MSBD5004 Math Methods for Data Analysis

- Introduce Math Tools for machine learning
- Supervised Learning (Regression)

Given a training data set $\{(x_i, y_i)\}_{i=1,2,\cdots,m}$ (input x_i, label y_i)

$x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$

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

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

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:

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

    So $L: \mathcal{F} \to \mathbb{R}$ — function of function / functional
- $\min_{f \in \mathcal{F}} L(f)$ — 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:

Addition $+ : V \times V \to V$ (i.e., $\forall x, y \in V$, $x+y \in V$)

Scalar multiplication $\cdot : \mathbb{R} \times V \to V$ (i.e., $\forall \alpha \in \mathbb{R}, x \in V$, $\alpha \cdot x \in V$)

that satisfy the following:

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

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

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

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

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

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

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

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

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 $\{v_1, \cdots, v_p\} \subseteq V$, a linear combination of $\{v_1, \cdots, v_p\}$ is a vector $w \in V$ in the form of

$$w = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_p v_p,$$

where $\alpha_1, \cdots \alpha_p \in \mathbb{R}$ are coefficients.

- linear dependence / independence:

  A set of vectors $\{v_1, \cdots, v_p\} \subseteq V$ are linearly independent

  $$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_p v_p = 0 \iff \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$$

- 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 is closed under "+" and "$\cdot$")

  — $\{0\}$ is a subspace of V.
  ↑ 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

  $$\text{span}\{S\} = \left\{ \sum_{i=1}^{k} \alpha_i v_i \;\middle|\; \alpha_i \in \mathbb{R},\ v_i \in S,\ k \in \mathbb{N} \right\}$$

  $\mathbb{N}$ — 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

  $$\begin{cases} \text{span}\{B\} = V \text{ and,} \\ \text{The elements in } B \text{ are linearly independent.} \end{cases}$$

  — Every vector space has at least one basis.

  — All bases of the same vector space has the same cardinality.

  — The dimension of V is

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

- Examples of vector spaces

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

$+$: standard number addition

$\cdot$: $\quad\quad\quad\quad$ multiplication

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

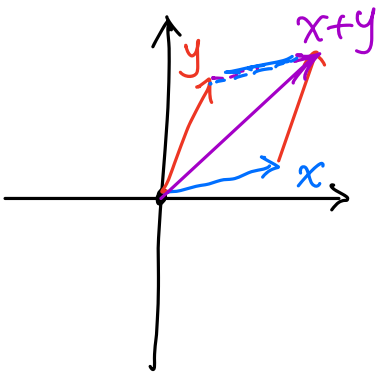
Ex. 2: $\mathbb{R}^n$ is a vector space with

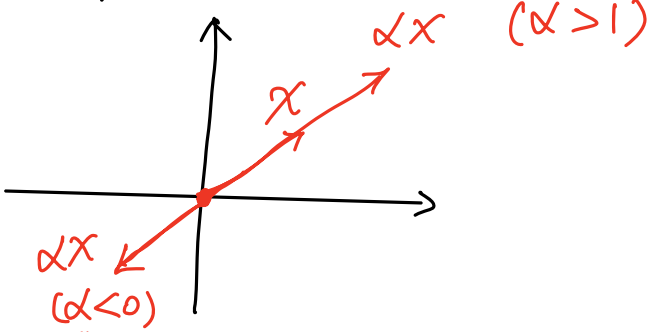
$+$: $\forall\ x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $\quad x+y = \begin{pmatrix} x_1+y_1 \\ \vdots \\ x_n+y_n \end{pmatrix}$

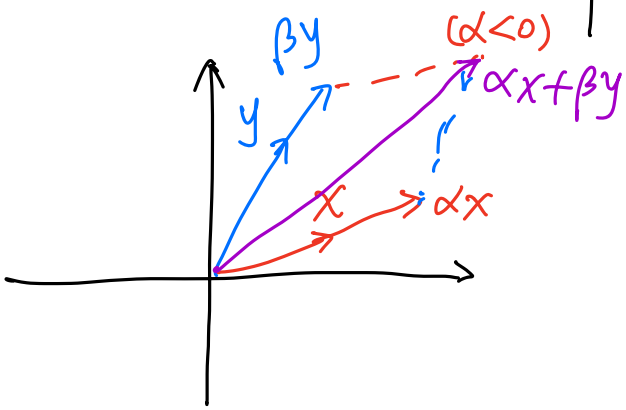
$\cdot$: $\forall\ \alpha \in \mathbb{R}$, $x \in \mathbb{R}^n$ $\quad \alpha \cdot x = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{pmatrix}$

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

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







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

- $\{e_1, e_2, \cdots, e_n\}$ 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

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

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

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

is a vector space.

In this vectors

— 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 $E_{ij} = \begin{bmatrix} & & \\ & 1 & \\ & & \end{bmatrix}$ ((i,j)-entry; all other entries are 0.)

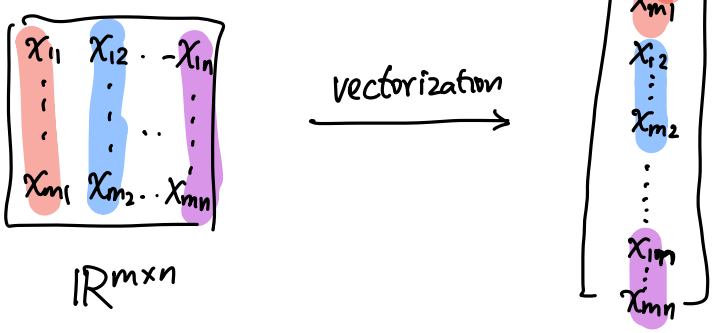
is a basis of $\mathbb{R}^{m\times n}$

— $\dim(\mathbb{R}^{m\times n}) = mn$

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

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

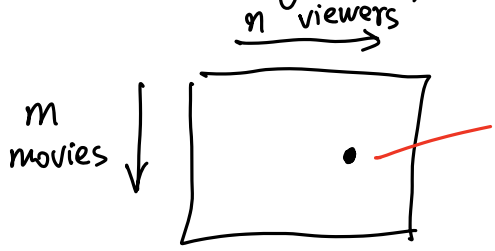
(vectorization)



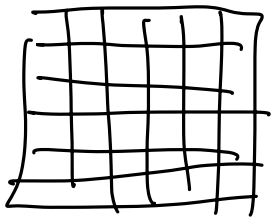
$\mathbb{R}^{mn}$

— $\mathbb{R}^{m\times n}$ can be used to represent

- Black-white digital images of resolution $m\times n$ pixels
- In recommender systems, ratings of movies by viewers



- Numerical tabular data of m rows and n column



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

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

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

is a vector space

On this vector space

— $0 = [0]_{i=1,j=1,k=1}^{m\ n\ l}$

— $\{E_{ijk} \mid \begin{matrix} i=1,\cdots,m \\ j=1,\cdots,n \\ k=1,\cdots,l \end{matrix}\}$ is a basis of $\mathbb{R}^{m\times n\times l}$

$E_{ijk}\in\mathbb{R}^{m\times n\times l}$ with only (i,j,k)-entry 1. and others 0.

— $\dim(\mathbb{R}^{m\times n\times l}) = mnl$

— $\mathbb{R}^{m\times n\times l}$ is the same as $\mathbb{R}^{mnl}$

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

- color image with $m\times n$ 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 $C[a,b] = \{ f \mid f \text{ is a continous function on } [a,b] \}$

with

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

define $f+g$ by $(f+g)(t) = f(t) + g(t) \quad \forall t \in [a,b]$

$\cdot : \forall \alpha \in \mathbb{R}$, $f \in C[a,b]$ define $\alpha \cdot f$ by

$(\alpha f)(t) = \alpha \cdot f(t) \quad \forall t \in [a,b]$

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 numbe in $[a,b]$ to 0.

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

$$x_i \longrightarrow \boxed{?} \longrightarrow y_i \quad \text{with} \quad x_i \in [a,b],\ y_i \in \mathbb{R}, \quad i = 1, \cdots, m$$

Ex. 7. The infinite sequence set

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

with:

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

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

It forms a vector space.

- $0 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$
- $\{ e_i \mid i \in \mathbb{N} \}$ is a basis of l_∞
- $\dim(l_\infty) = +\infty$

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

§ 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

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

Shift invariant

Therefore, we only need to define "distance of a vector" to 0

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

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

It should satisfy:

① $\|x\| \geq 0$ $\forall x \in V$ and $\|x\| = 0 \iff x = 0$

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

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

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

① $\|x\| \geq 0$ $\forall x \in V$ and $\|x\| = 0 \iff x = 0$

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

③ $\|x+y\| \leq \|x\| + \|y\|$ $\forall x, y \in V$.
(Triangle inequality)

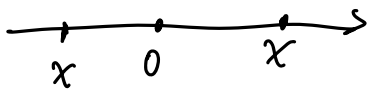
With a norm $\|\cdot\|$,

distance $(x, y) = \|x - y\|$

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

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

Then we can check it is a norm.



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

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)

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$
$$= \left(\sum_{i=1}^n |x_i|^2\right)^{1/2}$$

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

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

- Infinity norm: $\|x\|_\infty = \lim_{p \to +\infty} \|x\|_p$

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

  Can prove ↑

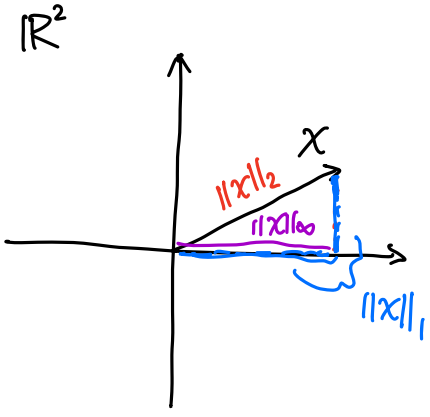
Let $i_0 = \arg\max_{i=1}^n |x_i|$

$$|x_{i_0}| \leq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$$
$$\leq \underbrace{\left(n\,|x_{i_0}|^p\right)^{1/p}}_{n^{1/p}\,|x_{i_0}|}$$

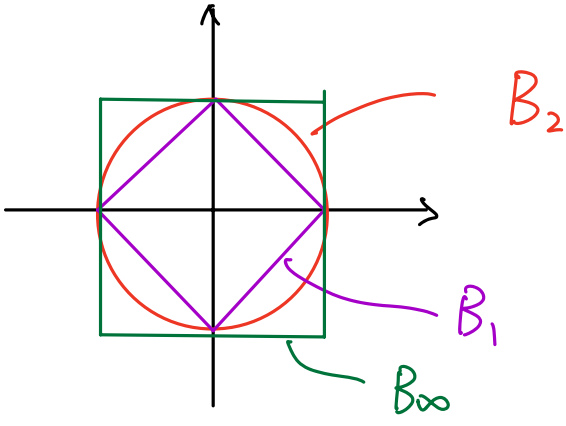
Let $p \to +\infty$

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

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



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, 1-norm minization (LASSO)

—— for vectors following Gaussian distribution, 2-norm minization

—— for vector with ± 1 entries, ∞-norm minization

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

  — $p=1$: $\|A\|_{1,vec} = \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|$

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

  (also known as Frobenius norm, denoted by $\|\cdot\|_F$)

  (i.e, $\|A\|_F = \|A\|_{2,vec}$)

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

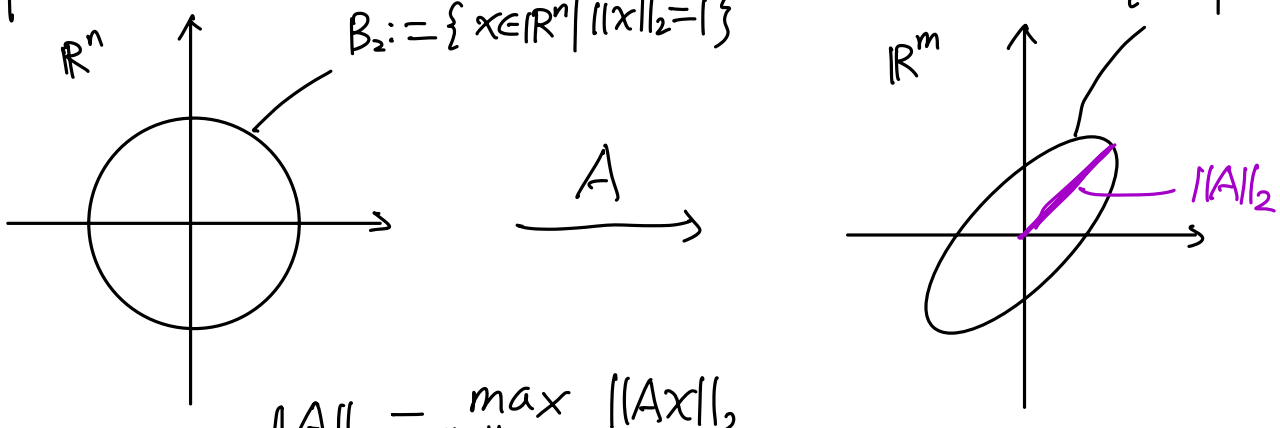
- $\mathbb{R}^{m\times n}$ can be viewed as linear transformations $\mathbb{R}^n \to \mathbb{R}^m$

  we can define the matrix p-norm

$$\|A\|_p = \max_{x\neq 0}\frac{\|Ax\|_p}{\|x\|_p} \overset{\text{can prove}}{=\!=} \max_{x:\|x\|_p=1}\|Ax\|_p$$

$x:\|x\|_p=1 \Leftrightarrow x\in B_p$

— $p=2$:



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

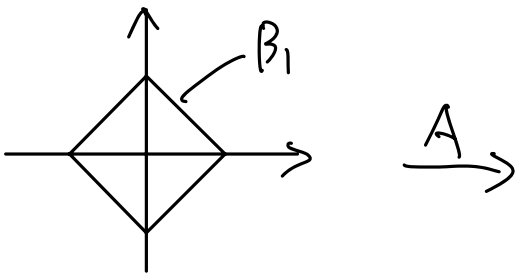
$\Leftrightarrow \|A\|_2^2 = \max_{\|x\|_2^2=1} \|Ax\|_2^2 = \max_{x^Tx=1} x^TA^TAx$ $\quad \boxed{\|x\|_2^2 = x^Tx}$

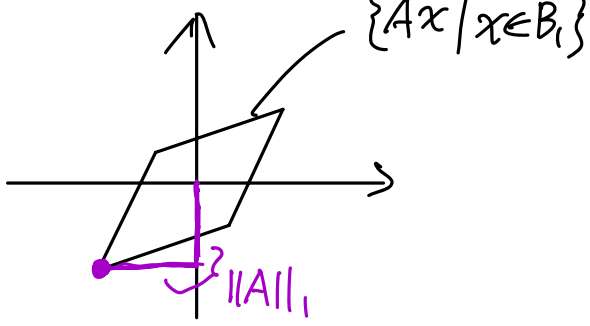
$=$ the max eigenvalue of A^TA

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

$\|A\|_1 = \max_{j=1}^{n} \|a_j\|_1$ if $A = [a_1\ a_2\ \cdots\ a_n]$ where $a_i \in \mathbb{R}^m$.

