error}}{|x - x_0|} = 0$

i.e., $\text{error} = o(|x - x_0|)$
little o notation.

Consider $f: V \rightarrow \mathbb{R}$, where V is a Hilbert space, and $x^{(0)} \in V$

Then, we find a function $g: V \rightarrow \mathbb{R}$ s.t.

- ① g is affine and bounded
- ② g passes thru $(x^{(0)}, f(x^{(0)}))$, i.e., $g(x^{(0)}) = f(x^{(0)})$
- ③ The error of g to approximate f satisfies $\text{error} = o(\|x^{(0)} - x\|)$
 i.e., $\lim_{x \rightarrow x^{(0)}} \frac{|f(x) - g(x)|}{\|x - x^{(0)}\|} = 0$

To get an explicit form of g :

$$\begin{aligned} ① \Rightarrow g(x) &= \langle v, x \rangle + a, \text{ where } v \in V, a \in \mathbb{R} \\ &= \langle v, x - x^{(0)} \rangle + \langle v, x^{(0)} \rangle + a \end{aligned}$$

$$② \Rightarrow g(x^{(0)}) = \langle v, x^{(0)} \rangle + a = f(x^{(0)})$$

$$\Rightarrow g(x) = f(x^{(0)}) + \langle v, x - x^{(0)} \rangle$$

$$(3) \lim_{\|x-x^{(0)}\| \rightarrow 0} \frac{|f(x) - (f(x^{(0)}) + \langle v, x - x^{(0)} \rangle)|}{\|x - x^{(0)}\|} = 0$$

Definition: Let V be a Hilbert space. Let $f: V \rightarrow \mathbb{R}$. Then f is said Frechet differentiable at $x^{(0)}$ if

$$\exists v \in V \text{ s.t.}$$

$$\lim_{\|x-x^{(0)}\| \rightarrow 0} \frac{|f(x) - (f(x^{(0)}) + \langle v, x - x^{(0)} \rangle)|}{\|x - x^{(0)}\|} = 0.$$

If f is differentiable at $x^{(0)}$, then v is called the gradient of f at $x^{(0)}$, denoted by $\nabla f(x^{(0)})$.

— When $V = \mathbb{R}$, then $\nabla f(x^{(0)}) = f'(x^{(0)})$.

Example 1: $f(x) = \|x\|^2 \equiv \langle x, x \rangle$, where $\|\cdot\|$ is the norm induced by the inner product.

At $x^{(0)} \in V$,

$$\begin{aligned} f(x) &= \|x\|^2 = \|x^{(0)} + (x - x^{(0)})\|^2 \\ &= \|x^{(0)}\|^2 + 2 \langle x^{(0)}, x - x^{(0)} \rangle + \|(x - x^{(0)})\|^2 \\ &= f(x^{(0)}) + \langle 2x^{(0)}, x - x^{(0)} \rangle + \|(x - x^{(0)})\|^2 \end{aligned}$$

$$\begin{aligned} &\lim_{\|x-x^{(0)}\| \rightarrow 0} \frac{|f(x) - (f(x^{(0)}) + \langle 2x^{(0)}, x - x^{(0)} \rangle)|}{\|x - x^{(0)}\|} \\ &= \lim_{\|x-x^{(0)}\| \rightarrow 0} \frac{\|x - x^{(0)}\|^2}{\|x - x^{(0)}\|} = 0 \end{aligned}$$

$$\text{So, } \nabla f(x^{(0)}) = 2x^{(0)}$$

$$\text{i.e. } \boxed{\nabla f(x) = 2x}$$

Special cases: — $V = \mathbb{R}^n$, $f(x) = \|x\|_2^2$, then $\nabla(\|x\|_2^2) = 2x$

- $V = \mathbb{R}^n$, $f(x) = \|x\|_A^2$, then $\nabla(\|x\|_A^2) = 2x$
 with $\langle \cdot, \cdot \rangle_A$

- $V = \mathbb{R}^n$, $g(x) = \|x\|_A^2$, then $\nabla(\|x\|_A^2) \neq 2x$
 with $\langle \cdot, \cdot \rangle$

- $V = \mathbb{R}^{m \times n}$ with $\langle \cdot, \cdot \rangle$ $f(x) = \|X\|_F^2$, then $\nabla(\|X\|_F^2) = 2X$

- $V = L^2(0,1)$ $F(f) = \|f\|^2 = \int_0^1 |f(t)|^2 dt$
 $\langle f, g \rangle = \int_0^1 f(t)g(t) dt$
 then $\nabla F(f) = 2f$

Example 2: $f(x) = \langle a, x \rangle$ for some $a \in V$.

At $x^{(0)} \in V$,

$$\begin{aligned} f(x) &= \langle a, x \rangle = \langle a, x^{(0)} \rangle + \langle a, x - x^{(0)} \rangle \\ &= f(x^{(0)}) + \langle a, x - x^{(0)} \rangle \\ \lim_{\|x - x^{(0)}\| \rightarrow 0} \frac{|f(x) - (f(x^{(0)}) + \langle a, x - x^{(0)} \rangle)|}{\|x - x^{(0)}\|} &= \lim_{\|x - x^{(0)}\| \rightarrow 0} \frac{0}{\|x - x^{(0)}\|} \\ &= 0 \end{aligned}$$

So, $\nabla f(x^{(0)}) = a$

i.e.,

$$\boxed{\nabla f(x) = a}$$

Similarly, $g(x) = \langle a, x \rangle + b$, where $a \in V$ and $b \in \mathbb{R}$

$$\boxed{\nabla g(x) = a}$$

Example 3: $f(x) = \|x - a\|^2$, where $a \in V$, and $\|\cdot\|$ is the norm induced by the inner product.

At $x^{(0)} \in V$,

$$\begin{aligned} f(x) &= \|x - a\|^2 = \|(x^{(0)} - a) + (x - x^{(0)})\|^2 \\ &= \|x^{(0)} - a\|^2 + 2 \langle x^{(0)} - a, x - x^{(0)} \rangle + \|x - x^{(0)}\|^2 \end{aligned}$$

$$= f(x^{(0)}) + \langle 2(x^{(0)} - a), x - x^{(0)} \rangle + \|x - x^{(0)}\|^2$$

$$\lim_{\|x - x^{(0)}\| \rightarrow 0} \frac{|f(x) - (f(x^{(0)}) + \langle 2(x^{(0)} - a), x - x^{(0)} \rangle)|}{\|x - x^{(0)}\|}$$

$$= \lim_{\|x - x^{(0)}\| \rightarrow 0} \frac{\|x - x^{(0)}\|^2}{\|x - x^{(0)}\|} = 0$$

$$\text{So, } \nabla f(x^{(0)}) = 2(x^{(0)} - a)$$

or, $\boxed{\nabla f(x) = 2(x - a)}$

properties of Frechet differentiation

① Frechet differentiation is linear in f , i.e.

$\forall \alpha, \beta \in \mathbb{R}, f, g : V \rightarrow \mathbb{R}$,

$$\nabla(\alpha f + \beta g)(x) = \alpha \nabla f(x) + \beta \nabla g(x),$$

provided $\nabla f(x)$ and $\nabla g(x)$ exist.

$$\boxed{\begin{matrix} f & g \\ V \rightarrow \mathbb{R} & \mathbb{R} \rightarrow \mathbb{R} \end{matrix}}$$

② Chain rule

Let $f : V \rightarrow \mathbb{R}$ and $g : \mathbb{R} \rightarrow \mathbb{R}$ (Then $g \circ f : V \rightarrow \mathbb{R}$)

Then $\nabla(g \circ f)(x) = g'(f(x)) \cdot \nabla f(x)$,

provided g and f are both differentiable at $f(x)$ and x resp.

Sketch of proof.

$$0 \leq \lim_{\|y - x\| \rightarrow 0} \frac{|g(f(y)) - (g(f(x)) + \langle g'(f(x)) \cdot \nabla f(x), y - x \rangle)|}{\|y - x\|}$$

$$\leq \lim_{\|y - x\| \rightarrow 0} \left(\frac{|g(f(y)) - (g(f(x)) + g'(f(x))(f(y) - f(x))|}{\|y - x\|} \right)$$

$$+ \left(\frac{|g'(f(x))| \cdot |f(y) - (f(x) + \langle \nabla f(x), y - x \rangle)|}{\|y - x\|} \right) = 0$$

For I_2 : $\lim_{\|y-x\|\rightarrow 0} I_2 = |g'(f(x))| \cdot \lim_{\|y-x\|\rightarrow 0} \frac{|f(y) - (f(x) + \langle \nabla f(x), y-x \rangle)|}{\|y-x\|} = 0$

(we can prove $|g'(f(x))| < +\infty$ using differentiability of g at $f(x)$)

For I_1 :

$$\lim_{\|y-x\|\rightarrow 0} I_1 = \lim_{\|y-x\|\rightarrow 0} \left(\frac{\left| g(f(y)) - \left(g(f(x)) + g'(f(x))(f(y) - f(x)) \right) \right|}{\|f(y) - f(x)\|} \cdot \frac{\|f(y) - f(x)\|}{\|y-x\|} \right) \quad I_3$$

— For I_3 : $\lim_{\|y-x\|\rightarrow 0} I_3 = 0$ by the differentiability of g at $f(x)$.

— For I_4 : $I_4 < +\infty$ by the differentiability of f at x .

($I_4 < \|\nabla f(x)\| + 1$ for sufficiently small $\|y-x\|$)

$$\Rightarrow \lim_{\|y-x\|\rightarrow 0} I_1 = 0$$

④

Example 4: $f(x) = \|x\|$, where $\|\cdot\|$ is the norm induced by the inner product on V .

Define $f_1(x) = \|x\|^2$ for $x \in V$

$f_2(t) = \sqrt{t}$ for $t \in \mathbb{R}$

$$\text{Then } f(x) = \sqrt{\|x\|^2} = f_2(f_1(x)) \quad \text{i.e. } f = f_2 \circ f_1$$

— When $x \neq 0$, f_1 and f_2 are differentiable at x and $f_1(x)$ resp.

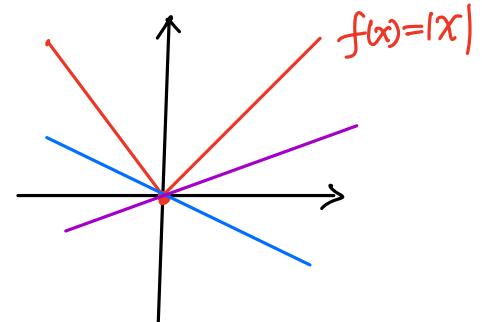
By the chain rule, $\nabla f(x) = \nabla(f_2 \circ f_1)(x) = f'_2(f_1(x)) \cdot \nabla f_1(x)$

Since $\nabla f_1(x) = \nabla(\|x\|^2) = 2x$

$$f'_2(t) = (\sqrt{t})' = \frac{1}{2\sqrt{t}}$$

$$\Rightarrow \nabla f(x) = \frac{1}{2\sqrt{\|x\|^2}} \cdot 2x = \frac{x}{\|x\|}$$

- When $x=0$, we can not apply the chain rule
we can prove f is non-differentiable at $x=0$.
(e.g., $V=\mathbb{R}$, $f(x)=|x|$)



Special cases:

$$- V=\mathbb{R}, \quad f(x)=|x|, \quad \nabla(|x|) = \frac{x}{\|x\|} = \begin{cases} 1 & \text{if } x>0 \\ -1 & \text{if } x<0 \end{cases} \quad \text{for } x \neq 0$$

$$- V=\mathbb{R}^n, \quad f(x)=\|x\|_2, \quad \nabla(\|x\|_2) = \frac{x}{\|x\|_2} = \frac{1}{\|x\|_2} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

:

- ③ For functions $f: \mathbb{R}^n \rightarrow \mathbb{R}$, where \mathbb{R}^n is with the Euclidean inner product, if f is differentiable at $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$, ~~then~~

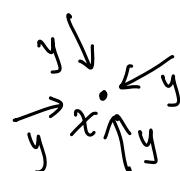
then

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}$$

(Frechet differentiation is consistent with the standard differentiation in multi-variate calculus)