- $p=\infty$: $\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty$ can prove $= \max_{i=1}^{m} \|a^{(i)}\|_1$ 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

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

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

$\|\cdot\|_p \qquad \|\cdot\|_q$

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

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

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

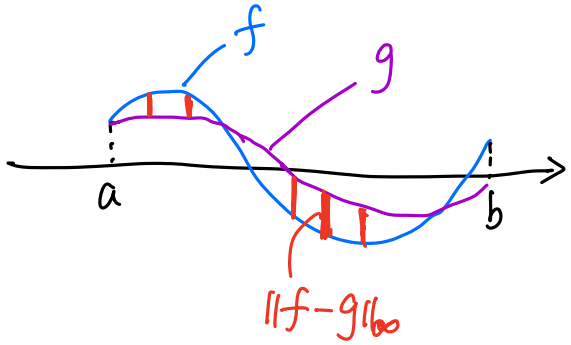
Ex. 4. $C[a,b] = \{f : [a,b] \to \mathbb{R} \mid f \text{ is continuous}\}$ is a vector space.

$\forall f \in C[a,b]$, 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. $l_\infty = \left\{ a=\begin{pmatrix} a_1 \\ a_2 \\ a_i \\ \vdots \end{pmatrix} \mid \exists\, C>0 \text{ s.t. } |a_i|\leq C \ \forall i \right\}$

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

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

"sup" can not be replaced by "max",

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

$\max_i |a_i|$ doesn't exist

because $a_i < a_{i+1}$ $\forall i$

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

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

we can prove that

$\|\cdot\|_p$ is a norm on l_p.

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

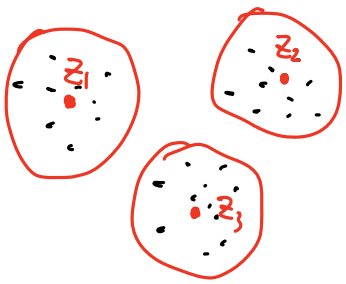
Clustering:

Suppose we are given N vectors in $\mathbb{R}^n$

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

Group them into different K groups.

($\mathbb{R}^n$ can be replaced with other vector spaces)



Applications:
- Image Clustering
- Text data Clustering
- Recommender System

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

G_j — be the groups, $G_j = \{ i \mid c_i = j \}$
$j = 1, 2, \cdots, K$

and

z_j — be the representative vector in G_j

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

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

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

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

we want J small.

So we solve

$$\min J$$

$$\Updownarrow$$

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

Alternating minimizations.

Step 0: Initialize $z_1, z_2, \cdots, z_K$

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

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

Step 2: Fix $G_1, \cdots, G_K$, Solve

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

Repeat

To solve subproblem (1):

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

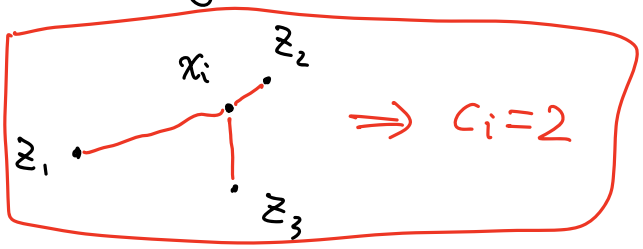
Then (1) $\Leftrightarrow \min_{c_1, c_2, \cdots, c_N} \sum_{i=1}^{N} \|x_i - z_{c_i}\|_2^2$

$\Leftrightarrow \min_{c_1, c_2, \cdots, c_N} \|x_1 - z_{c_1}\|_2^2 + \|x_2 - z_{c_2}\|_2^2 + \cdots + \|x_N - z_{c_N}\|_2^2$

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

$\Leftrightarrow \min_{c_i \in \{1,2,\cdots,K\}} \|x_i - z_{c_i}\|_2^2 \quad i = 1, 2, \cdots, N$

$\Leftrightarrow c_i = \text{argmin} \{\|x_i - z_1\|_2^2, \|x_i - z_2\|_2^2, \cdots, \|x_i - z_K\|_2^2\}$



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

the we use c_i to define G_j:

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

To solve subproblem (2):

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

$\Updownarrow$

$$\min_{z_1, z_2, \cdots z_K} \sum_{i \in G_1} \|x_i - z_1\|_2^2 + \sum_{i \in G_2} \|x_i - z_2\|_2^2 + \cdots + \sum_{i \in G_K} \|x_i - z_K\|_2^2$$

$\Updownarrow$

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

$\Updownarrow$

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

$|G_j|$ — number of elements in G_j

— Mean of vectors in G_j

for $j = 1, 2, \cdots, K$.

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

$\Leftrightarrow$ take derivative and set it 0.

$$\sum_{i \in G_j} 2(z_j - x_i) = 0$$

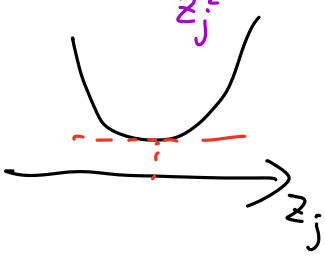
$$\Updownarrow$$

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

$$\Updownarrow$$

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

with a positive coefficient before z_j^2



z_j is the mean of vectors in G_j

Algorithm: (K-means algorithm)

Step 0: Initialize $z_1, z_2, \cdots, z_K$

→ Step 1: $C_i = \arg\min_{j \in \{1,2,\cdots,K\}} \{\|x_i - z_j\|_2\}$, $i = 1, \cdots, N$

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

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

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

(, (z_j is the mean of vectors in Group j))

Repeat

Replace 2-norm with 1-norm.

$$\min_{\substack{G_1, G_2, \cdots, G_K \\ z_1, z_2, \cdots z_K}} \sum_{j=1}^{K} \sum_{i \in G_j} \|x_i - z_j\|_1$$

Alternating minimizations.

Step 0: Initialize $z_1, z_2, \cdots, z_K$

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

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

Step 2: Fix $G_1, \cdots, G_K$. Solve

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

Repeat

To solve subproblem (1):

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

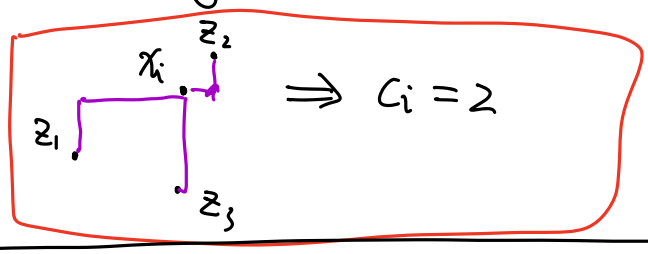
Then (1) $\Leftrightarrow$ $\min_{c_1,c_2,\cdots,c_N} \sum_{i=1}^{N} \|x_i - z_{c_i}\|_1$

$\Leftrightarrow$ $\min_{c_1,c_2,\cdots,c_N} \|x_1 - z_{c_1}\|_1 + \|x_2 - z_{c_2}\|_1 + \cdots + \|x_N - z_{c_N}\|_1$

$\Leftrightarrow$ $\min_{c_i} \|x_i - z_{c_i}\|_1$, $\quad i = 1, 2, \cdots, N$

$\Leftrightarrow$ $\min_{c_i \in \{1,2,\cdots,K\}} \|x_i - z_{c_i}\|_1 \quad i = 1, 2, \cdots, N$

$\Leftrightarrow$ $c_i = \operatorname{argmin} \{\|x_i - z_1\|_1, \|x_i - z_2\|_1, \cdots, \|x_i - z_K\|_1\}$



Assign x_i to the nearest representative in 1-norm distance

To solve subproblem (2):

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

$$\Updownarrow$$

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

$$\Updownarrow$$

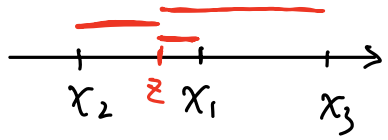
$$\min_{z_j} \sum_{i\in G_j} \|x_i - z_j\|_1, \qquad j = 1, \cdots, K$$

$$\Updownarrow$$

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

$|x_1 - z| + |x_2 - z| + |x_3 - z|$
$= |x_3 - x_2| + |z - x_1|$ 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 $z_1, z_2, \cdots, z_K$

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

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

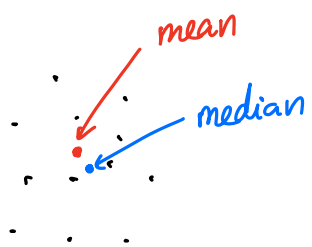
$G_j = \{i \mid C_i = j\}$, $j = 1, \cdots, K$.

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

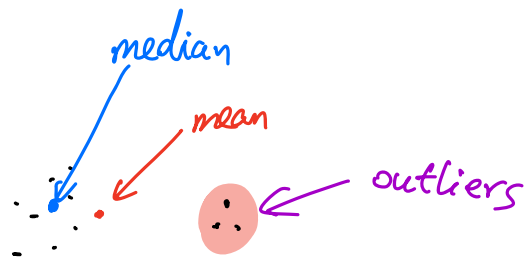
(z_j is the mean of vectors in Group j)

Repeat

Comparison of K-means and K-medians



Both mean and median are good for a representative vec.



Median seems better than the mean.

— Mean is sensitive to outliers

— Median is robust.

— In ML algorithms,

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

Approximation of vectors on vector spaces.

— Iterative alg.

— Approximate of functions on vector spaces (Calcus 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 $x^{(k)} \to x$, if $\lim_{k\to+\infty} \|x^{(k)} - x\| = 0$

i.e., $x^{(k)} \to x \iff \lim_{k\to+\infty} \|x^{(k)} - x\| = 0$

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

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

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

(constant of k)

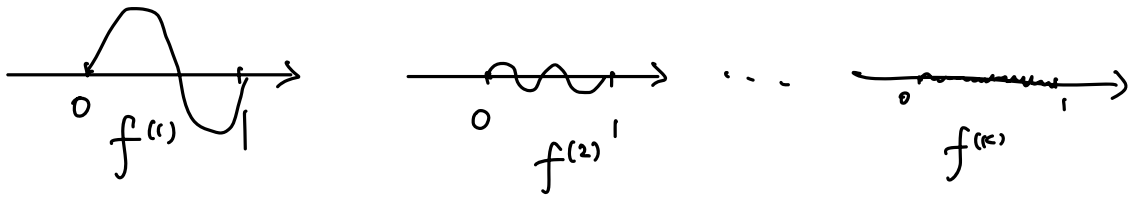
i.e., $\lim_{k\to+\infty} \|x^{(k)} - x\|_2 = 0$ $\quad x^{(k)} \to x$.

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

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

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

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



Ex 3. Consider infinite sequences:

$$a^{(k)} = \begin{pmatrix} 1/k \\ \vdots \\ 1/k \\ 0 \\ 0 \\ \vdots \end{pmatrix} \left.\vphantom{\begin{pmatrix} 1/k \\ \vdots \\ 1/k \\ 0 \\ 0 \\ \vdots \end{pmatrix}}\right\} k\text{-terms} \in l_1, l_2, l_\infty$$

$\forall k$:

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

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

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

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

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

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

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

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

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

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

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

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

So, $a^{(k)} \not\to a$ 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.

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

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

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

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

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

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

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

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

If this happen, 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 $\lim_{k\to+\infty} x^{(k)} \notin V$

- Calculus of functions on vector spaces.

  $f(\lim x^{(k)}) = f(x)$ — not defined.

$\Rightarrow$ not in V

Completeness:

Cauchy sequence:

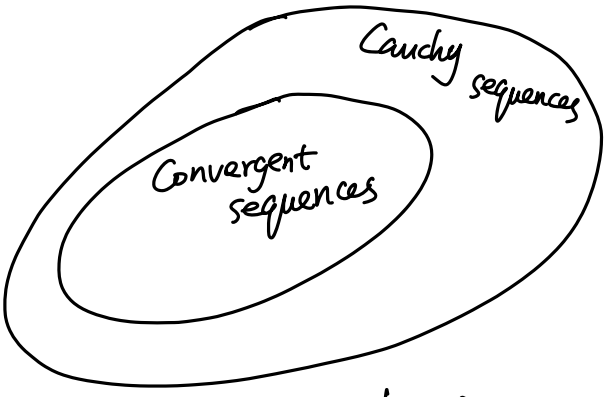
$\{x^{(k)}\}_{k\in\mathbb{N}}$ is a Cauchy sequence if:

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

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

$x^{(k)}$ $x^{(l)}$

x



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

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

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

$\Updownarrow$

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

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

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

We call a 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.

- $V=\{a \mid \|a\|_1<+\infty, a\in l_\infty\}$ 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 $\|f\|_p=\left(\int_a^b |f(t)|^p dt\right)^{1/p}$

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

Applications:

- Supervised learning:

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

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

  An iter alg. generate

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

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

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

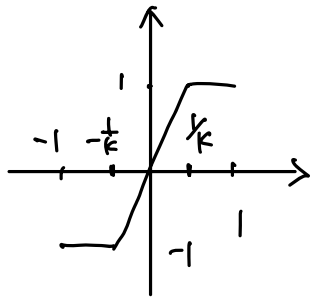
$(C[a,b], \|\cdot\|_1)$ is incomplete.

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

$\Rightarrow$ f is discontinuous.