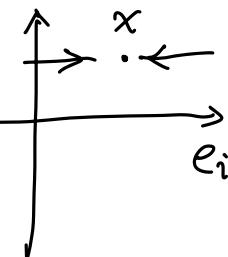
proof. Since f is differentiable at $x \in \mathbb{R}^n$,

$$\lim_{\|y-x\| \rightarrow 0} \frac{|f(y) - (f(x) + \langle \nabla f(x), y-x \rangle)|}{\|y-x\|_2} = 0$$



Choose $y = x + t \cdot e_i$, where $t \in \mathbb{R}$

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{pmatrix} \leftarrow i\text{-th entry}$$



and $t \rightarrow 0$ (i.e., $g(t) = f(x+te_i)$)

$$0 = \lim_{t \rightarrow 0} \frac{|f(x+te_i) - (f(x) + t \langle \nabla f(x), e_i \rangle)|}{|t-0|}$$

$$0 = \lim_{t \rightarrow 0} \frac{|g(t) - (g(0) + \langle \nabla f(x), e_i \rangle \cdot (t-0))|}{|t-0|}$$

i-th component of $\nabla f(x)$

So, $g'(0) = \langle \nabla f(x), e_i \rangle$

$$\frac{d}{dt} g(t) \Big|_{t=0} = \frac{d}{dt} f(x+te_i) \Big|_{t=0} = \frac{\partial f}{\partial x_i}(x)$$

So, $[\nabla f(x)]_i = \frac{\partial f}{\partial x_i}(x). \quad \forall i$ □

Example: $f(x) = \|x\|_2$, where $x \in \mathbb{R}^n$

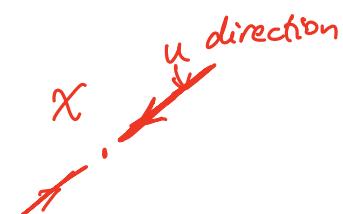
$$\begin{aligned} \frac{\partial f}{\partial x_i}(x) &= \frac{\partial \|x\|_2}{\partial x_i} = \frac{\partial \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}}{\partial x_i} \\ &= \frac{1}{2} \cdot \frac{1}{\sqrt{x_1^2 + x_2^2 + \dots + x_n^2}} \cdot \frac{\partial (x_1^2 + x_2^2 + \dots + x_n^2)}{\partial x_i} \\ &= \frac{1}{2} \frac{1}{\|x\|_2} \cdot \frac{\partial x_i^2}{\partial x_i} = \frac{1}{2} \frac{1}{\|x\|_2} \cdot 2x_i \\ &= \frac{x_i}{\|x\|_2} \quad \text{if } \|x\|_2 \neq 0 \end{aligned}$$

So, $\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix} = \frac{1}{\|x\|_2} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \frac{x}{\|x\|_2}$ □

④ Let $f: V \rightarrow \mathbb{R}$, Assume f is differentiable at $x \in V$.

Let $u \in V$,

Then $\langle \nabla f(x), u \rangle = \underbrace{\frac{d}{dt} f(x+tu)}_{\text{direction}}$



↑ directional derivative of f
along u .

proof. the same as ③ by replacing e_i with u_j .

⑤ Taylor's expansion.

For $f: \mathbb{R} \rightarrow \mathbb{R}$, Taylor's expansion

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + o(|x-x_0|)$$

For $f: V \rightarrow \mathbb{R}$, because: if f is diff at $x^{(0)} \in V$,

then $\lim_{\|x-x^{(0)}\| \rightarrow 0} \frac{|f(x) - (f(x^{(0)}) + \langle \nabla f(x^{(0)}), x-x^{(0)} \rangle)|}{\|x-x^{(0)}\|} = 0$



$$f(x) = f(x^{(0)}) + \langle \nabla f(x^{(0)}), x-x^{(0)} \rangle + o(\|x-x^{(0)}\|)$$



Taylor expansion

Differentiation of functions on Banach spaces

Let V be a Banach space with a norm $\|\cdot\|$

Let $f: V \rightarrow \mathbb{R}$

We use affine approximation at $x^{(0)} \in V$

We find a function $g: V \rightarrow \mathbb{R}$ s.t.

① g is affine and bounded

② $g(x^{(0)}) = f(x^{(0)})$

③ $|f(x) - g(x)| = o(\|x-x^{(0)}\|)$ (i.e., $\lim_{\|x-x^{(0)}\| \rightarrow 0} \frac{|f(x)-g(x)|}{\|x-x^{(0)}\|} = 0$)

④ $\Rightarrow g(x) = Lx + a$, where $L: V \rightarrow \mathbb{R}$ linear and bounded
 $a \in \mathbb{R}$

$$\textcircled{2} \Rightarrow g(x^{(0)}) = L(x^{(0)}) + a = f(x^{(0)})$$

$$\begin{aligned}\Rightarrow g(x) &= L(x+a) = L((x-x^{(0)})+x^{(0)}) + a \\ &= L(x-x^{(0)}) + L(x^{(0)}) + a \\ &= f(x^{(0)}) + L(x-x^{(0)})\end{aligned}$$

$$\textcircled{3} \quad \lim_{\|x-x^{(0)}\| \rightarrow 0} \frac{|f(x) - (f(x^{(0)}) + L(x-x^{(0)}))|}{\|x-x^{(0)}\|} = 0$$

Definition: Let V be Banach space with $\|\cdot\|$, and $x^{(0)} \in V$.

Let $f: V \rightarrow \mathbb{R}$. Then f is differentiable if:

\exists a linear and bounded function $L: V \rightarrow \mathbb{R}$ s.t.

$$\lim_{\|x-x^{(0)}\| \rightarrow 0} \frac{|f(x) - (f(x^{(0)}) + L(x-x^{(0)}))|}{\|x-x^{(0)}\|} = 0$$

The linear and bounded function L is called the differentiation of f , and it is denoted by $Df(x^{(0)}) = L$

§ 4.4. Case Study: Optimization and Gradient Descent

$\min_{x \in V} f(x)$, where V is a Hilbert space
 (OPT) and $f: V \rightarrow \mathbb{R}$

§ 4.4.1. Solvability and Optimality

- Solvability of (OPT)

- We say $x^{(*)} \in V$ is a solution of (OPT) if

$$f(x^{(*)}) \leq f(x) \quad \forall x \in V$$

We also call $x^{(*)}$ a global minimizer of f ,

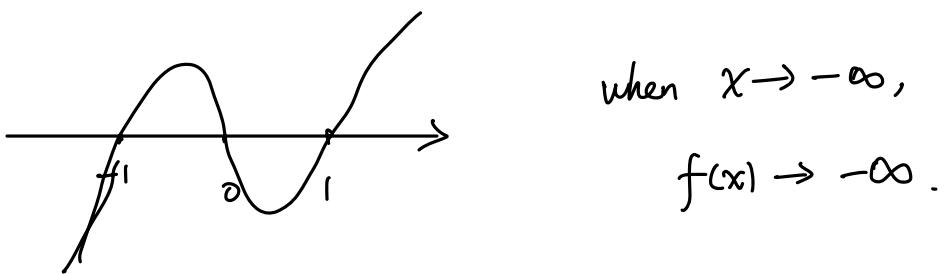
denoted by
$$x^{(*)} = \arg \min_{x \in V} f(x)$$

(0-th order optimality condition)

- The existence of a solution of (OPT) is not guaranteed automatically.

Some examples:

- $f: \mathbb{R} \rightarrow \mathbb{R}$ $f(x) = x(x-1)(x+1)$



when $x \rightarrow -\infty$,

$$f(x) \rightarrow -\infty.$$

- $f: V \rightarrow \mathbb{R}$, $f(x) = \langle a, x \rangle$ for some $a \in V$.

Then set $x = ca$ for $c \in \mathbb{R}$.

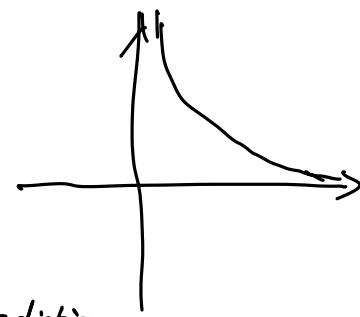
$$\text{So, } f(x) = f(ca) = \langle ca, a \rangle = c \|a\|^2 \rightarrow -\infty \text{ as } c \rightarrow -\infty$$

- $f: \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = \frac{1}{x} \quad x > 0$

$$\text{So, } f(x) > f(x+1) \quad \forall x > 0.$$

If we have $x^{(\star)} = \arg \min_{x \in \mathbb{R}} f(x)$,

then $f(x^{(\star)}) > f(x^{(\star)} + 1)$ contradiction.



- characterization of the solution of (DPT).

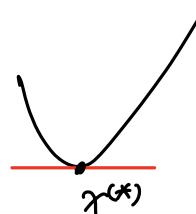
- 0-th order optimality condition

$$x^{(\star)} = \arg \min_{x \in V} f(x) \iff \boxed{f(x^{(\star)}) \leq f(x) \quad \forall x \in V}$$

(impossible to check numerically)

- 1-st order optimality condition

Thm: Assume $f: V \rightarrow \mathbb{R}$ is differentiable at $x^{(\star)} \in V$.



Then

$$x^{(\star)} = \arg \min_{x \in V} f(x) \implies \nabla f(x^{(\star)}) = 0$$

Proof. By expansion,

$$f(x) = f(x^{(\star)}) + \langle \nabla f(x^{(\star)}), x - x^{(\star)} \rangle + o(\|x - x^{(\star)}\|)$$

Suppose $\nabla f(x^{(\star)}) \neq 0$

Choose



$$\tilde{x} = x^{(*)} - t \cdot \nabla f(x^{(*)})$$

with $t > 0$

So,

$$\begin{aligned} f(\tilde{x}) &= f(x^{(*)}) + \langle \nabla f(x^{(*)}), -t \cdot \nabla f(x^{(*)}) \rangle + o(|t| \|\nabla f(x^{(*)})\|) \\ &= f(x^{(*)}) - t \|\nabla f(x^{(*)})\|^2 + o(|t| \|\nabla f(x^{(*)})\|) \end{aligned}$$

Because

$$\lim_{|t| \rightarrow 0} \frac{o(|t| \|\nabla f(x^{(*)})\|)}{|t| \|\nabla f(x^{(*)})\|} = 0,$$

$$\Rightarrow \forall C > 0, \exists t \text{ s.t. } o(|t| \|\nabla f(x^{(*)})\|) < C \cdot |t| \|\nabla f(x^{(*)})\|$$

We choose $C = \frac{1}{2} \|\nabla f(x^{(*)})\|^2$

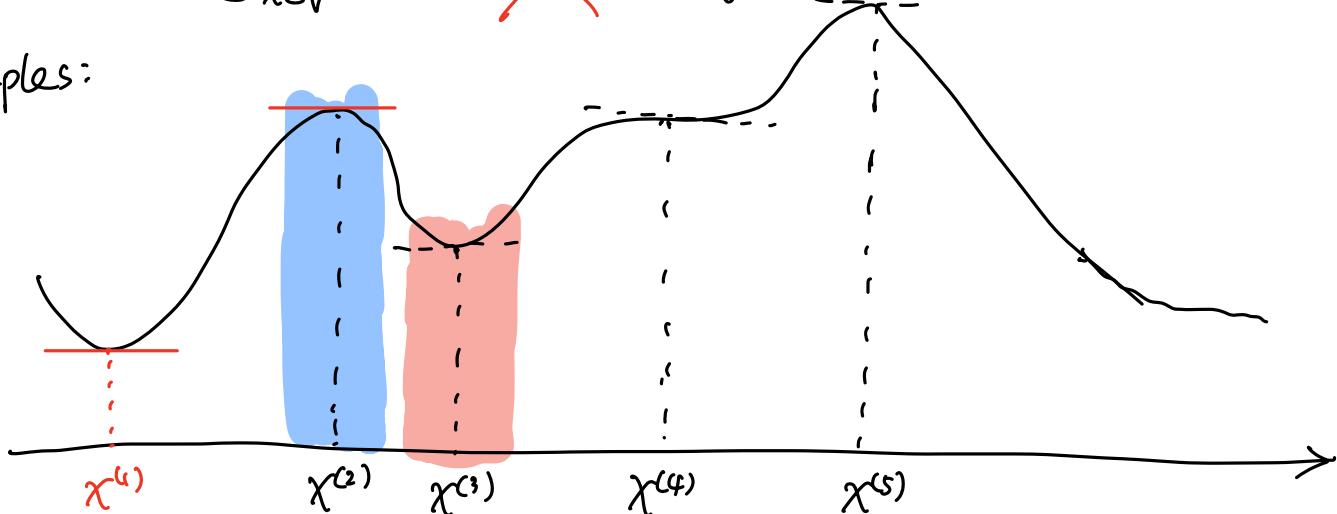
Then $o(|t| \|\nabla f(x^{(*)})\|) < \frac{1}{2} |t| \|\nabla f(x^{(*)})\|^2$

Then $f(\tilde{x}) = f(x^{(*)}) - t \|\nabla f(x^{(*)})\|^2 + \frac{1}{2} |t| \cdot \|\nabla f(x^{(*)})\|^2$
 $< f(x^{(*)})$ contradiction \otimes

The reverse is not true in general, i.e.,

$$x^{(*)} = \arg \min_{x \in V} f(x) \quad \cancel{\text{---}} \quad \nabla f(x^{(*)}) = 0$$

Examples:



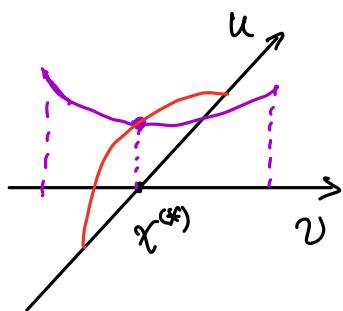
All $x^{(i)}$ satisfies $\nabla f(x^{(i)}) = 0$

But only $x^{(1)} = \arg \min_{x \in V} f(x)$

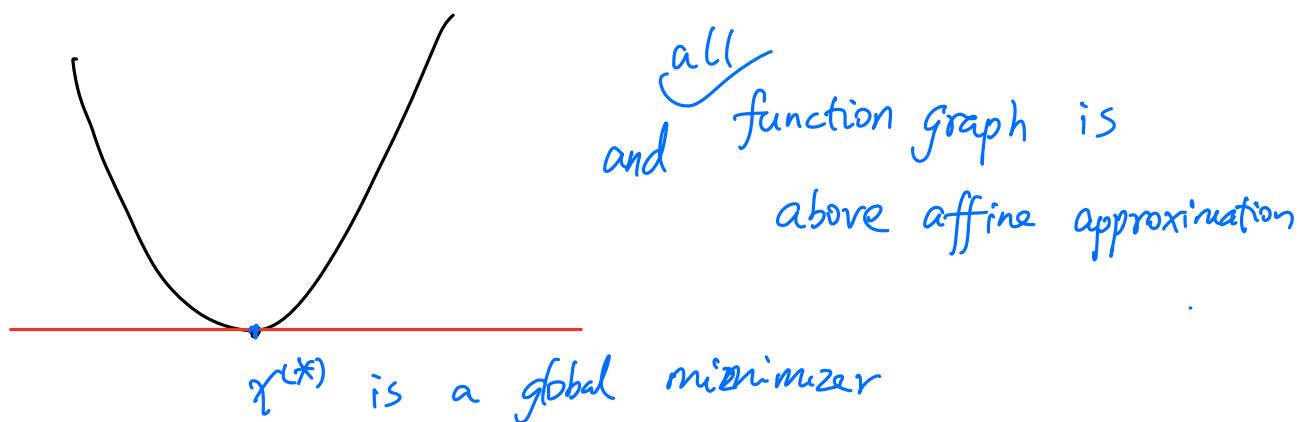
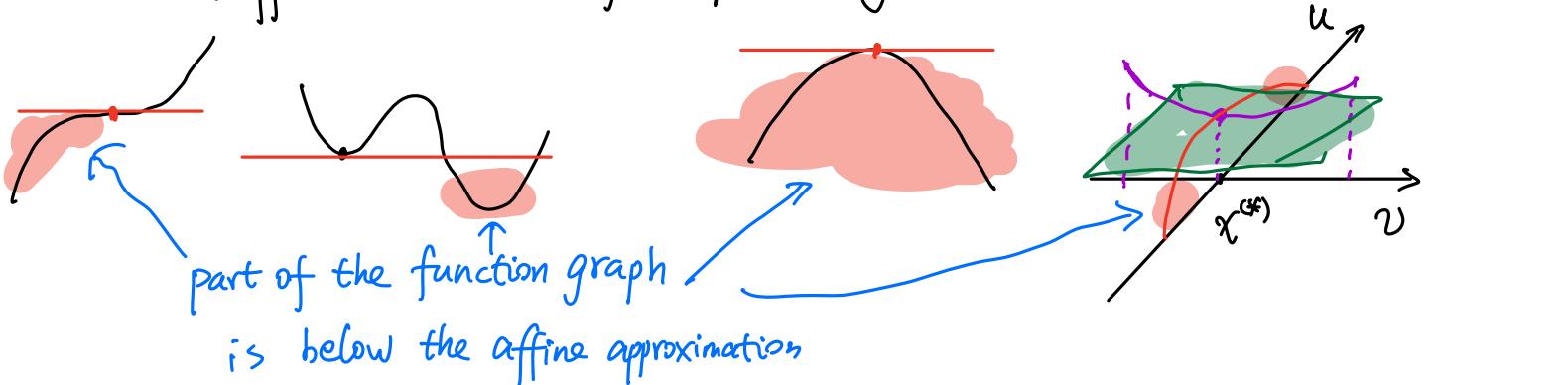
Actually, $\nabla f(x^{(*)}) = 0$, $x^{(*)}$ can be:

- Global minimizer (e.g., $x^{(1)}$)

- local minimizer (e.g., $x^{(3)}$)
 $f(x^{(3)}) \leq f(x) \quad \forall x \in V$ satisfying $\|x - x^{(3)}\| \leq \varepsilon$
for some ε .
- local maximizer (e.g., $x^{(2)}$)
 $f(x^{(2)}) \geq f(x) \quad \forall x \in V$ satisfying $\|x - x^{(2)}\| \leq \varepsilon$.
for some ε .
- Global maximizer (e.g., $x^{(5)}$)
 $f(x^{(5)}) \geq f(x) \quad \forall x \in V$.
- No information on optimality (e.g., $x^{(4)}$)
- Saddle point (only for V with $\dim(V) \geq 2$)
 - $\exists u, v \in V$
 - s.t. $\begin{cases} f(x^{(*)}) \geq f(x^{(*)}) + t u \\ f(x^{(*)}) \leq f(x^{(*)}) + t v \end{cases}$
 - $\forall t \in \mathbb{R} : |t| \leq \varepsilon$
for some ε .



- Sufficient condition for optimality.



Theorem: Let $f: V \rightarrow \mathbb{R}$, where V is a Hilbert space.

(*) Assume f is differentiable and

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle \quad \forall x, y \in V$$

Then,

$$x^{(*)} = \underset{x \in V}{\operatorname{argmin}} f(x) \iff \nabla f(x^{(*)}) = 0$$

(Affine approximation always support the function)

Proof. " \Rightarrow " proved in the previous theorem.

" \Leftarrow ": Since $\nabla f(x^{(*)}) = 0$,

$$\begin{aligned} f(x) &\geq f(x^{(*)}) + \langle \nabla f(x^{(*)}), x - x^{(*)} \rangle \\ &= f(x^{(*)}) + \langle 0, x - x^{(*)} \rangle \\ &= f(x^{(*)}) \end{aligned}$$

$\forall x \in V$

$$\text{So, } x^{(*)} = \underset{x \in V}{\operatorname{argmin}} f(x)$$



The condition (*) in the theorem:

— (*) is commonly known as
Differentiable + Convex

— Which functions satisfy (*)?

Ex. 1: $f(x) = \|x\|^2$, where $x \in V$ — Hilbert space
and $\|x\|^2 = \langle x, x \rangle$

In this case, $\nabla f(x) = 2x$, and

$$\begin{aligned} f(y) &= \|y\|^2 = \|x + (y-x)\|^2 \\ &= \|x\|^2 + \langle 2x, y-x \rangle + \|y-x\|^2 \\ &= f(x) + \langle \nabla f(x), y-x \rangle + \|y-x\|^2 \\ \Rightarrow f(y) &\geq f(x) + \langle \nabla f(x), y-x \rangle \quad \forall x, y \in V \end{aligned}$$

Ex. 2: Let $f_1, f_2, \dots, f_n: V \rightarrow \mathbb{R}$ all satisfy (\star)

Then, $f = \sum_{i=1}^n c_i f_i$, where $c_i \geq 0$, $i=1, \dots, n$
also satisfy (\star)

proof. f_i satisfies (\star) :

$$f_i(y) \geq f_i(x) + \langle \nabla f_i(x), y-x \rangle \quad \forall x, y \in V$$

Since $c_i \geq 0$

$$c_i f_i(y) \geq c_i f_i(x) + \langle c_i \cdot \nabla f_i(x), y-x \rangle$$

Sum them over i from 1 to n ,

$$\sum_{i=1}^n c_i f_i(y) \geq \sum_{i=1}^n c_i f_i(x) + \left\langle \sum_{i=1}^n c_i \cdot \nabla f_i(x), y-x \right\rangle$$

$$\text{i.e., } f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle \quad \blacksquare$$

§4.4.2. Gradient Descent

Goal: Solve $\min_{x \in V} f(x)$ numerically, where $f: V \rightarrow \mathbb{R}$
differentiable.
 $(\nabla f(x^{(k)}) \neq 0)$

Assume we have an estimation $x^{(k)} \in V$ of the solution $x^{(*)}$,
we want to find a better estimation $x^{(k+1)}$.

- We use **affine approximation locally**.

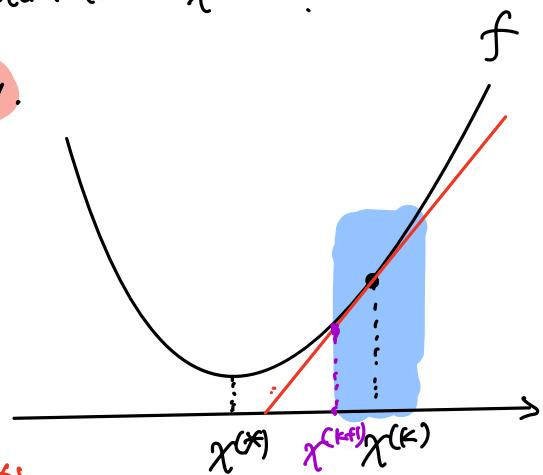
Instead of

$$\min_{x \in V} f(x)$$

approximate it by

$$\|x - x^{(k)}\| \leq \text{Constant} \quad f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle$$

approximated by
affine approximation
at $x^{(k)}$



We define

$$x^{(k+1)} = \arg \begin{cases} \min & f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle \\ \text{s.t.} & \|x - x^{(k)}\| \leq \alpha_k \cdot \|\nabla f(x^{(k)})\|, \end{cases} \quad (*)$$

Notice that ~~the~~ where $\alpha_k > 0$ is a constant.

$$(*) \Leftrightarrow \begin{cases} \min & \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle \\ \text{s.t.} & \|x - x^{(k)}\| \leq \alpha_k \cdot \|\nabla f(x^{(k)})\| \end{cases}$$

By Cauchy-Schwartz,

$$\begin{aligned} \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle &\stackrel{\textcircled{1}}{\geq} - \|\nabla f(x^{(k)})\| \cdot \|x - x^{(k)}\| \\ &\stackrel{\textcircled{2}}{\geq} - \|\nabla f(x^{(k)})\| \cdot \alpha_k \cdot \|\nabla f(x^{(k)})\| \\ &= - \alpha_k \|\nabla f(x^{(k)})\|^2 \end{aligned} \quad (**)$$

Both ① and ② can be "=" by choosing some x .