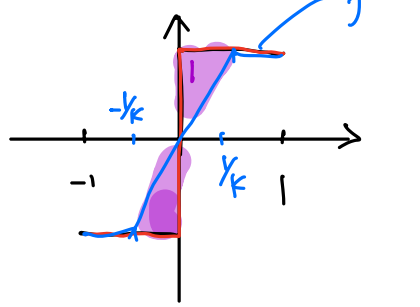
Explicit Example:

$f^{(k)}$



$f^{(k)} \in C[-1,1]$

f



$f \notin C[-1,1]$

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

$\to 0$ as $k \to +\infty$

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

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

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

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

— Iterative algorithm.

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

Check the convergence ?

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

— Use Cauchy sequence.

We check $\boxed{\|x^{(k)} - x^{(l)}\| \leq \varepsilon}$ — for large k, l.

practically, we just check

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

§ 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,
> $\exists\ C_1, C_2 > 0$ 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.

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

proof: "$\Rightarrow$"

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

Due to the norm equivalence,

$C_1\|x^{(k)} - x\| \leq \|x^{(k)} - x\|_B \leq C_2\|x^{(k)} - x\|_A$

$$C_1\|x-x\|_A \le \|x-x\|_B \le C_2\|x-x\|_A$$

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

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

i.e. $x^{(k)} \to x$ 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.

$$\|a\|_2 \le \|a\|_1 \le \sqrt{n}\,\|a\|_2 \qquad \forall a\in\mathbb{R}^n$$

$$\left.\begin{aligned}\|a\|_2^2 &= \sum_{i=1}^n |a_i|^2 \\ \|a\|_1^2 &= \Big(\sum_i |a_i|\Big)^2 = \sum_{i,j}|a_i||a_j| = \sum_{i=1}^n |a_i|^2 + \sum_{i\neq j}|a_i||a_j|\end{aligned}\right\} \|a\|_2^2 \le \|a\|_1^2$$

$$\|a\|_1^2 = \sum_{i=1}^n |a_i|^2 = \sum_{i=1}^n\sum_{j=1}^n |a_i||a_j| \le \sum_{i=1}^n\sum_{j=1}^n \left(\frac{1}{2}(|a_i|^2+|a_j|^2)\right)$$

$$= n\left(\sum_{i=1}^n |a_i|^2\right) = n\cdot\|a\|_2^2$$

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

$$\|a\|_\infty \le \|a\|_2 \le \sqrt{n}\,\|a\|_\infty \qquad \forall a\in\mathbb{R}^n$$

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

$$\|a\|_\infty \le \|a\|_1 \le n\,\|a\|_\infty \qquad \forall a\in\mathbb{R}^n$$

– The convergence speed depends on norms.

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

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

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

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

To achieve an ε- precision

2-norm: $\|x^{(k)}-0\|_2 = \frac{1}{k} \le \varepsilon \implies \boxed{k \ge \frac{1}{\varepsilon}}$

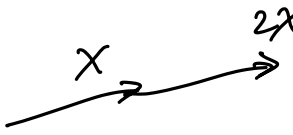
1-norm: $\|x^{(k)}-0\|_1 = \frac{\sqrt{2}}{k} \le \varepsilon \implies \boxed{k \ge \frac{\sqrt{2}}{\varepsilon}}$

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.

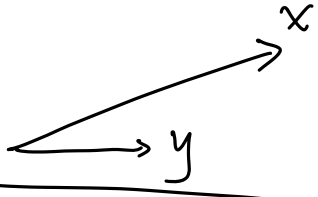
How two vectors are corelated/aligned?

- 

$\|x-2x\| = \|x\|$ ← not small

- x, y pictures of the same scene — at noon / at night

$$x \approx 2y$$



$\|x-y\|$ big

we need correlation of two vectors

§. 3.1 Inner product.

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

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

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

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

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

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

- 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.

$\langle x, y \rangle_A = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i y_j = x^T A y$ where $A \in \mathbb{R}^{n \times n}$ is a symmetric positive definite (SPD) matrix,

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

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

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

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

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

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

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

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

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 $\langle \cdot, \cdot \rangle$

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

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

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 $C[a,b] = \{ f \mid f: [a,b] \to \mathbb{R} \text{ is a continuous function} \}$

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

§3.2. Properties of inner products

Cauchy-Schwartz inequality (C-S)

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

then $\forall\ x, y \in V$, $\quad |\langle x, y \rangle|^2 \le \langle x, x \rangle \langle y, y \rangle$

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

$$\left.\begin{aligned} &|\langle x, y \rangle|^2 = |\langle x, 0 \rangle|^2 = 0 \\ &\langle x, x \rangle \langle y, y \rangle = \langle x, x \rangle \langle 0, 0 \rangle = 0 \end{aligned}\right\} \Rightarrow |\langle x, y \rangle|^2 = 0 \le 0 = \langle x, x \rangle \langle y, y \rangle$$

- If $y \neq 0$, then

Let $\lambda \in \mathbb{R}$ be arbitrary, consider

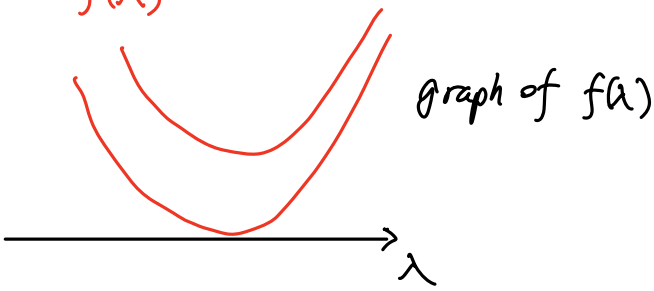
$$0 \le \langle x + \lambda y, x + \lambda y \rangle = \langle x, x \rangle + \lambda \langle x, y \rangle + \lambda \langle y, x \rangle + \lambda^2 \langle y, y \rangle$$

$$= \lambda^2 \langle y, y \rangle + \lambda \cdot 2 \langle x, y \rangle + \langle x, x \rangle$$

$$f(\lambda)$$

Since $y \neq 0$, $\langle y, y \rangle > 0$

$f(\lambda) \ge 0$



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

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

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

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

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

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

$\Rightarrow$ "="

If $y=\alpha x$, Similar.

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

- If $y=0$, then $y = 0\cdot x$ ✓
- 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

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

the norm induced by the inner product

the default norm on inner prod. space

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

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

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

② $\|\alpha x\| = \langle \alpha x, \alpha x\rangle^{1/2} = \left(\alpha^2\langle x,x\rangle\right)^{1/2} = |\alpha|\cdot\langle x,x\rangle^{1/2}$

$= |\alpha|\cdot\|x\|$

③ $\forall\, x,y\in V$ WTS: $\|x+y\| \leq \|x\| + \|y\|$

$\|x+y\|^2 = \langle x+y, x+y\rangle$

$= \langle x,x\rangle + 2\langle x,y\rangle + \langle y,y\rangle$

$= \|x\|^2 + 2\langle x,y\rangle + \|y\|^2$

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

C-S $\leq \|x\|^2 + 2\|x\|\cdot\|y\| + \|y\|^2$

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

$$\Rightarrow \quad \|x+y\| \leq \|x\| + \|y\| \qquad \blacksquare$$

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

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

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

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

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

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

inner prod $\Downarrow$ norm

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

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

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

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

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

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

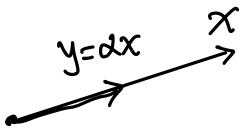
$$-\|x\|\,\|y\| \leq \langle x,y\rangle \leq \|x\|\,\|y\|$$

Assume $x,y \neq 0$

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

- 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 $\alpha > 0$.

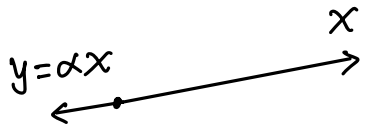


So, $\angle(x, y) = 0$

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

$|\langle x, y\rangle|$

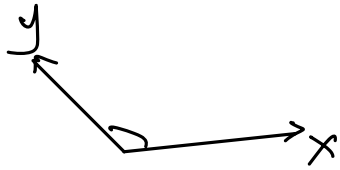
"$=$" holds in C-S $\Rightarrow$ $x = \alpha y$ or $y = \alpha x$ with $\alpha < 0$

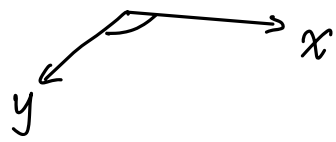


So, $\angle(x, y) = \pi$

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

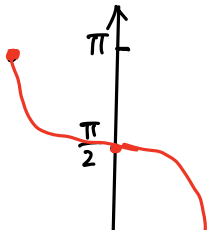
$\angle(x, y) \in (0, \pi)$



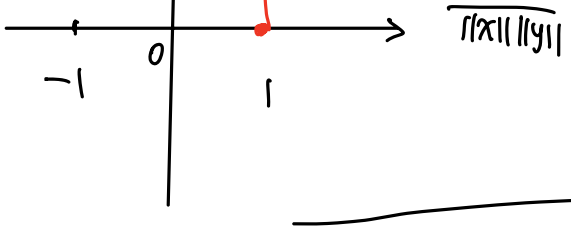


So, the angles should satisfy

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



$\langle x, y\rangle$



we define $\boxed{\angle(x,y) = \arccos \frac{\langle x,y\rangle}{\|x\|\|y\|}}$

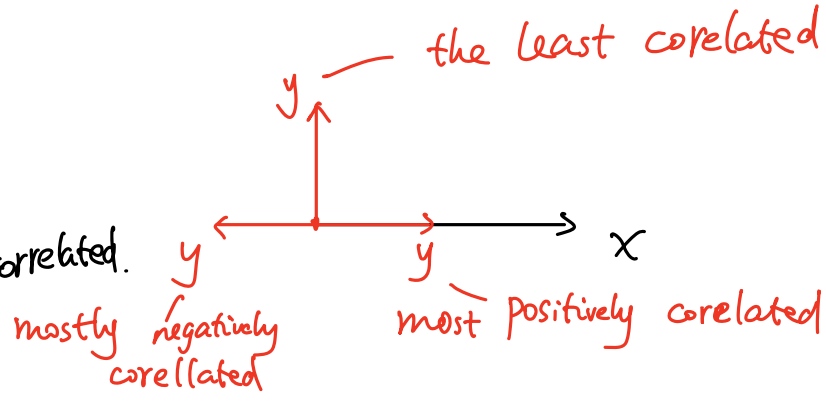
↑

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

$$\begin{cases} \angle(x,y) = \frac{\pi}{2} \\ x,y \text{ are the least correlated.} \end{cases}$$

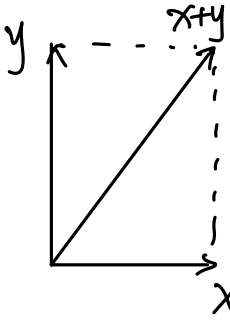


we call x, y are orthogonal, denoted $x \perp y$

- Pythagoras' thm

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

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



proof. $\|x+y\|^2 = \langle x+y, x+y\rangle = \langle x,x\rangle + 2\langle x,y\rangle + \langle y,y\rangle$

$= \|x\|^2 + \|y\|^2 + 2\langle x,y\rangle$

Then

$$x \perp y \iff \langle x,y\rangle = 0 \iff \|x+y\|^2 - \|x\|^2 - \|y\|^2 = 2\langle x,y\rangle = 0$$

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

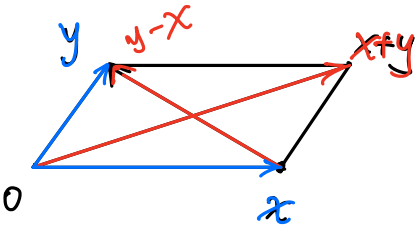
- Parrallelogram Law:

Let V be ~~a~~ 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.

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



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

If

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

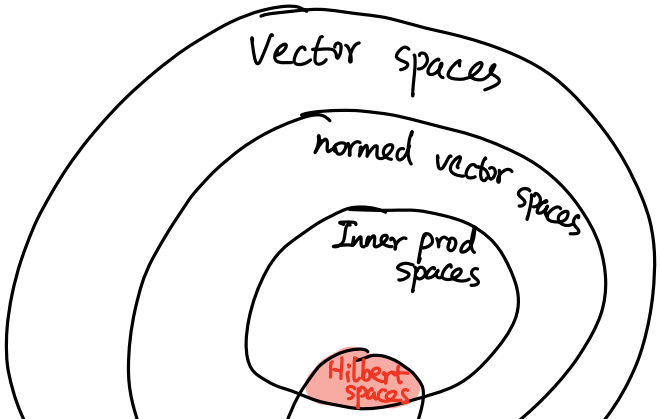
then $\exists\ \langle \cdot, \cdot \rangle$ on V such that

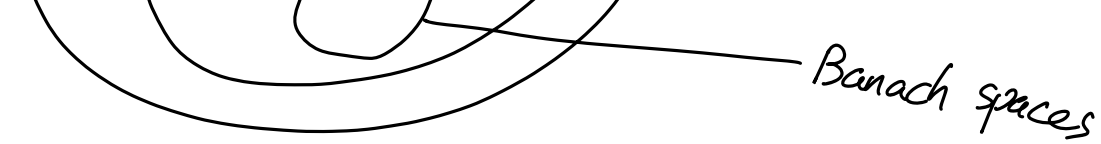
$$\|x\| = (\langle x, x\rangle)^{1/2} \quad \forall\, x \in V.$$

vector space $\xrightarrow{\text{inner product}}$ inner product space (with ~~a~~ the induced norm)

$\Rightarrow$ An inner prod space is alway a normed vector space

Complete Inner Product spaces are called Hilbert 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. $l_2 = \left\{ \begin{pmatrix} a_1 \\ a_2 \\ \vdots \end{pmatrix} \middle| \sum_{i=1}^{+\infty} a_i^2 < +\infty \right\}$ with inner prod

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

is a Hilbert space.