- ① becomes "=" when $x - x^{(k)} = -c \cdot \nabla f(x^{(k)})$

for some $c \geq 0$

- ② becomes "=" when $\|x - x^{(k)}\| = \alpha_k \cdot \|\nabla f(x^{(k)})\|$

$$\begin{aligned} \text{i.e., } \|x - x^{(k)}\| &= c \cdot \|\nabla f(x^{(k)})\| \\ &= \alpha_k \|\nabla f(x^{(k)})\| \end{aligned}$$

$$\Rightarrow c = \alpha_k$$

Altogether, when x satisfies $x - x^{(k)} = -\alpha_k \nabla f(x^{(k)})$,

we have $\begin{cases} \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle = -\alpha_k \|\nabla f(x^{(k)})\|^2 \\ \text{and } \|x - x^{(k)}\| \leq \alpha_k \|\nabla f(x^{(k)})\| \end{cases}$

By (**) :

$$\begin{cases} \min & \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle \\ \text{s.t.} & \|x - x^{(k)}\| \leq \alpha_k \cdot \|\nabla f(x^{(k)})\| \end{cases} \geq -\alpha_k \|\nabla f(x^{(k)})\|$$

So, $x^{(k+1)} - x^{(k)} = -\alpha_k \nabla f(x^{(k)})$

i.e.,

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

Gradient Descent Algorithm.
(GD)

- $\alpha_k > 0$ is a.k.a. step size / learning rate.
- How to choose α_k is crucial to the speed of GD.
There are many strategies.

- Convergence of GD:

- GD converges with a sufficiently small α_k .
- If GD converges, then we have

$$\lim_{k \rightarrow \infty} \nabla f(x^{(k)}) = 0 \quad (\text{Roughly, } \nabla f(x^\infty) = 0)$$

- If f satisfies (), $x^\infty = \arg \min_{x \in V} f(x)$
- Otherwise, x^∞ is not guaranteed to be a global minimizer.

§ 4.4.3 Examples of Optimization

Least Squares (LS):

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2,$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ are given

(LS) arises in:

— Linear regression: $\min_{\beta \in \mathbb{R}^{n+1}} \frac{1}{2} \|X\beta - y\|_2^2$

— Ridge regression: $\min_{\beta \in \mathbb{R}^{n+1}} \|X\beta - y\|_2^2 + \lambda \|\beta\|_2^2$
 $\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}$

$$\|X\beta - y\|_2^2 + \lambda \|a\|_2^2 = \left\| \frac{X\beta - y}{\sqrt{\lambda} a} \right\|_2^2$$

$$= \left\| \underbrace{\begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_x - \underbrace{\begin{bmatrix} y \\ 0 \end{bmatrix}}_b \right\|_2^2$$

- Kernel Ridge Regression: Can be reformulated as LS.

Let $f(x) = \frac{1}{2} \|Ax - b\|_2^2$ (we want to $\min_{x \in \mathbb{R}^n} f(x)$)

- $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable. Let's find ∇f .

Use definition of gradient

$$\begin{aligned} f(y) &= \frac{1}{2} \|Ay - b\|_2^2 = \frac{1}{2} \|(Ax - b) + (Ay - Ax)\|_2^2 \\ &= \frac{1}{2} \|Ax - b\|_2^2 + \langle Ax - b, Ay - Ax \rangle + \frac{1}{2} \|Ay - Ax\|_2^2 \\ &= f(x) + \langle Ax - b, A(y - x) \rangle + \frac{1}{2} \|A(y - x)\|_2^2 \\ &= f(x) + \langle A^T(Ax - b), y - x \rangle + \frac{1}{2} \|A(y - x)\|_2^2 \\ &\quad \left| \frac{f(y) - (f(x) + \langle A^T(Ax - b), y - x \rangle)}{\|y - x\|_2} \right| \end{aligned}$$

$$\begin{aligned} &\langle u, Av \rangle \\ &= u^T A v \\ &= (A^T u)^T v \\ &= \langle A^T u, v \rangle \end{aligned}$$

$$\begin{aligned} &= \lim_{\|y-x\|_2 \rightarrow 0} \frac{\frac{1}{2} \|A(y-x)\|_2^2}{\|y-x\|_2} \\ &\leq \lim_{\|y-x\|_2 \rightarrow 0} \frac{\frac{1}{2} \|A\|_2^2 \cdot \|y-x\|_2^2}{\|y-x\|_2} = 0 \end{aligned}$$

$$\Rightarrow \nabla f(x) = A^T(Ax - b)$$

- $\forall y, x \in \mathbb{R}^n$,

$$f(y) - (f(x) + \langle \nabla f(x), y - x \rangle)$$

$$= f(y) - \left(f(x) + \langle A^T(Ax-b), y-x \rangle \right)$$

$$= \frac{1}{2} \|A(y-x)\|_2^2 \geq 0$$

Therefore,

$$x^{(*)} = \arg \min_{x \in \mathbb{R}^n} \left(\frac{1}{2} \|Ax-b\|_2^2 \right) \iff A^T(Ax^{(*)}-b)=0$$

$$\iff \underset{n \times n}{A^T A} x^{(*)} = A^T b \quad \text{normal equation of LS}$$

- Geometry:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax-b\|_2^2$$

$$\Updownarrow y = Ax$$

$$\begin{cases} \min_{y \in \mathbb{R}^m} \frac{1}{2} \|y-b\|_2^2 \\ \text{s.t. } y \in \text{Ran}(A) \end{cases} \quad \begin{array}{l} \text{(Projection of)} \\ b \text{ onto} \\ \text{Ran}(A) \end{array}$$

$$\text{Choose } z = A^T(b-Ax^{(*)})$$

Then

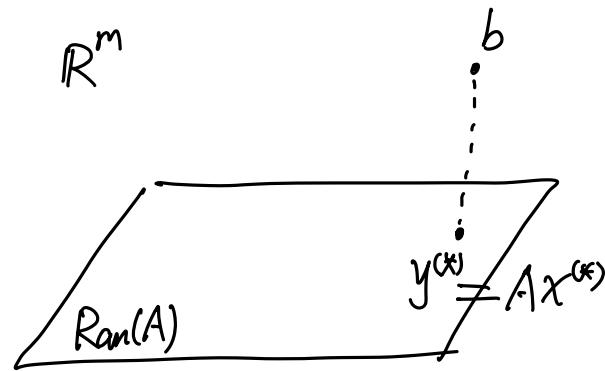
$$\langle A^T(b-Ax^{(*)}), A^T(b-Ax^{(*)}) \rangle = 0$$

$$\|A^T(b-Ax^{(*)})\|_2^2 = 0$$

$$A^T(b-Ax^{(*)}) = 0$$

$$\boxed{A^T A x^{(*)} = A^T b}$$

\mathbb{R}^m



Because $y^{(*)}$ is projection,

$$y^{(*)} = b - A x^{(*)} \perp \text{Ran}(A)$$

i.e., $\forall z \in \mathbb{R}^n$,

$$\langle b - A x^{(*)}, A z \rangle = 0$$

\Updownarrow

$$\langle A^T(b-Ax^{(*)}), z \rangle = 0 \quad \forall z \in \mathbb{R}^n$$

- Algorithms for LS

- Solve the normal equation

$A^T A x^{(*)} = A^T b$ by Cholesky decomposition
(LAPACK)

— Gradient Descent:

$$\chi^{(k+1)} = \chi^{(k)} - \alpha_k \cdot A^T(A\chi^{(k)} - b)$$

- How to choose the step size α_k ?

We can choose an optimal α_k by exact line search.

Define $\chi(\alpha) = \chi^{(k)} - \alpha A^T(A\chi^{(k)} - b)$, where $\alpha \in \mathbb{R}$

We solve $\alpha_k = \arg \min_{\alpha \in \mathbb{R}} f(\chi(\alpha)) \stackrel{\text{def}}{=} g(\alpha)$.

Then $g: \mathbb{R} \rightarrow \mathbb{R}$ is differentiable and satisfies

$$g(\beta) \geq g(\alpha) + g'(\alpha) \cdot (\beta - \alpha) \quad \forall \alpha, \beta \in \mathbb{R}$$

So, α_k satisfies

$$g'(\alpha_k) = 0$$

:

$$\alpha_k = \frac{\|A^T(A\chi^{(k)} - b)\|_2^2}{\|AA^T(A\chi^{(k)} - b)\|_2^2}$$

Full alg:

```
Initialize  $\chi^{(0)} \in \mathbb{R}^n$ .  
for  $k = 0, 1, 2, \dots$   
   $g^{(k)} = A^T(A\chi^{(k)} - b)$   
   $\alpha_k = \frac{\|g^{(k)}\|_2^2}{\|Ag^{(k)}\|_2^2}$   
   $\chi^{(k+1)} = \chi^{(k)} - \alpha_k g^{(k)}$   
end for
```

It is also known as
the Steepest Descent
algorithm.

Ch 5. Linear Transformations / Operators and Differentiation

§ 5.1. Linear transformations / operators

Let V_1, V_2 be two vector spaces

A mapping $L: V_1 \rightarrow V_2$ is a linear operator (or linear transformation)

if

$$L(\alpha x + \beta y) = \alpha \cdot L(x) + \beta \cdot L(y)$$

by the definition, we can obtain a property of L : $L(0) = L(0 \cdot x) = 0 \cdot L(x) = 0$ $\forall x, y \in V, \alpha, \beta \in \mathbb{R}$

Example 1: Let $A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}$ $(A: \mathbb{R}^n \rightarrow \mathbb{R}^m)$

- Define a mapping $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$ by

$$L(x) = Ax \quad \forall x \in \mathbb{R}^n$$

Then L is a linear transformation because

$$\begin{aligned} L(\alpha x + \beta y) &= A(\alpha x + \beta y) = \alpha \cdot Ax + \beta \cdot Ay \quad \forall x, y \in \mathbb{R}^n \\ &= \alpha \cdot L(x) + \beta \cdot L(y) \quad \alpha, \beta \in \mathbb{R} \end{aligned}$$

- Reversely, any linear transformation $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$ must be in the form of $L(x) = Ax \quad \forall x \in \mathbb{R}^n$ for some $A \in \mathbb{R}^{m \times n}$

Proof. If $x \in \mathbb{R}^n$, $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 e_1 + x_2 e_2 + \dots + x_n e_n$ where $e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{pmatrix}_{i\text{-th entry}} \in \mathbb{R}^n$.

Then $L(x) = L(x_1 e_1 + x_2 e_2 + \dots + x_n e_n)$

L is a linear trans

$$\begin{aligned} &= x_1 \cdot L(e_1) + x_2 \cdot L(e_2) + \dots + x_n \cdot L(e_n) \\ &= [L(e_1) \ L(e_2) \ \dots \ L(e_n)] \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\ &\stackrel{\text{def}}{=} A_{m \times n} x \\ &= Ax \quad \blacksquare \end{aligned}$$

Example 2: Discrete Fourier Transform (DFT) is a linear transformation

$\mathbb{C}^n \rightarrow \mathbb{C}^n$ defined as follows:

Let $w_n = e^{-\frac{2\pi}{n} i}$ where $i = \sqrt{-1}$

$\Rightarrow w_n^k = e^{j\theta} = \cos \theta + j \sin \theta, k = -1$
 $\left(= \cos\left(-\frac{2\pi}{n}\right) + i \sin\left(-\frac{2\pi}{n}\right)\right)$

$$\begin{aligned} e^{-2\pi i} &= \cos(-2\pi) + i \sin(-2\pi) \\ &= 1 + 0i = 1 \end{aligned}$$

and define

$$F_n = \frac{1}{\sqrt{n}} \begin{bmatrix} w_n^{0,0} & w_n^{0,1} & w_n^{0,2} & \dots & w_n^{0,(n-1)} \\ w_n^{1,0} & w_n^{1,1} & w_n^{1,2} & \dots & w_n^{1,(n-1)} \\ w_n^{2,0} & w_n^{2,1} & w_n^{2,2} & \dots & w_n^{2,(n-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ w_n^{(n-1),0} & w_n^{(n-1),1} & w_n^{(n-1),2} & \dots & w_n^{(n-1),(n-1)} \end{bmatrix} \in \mathbb{C}^{n \times n}$$

Then for any $x \in \mathbb{C}^n$, its DFT is

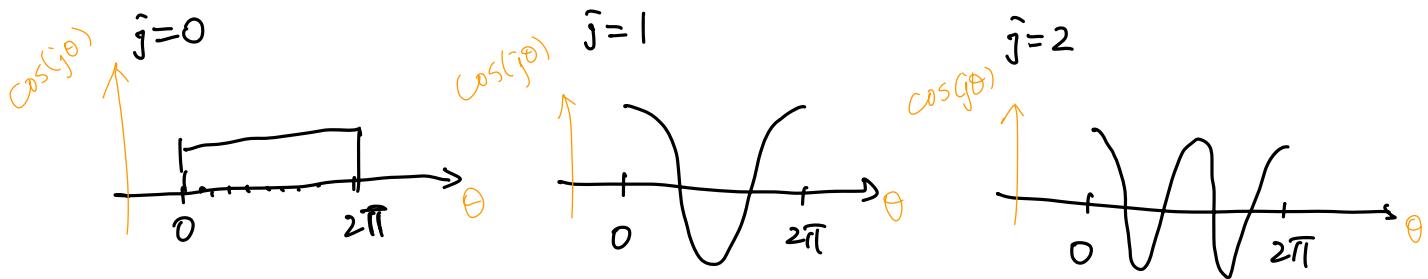
$$\hat{x} = F_n x$$

in this DFT we just decompose the signal in terms of signals with different frequencies

We can plot the j -th column of F_n :

Real part $\text{Re} \left(\begin{pmatrix} w_{1j} \\ w_{2j} \\ w_{3j} \\ \vdots \\ w_{nj} \end{pmatrix} \right) = \begin{pmatrix} \cos 0 \\ \cos(j \frac{2\pi}{n}) \\ \cos(2j \cdot \frac{2\pi}{n}) \\ \vdots \\ \cos((n-1)j \cdot \frac{2\pi}{n}) \end{pmatrix}$

discretization of $\cos(j\theta)$, where $\theta = 0, \frac{2\pi}{n}, 2 \cdot \frac{2\pi}{n}, \dots, (n-1) \frac{2\pi}{n}$



[j increases, frequency increases]

Example 3: Convolution with circular boundary condition

Let $a \in \mathbb{R}^n$ be a kernel of the convolution.

For any $x \in \mathbb{R}^n$, define $a \otimes: \mathbb{R}^n \rightarrow \mathbb{R}^n$ by

该条件叫做卷积的
因果性。

$$(a \otimes x)_k = \sum_{i+j=k} a_i x_j$$

$k = 0, 1, \dots, n-1$

$$\sum_{j=0}^{n-1} a_{(k-j) \bmod n} x_j$$

Then

$$a \otimes x = \begin{bmatrix} x_0 & x_1 & \cdots & x_{n-1} \end{bmatrix} \begin{bmatrix} a_0 & a_{n-1} & \cdots & a_2 & a_1 \\ a_1 & a_0 & a_{n-1} & \cdots & a_2 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{n-1} & \cdots & \cdots & a_2 & a_0 \end{bmatrix}$$

$\bmod n$ — circular boundary condition
What if $k < 0$? for example:
when $k-i = -1$ (negative)
then we treat it as $n-1$
(i.e. $\bmod n$ is equivalent to $0, 1, \dots, n-1$)

$$a \otimes x = \sum_{k=0}^{n-1} a_k x_{(k-i) \bmod n}$$

So, $a \otimes$ is a linear transformation $\mathbb{R}^n \rightarrow \mathbb{R}^n$.

Example 4: Let $f: V \rightarrow \mathbb{R}$ be a linear function on V .

Then f is a linear transformation $V \rightarrow \mathbb{R}$

Example 5: Let $a \in V$. Define $L: \mathbb{R} \rightarrow V$ by

$$L(x) = x \cdot a \quad \forall x \in \mathbb{R}$$

Then L is a linear transformation because

$$L(\alpha x + \beta y) = (\alpha x + \beta y) \cdot a = \alpha \cdot (x \cdot a) + \beta \cdot (y \cdot a) = \alpha \cdot L(x) + \beta \cdot L(y)$$

Example 6: Let $V_1 = \{f \mid f \text{ and } f' \text{ are continuous on } [a, b]\}$

$$V_2 = \{f \mid f \text{ is continuous on } [a, b]\}$$

The differentiation operator $D: V_1 \rightarrow V_2$ defined by

$$Df = f', \quad \forall f \in V_1$$

It is a linear transformation because

$$(continuous) D(\alpha f + \beta g) = (\alpha f + \beta g)' = \alpha \cdot f' + \beta \cdot g' = \alpha \cdot Df + \beta \cdot Dg$$

Example 7: Fourier Transform $\mathcal{F}: L^2(\mathbb{R}) \rightarrow L^2(\mathbb{R})$ defined by

$$(\mathcal{F}f)(\omega) = \int_{-\infty}^{+\infty} f(x) e^{-2\pi i \omega x} dx$$

is a linear transformation

Example 8: Given a function $g: \mathbb{R} \rightarrow \mathbb{R}$ (smooth enough),

define convolution with kernel g by:

$$(g * f)(x) = \int_{-\infty}^{+\infty} g(x-y) f(y) dy \quad \forall f \in L^2(\mathbb{R})$$

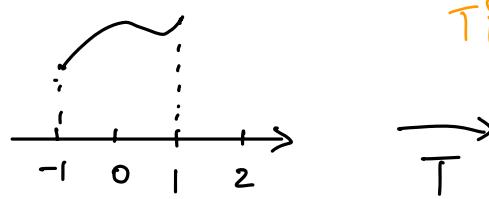
Then

$g *: L^2(\mathbb{R}) \rightarrow L^2(\mathbb{R})$ is a linear transformation

Example 9: Consider $C[-1, 1]$ and $C[0, 2]$

Define translation $T: C[-1, 1] \rightarrow C[0, 2]$ by

$$\forall f \in C[-1, 1], \quad (Tf)(t) = f(t-1) \quad \forall t \in [0, 2]$$



$$T\{f(t)\} = f(t-1)$$

$$\checkmark T\{f_1(t) + f_2(t)\} = f_1(t-1) + f_2(t-1) \\ = T\{f_1(t)\} + T\{f_2(t)\}$$

$$\checkmark T\{\alpha f(t)\} = \alpha f(t-1) = \alpha \cdot T\{f(t)\}$$

example 7.8 也可證

Example 10: Let $a_1, a_2, \dots, a_k \in H$, where H is a Hilbert space.

Define $L: H \rightarrow \mathbb{R}^k$ by

$$L(x) = \begin{pmatrix} \langle a_1, x \rangle \\ \langle a_2, x \rangle \\ \vdots \\ \langle a_k, x \rangle \end{pmatrix} \in \mathbb{R}^k$$

since $\langle \cdot, \cdot \rangle$ has a linear property as we have shown before

$$= \begin{pmatrix} L(a_1x + \beta y) \\ \langle a_1, \alpha x + \beta y \rangle \\ \vdots \\ \langle a_k, \alpha x + \beta y \rangle \end{pmatrix}$$

$$= \begin{pmatrix} \langle a_1, x \rangle \\ \langle a_1, y \rangle \\ \vdots \\ \langle a_k, x \rangle \\ \langle a_k, y \rangle \end{pmatrix} + \beta \begin{pmatrix} \langle a_1, y \rangle \\ \vdots \\ \langle a_k, y \rangle \end{pmatrix}$$

Then L is a linear transformation.

- Bounded linear operators

Let $A: V_1 \rightarrow V_2$ be a linear operator,

where V_1 and V_2 are normed vector spaces.

To define "norm" of linear operators, we need to set up a vector space of linear operators. 由 LO 构成向量空间 —— 我称之为 V_0 .

- $\forall A, B: V_1 \rightarrow V_2$ linear, define $A+B$ by

$$(A+B)(x) = A(x) + B(x) \quad \forall x \in V_1$$

We can check $A+B: V_1 \rightarrow V_2$ is a linear operator also

$$\begin{aligned} (A+B)(\alpha x + \beta y) &= A(\alpha x + \beta y) + B(\alpha x + \beta y) \\ &\stackrel{\text{since } A \text{ is a linear operator/B is...}}{=} \alpha \cdot A(x) + \beta \cdot A(y) + \alpha \cdot B(x) + \beta \cdot B(y) \\ &= \alpha \cdot (A(x) + B(x)) + \beta \cdot (A(y) + B(y)) \\ &\stackrel{\textcircled{1}}{=} \alpha \cdot (A+B)(x) + \beta \cdot (A+B)(y) \end{aligned}$$

- $\forall \alpha \in \mathbb{R}$ and $\forall A: V_1 \rightarrow V_2$ linear, define $\alpha \cdot A$ by

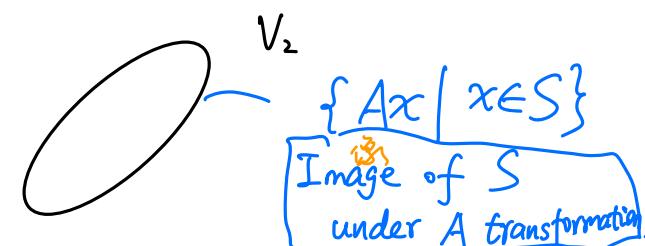
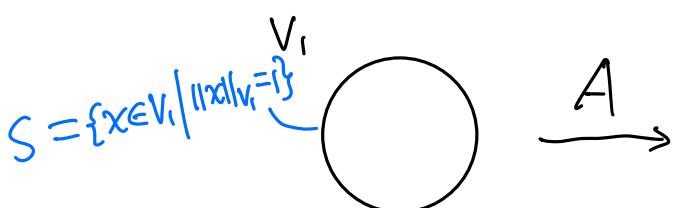
$$(\alpha \cdot A)(x) = \alpha \cdot A(x) \quad \forall x \in V_1$$

Then $\alpha \cdot A: V_1 \rightarrow V_2$ (we can check $\alpha \cdot A$ is a linear operator)

We define norm of a linear operators $V_1 \rightarrow V_2$ to describe how "large" a linear operator is.

Let $A: V_1 \rightarrow V_2$ be a linear operator, and

V_1, V_2 are normed vector spaces.



Define

$$\|A\| = \sup_{\|x\|_{V_1}=1} \|Ax\|_{V_2} = \sup_{\substack{x \neq 0 \\ x \in V_1}} \frac{\|Ax\|_{V_2}}{\|x\|_{V_1}}$$

Indeed, $\|A\|$ is a norm of A .