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

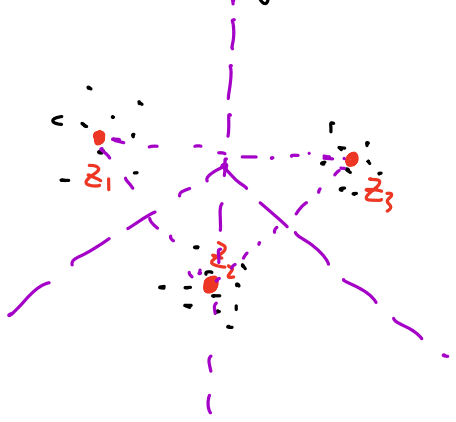
with inner product $\langle f, g \rangle = \int_a^b f(t) g(t)\, dt$

is NOT a Hilbert space.

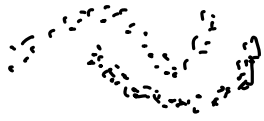
The completion of $C[a,b]$ with $\langle \cdot, \cdot \rangle$ and its induced norm is $L^2(a,b)$

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

— K-means works well for clusters whose boudaries are (linear)



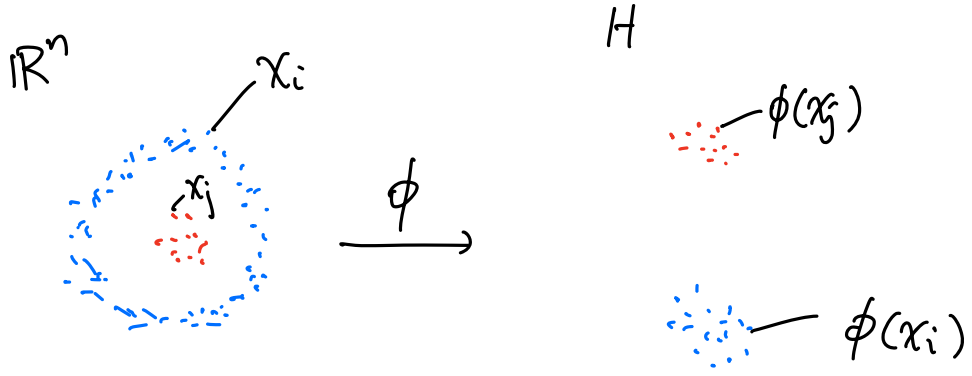
— K-means will not work for



— 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.



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

An explicit example of ϕ:

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

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

Given data points $x_1, x_2, \cdots, x_N \in \mathbb{R}^n$,
we want to group them into K groups.

Let $\phi : \mathbb{R}^n \to H$ ← feature space; ← some Hilbert space.

↑ feature transformation

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 $\phi(x_1), \phi(x_2), \cdots, \phi(x_N)$

Step 0: Initialize $z_1, z_2, \cdots, z_K$ in H

Step 1: $c_i = \arg\min_{j \in \{1,2,\cdots,K\}} \{ \| \phi(x_i) - z_j \|_H^2 \}$, $i = 1, \cdots, N$ (1)

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

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

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

(z_j is the mean of vectors in Group j)

Repeat

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

Step 0: Initialize $G_1, \cdots, G_K$

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

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

Repeat

We ~~a~~re-arrange the function in (2):

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

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

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

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

$$= \langle \phi(x_i), \phi(x_i) \rangle_H + \frac{1}{|G_j|^2} \sum_{\ell_1 \in G_j} \sum_{\ell_2 \in G_j} \langle \phi(x_{\ell_1}), \phi(x_{\ell_2}) \rangle_H$$
$$- \frac{2}{|G_j|} \sum_{\ell \in G_j} \langle \phi(x_i), \phi(x_\ell) \rangle_H$$

All terms involving ϕ, H are in the form

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

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

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

Such that $K(x, y) = \langle \phi(x), \phi(y) \rangle_H$ for some $\phi : \mathbb{R}^n \to H$

(Kernel function) $\langle \cdot, \cdot \rangle_H$ on H.

The resulting algorithm is:

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

Step 0: Initialize $G_1, \cdots, G_K$

→ Step 1:

$$c_i = \underset{j \in \{1, \cdots, K\}}{\operatorname{argmin}} \left\{ K(x_i, x_i) + \frac{1}{|G_j|^2} \sum_{\ell_1 \in G_j} \sum_{\ell_2 \in G_j} K(x_{\ell_1}, x_{\ell_2}) - \frac{2}{|G_j|} \sum_{\ell \in G_j} K(x_i, x_\ell) \right\}, \quad i = 1, \cdots, N$$

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

— Repeat

Kernel K-means

How to choose the kernel K ?

— Necessary conditions for K:

① $K(x, y) = \langle \phi(x), \phi(y) \rangle_H = \langle \phi(y), \phi(x) \rangle_H = K(y, x)$

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

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

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

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

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

Define a Kernel matrix $K = \left[K(y_i, y_j) \right]_{i=1,\, j=1}^{m \quad m} \in \mathbb{R}^{m \times m}$

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

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

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

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

①② together: A necessary condition for a good kernel is:

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

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

is SPSD

The kernel function is SPSD

- Sufficient condition

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

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

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

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

The we have: $K(x,y) = k(y,x)$

$\forall\, y_1, \cdots, y_m \in \mathbb{R}^n$, $K := [K(x_i, x_j)]_{i,j=1}^m$ is SPSD

non-linear if one vector is fixed

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

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

$\forall\, y_1, \cdots, y_m \in \mathbb{R}^n$ $[\langle y_i, y_j \rangle]_{i,j=1}^m$ is SPSD

(b) 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 $\langle \cdot, \cdot \rangle$ $\longrightarrow$ non-linear inner product $K(\cdot, \cdot)$

2. No feature transformation $\longrightarrow$ $x_i \to \phi(x_i)$ in H

K-means in $\mathbb{R}^n$ $\longrightarrow$ K-means in H

Some popular kernel functions

① $K(x,y) = \langle x, y \rangle_A$ $\qquad \phi(x) = A^{1/2} x$, $H = \mathbb{R}^n$

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 $\alpha = 2$

$$K(x,y) = (x^T y + 1)^2 = (x_1 y_1 + x_2 y_2 + 1)^2$$
$$= x_1^2 y_1^2 + x_2^2 y_2^2 + 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2$$
$$= \left\langle \begin{pmatrix} x_1^2 \\ x_2^2 \\ 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ \sqrt{2}x_1x_2 \end{pmatrix}, \begin{pmatrix} y_1^2 \\ y_2^2 \\ 1 \\ \sqrt{2}y_1 \\ \sqrt{2}y_2 \\ \sqrt{2}y_1y_2 \end{pmatrix} \right\rangle \equiv \langle \phi(x), \phi(y) \rangle_{\mathbb{R}^6}$$

So, $\phi(x) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ \sqrt{2}x_1x_2 \end{pmatrix}$ and $H = \mathbb{R}^6$

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

The kernel is used in 1-layer neural network.

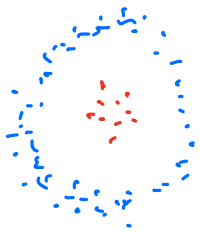
④ $K(x,y) = e^{-\frac{\|x-y\|_2^2}{\sigma^2}}$, where $\sigma > 0$ is a parameter

(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 $\|\phi(x_i)\|_H^2 = \langle \phi(x_i), \phi(x_i) \rangle_H = K(x_i, x_i) = 1 \quad \forall\ i$

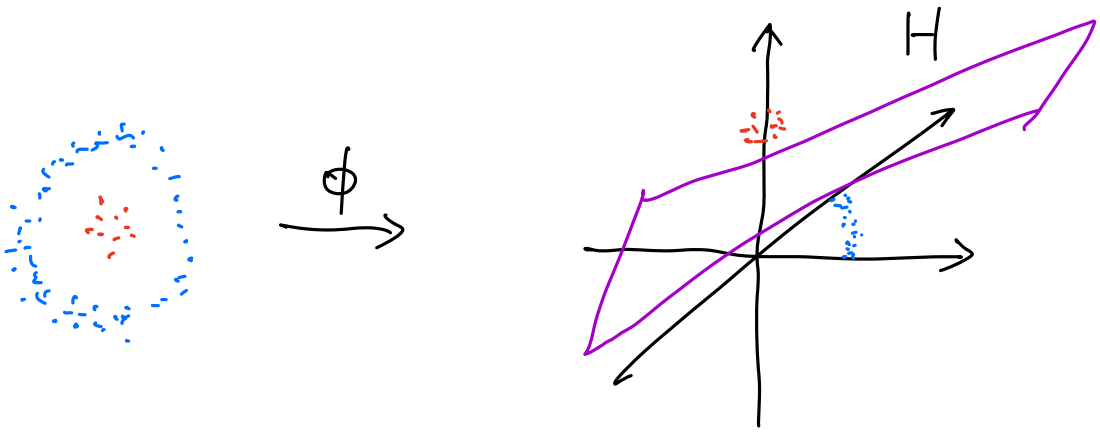
  i.e., all $\phi(x_i)$, $i = 1, \cdots, N$ are on the unit sphere on H.

- $K(x_i, x_j) \begin{cases} \approx 0 & \text{if } \|x_i - x_j\|_2 \text{ is large} \\ \approx 1 & \text{if } \|x_i - x_j\|_2 \text{ is small.} \end{cases}$

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

$\phi(x_i) \perp\!\!\!\perp \phi(x_j)$ 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 $x_1, x_2, \cdots, x_N \in \mathbb{R}^n$ and certain pairs of them are similar / dissimilar

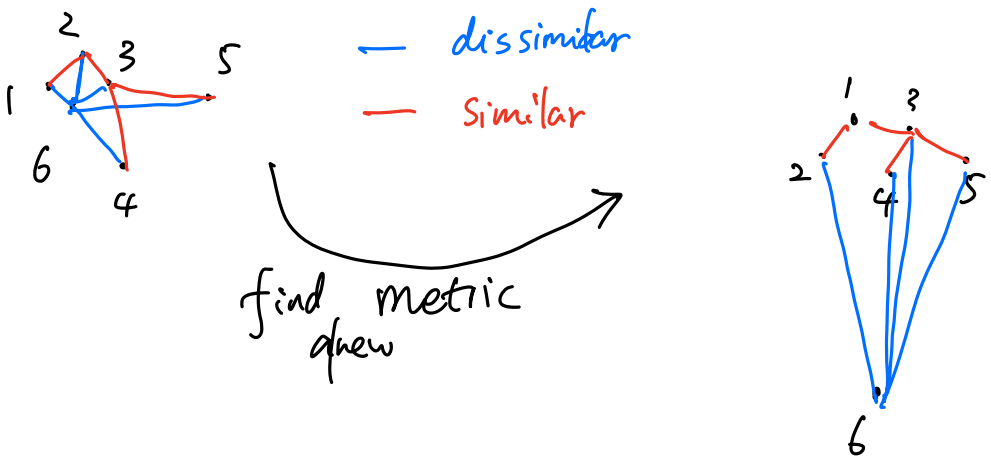
S: $(x_i, x_j) \in S$ if x_i and x_j are similar

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

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

- norms induced by inner products

  Given an SPD matrix $A \in \mathbb{R}^{n\times n}$, $\langle x, y\rangle_A = x^T A y$ $\forall x, y \in \mathbb{R}^n$

  and $\|x\|_A = (x^T A x)^{1/2}$ $\forall x \in \mathbb{R}^n$

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

Then find a metric $\Longleftrightarrow$ 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

$$\overline{\{\text{All SPD matrices}\}} = \{\text{All SPSD matrices}\}$$

$$\{A \mid A = A^T,\ x^T A x > 0\ \forall x \neq 0\}$$

symmetric positive semi-definite

$$\{A \mid A = A^T,\ x^T A x \geq 0\ \forall x\}$$

- Instead of SPD matrix, we find an SPSD matrix $A \in \mathbb{R}^{n\times n}$ and define $\|x-y\|_A = \left((x-y)^T A (x-y)\right)^{1/2}$

- $\|x\|_A$ is not a norm, because

$$\|x\|_A = 0 \Leftrightarrow x^T A x = 0 \not\Leftarrow x = 0$$

- $\|x\|_A$ is still good enough, because
  - ① $\|x\|_A \geq 0 \quad \forall x \in \mathbb{R}^n$. $\quad \|x\|_A = 0 \not\Leftarrow x = 0$
  - 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.

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

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

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

Then we solve

$$\begin{cases} \min\limits_{A \in \mathbb{R}^{n \times n}} & \sum\limits_{(x_i, x_j) \in S} \|x_i - x_j\|_A^2 \\ \text{s.t.} & A \text{ is SPSD} \\ & \|x_i - x_j\|_A^2 \geq 1 \quad \forall (x_i, x_j) \in D \end{cases}$$

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 $f: V \to \mathbb{R}$ 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 $\mathbb{R}^n$

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

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

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

For example: on $\mathbb{R}^2$, $f\begin{pmatrix} 0 \\ 1 \end{pmatrix} = 1$, $f\begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1$

But $f\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) = f\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 1$

$\neq$

$f\begin{pmatrix} 0 \\ 1 \end{pmatrix} + f\begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1 + 1 = 2$

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

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

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

is linear,

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

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

$$F(\alpha f + \beta g) = \int_a^b (\alpha f + \beta g)(t)\,dt = \alpha \int_a^b f(t)\,dt + \beta \int_a^b g(t)\,dt$$
$$= \alpha \cdot F(f) + \beta \cdot F(g)$$

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

$$f(\alpha x + \beta y) = \langle a, \alpha x + \beta y \rangle = \alpha \langle a, x \rangle + \beta \langle a, y \rangle$$
$$= \alpha f(x) + \beta f(y).$$

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

— $\|-x\| = \|x\|$

— If $\|\cdot\|$ is linear, then

$$\|-x\| = \|(-1)\cdot x + 0 \cdot x\| = (-1)\cdot \|x\| + 0 \cdot \|x\|$$
$$= -\|x\|. \quad \text{Contradiction.}$$

Properties of linear functions.