Proof:

① $\|A\| \geq 0$ obvious.

by definition of norm at Pg - ①

$$\begin{aligned} \|A\| = 0 &\iff \sup_{\|x\|_{V_1}=1} \|Ax\|_{V_2} = 0 \iff Ax = 0 \quad \forall x: \|x\|_{V_1}=1 \\ &\iff A \cdot \frac{y}{\|y\|_{V_1}} = 0 \quad \forall y \neq 0, y \in V_1 \\ &\iff A y = 0 \quad \forall y \neq 0, y \in V_1 \\ &\iff A y = 0 \quad \forall y \in V_1 \quad \text{i.e., } A \text{ is a zero operator} \\ &\iff A = 0 \quad \text{i.e., it maps } V_1 \rightarrow 0 \end{aligned}$$

(*) in Pg

1. $\|\cdot\|_{V_1}$ is a norm by defn

(*) in Pg

$\|(\alpha A)(x)\|_{V_2} = \sup_{\|x\|_{V_1}=1} \|\alpha A(x)\|_{V_2}$ by def of norm Pg - ②

$= |\alpha| \cdot \sup_{\|x\|_{V_1}=1} \|A(x)\|_{V_2} = |\alpha| \cdot \|A\|$ by property of "sup"

$$\begin{aligned} ② \quad \|\alpha A\| &= \sup_{\|x\|_{V_1}=1} \|(\alpha A)(x)\|_{V_2} \leq \sup_{\|x\|_{V_1}=1} (|\alpha| \|A(x)\|_{V_2}) \\ &= |\alpha| \cdot \sup_{\|x\|_{V_1}=1} \|A(x)\|_{V_2} = |\alpha| \cdot \|A\| \end{aligned}$$

$$\begin{aligned} ③ \quad \|A+B\| &= \sup_{\|x\|_{V_1}=1} \|(A+B)(x)\|_{V_2} = \sup_{\|x\|_{V_1}=1} \|A(x) + B(x)\|_{V_2} \\ &\leq \sup_{\|x\|_{V_1}=1} (\|A(x)\|_{V_2} + \|B(x)\|_{V_2}) \leq \sup_{\|x\|_{V_1}=1} \|A(x)\|_{V_2} + \sup_{\|x\|_{V_1}=1} \|B(x)\|_{V_2} \\ &= \|A\| + \|B\| \end{aligned}$$

④ A, B linear

by def of norm Pg - ③

def of $\|A\|, \|B\|$

Additionally, $\|\cdot\|$ satisfies: consistency.

$$④ \quad \boxed{\|Ax\|_{V_2} \leq \|A\| \cdot \|x\|_{V_1}, \quad \forall x \in V_1}$$

$$\text{Proof. } \|A\| = \sup_{\substack{y \in V_1 \\ y \neq 0}} \frac{\|Ay\|_{V_2}}{\|y\|_{V_1}} \geq \frac{\|Ax\|_{V_2}}{\|x\|_{V_1}} \quad \text{if } x \neq 0$$

$$\Rightarrow \|Ax\|_{V_2} \leq \|A\| \cdot \|x\|_{V_1}$$

If $x=0$, obvious ($\|A\vec{0}\|_{V_2} = \|\vec{0}\| = 0 = \|A\| \cdot \|\vec{0}\|_{V_1}$)

⑤ Let $B: V_1 \rightarrow V_2$ linear
 $A: V_2 \rightarrow V_3$

$$(V_1) \xrightarrow{B} (V_2) \xrightarrow{A} (V_3)$$

Then we define AB by $(AB)(x) = A(Bx) \quad \forall x \in V_1$

(i.e., $AB = A \circ B$)

for 4

$$\begin{aligned}
 & AB(x+\beta y) \\
 &= A(B(x+\beta y)) \\
 &= A(\alpha Bx + \beta (By)) \\
 &= \alpha(AB)x + \beta(AB)y
 \end{aligned}$$

Then $AB: V_1 \rightarrow V_3$ is a linear operator. (prove it!)

We have $\|AB\| \leq \|A\| \cdot \|B\|$.

proof. $\|AB\| = \sup_{\|x\|_{V_1}=1} \|(AB)x\|_{V_3}$ ABx
 $\leq \sup_{\|x\|_{V_1}=1} \|(Ax)(Bx)\|_{V_3}$ def of ||AB||
we just discussed about in last page
 $\leq \sup_{\|x\|_{V_1}=1} \|(Ax)\|_{V_2} \cdot \|(Bx)\|_{V_3}$ property of "sup"
 $= \|A\| \cdot \left(\sup_{\|x\|_{V_1}=1} \|(Bx)\|_{V_2} \right)$
 $= \|A\| \cdot \|B\| \quad \square$

Example 1: $A \in \mathbb{R}^{m \times n}$ is a linear transformation $\mathbb{R}^n \rightarrow \mathbb{R}^m$

$$\|A\| = \sup_{\|x\|_{\mathbb{R}^n}=1} \|Ax\|_{\mathbb{R}^m}$$

- If $\|\cdot\|_{\mathbb{R}^n} = \|\cdot\|_2$
 $\|\cdot\|_{\mathbb{R}^m} = \|\cdot\|_2$, then $\|A\| = \|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$

$$= (\max \text{ eigenvalue of } A^T A)^{\frac{1}{2}}$$

- $\|\cdot\|_1 \dots \|A\|_1$
- $\|\cdot\|_\infty \dots \|A\|_\infty$

Example 2: DFT $F_n: \mathbb{C}^n \rightarrow \mathbb{C}^n$

Define $F_n^H = \overline{F_n^T}$ H: conjugate transpose (共轭转置)

Then, it has been proved $F_n^H \cdot F_n = F_n \cdot F_n^H = I$

(i.e., F_n is unitary) 酉矩阵

(and F_n^H is the inverse discrete Fourier transform (离散傅立叶逆变换))

So, $\|F_n\|_2 = \sup_{\|x\|_2=1} \|F_n x\|_2 = \left(\sup_{\|x\|_2=1} \|F_n x\|_2^2 \right)^{\frac{1}{2}}$

$$= \left(\sup_{\|x\|_2=1} (x^H F_n^H F_n x) \right)^{\frac{1}{2}} = \left(\sup_{\|x\|_2=1} (x^H x) \right)^{\frac{1}{2}}$$

$$= \left(\sup_{\|x\|_2=1} \|x\|_2^2 \right)^{\frac{1}{2}} = 1$$

For $z \in \mathbb{C}^n$,

$$\|z\|_2^2 = z^H z$$

$$\|F_n x\|_2 = \dots = \|x\|_2 \quad (?)$$

Example 3: Fourier transform $\mathcal{F} : L^2(\mathbb{R}) \rightarrow L^2(\mathbb{R})$

$$(\mathcal{F}f)(\xi) = \int_{-\infty}^{+\infty} f(x) e^{-2\pi i \xi x} dx$$

(不直观的) Plancherel identity: $\|\mathcal{F}f\|_2 = \|f\|_2$

"the fourier transform doesn't change the 'size' of function"

Then $\|\mathcal{F}\| = \sup_{\|f\|_2=1} \|\mathcal{F}f\|_2 = 1$

Example 4: Let $T : C[-1, 1] \rightarrow C[0, 2]$ defined by

$$(Tf)(t) = f(t-1) \quad \forall t \in [0, 2]$$

... (简单地理解为 | 简单地理解为 intuitive intuive)
不考虑范数。 虽然 T 没有改变范数 norm.

$$\|T\| = 1$$

Example 5: Let $V_1 = \{f \mid f \text{ and } f' \text{ are continuous on } [0, 1]\}$

$$V_2 = \{f \mid f \text{ is continuous on } [0, 1]\}$$

The differentiation operator $D : V_1 \rightarrow V_2$ defined by

$$Df = f'$$

On, V_1 and V_2 , we use $\|\cdot\|_\infty$ (i.e., $\|f\|_\infty = \max_{t \in [0, 1]} |f(t)|$)

Let's calculate $\|D\|$

Consider $f_k(t) = \sin(2\pi kt)$, $t \in [0, 1]$ 包含了所有整数倍 $k=1, 2, \dots$
 K is an integer

Then $(Df_k)(t) = f'_k(t) = 2\pi k \cos(2\pi kt)$

$$\text{So, } \|f_k\|_\infty = 1, \quad \|Df_k\|_\infty = 2\pi k$$

So,

$$\|D\| \geq \lim_{k \rightarrow \infty} \|Df_k\|_\infty = \lim_{k \rightarrow \infty} 2\pi k = +\infty \quad \text{i.e. } D \text{ is unbounded.}$$

Bounded linear operators $V_1 \rightarrow V_2$, where V_1, V_2 are normed vector spaces.

$$\mathcal{L}(V_1, V_2) = \{ A \mid A \text{ is a linear operator } V_1 \rightarrow V_2 \text{ with } \|A\| < +\infty \}$$

(All linear and bounded operators $V_1 \rightarrow V_2$)

— $\mathcal{L}(V_1, V_2)$ is closed under "+" and ". ", because

• If $A, B \in \mathcal{L}(V_1, V_2)$, then $\|A\| < \infty$, $\|B\| < \infty$, and

$$\|A+B\| \leq \|A\| + \|B\| < \infty \quad \text{if } \|\cdot\|_{V_0} \text{ is norm. 由门泰原定理}$$

$$\text{So, } A+B \in \mathcal{L}(V_1, V_2)$$

since: ② $\mathcal{L}(V_1, V_2)$ is closed under "+" and "·".

① $\mathcal{L}(V_1, V_2)$ is a vector space
因为 $\mathcal{L}(V_1, V_2) \subset \mathcal{L}(V_0, \text{given by defn})$ 且 $\mathcal{L}(V_0, \text{given by defn})$ 是向量空间.

因此 $\mathcal{L}(V_1, V_2)$ 也是向量空间. 由 ①+② 得证.

and since we have a norm on it. Thus,

$\mathcal{L}(V_1, V_2)$ is a normed vector space

Further, if V_1, V_2 are Banach spaces,

then $\mathcal{L}(V_1, V_2)$ is a Banach space.

Some special cases:

$$-\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m) = \mathbb{R}^{m \times n} \quad (\text{if } \mathbb{R}^n \text{ is normed by } \|\cdot\|_1)$$

$$-\mathcal{L}(V_1, V_2) = \{A \mid A \text{ is a linear operator } V_1 \rightarrow V_2\}$$

if V_1, V_2 are finite-dimensional normed vector spaces

11.4
11.11

Bounded Linear operators on Hilbert Spaces

Let $(V_1, \langle \cdot, \cdot \rangle_{V_1}), (V_2, \langle \cdot, \cdot \rangle_{V_2})$ be two Hilbert spaces

Let $A \in \mathcal{L}(V_1, V_2)$

We can define the adjoint operator of A , denoted by A^* , by

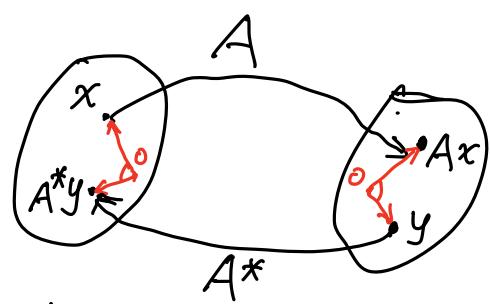
$$\textcircled{1} \quad A^* \in \mathcal{L}(V_2, V_1) \quad \text{伴隨算子}$$

$$\textcircled{2} \quad \forall x \in V_1, y \in V_2$$

$$\langle Ax, y \rangle_{V_2} = \langle x, A^*y \rangle_{V_1}$$

且 A 是线性且有界的算子 \Rightarrow A^* 也是线性且有界的算子

with standard innerprod $\langle \cdot, \cdot \rangle$



Example 1: Consider $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m) = \mathbb{R}^{m \times n}$

$$\forall x \in \mathbb{R}^n, y \in \mathbb{R}^m$$

All linear (and bounded) operators maps $x \in \mathbb{R}^n$ to \mathbb{R}^m .
omitted since $\mathbb{R}^n, \mathbb{R}^m$ are both finite-dim and thus bounded.

$$\langle Ax, y \rangle_{\mathbb{R}^m} = y^T A x = (y^T A) x = (A^T y)^T x = \langle x, A^T y \rangle_{\mathbb{R}^n}$$

and $A^T \in \mathbb{R}^{n \times m} = \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n)$

$$\text{So, } \boxed{A^* = A^T}$$

The adjoint is an extension of the matrix transpose. *(f is generalization)*

Example 2: Let V be a Hilbert space.

Consider $\mathcal{L}(V, \mathbb{R})$, i.e., linear and bounded functions on V .

Let $f \in \mathcal{L}(V, \mathbb{R})$. Then $f(x) = \langle a, x \rangle_V$ for some $a \in V$.

now We find $f^* \in \mathcal{L}(\mathbb{R}, V)$

For any $x \in V, y \in \mathbb{R}$

$$\langle f(x), y \rangle_{\mathbb{R}} = y \cdot f(x) = y \langle a, x \rangle_V = y \langle x, a \rangle_V = \langle x, y \cdot a \rangle_V$$

Define $g(y) = y \cdot a$. Then $g: \mathbb{R} \rightarrow V$

linear — g is linear (since $g(dx+by) = \dots$)

bounded — $\|g\| = \sup_{|y|=1} \|y \cdot a\| = \sup_{|y|=1} (|y| \|a\|) = \|a\| < +\infty$

So, $g \in \mathcal{L}(\mathbb{R}, V)$

Thus, $f^* = g$, i.e., $f^*(y) = y \cdot a$

Example 3: Let $a = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} \in \mathbb{R}^n$. The circular convolution with kernel a ,

$a \otimes \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, is defined by:

$$\forall x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{pmatrix} \in \mathbb{R}^n \quad [a \otimes x]_k = \sum_{i=0}^{n-1} a_i x_{(k-i) \bmod n}$$

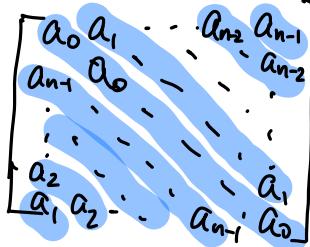
$k = 0, 1, \dots, n-1$

Since $a \otimes \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) = \mathbb{R}^{n \times n}$, it can be written as a matrix prod.

$$a \otimes x = \underbrace{\begin{bmatrix} a_0 & a_{n-1} & \cdots & a_2 & a_1 \\ a_1 & a_0 & a_{n-1} & \cdots & a_2 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_{n-1} & \cdots & \cdots & a_1 & a_0 \end{bmatrix}}_{\text{def } C_a} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} \equiv C_a x, \text{ where } C_a \in \mathbb{R}^{n \times n} \text{ is a circulant matrix with the first column } a.$$

So, the adjoint $(a \otimes)^*$ of $a \otimes$ is

$$\forall y \in \mathbb{R}^n, (a \otimes)^* y = C_a^T y = \begin{bmatrix} a_0 & a_1 & \dots & a_{n-2} & a_{n-1} \\ a_{n-1} & a_0 & \dots & a_{n-3} & a_{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_2 & a_1 & \dots & a_{n-1} & a_0 \\ a_1 & a_2 & \dots & a_0 & a_1 \end{bmatrix} y = \begin{pmatrix} a_0 \\ a_{n-1} \\ \vdots \\ a_2 \\ a_1 \end{pmatrix} \otimes y$$



$= \begin{pmatrix} a_0 \\ a_{-1} \bmod n \\ a_{-2} \bmod n \\ \vdots \\ a_{-(n-2)} \bmod n \\ a_{-(n-1)} \bmod n \end{pmatrix} \otimes y = a(-\cdot) \otimes y$

denoted by $a(-\cdot)$, the flipped a .

Thus, $(a \otimes)^* = a(-\cdot) \otimes$ — flipped convolution.

Example 5: Continuous convolution with a kernel function $g: \mathbb{R} \rightarrow \mathbb{R}$

4

$\overbrace{C_g: L^2(\mathbb{R}) \rightarrow L^2(\mathbb{R})}$

is defined by

$$(C_g f)(x) = \int_{-\infty}^{+\infty} g(x-y) f(y) dy$$

+
what we want to see in a convolution?

we have showed C_g is linear
@ P93 - ex8

Then $C_g \in L(L^2(\mathbb{R}), L^2(\mathbb{R}))$ if g is smooth enough:

$$\|C_g f\|_2 \leq \dots \leq (\text{constant of } g) \cdot \|f\|_2$$

Young's inequality \rightarrow (Skipped. 只用到 Young's inequality).

Now we try to find $C_g^* \in L^2(\mathbb{R})$

$\forall f \in L^2(\mathbb{R}), h \in L^2(\mathbb{R})$

$$\begin{aligned} \langle C_g f, h \rangle_{L^2(\mathbb{R})} &= \int_{-\infty}^{+\infty} (C_g f)(x) \cdot h(x) dx = \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} g(x-y) f(y) dy \right) h(x) dx \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x-y) f(y) h(x) dx dy \\ &= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} g(x-y) h(x) dx \right) \cdot f(y) dy \\ &= \int_{-\infty}^{+\infty} (C_{g(-\cdot)} h)(y) f(y) dy = \langle f, (C_{g(-\cdot)} h) \rangle_{L^2(\mathbb{R})} \end{aligned}$$

also "flipping" $g(x-y) \leftrightarrow g(y-x)$, the $\int_{-\infty}^{+\infty} g(y-x) h(x) dx$ is a convolution with kernel g .

here g has been already flipped so we denote it as $g(-\cdot)$ $\times g(-\cdot)$ is FTS.

So, $C_g^* = C_{g(-\cdot)}$

i.e., the adjoint of convolution with kernel g is
the convolution with kernel $g(-\cdot)$ (flipping of g)

Example 5: Let $T: L^2(-1,1) \rightarrow L^2(0,2)$ by translation

$$\forall f \in L^2(-1,1) \quad (Tf)(t) = f(t-1) \quad \forall t \in (0,2)$$

Then • T is linear

• T is bounded

$$\|T\| = \sup_{\|f\|_2=1} \|Tf\|_2 = \sup_{\|f\|_2=1} \|f\|_2 = \text{and } \|T\| = \sup_{\|f\|_2=1} \|f\|_2 = \text{respectively.}$$

$$\Rightarrow T \in \mathcal{L}(L^2(-1,1), L^2(0,2))$$

Now we try to find $T^*: \mathcal{L}(L^2(0,2), L^2(-1,1))$

$$\begin{aligned} \langle Tf, g \rangle & \xrightarrow{\forall g \in L^2(0,2)} = \int_0^2 (Tf)(t) \cdot g(t) dt = \int_0^2 f(t-1) \cdot g(t) dt \\ & \xleftarrow{\substack{s=t-1 \\ \therefore t=s+1}} \int_{-1}^1 f(s) g(s+1) ds = \langle f, \tilde{T}g \rangle, \end{aligned}$$

$$\text{where } (\tilde{T}g)(s) = g(s+1)$$

— \tilde{T} is linear

— \tilde{T} is bounded because

$$\|\tilde{T}\| = \sup_{\|g\|_2=1} \|\tilde{T}g\|_2 = \sup_{\|g\|_2=1} \|g\|_2 = 1$$

$$\tilde{T} \in \mathcal{L}(L^2(0,2), L^2(-1,1))$$

$$\text{So, } T^* = \tilde{T}$$

Example 6: Let $a_1, \dots, a_k \in H$, where H is a Hilbert space.

Define $L: H \rightarrow \mathbb{R}^k$ by

$$L(x) = \begin{pmatrix} \langle a_1, x \rangle \\ \langle a_2, x \rangle \\ \vdots \\ \langle a_k, x \rangle \end{pmatrix} \in \mathbb{R}^k$$

Then $L \in \mathcal{L}(H, \mathbb{R}^k)$, because

— L is linear

L is bounded, because

$$\|L\| = \sup_{\|x\|=1} \|Lx\|_2 = \sup_{\|x\|=1} \left(\sum_{i=1}^k \langle a_i, x \rangle^2 \right)^{\frac{1}{2}}$$

Cauchy-Schwartz $\leq \sup_{\|x\|=1} \left(\sum_{i=1}^k \|a_i\|^2 \|x\|^2 \right)^{\frac{1}{2}} = \left(\sum_{i=1}^k \|a_i\|^2 \right)^{\frac{1}{2}} < +\infty$

Let's find L^* :

$\forall x \in H, u \in \mathbb{R}^k$,

$$\langle Lx, u \rangle_{\mathbb{R}^k} = \sum_{i=1}^k (\langle x, a_i \rangle \cdot u_i) = \left\langle x, \sum_{i=1}^k u_i \cdot a_i \right\rangle_H$$

Define $S: \mathbb{R}^k \rightarrow H$ by $S(u) = \sum_{i=1}^k u_i \cdot a_i$

Then

- S is linear (obvious)

- S is bounded because

$$\begin{aligned} \|S\| &= \sup_{\|u\|_2=1} \|Su\|_H = \sup_{\|u\|_2=1} \left\| \sum_{i=1}^k u_i a_i \right\|_H \leq \sup_{\|u\|_2=1} \left(\sum_{i=1}^k |u_i| \cdot \|a_i\|_H \right) \\ &= \sup_{\|u\|_2=1} \left\langle \begin{pmatrix} |u_1| \\ |u_2| \\ \vdots \\ |u_k| \end{pmatrix}, \begin{pmatrix} \|a_1\|_H \\ \|a_2\|_H \\ \vdots \\ \|a_k\|_H \end{pmatrix} \right\rangle_{\mathbb{R}^k} \stackrel{\text{由 } \langle \cdot, \cdot \rangle_H \text{ 为内积}}{\leq} \left(\sum_{i=1}^k |u_i|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^k \|a_i\|_H^2 \right)^{\frac{1}{2}} \\ &\quad \leq \left(\sum_{i=1}^k |u_i|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^k \|a_i\|_H^2 \right)^{\frac{1}{2}} = \left(\sum_{i=1}^k \|a_i\|_H^2 \right)^{\frac{1}{2}} < +\infty \end{aligned}$$

So, $S \in \mathcal{L}(\mathbb{R}^k, H)$

Thus, $L^* = S$

§ 5.2. Differentiation of operators

Let V_1, V_2 be two Banach spaces with $\|\cdot\|_{V_1}, \|\cdot\|_{V_2}$

Let $F: V_1 \rightarrow V_2$ be a mapping (not necessarily linear)

Then at $x^{(0)} \in V_1$, we find an affine approximation to F and bounded

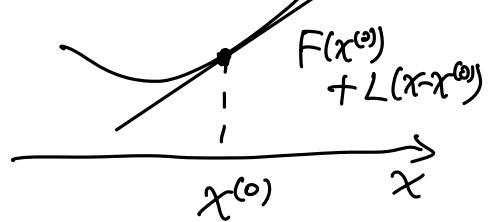
At $x^{(0)} \in V_1$, the affine and bounded approximation passing through

F

$(x^{(0)}, F(x^{(0)}))$ is

$$F(x) \approx F(x^{(0)}) + L(x - x^{(0)}),$$

where $L \in \mathcal{L}(V_1, V_2)$



If this approximation has an error $O(\|x - x^{(0)}\|_{V_1})$, then we call F is differentiable at $x^{(0)}$, and L is its differentiation at $x^{(0)}$, denoted by $DF(x^{(0)}) = L$

Definition: $F: V_1 \rightarrow V_2$ is differentiable at $x^{(0)} \in V_1$ if

$$\lim_{\|x - x^{(0)}\|_{V_1} \rightarrow 0} \frac{\|F(x) - (F(x^{(0)}) + L(x - x^{(0)}))\|_{V_2}}{\|x - x^{(0)}\|_{V_1}} = 0$$

for some $L \in \mathcal{L}(V_1, V_2)$.

In this case, L is called the differentiation of F at $x^{(0)}$, denoted by $DF(x^{(0)}) = L$.

$F(x^{(0)})$ 在 \mathbb{R} 上可导 / "可微"

Example 1: If $f: V \rightarrow \mathbb{R}$, where V is a Hilbert space, is differentiable at $x^{(0)}$, then,

$$\lim_{\|x - x^{(0)}\| \rightarrow 0} \frac{|f(x) - (f(x^{(0)}) + \langle \nabla f(x^{(0)}), x - x^{(0)} \rangle_H)|}{\|x - x^{(0)}\|} = 0$$

So, $Df(x^{(0)}) \in \mathcal{L}(V, \mathbb{R})$ and

$$Df(x^{(0)})(y) = \langle \nabla f(x^{(0)}), y \rangle \quad \forall y \in V$$

Example 2: Let $A \in \mathcal{L}(V_1, V_2)$. Then,

$$DA(x^{(0)}) = A \quad \forall x^{(0)} \in V_1$$

$$\begin{aligned} A(x) &= A(x + x^{(0)} - x^{(0)}) \\ &\approx Ax^{(0)} + A(x - x^{(0)}) \end{aligned}$$

(intuitively)

Example 3: Let $F_1, \dots, F_k : H \rightarrow \mathbb{R}$ be k differentiable functions on a Hilbert space H . Let $F : H \rightarrow \mathbb{R}^k$ by

$$F(x) = \begin{bmatrix} F_1(x) \\ F_2(x) \\ \vdots \\ F_k(x) \end{bmatrix} \in \mathbb{R}^k$$

Let's find $DF(x)$.