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

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

  Choose $\alpha = 0$. $\boxed{f(0) = 0}$

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

$\boxed{\text{Linearity} \iff \text{Homogeneity} + \text{Additivity}}$

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

$= \alpha_1 f(x_1) + f(\alpha_2 x_2 + \cdots + \alpha_k x_k)$

$\vdots$

$= \alpha_1 f(x_1) + \alpha_2 f(x_2) + \cdots + \alpha_k f(x_k)$

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 $f : H \to \mathbb{R}$,
> $\exists\, a \in H$ 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 $f : \mathbb{R}^n \to \mathbb{R}$
> $\exists$ 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

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

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

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

$$\begin{aligned} f(x) &= f(x_1 e_1 + x_2 e_2 + \cdots + x_n e_n) \\ &= x_1 \cdot f(e_1) + x_2 \cdot f(e_2) + \cdots + x_n \cdot f(e_n) \\ &= \left\langle \begin{pmatrix} f(e_1) \\ f(e_2) \\ \vdots \\ f(e_n) \end{pmatrix}, \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \right\rangle = \langle a, x \rangle \end{aligned}$$

a ||| def

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

$$f(e_i) = \underbrace{\langle a, e_i \rangle}_{a_i} = \underbrace{\langle b, e_i \rangle}_{b_i} \quad \forall\, i = 1, \cdots, n$$

$$\Rightarrow \quad a_i = b_i \quad i = 1, \cdots, n \quad \Rightarrow \quad a = b \qquad \boxtimes$$

It can be extended to general Hilbert spaces.

Riesz Representation Theorem:

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

$$f(x) = \frac{x_1 + x_2 + \cdots + x_n}{n} = \left\langle \begin{pmatrix} 1/n \\ 1/n \\ \vdots \\ 1/n \end{pmatrix}, \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \right\rangle$$

Ex. 2: Let H be a Hilbert space.

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

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

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

The trace function

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

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

Remark:

- f is bounded $\iff$ $\exists\ C > 0$ s.t. $|f(x)| \le C \cdot \|x\|$ $\forall\ x \in H$.
- 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

functional on incomplete inner prod space.

Ex: $C^1[-1,1] = \{f \mid f, f' \in C[-1,1]\}$ with $\langle f,g\rangle = \int_{-1}^{1} f(t)g(t)\,dt$ (This space is incomplete)

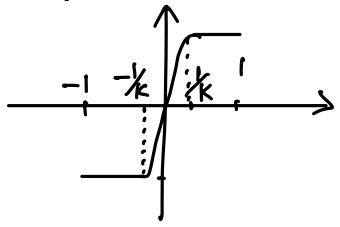
Define $F: C^1[-1,1] \to \mathbb{R}$ by

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

— F is linear because

$$F(\alpha f+\beta g) = (\alpha f+\beta g)'(0) = \alpha f'(0) + \beta g'(0)$$
$$= \alpha F(f) + \beta F(g)$$

— F is unbounded.

f_k: 

$\in C^1[-1,1]$

$$F(f_k) = f_k'(0) = k \to +\infty \quad \text{as} \quad k \to +\infty$$

but $\|f_k\| = \left(\int_{-1}^{1} f_k^2(t)\,dt\right)^{1/2} \le \left(\int_{-1}^{1} 1\,dt\right)^{1/2} = \sqrt{2}$

So, $$\frac{F(f_k)}{\|f_k\|} \ge \frac{k}{\sqrt{2}} \to +\infty \quad \text{as} \quad k \to +\infty$$

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

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

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

$$|G(f)| = \left|\int_{-1}^{1} f(t)\,dt\right| = \left|\int_{-1}^{1} f(t)\cdot 1\,dt\right| = |\langle f, 1\rangle|$$
$$\overset{C\text{-}S}{\le} \|f\|\cdot\|1\| = \sqrt{2}\cdot\|f\| < +\infty$$

— 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:

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

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

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

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

$= \langle 1, f \rangle$

Hyperplane

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

Let $a \in H$

- Consider

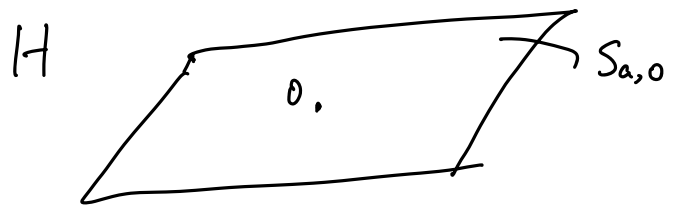
$$S_{a,0} = \{x \in H \mid \langle a, x \rangle = 0\} \subseteq H$$

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

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

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

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



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

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

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

① $\forall\, x \in S_{a,b}$

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

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

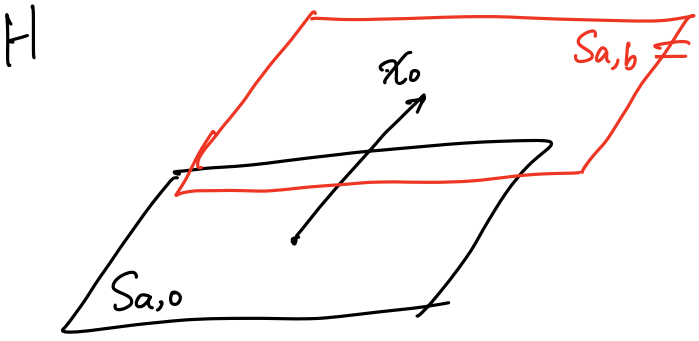
$\Rightarrow x \in x_0 + S_{a,0} \Rightarrow \boxed{S_{a,b} \subseteq x_0 + S_{a,0}}$

② $\forall\, x \in S_{a,0}$

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

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

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



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

Also, its co-dimension is 1

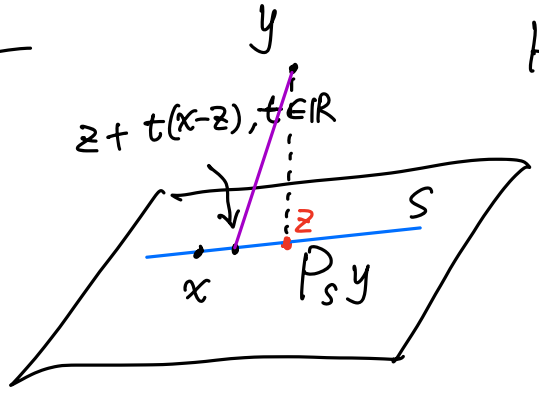
(It is defined in terms of 1 equation only)

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,

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



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

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

$\Updownarrow$

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

$t\langle z-y, x-z\rangle$

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 on the line is $\boxed{z+t(x-z),\quad t\in\mathbb{R}}$

$$\text{Since } \langle a, z+t(x-z)\rangle = \langle a,z\rangle + t\langle a, x-z\rangle$$
$$= \langle a,z\rangle + t\langle a,x\rangle - t\langle a, z\rangle$$
$$= b + t\cdot b - t\cdot b = b$$
$$\Rightarrow \quad z+t(x-z) \in S \quad \forall t\in\mathbb{R}.$$

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

$$\|z-y\|^2 \le \|(z+t(x-z))-y\|^2 = \|(z-y)+t(x-z)\|^2$$
$$= \langle (z-y)+t(x-z),\ (z-y)+t(x-z)\rangle$$
$$= \|z-y\|^2 + t^2\|x-z\|^2 + 2t\langle z-y, x-z\rangle$$
$$\Rightarrow \quad 2t\langle z-y, x-z\rangle \ge -t^2\|x-z\|^2$$

- If $t>0$, then
$$\langle z-y, x-z\rangle \ge -\frac{t}{2}\|x-z\|^2$$
Let $t \to 0_+$
$$\boxed{\langle z-y, x-z\rangle \ge 0}$$

- If $t<0$, then
$$\langle z-y, x-z\rangle \le -\frac{t}{2}\|x-z\|^2$$
Let $t \to 0_-$
$$\boxed{\langle z-y, x-z\rangle \le 0}$$

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

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

then: $\forall\, x\in S$,

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

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

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

① $\langle a, z\rangle = \langle a, y\rangle - \left\langle a, \frac{\langle a,y\rangle - b}{\|a\|^2} a\right\rangle$

$$= \langle a, y\rangle - \frac{\langle a,y\rangle - b}{\|a\|^2}\langle a, a\rangle$$
$$= b$$
$$\Rightarrow \quad z \in S$$

② $\forall x \in S$,

$$\langle z-y, x-z\rangle = \left\langle -\frac{\langle a,y\rangle - b}{\|a\|^2} a, \; x-z\right\rangle$$
$$= -\frac{\langle a,y\rangle - b}{\|a\|^2}\langle a, x-z\rangle$$
$$= -\frac{\cdots}{\cdots}\left(\langle a, x\rangle - \langle a, z\rangle\right)$$
$$= -\frac{\cdots}{\cdots}(b-b) = 0$$

$\Rightarrow$ 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:

$$\langle z_1 - y, z_2 - z_1\rangle + \langle y - z_2, z_2 - z_1\rangle = 0$$
$$\Leftrightarrow \quad \langle z_1 - y + y - z_2, z_2 - z_1\rangle = 0$$

$$\Leftrightarrow \quad \langle z_1 - z_2,\ z_2 - z_1 \rangle = 0$$
$$\Leftrightarrow \quad -\|z_1 - z_2\|^2 = 0 \quad \Rightarrow \quad z_1 = z_2 \qquad \blacksquare$$

In summary,

Let H be a Hilbert space, and

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

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,

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

(i.e. $\Leftrightarrow \quad y - P_S y \perp S_{a,b}$)

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

Properties:

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

$$f(\alpha x + \beta y) = g(\alpha x + \beta y) + b \qquad (g \text{ is linear})$$
$$= \alpha g(x) + \beta g(y) + b$$

if $\alpha + \beta = 1$

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

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

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 $f: H \to \mathbb{R}$ is bounded then

f is affine $\iff f(x) = \langle a, x\rangle + b$. 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), $\cdots$, (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 $f: \mathbb{R}^n \to \mathbb{R}$ s.t.

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

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

- The set of all functions $\mathbb{R}^n \to \mathbb{R}$ 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 $\underline{\Phi}$ of all functions.

- Which $\underline{\Phi}$?
  - How to choose $\underline{\Phi}$ is fundamental.
  - $\underline{\Phi}$ can not be too big.
  - $\underline{\Phi}$ can not be too small. —— weak approximation power.

- Linear regression

$$\mathcal{F} = \{\text{affine functions } \mathbb{R}^n \to \mathbb{R}\}$$

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

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

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

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

- Least squares: ($x_i^T a$)

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

- Write

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

Then

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

So,

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

- We have $n+1$ unknowns

... N linear equations

So $\boxed{N \geq n+1}$ 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 $\quad n \sim 10M$

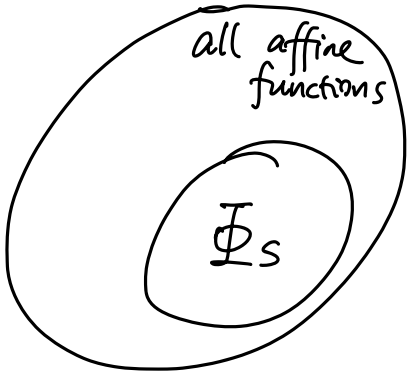
y_i label

N number of available images $\quad N \sim 100K$

In this case $\boxed{N << n}$

~~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

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

Then we solve

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

regularization parameter.

$\Updownarrow$ by optimization theory

(Ridge regression) $\min_{a \in \mathbb{R}^n} \|X\begin{pmatrix} a \\ b \end{pmatrix} - y\|_2^2 + \lambda \|a\|_2^2$,

regularization term

$b \in \mathbb{R}$ data-fitting regularization

where $\lambda > 0$ is a constant depending on C and others.

Larger λ, small $\|\cdot\|_2$, looser fitting to the data.

Smaller λ, larger $\|\cdot\|_2$, tighter fitting to the data.

- LASSO regression

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

We solve

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

$$\Updownarrow$$

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

$$\Updownarrow \text{ by Lagrangian theory}$$

$$\min_{\substack{a \in \mathbb{R}^n \\ b \in \mathbb{R}}} \underbrace{\left\| X \begin{pmatrix} a \\ b \end{pmatrix} - y \right\|_2^2}_{\text{data-fitting}} + \lambda \underbrace{\|a\|_1}_{\text{regularization}} \qquad (\text{LASSO regression})$$

λ: regularization parameter

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

$$x = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix}$$

$$\langle a, x \rangle = \sum_{i=1}^{n} a_i \xi_i = \sum_{i \in I} a_i \xi_i \quad \text{if } a \text{ is sparse}$$

$I = \{ i \mid a_i \neq 0 \}$, $|I| \ll n$.

The regression involves only ξ_i, $i \in I$

Linear model has some limitations.

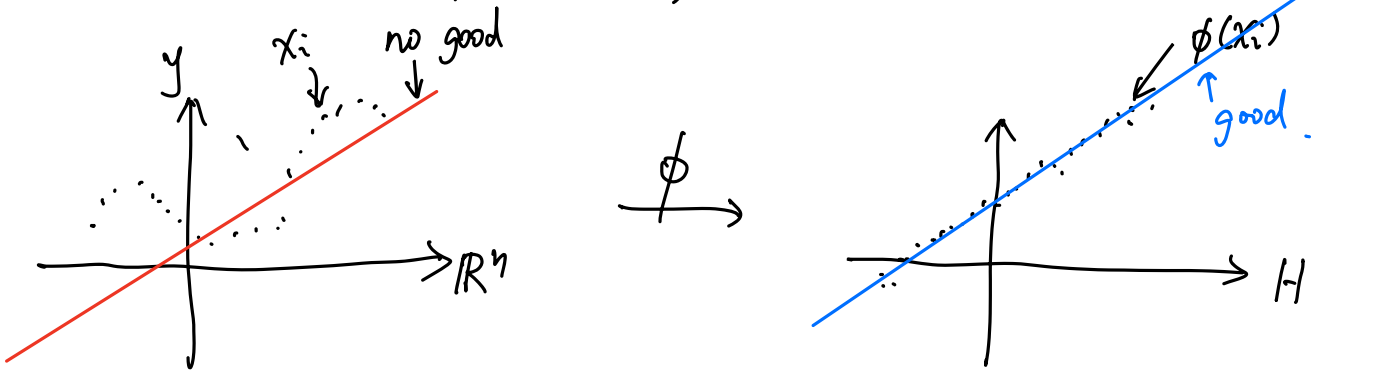
§4.2.2 Kernel Ridge regression

To improve the linear regression, we use kernel method.