Since F_i is differentiable at $\forall x \in H$

所以 $F_i(y) = F_i(x) + \langle \nabla F_i(x), y-x \rangle + R_i(x, y)$
 and $\lim_{\|y-x\|_H \rightarrow 0} \frac{|R_i(x, y)|}{\|y-x\|_H} = 0$, $i=1, 2, \dots, k$.

Then,

$$F(y) = F(x) + \begin{bmatrix} \langle \nabla F_1(x), y-x \rangle \\ \langle \nabla F_2(x), y-x \rangle \\ \vdots \\ \langle \nabla F_k(x), y-x \rangle \end{bmatrix} + \begin{bmatrix} R_1(x, y) \\ R_2(x, y) \\ \vdots \\ R_k(x, y) \end{bmatrix}$$

is a linear operator applied to $y-x$, denoted by $L(y-x)$, where $L \in L(H, \mathbb{R}^k)$

• $R(x, y)$ satisfies

$$\|R(x, y)\|_2 \leq \|R(x, y)\|_1 = \sum_{i=1}^k |R_i(x, y)|$$

$$0 \leq \lim_{\|y-x\|_H \rightarrow 0} \frac{\|R(x, y)\|_2}{\|y-x\|_H} \leq \lim_{\|y-x\|_H \rightarrow 0} \frac{\sum_{i=1}^k |R_i(x, y)|}{\|y-x\|_H} = \sum_{i=1}^k \left(\lim_{\|y-x\|_H \rightarrow 0} \frac{|R_i(x, y)|}{\|y-x\|_H} \right) = 0$$

$$\Rightarrow \lim_{\|y-x\|_H \rightarrow 0} \frac{\|R(x, y)\|_2}{\|y-x\|_H} = 0$$

So,

$$DF(x)(y) = \begin{bmatrix} \langle \nabla F_1(x), y \rangle \\ \langle \nabla F_2(x), y \rangle \\ \vdots \\ \langle \nabla F_k(x), y \rangle \end{bmatrix} \quad \forall y \in H. \quad (\star-1)$$

(按需求 D).

an remainder

$$DF(x) \in \mathcal{L}(H, \mathbb{R}^k)$$

— A special case: $H = \mathbb{R}^n$. Then $F: \mathbb{R}^n \rightarrow \mathbb{R}^k$
and $F_i: \mathbb{R}^n \rightarrow \mathbb{R}$

So, since $\nabla F_i(x) = \begin{pmatrix} \frac{\partial F_i}{\partial x_1}(x) \\ \frac{\partial F_i}{\partial x_2}(x) \\ \vdots \\ \frac{\partial F_i}{\partial x_n}(x) \end{pmatrix} \in \mathbb{R}^n$ (*-2)

$\left. \begin{array}{l} \text{Then} \\ DF(x) \\ \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^k) \\ = \mathbb{R}^{k \times n} \end{array} \right\}$

Thus,

$DF(x)(y) = \begin{bmatrix} \nabla F_1(x)^T y \\ \nabla F_2(x)^T y \\ \vdots \\ \nabla F_k(x)^T y \end{bmatrix} = \begin{bmatrix} \nabla F(x)^T \\ \nabla F_2(x)^T \\ \vdots \\ \nabla F_k(x)^T \end{bmatrix} y$

注意到 $()$ 因为 $DF(x)(y)$ 可以看成 \mathbb{R}^n 空间中 y 的线性组合，所以其算子就是该矩阵的转置。*

So,

$DF(x) = \begin{bmatrix} \nabla F_1(x)^T \\ \nabla F_2(x)^T \\ \vdots \\ \nabla F_k(x)^T \end{bmatrix} = \begin{bmatrix} \frac{\partial F_1}{\partial x_1}(x) & \frac{\partial F_1}{\partial x_2}(x) & \cdots & \frac{\partial F_1}{\partial x_n}(x) \\ \frac{\partial F_2}{\partial x_1}(x) & \frac{\partial F_2}{\partial x_2}(x) & \cdots & \frac{\partial F_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_k}{\partial x_1}(x) & \frac{\partial F_k}{\partial x_2}(x) & \cdots & \frac{\partial F_k}{\partial x_n}(x) \end{bmatrix} \in \mathbb{R}^{k \times n}$

$\in \mathbb{R}^{k \times n}$, so, as an linear operator.

Jacobian Matrix

(Jacobian matrix is the differentiation of $F: \mathbb{R}^n \rightarrow \mathbb{R}^k$ where $\mathbb{R}^n, \mathbb{R}^k$ are with the standard inner product)

Chain rule: Let $F: V_1 \rightarrow V_2$ and $G: V_2 \rightarrow V_3$
then $G \circ F: V_1 \rightarrow V_3$ and

$D(G \circ F) = DG \circ DF$

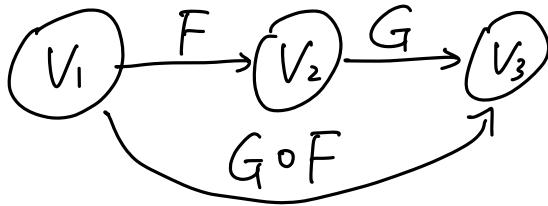
$$D(G \circ F)(x) = DG(F(x)) \circ DF(x)$$

$\in L(V_1, V_3)$

$\in L(V_2, V_3)$

$\in L(V_1, V_2)$

if G and F are differentiable at $F(x)$ and x respectively.



proof: Similar to the chain rule before.

Example 4: $f(x) = f_1(x) \cdot f_2(x)$, where $f_1, f_2, f: V \rightarrow \mathbb{R}$ (V is a Hilbert space)

Define $F: V \rightarrow \mathbb{R}^2$, by $F(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix}$

$G: \mathbb{R}^2 \rightarrow \mathbb{R}$ by $G\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \alpha_1 \cdot \alpha_2$

Then $f(x) = G(F(x))$

By the chain rule,

$$Df(x) = D(G \circ F)(x) = DG(F(x)) \circ DF(x)$$

- For $DG(x)$: Since $G: \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$\nabla G(\alpha) = \begin{pmatrix} \frac{\partial G}{\partial \alpha_1}(\alpha) \\ \frac{\partial G}{\partial \alpha_2}(\alpha) \end{pmatrix} = \begin{pmatrix} \alpha_2 \\ \alpha_1 \end{pmatrix} \Rightarrow \nabla G(F(x)) = \begin{pmatrix} f_2(x) \\ f_1(x) \end{pmatrix}$$

So, $DG(\alpha)(\beta) = \langle \nabla G(\alpha), \beta \rangle$ (From Example 1) $\forall \beta \in \mathbb{R}^2$

由函数
和函数的映射是单射
(为什么? D 是双射)
导数的对称性
 $\nabla G(\alpha) = (\alpha_2, \alpha_1)$

- For $DF(x)$: From Example 2,

$$DF(x)(z) = \begin{bmatrix} \langle \nabla f_1(x), z \rangle \\ \langle \nabla f_2(x), z \rangle \end{bmatrix} \quad \forall z \in V$$

Hilbert Space

Thus,

$$\begin{aligned} \forall y \in V, \quad Df(x)(y) &= \left(DG(F(x)) \circ DF(x) \right)(y) \\ &= DG(F(x)) \left(DF(x)(y) \right) \\ &\stackrel{\text{exp-2}}{=} DG(F(x)) \left(\langle \nabla f_1(x), y \rangle \right) \\ &\stackrel{\text{exp-1}}{=} \left\langle \begin{pmatrix} f_2(x) \\ f_1(x) \end{pmatrix}, \begin{pmatrix} \nabla f_1(x), y \\ \nabla f_2(x), y \end{pmatrix} \right\rangle = \boxed{H \text{为} \nabla f \text{的梯度, } D = H \text{ 为} \nabla f \text{ 的微分} \Rightarrow \text{由定理 19.80}} \\ &= f_2(x) \cdot \langle \nabla f_1(x), y \rangle + f_1(x) \cdot \langle \nabla f_2(x), y \rangle \\ &= \left\langle f_2(x) \cdot \nabla f_1(x) + f_1(x) \cdot \nabla f_2(x), y \right\rangle \end{aligned}$$

So, by Example 1,

$$\boxed{\nabla f(x) = f_1(x) \cdot \nabla f_2(x) + f_2(x) \cdot \nabla f_1(x)}$$

Example 5: Let $A \in \mathcal{L}(V, W)$, where V, W are Hilbert spaces

Define $f: V \rightarrow \mathbb{R}$ by

$$f(x) = \frac{1}{2} \|Ax - b\|_W^2 \quad \text{where } b \in W$$

Let's find $Df(x)$ and $\nabla f(x)$

Define $F(x) = Ax$ So, $F \in \mathcal{L}(V, W)$

$$G(y) = \frac{1}{2} \|y - b\|_W^2 \quad G: W \rightarrow \mathbb{R}$$

Then $G \circ F: V \rightarrow \mathbb{R}$ and $G(F(x)) = f(x)$.

By the chain rule,

$$\begin{aligned} \forall y \in V, \quad D(G \circ F)(x)(y) &= \left(DG(F(x)) \circ DF(x) \right)(y) \\ &= DG(F(x)) \left(DF(x)(y) \right) \end{aligned}$$

by example-2
— For DF : $DF(x) = A$, since $F \in \mathcal{L}(V, W)$

by example-1
— For DG : $DG(u)(v) = \langle \nabla G(u), v \rangle_V$ $\forall v \in V$

$$\begin{aligned} &\stackrel{\text{由例 17-2 时得证}}{=} \langle \nabla G(u), v \rangle_V \\ &\stackrel{\text{即 } \nabla G(u) = \nabla f(u)}{=} \langle u - b, v \rangle_V \\ &\stackrel{\text{即 } \nabla f(u) = u - b}{=} \langle u - b, v \rangle_V \end{aligned}$$

$$\forall y \in V,$$

$Df(x)(y) = D(G \circ F)(x)(y) = \langle Ax - b, Ay \rangle_v$
伴隨定義 P98
 $= \langle A^*(Ax - b), y \rangle_w$

and $\boxed{\nabla f(x) = A^*(Ax - b)}$

In particular, if $V = \mathbb{R}^n$, $W = \mathbb{R}^m$, then $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$

P99-例題 example-1

$$f(x) = \frac{1}{2} \|Ax - b\|_2^2, \quad \text{so, } \boxed{\nabla f(x) = A^T(Ax - b)}$$

Rules for differentiation.

① Linearity

Let $F, G: V_1 \rightarrow V_2$

$$\boxed{D(\alpha F + \beta G) = \alpha DF + \beta DG}$$

Then $D(\alpha F + \beta G)(x) = \alpha \cdot DF(x) + \beta \cdot DG(x)$

$\alpha, \beta \in \mathbb{R}$
 $x \in V$

if $DF(x)$ and $DG(x)$ exists.

② Chain Rule:

Let $F: V_1 \rightarrow V_2$, $G: V_2 \rightarrow V_3$.

Then $D(G \circ F)(x) = DG(F(x)) \circ DF(x)$

$$\boxed{D(G \circ F) = DG \circ DF}$$

Special cases:

- $V_2 = V_3 = \mathbb{R}$, $V_1 = V$

RP $F: V \rightarrow \mathbb{R}$, $G: \mathbb{R} \rightarrow \mathbb{R}$, and $G \circ F: V \rightarrow \mathbb{R}$

So, $D(G \circ F)(x)(y) = DG(F(x))(DF(x)(y))$

$$= G'(F(x)) \cdot DF(x)(y)$$

$$= G'(F(x)) \cdot \langle \nabla F(x), y \rangle = \langle G'(F(x)) \cdot \nabla F(x), y \rangle$$

這個結果是
Hilbert space.
若 V 是 Hilbert space, $F: V \rightarrow \mathbb{R}$,
且 F 在 V 上可微, 那麼 F 在 V 上的導數
是 V 到 \mathbb{R} 的線性映射。

Thus, $\boxed{\nabla(G \circ F)(x) = G'(F(x)) \cdot \nabla F(x)}$