① Transform: $\phi: \mathbb{R}^n \to H$ ← Hilbert space ← feature 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 $f: H \to \mathbb{R}$ be affine and bounded

Then $f(\phi(x)) = g(\phi(x)) + b$

$= \langle \phi(x), a \rangle_H + b$ for some $a \in H$, $b \in \mathbb{R}$

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

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

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

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

Then $f(\phi(x)) = \langle \phi(x), a \rangle_H + b$

↑ affine and bounded function

$= \left\langle \begin{pmatrix} \phi(x) \\ 1 \end{pmatrix}, \begin{pmatrix} a \\ b \end{pmatrix} \right\rangle_{\tilde{H}}$

$= \langle \tilde{\phi}(x), \tilde{a} \rangle_{\tilde{H}}$, where $\tilde{a} \in \tilde{H}$.

on H

a linear and bounded function on $\widetilde{H}$

- The task is:

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

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

$$\updownarrow$$

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

- We solve

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

Using least squares, we solve

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

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

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

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

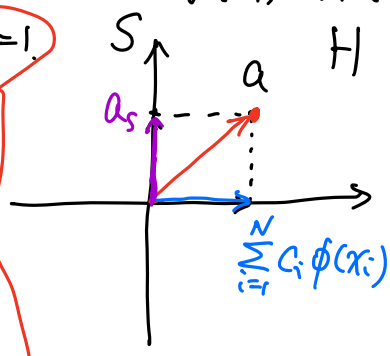
$$a = \sum_{i=1}^{N} c_i \phi(x_i),$$

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 $\langle a_S, \phi(x_i)\rangle_H = 0$ $\forall i=1,\cdots,N$.



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

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

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

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

- For $a - P_S a$: By the explicit formular 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.

$$\sum_{i=1}^N \left(\langle a, \phi(x_i)\rangle_H - y_i\right)^2 + \lambda \|a\|_H^2$$

$$= \sum_{i=1}^N \left(\langle a_S + \sum_{j=1}^N c_j \phi(x_j), \phi(x_i)\rangle_H - y_i\right)^2 + \lambda \left\| a_S + \sum_{j=1}^N c_j \phi(x_j)\right\|_H^2$$

$$= \sum_{i=1}^N \left(\cancel{\langle a_S, \phi(x_i)\rangle} + \sum_{j=1}^N c_j \langle \phi(x_j), \phi(x_i)\rangle_H - y_i\right)^2$$

$$+ \lambda\left(\|a_S\|_H^2 + 2\cancel{\sum_{j=1}^N c_j \langle a_S, \phi(x_j)\rangle_H} + \sum_{j=1}^N \sum_{i=1}^N c_i c_j \langle \phi(x_i), \phi(x_j)\rangle_H\right)$$

$\langle a_S, \phi(x_i)\rangle_H = 0$, $i=1,\cdots,N$

$$\geq \sum_{i=1}^N \left(\sum_{j=1}^N c_j \langle \phi(x_j), \phi(x_i)\rangle_H - y_i\right)^2 + \lambda \sum_{j=1}^N \sum_{i=1}^N c_i c_j \langle \phi(x_i), \phi(x_j)\rangle_H$$

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

$\in \mathbb{R}^{N\times N}$

$$= \sum_{i=1}^{N}\left(\sum_{j=1}^{N} K_{ji}c_j - y_i\right)^2 + \lambda \sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j K_{ij} + \lambda \|a_s\|_H^2$$

$$= \sum_{i=1}^{N}\left((K^T c)_i - y_i\right)^2 + \lambda c^T K c + \lambda \|a_s\|_H^2$$

$$= \|K^T c - y\|_2^2 + \lambda c^T K c + \lambda \cdot \|a_s\|_H^2$$

↑ 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,

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

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

and

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

To solve ②: Because $\lambda > 0$,

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

$$\iff a_s^* = 0$$

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

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

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

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 $\phi, H, \langle \cdot, \cdot \rangle_H$

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

$$\| \text{def}$$

$$K(x_i, x_j)$$

Full Kernel Ridge regression Alg:

① Choose a kernel function

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

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

$K(x, y) = (x^T y + 1)^\alpha$ (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

$$f(x) = \langle a^*, \phi(x) \rangle_H = \left\langle \sum_{i=1}^N c_i^* \phi(x_i), \phi(x) \right\rangle_H$$

$$= \sum_{i=1}^N c_i^* \langle \phi(x_i), \phi(x) \rangle_H$$

$$= \sum_{i=1}^N c_i^* K(x_i, x)$$

§4.2.3. Linear Classification

Classification: Given

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

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

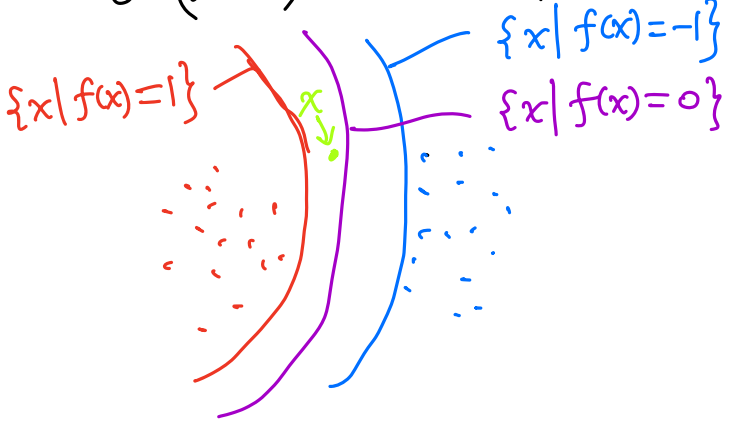
$y_1, y_2, \cdots, y_N \in \{-1, 1\}$

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

Find a function $f: \mathbb{R}^n \to \mathbb{R}$ 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.



- class −1
- class +1

Which function class should f be in?

— linear classifiers: Use affine functions.

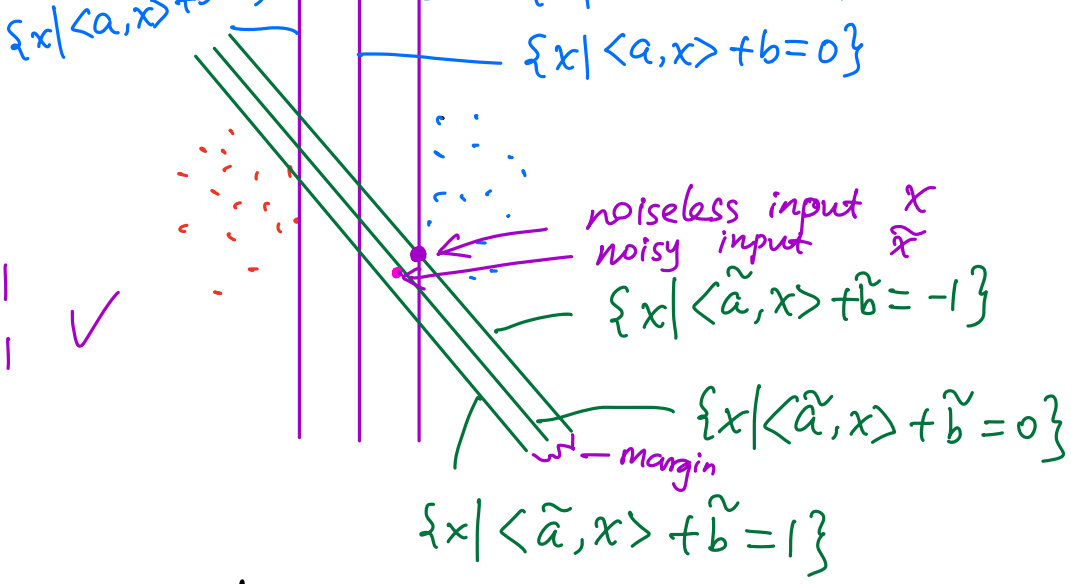
— Any affine function $f: \mathbb{R}^n \to \mathbb{R}$ must be in

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

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

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

The solution is not unique.



For x:

(a, b) -1

$(\tilde{a}, \tilde{b})$ -1 ✓

For $\tilde{x}$:

(a, b) -1

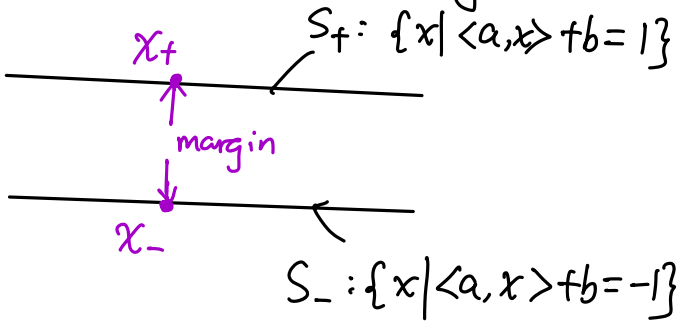
$(\tilde{a}, \tilde{b})$ $+1$

— which solution is better?

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

— larger margin, better classifier.

— Support vector machine (SVM): maximize the margin



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

Notice that: $x_- = P_{S_-} x_+$

$x_+ = P_{S_+} x_-$

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

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

$x_- \in S_- \Rightarrow \langle a, x_- \rangle + b = -1$

$\Rightarrow \langle a, x_- \rangle - (1-b) = -2$

So, $\|x_+ - x_-\|_2 = \frac{|-2|}{\|a\|_2^2} \|a\|_2 = \frac{2}{\|a\|_2}$

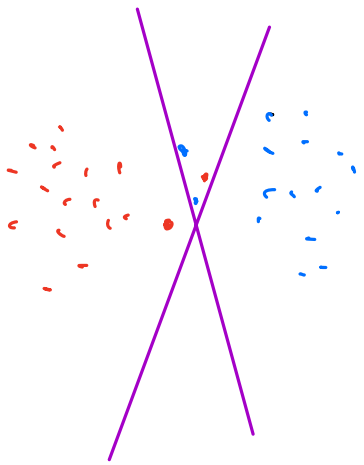
So, in SVM:

$$\begin{cases} \max\limits_{\substack{a\in\mathbb{R}^n \\ b\in\mathbb{R}}} \ \dfrac{2}{\|a\|_2} \\ \text{s.t} \ \begin{cases} \langle a, x_i\rangle + b \geq 1 & \text{if } y_i = 1 \\ \langle a, x_i\rangle + b \leq -1 & \text{if } y_i = -1 \end{cases} \quad i=1,\cdots,N \end{cases}$$

$$\Updownarrow$$

$$\begin{cases} \min\limits_{\substack{a\in\mathbb{R}^n \\ b\in\mathbb{R}}} \ \|a\|_2^2 \\ \text{s.t.} \quad y_i\left(\langle a, x_i\rangle + b\right) \geq 1, \quad i=1,\cdots,N \end{cases} \qquad \text{(SVM-1)}$$

- (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: $h : \mathbb{R} \to \mathbb{R}\cup\{+\infty\}$

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

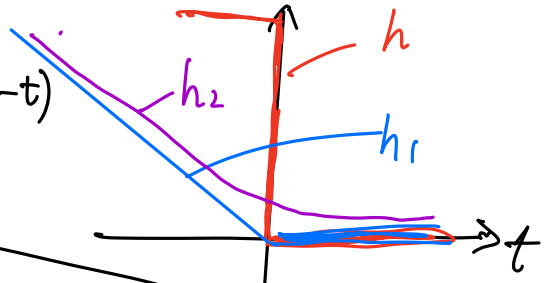
Then

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

where $\lambda > 0$

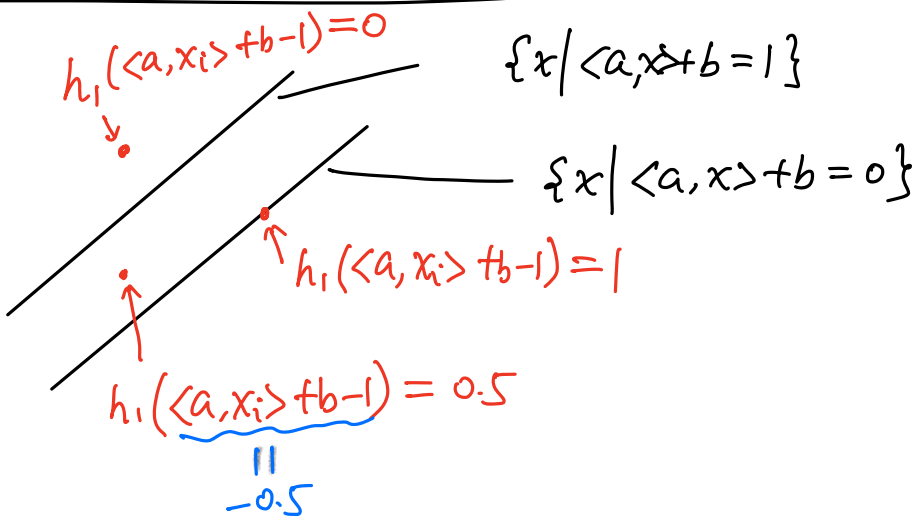
To avoid $+\infty$ function value for any (a,b), we approximate h by some function that takes only finite values.

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



Soft-Margin SVM

$$\min_{\substack{a \in \mathbb{R}^n \\ b \in \mathbb{R}}} \sum_{i=1}^{N} h_1\left(y_i(\langle a, x_i \rangle + b) - 1 \right) + \lambda \|a\|_2^2 \qquad (SVM-2)$$



- (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

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

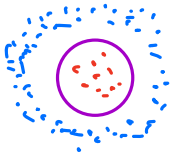
- h_2 is the logistic loss function
- h_2 has some statistical meaning.

So we solve

$$\min_{\substack{a\in\mathbb{R}^n \\ b\in\mathbb{R}}} \sum_{i=1}^{N} h_2\left(y_i(\langle a, x_i\rangle + b) - 1\right) + \lambda \|a\|_2^2$$

— logistic regression

All SVMs are linear. They will not work if



We choose non-linear $f: \mathbb{R}^n \to \mathbb{R}$

Kernel SVM:

f: ① $\phi: \mathbb{R}^n \to H$ $\quad x_i \to \phi(x_i)$

② SVM on H:

$f(x_i) = \langle a, \phi(x_i)\rangle_H$, for some $a \in H$

and we solve

$\tilde{h} = h, h_1, h_2$

(K-SVM)
$$\min_{a\in H} \sum_{i=1}^{N} \tilde{h}\left(y_i \cdot \langle a, \phi(x_i)\rangle - 1\right) + \lambda \|a\|_H^2,$$

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

$a \longleftrightarrow (a_S, c)$

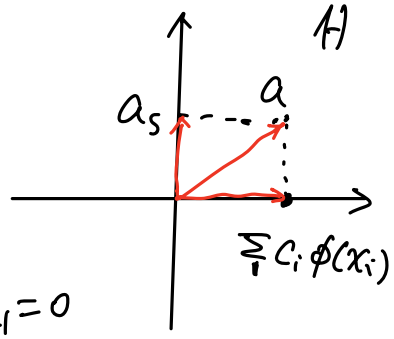
Representer Thm

Since $\forall a \in H$, 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

$\langle a_S, \phi(x_i)\rangle_H = 0$, $i = 1, \cdots, N$.



Then, the objective function in (K-SVM) is

$$\sum_{i=1}^{N} \tilde{h}\left(y_i \cdot \langle a, \phi(x_i)\rangle_H - 1\right) + \lambda \|a\|_H^2$$

$$= \sum_{i=1}^{N} \tilde{h}\left(y_i \cdot \langle a_S + \sum_{j=1}^{N} c_j \phi(x_j), \phi(x_i)\rangle_H - 1\right) + \lambda \left\| a_S + \sum_{j=1}^{N} c_j \phi(x_j)\right\|_H^2$$

$$= \sum_{i=1}^{N} \tilde{h}\left(y_i \cdot \left(\sum_{j=1}^{N} c_j \langle \phi(x_j), \phi(x_i)\rangle_H\right) - 1\right)$$

$$+ \lambda \left(\|a_S\|_H^2 + \sum_{j_1=1}^{N}\sum_{j_2=1}^{N} c_{j_1} c_{j_2} \langle \phi(x_{j_1}), \phi(x_{j_2})\rangle_H \right)$$

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

$$= \sum_{i=1}^{N} \tilde{h}\left(y_i \left(K^T c\right)_i - 1\right) + \lambda\, c^T K c \quad + \lambda \|a_S\|_H^2$$

Then $a = a_S + \sum_{j=1}^{N} c_j \phi(x_j)$

$$\text{(K-SVM)} \iff \begin{cases} \min\limits_{\substack{a_S \in H \\ c\in\mathbb{R}^N}} \sum\limits_{i=1}^{N} \tilde{h}\left(y_i \left(K^T c\right)_i - 1\right) + \lambda c^T K c + \lambda \|a_S\|_H^2 \\ \text{s.t.} \quad \langle a_S, \phi(x_j)\rangle_H = 0, \quad j = 1, \cdots, N \end{cases}$$

$$\iff \min_{a_S \in H} \|a_S\|_H^2 \quad \text{s.t.} \quad \langle a_S, \phi(x_j)\rangle_H = 0, \quad j = 1, \cdots, N$$

and

$$\min_{c\in\mathbb{R}^N} \sum_{i=1}^{N} \tilde{h}\left(y_i \left(K^T c\right)_i - 1\right) + \lambda c^T K c$$

$$\iff a_S^* = 0$$

and

$$c^* = \arg\min_{c\in\mathbb{R}^N} \sum_{i=1}^{N} \tilde{h}\left(y_i \left(K^T c\right)_i - 1\right) + \lambda c^T K c$$

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

② Form the kernel matrix $K = [K(x_i, x_j)]_{i,j=1}^N \in \mathbb{R}^{N\times N}$

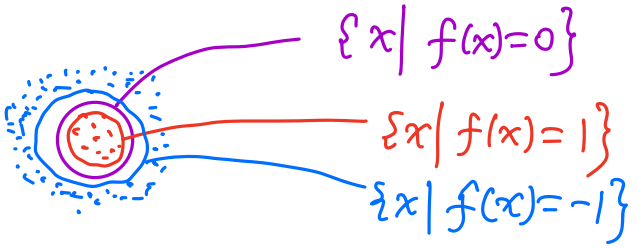
③ Solve

$$c^* = \arg\min_{c\in\mathbb{R}^N} \sum_{i=1}^N \tilde{h}\left(y_i\left((K^T c)_i - 1\right)\right) + \lambda c^T K c$$

④ The optimal $a^* = \sum_{j=1}^N c_j^* \phi(x_j)$, and the classification function is:

$$f(x) = \langle a^*, \phi(x)\rangle_H$$
$$= \left\langle \sum_{j=1}^N c_j^* \phi(x_j), \phi(x)\right\rangle_H = \sum_{j=1}^N c_j^* \langle \phi(x_j), \phi(x)\rangle_H$$
$$= \sum_{j=1}^N c_j^* K(x_j, x)$$

and $\operatorname{sgn}(f(x))$ is the class that x is in.



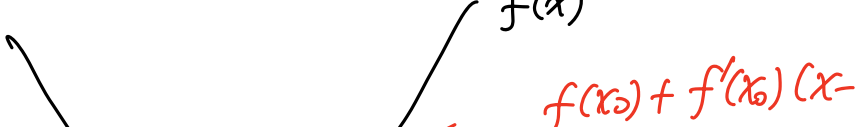
Linear functions have some limitations.
We still need non-linear functions

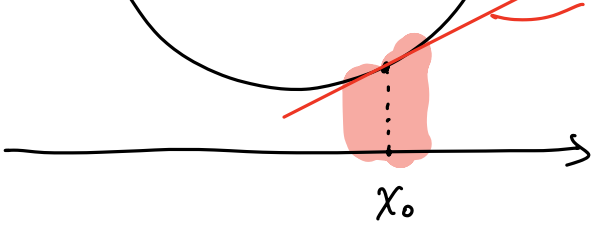
— functions by the kernel trick

— Neural network functions

— Optimization problem in Linear models

§4.3. Linear approximation and differentiation





Differentiation is a local affine approximation.

Recall for a function $f: \mathbb{R} \to \mathbb{R}$, its derivative at x_0 is

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

which is the same as

$$0 = \lim_{x \to x_0} \frac{|f(x) - (f(x_0) + f'(x_0)(x - x_0))|}{|x - x_0|}$$

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 $\lim_{x \to x_0} \frac{\text{error}}{|x - x_0|} = 0$

i.e., $\text{error} = o(|x - x_0|)$

little o notation.

Consider $f: V \to \mathbb{R}$, where V is a Hilbert space, and $x^{(0)} \in V$

Then, we find a function $g: V \to \mathbb{R}$ 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., $\lim_{x \to x^{(0)}} \frac{|f(x) - g(x)|}{\|x - x^{(0)}\|} = 0$

To get an explicit form of g:

① $\Rightarrow$ $g(x) = \langle v, x \rangle + a$, where $v \in V$, $a \in \mathbb{R}$

$= \langle v, x - x^{(0)} \rangle + \langle v, x^{(0)} \rangle + a$

② $\Rightarrow$ $g(x^{(0)}) = \langle v, x^{(0)} \rangle + a = f(x^{(0)})$

$$\Rightarrow \quad g(x) = f(x^{(0)}) + \langle v,\ x - x^{(0)} \rangle$$

③
$$\lim_{\|x-x^{(0)}\|\to 0} \frac{|f(x) - (f(x^{(0)}) + \langle v,\ x-x^{(0)}\rangle)|}{\|x - x^{(0)}\|} = 0$$

Definition: Let V be a Hilbert space. Let $f: V \to \mathbb{R}$. Then f is said Frechet differentiable at $x^{(0)}$ if

$$\exists\ v \in V \quad \text{s.t.}$$

$$\lim_{\|x-x^{(0)}\|\to 0} \frac{|f(x) - (f(x^{(0)}) + \langle v,\ x-x^{(0)}\rangle)|}{\|x - x^{(0)}\|} = 0 .$$

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 $x^{(0)} \in V$,

$$f(x) = \|x\|^2 = \|x^{(0)} + (x - x^{(0)})\|^2$$
$$= \|x^{(0)}\|^2 + 2\langle x^{(0)},\ x - x^{(0)}\rangle + \|x - x^{(0)}\|^2$$
$$= f(x^{(0)}) + \langle 2x^{(0)},\ x - x^{(0)}\rangle + \|x - x^{(0)}\|^2$$

$$\lim_{\|x-x^{(0)}\|\to 0} \frac{|f(x) - (f(x^{(0)}) + \langle 2x^{(0)},\ x-x^{(0)}\rangle)|}{\|x - x^{(0)}\|}$$

$$= \lim_{\|x-x^{(0)}\|\to 0} \frac{\|x - x^{(0)}\|^2}{\|x - x^{(0)}\|} = 0$$

So, $\nabla f(x^{(0)}) = 2x^{(0)}$

i.e. $\boxed{\nabla f(x) = 2x}$

Special cases: — $V = \mathbb{R}^n$, with $\langle \cdot, \cdot \rangle$, $f(x) = \|x\|_2^2$, then $\nabla(\|x\|_2^2) = 2x$

- $V=\mathbb{R}^n$ with $\langle\cdot,\cdot\rangle_A$, $f(x)=\|x\|_A^2$, then $\nabla(\|x\|_A^2)=2x$

- $V=\mathbb{R}^n$ with $\langle\cdot,\cdot\rangle$, $g(x)=\|x\|_A^2$, then $\nabla(\|x\|_A^2)\neq 2x$

- $V=\mathbb{R}^{m\times n}$ with $\langle\cdot,\cdot\rangle$, $f(X)=\|X\|_F^2$, then $\nabla(\|X\|_F^2)=2X$

- $V=L^2(0,1)$, $\langle f,g\rangle=\int_0^1 f(t)g(t)\,dt$

  $F(f)=\|f\|^2=\int_0^1|f(t)|^2\,dt$

  then $\nabla F(f)=2f$

Example 2: $f(x)=\langle a,x\rangle$ for some $a\in V$.

At $x^{(0)}\in V$,

$$f(x)=\langle a,x\rangle=\langle a,x^{(0)}\rangle+\langle a,x-x^{(0)}\rangle$$
$$=f(x^{(0)})+\langle a,x-x^{(0)}\rangle$$

$$\lim_{\|x-x^{(0)}\|\to 0}\frac{|f(x)-(f(x^{(0)})+\langle a,x-x^{(0)}\rangle)|}{\|x-x^{(0)}\|}=\lim_{\|x-x^{(0)}\|\to 0}\frac{0}{\|x-x^{(0)}\|}$$
$$=0$$

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 $x^{(0)}\in V$,

$$f(x)=\|x-a\|^2=\|(x^{(0)}-a)+(x-x^{(0)})\|^2$$
$$=\|x^{(0)}-a\|^2+2\langle x^{(0)}-a,x-x^{(0)}\rangle+\|x-x^{(0)}\|^2$$

$$= f(x^{(0)}) + \langle 2(x^{(0)}-a),\ x-x^{(0)}\rangle + \|x-x^{(0)}\|^2$$

$$\lim_{\|x-x^{(0)}\|\to 0} \frac{\left| f(x) - \left(f(x^{(0)}) + \langle 2(x^{(0)}-a),\ x-x^{(0)}\rangle\right)\right|}{\|x-x^{(0)}\|}$$

$$= \lim_{\|x-x^{(0)}\|\to 0} \frac{\|x-x^{(0)}\|^2}{\|x-x^{(0)}\|} = 0$$

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.

$\forall\ \alpha,\beta \in \mathbb{R}$, $f, g : V \to \mathbb{R}$,

$$\nabla(\alpha f + \beta g)(x) = \alpha \nabla f(x) + \beta \nabla g(x),$$

provided $\nabla f(x)$ and $\nabla g(x)$ exist.

$V \xrightarrow{f} \mathbb{R} \xrightarrow{g} \mathbb{R}$

② Chain rule

Let $f: V \to \mathbb{R}$ and $g: \mathbb{R} \to \mathbb{R}$ (Then $g\circ f : V \to \mathbb{R}$)

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.

$$0 \le \lim_{\|y-x\|\to 0} \frac{\left|g(f(y)) - \left(g(f(x)) + \langle g'(f(x))\cdot \nabla f(x),\ y-x\rangle\right)\right|}{\|y-x\|}$$

$$+ g'(f(x))(f(y)-f(x)) - g'(f(x))(f(y)-f(x))$$

$$\le \lim_{\|y-x\|\to 0} \left(\underbrace{\frac{\left|g(f(y)) - \left(g(f(x)) + g'(f(x))(f(y)-f(x))\right)\right|}{\|y-x\|}}_{I_1} + \underbrace{\frac{\left|g'(f(x))\left(f(y) - \left(f(x) + \langle \nabla f(x),\ y-x\rangle\right)\right)\right|}{\|y-x\|}}_{I_2} \right) = 0$$

I_1

I_2

For I_2: $\lim_{\|y-x\|\to 0} I_2 = |g'(f(x))| \cdot \lim_{\|y-x\|\to 0} \frac{|f(y)-(f(x)+\langle \nabla f(x), y-x\rangle)|}{\|y-x\|} = 0$

(we can prove $|g'(f(x))| < +\infty$ using differentiability of g at $f(x)$)

For I_1:

$$\lim_{\|y-x\|\to 0} I_1 = \lim_{\|y-x\|\to 0} \left(\underbrace{\frac{|g(f(y)) - (g(f(x)) + g'(f(x))(f(y)-f(x)))|}{|f(y)-f(x)|}}_{I_3} \cdot \underbrace{\frac{|f(y)-f(x)|}{\|y-x\|}}_{I_4} \right)$$

— For I_3: $\lim_{\|y-x\|\to 0} I_3 = 0$ 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 $\|y-x\|$)

$$\Rightarrow \boxed{\lim_{\|y-x\|\to 0} I_1 = 0}$$

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 $t \in \mathbb{R}$

Then $f(x) = \sqrt{\|x\|^2} = f_2(f_1(x))$ 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 $\nabla f_1(x) = \nabla(\|x\|^2) = 2x$

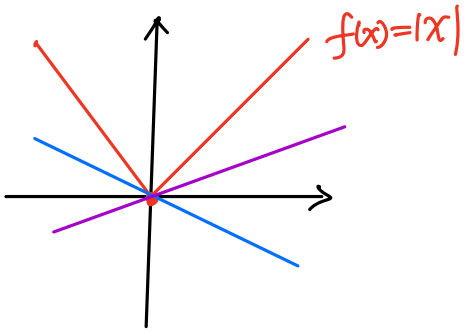
$f_2'(t) = (\sqrt{t})' = \frac{1}{2\sqrt{t}}$

$$\Rightarrow \quad \nabla f(x) = \frac{1}{2\sqrt{\|x\|^2}} \cdot 2x = \frac{x}{\|x\|}$$

— 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:

— $V=\mathbb{R}$, $f(x)=|x|$, $\nabla(|x|) = \frac{x}{|x|} = \begin{cases} 1 & \text{if } x>0 \\ -1 & \text{if } x<0 \end{cases}$ for $x\neq 0$

— $V=\mathbb{R}^n$, $f(x)=\|x\|_2$, $\nabla(\|x\|_2) = \frac{x}{\|x\|_2} = \frac{1}{\|x\|_2}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$

⋮

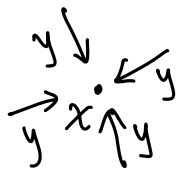
③ For functions $f: \mathbb{R}^n \to \mathbb{R}$, where $\mathbb{R}^n$ is with the Euclidean inner product, if f is differentiable at $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$, then

then $$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}$$

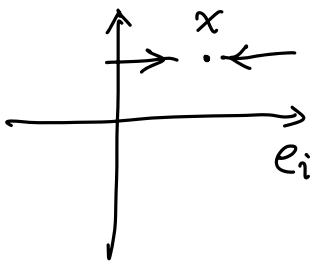
(Frechet differentiation is consistent with the standard differentiation in multi-variant calculus)

proof. Since f is differentiable at $x\in\mathbb{R}^n$,

$$\lim_{\|y-x\|\to 0} \frac{|f(y) - (f(x) + \langle \nabla f(x), y-x\rangle)|}{\|y-x\|_2} = 0$$



Choose $y = x + t\cdot e_i$, where $t\in\mathbb{R}$ $\quad e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ ← i-th entry



and $t \to 0$ (i.e., $g(t) = f(x+te_i)$)

$$0 = \lim_{t\to 0} \frac{\left| \overbrace{f(x+te_i)}^{g(t)} - \left(\overbrace{f(x)}^{g(0)} + t\langle \nabla f(x), e_i \rangle \right) \right|}{|t-0|}$$

$$0 = \lim_{t\to 0} \frac{\left| g(t) - \left(g(0) + \langle \nabla f(x), e_i \rangle \cdot (t-0) \right) \right|}{|t-0|}$$

So, $g'(0) = \langle \nabla f(x), e_i \rangle$ ← i-th component of $\nabla f(x)$.

$$\| \quad \frac{d}{dt} g(t)\Big|_{t=0} = \frac{d}{dt} f(x+te_i)\Big|_{t=0} = \frac{\partial f}{\partial x_i}(x)$$

So, $\left[\nabla f(x)\right]_i = \frac{\partial f}{\partial x_i}(x) \quad \forall i$ ⊠

Example: $f(x) = \|x\|_2$, where $x \in \mathbb{R}^n$

$$\frac{\partial f}{\partial x_i}(x) = \frac{\partial \|x\|_2}{\partial x_i} = \frac{\partial \sqrt{x_1^2+x_2^2+\cdots+x_n^2}}{\partial x_i}$$

$$= \frac{1}{2} \cdot \frac{1}{\sqrt{x_1^2+x_2^2+\cdots x_n^2}} \cdot \frac{\partial (x_1^2+x_2^2+\cdots+x_n^2)}{\partial x_i}$$

$$= \frac{1}{2} \frac{1}{\|x\|_2} \cdot \frac{\partial x_i^2}{\partial x_i} = \frac{1}{2} \cdot \frac{1}{\|x\|_2} \cdot 2x_i$$

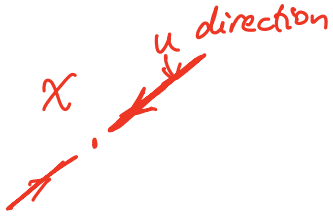
$$= \frac{x_i}{\|x\|_2} \quad \text{if} \quad \|x\|_2 \neq 0$$

So, $\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} = \frac{1}{\|x\|_2} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \frac{x}{\|x\|_2}$ ⊠

④ Let $f: V \to \mathbb{R}$, Assume f is differentiable at $x \in V$.

Let $u \in V$

Then $\langle \nabla f(x), u \rangle = \frac{d}{dt} f(x+tu)\Big|_{t=0}$



↑ directional derivative of f along u.

proof. the same as ③ by replacing e_i with u.

⑤ Taylor's expansion.

For $f: \mathbb{R} \to \mathbb{R}$, Taylor's expansion

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + o(|x-x_0|)$$

For $f: V \to \mathbb{R}$, because: if f is diff at $x^{(0)} \in V$,

then $$\lim_{\|x-x^{(0)}\| \to 0} \frac{|f(x) - (f(x^{(0)}) + \langle \nabla f(x^{(0)}), x-x^{(0)} \rangle)|}{\|x-x^{(0)}\|} = 0$$

$$\Updownarrow$$

$$\boxed{f(x) = f(x^{(0)}) + \langle \nabla f(x^{(0)}), x-x^{(0)} \rangle + o(\|x-x^{(0)}\|)}$$

↑ Taylor expansion

Differentiation of functions on Banach spaces

Let V be a Banach space with a norm $\|\cdot\|$

Let $f: V \to \mathbb{R}$

We use affine approximation at $x^{(0)} \in V$

We find a function $g: V \to \mathbb{R}$ 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. $\lim_{\|x-x^{(0)}\| \to 0} \frac{|f(x)-g(x)|}{\|x-x^{(0)}\|} = 0$)

① $\Rightarrow g(x) = Lx + a$, where $L: V \to \mathbb{R}$ linear and bounded. $a \in \mathbb{R}$

② $\Rightarrow g(x^{(0)}) = Lx^{(0)} + a = f(x^{(0)})$

$\Rightarrow g(x) = Lx + a = L((x-x^{(0)}) + x^{(0)}) + a$

$= L(x-x^{(0)}) + Lx^{(0)} + a$

$= f(x^{(0)}) + L(x-x^{(0)})$

③ $\lim_{\|x-x^{(0)}\|\to 0} \frac{|f(x) - (f(x^{(0)}) + L(x-x^{(0)}))|}{\|x-x^{(0)}\|} = 0$

Definition: Let V be Banach space with $\|\cdot\|$, and $x^{(0)} \in V$.
Let $f: V \to \mathbb{R}$. Then f is differentiable if:

$\exists$ a linear and bounded function $L: V \to \mathbb{R}$ s.t.

$$\lim_{\|x-x^{(0)}\|\to 0} \frac{|f(x) - (f(x^{(0)}) + L(x-x^{(0)}))|}{\|x-x^{(0)}\|} = 0$$

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

$\min_{x\in V} f(x)$ (OPT), where V is a Hilbert space and $f: V \to \mathbb{R}$

§4.4.1. Solvability and Optimality

- Solvability of (OPT)
  - We say $x^{(*)} \in V$ is a solution of (OPT) if

$$f(x^{(*)}) \leq f(x) \quad \forall\, x \in V$$

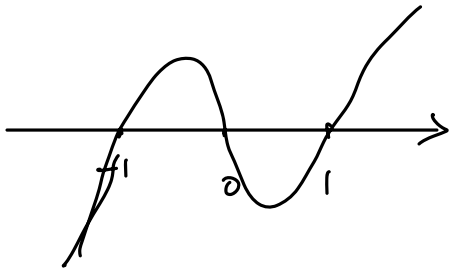
(0-th order optimality condition)

We also call $x^{(*)}$ a global minimizer of f, denoted by $x^{(*)} = \arg\min_{x\in V} f(x)$

  - The existence of a solution of (OPT) is not guaranteed automatically.

Some examples:

- $f: \mathbb{R} \to \mathbb{R}$ $\quad f(x) = x(x-1)(x+1)$



when $x \to -\infty$,

$f(x) \to -\infty$.

- $f: V \to \mathbb{R}$, $f(x) = \langle a, x \rangle$ for some $a \in V$.

  Then set $x = ca$ for $c \in \mathbb{R}$.

  So, $f(x) = f(ca) = \langle ca, a \rangle = c\|a\|^2 \to -\infty$

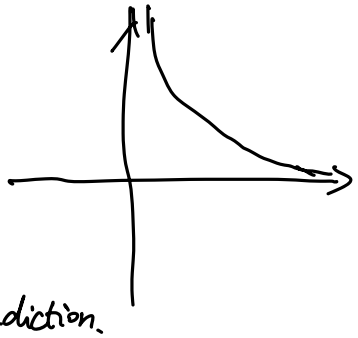
  as $c \to -\infty$

- $f: \mathbb{R} \to \mathbb{R}$, $f(x) = \frac{1}{x}$ $x > 0$

  So, $f(x) > f(x+1)$ $\forall x > 0$.

  If we have $x^{(*)} = \arg\min_{x \in \mathbb{R}} f(x)$,

  then $f(x^{(*)}) > f(x^{(*)}+1)$ Contradiction.



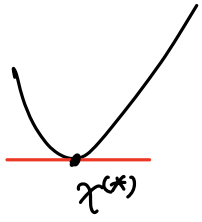
- characterization of the solution of (OPT).

  — 0-th order optimality condition

$$x^{(*)} = \arg\min_{x \in V} f(x) \iff \boxed{f(x^{(*)}) \leq f(x) \quad \forall\, x \in V}$$

(impossible to check numerically)

  — 1-st order optimality condition



Thm: Assume $f: V \to \mathbb{R}$ is differentible at $x^{(*)} \in V$.

Then

$$x^{(*)} = \arg\min_{x \in V} f(x) \implies \nabla f(x^{(*)}) = 0$$

Proof. By expansion,

$$f(x) = f(x^{(*)}) + \langle \nabla f(x^{(*)}),\, x - x^{(*)} \rangle + o\left(\|x - x^{(*)}\|\right)$$

Suppose $\nabla f(x^{(*)}) \neq 0$

Choose

$$\tilde{x} = x^{(*)} - t \cdot \nabla f(x^{(*)}) \quad \text{with} \quad t>0$$

So,

$$f(\tilde{x}) = f(x^{(*)}) + \langle \nabla f(x^{(*)}), -t \cdot \nabla f(x^{(*)}) \rangle + o\left(|t| \|\nabla f(x^{(*)})\|\right)$$
$$= f(x^{(*)}) - t \|\nabla f(x^{(*)})\|^2 + o\left(|t| \cdot \|\nabla f(x^{(*)})\|\right)$$

Because $$\lim_{|t|\to 0} \frac{o\left(|t| \cdot \|\nabla f(x^{(*)})\|\right)}{|t| \cdot \|\nabla f(x^{(*)})\|} = 0,$$

$$\Rightarrow \forall C>0, \quad \exists t \quad \text{s.t.} \quad o\left(|t| \cdot \|\nabla f(x^{(*)})\|\right) < C \cdot |t| \cdot \|\nabla f(x^{(*)})\|$$

We choose $C = \frac{1}{2}\|\nabla f(x^{(*)})\|$

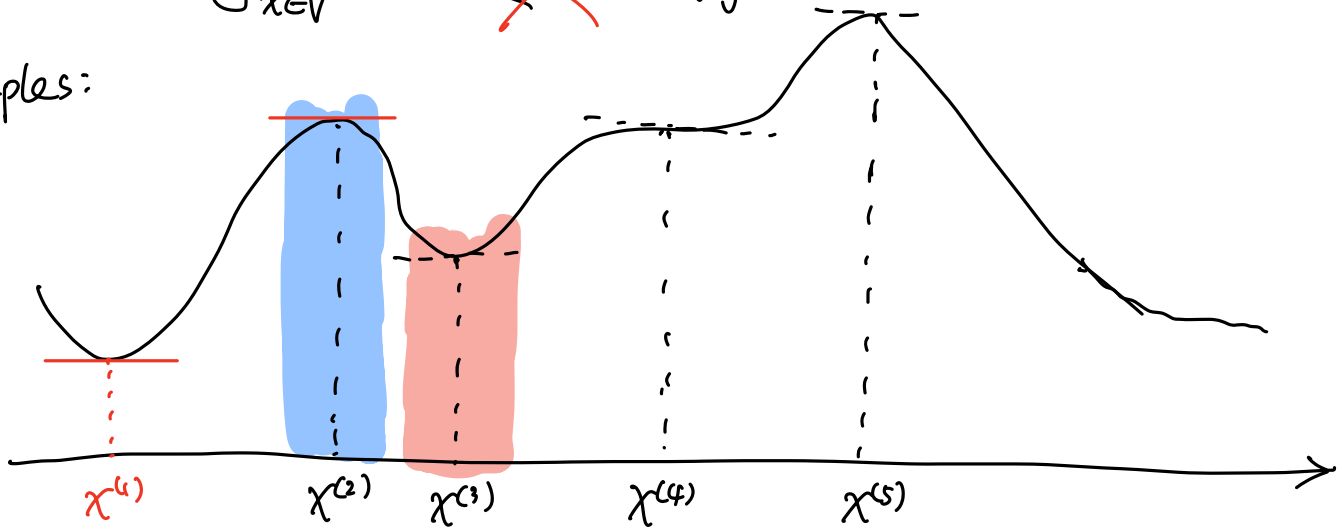
Then $$o\left(|t| \|\nabla f(x^{(*)})\|\right) < \frac{1}{2}|t| \cdot \|\nabla f(x^{(*)})\|^2$$

Then $$f(\tilde{x}) = f(x^{(*)}) - t \|\nabla f(x^{(*)})\|^2 + \frac{1}{2}|t| \cdot \|\nabla f(x^{(*)})\|^2$$
$$< f(x^{(*)}) \quad \text{contradition} \quad \otimes$$

The reverse is not true in general, i.e,

$$x^{(*)} = \arg\min_{x \in V} f(x) \not\Longrightarrow \nabla f(x^{(*)}) = 0$$

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., $x^{(3)}$)

  $f(x^{(3)}) \le f(x)$ $\forall\, x \in V$ satisfying $\|x - x^{(3)}\| \le \varepsilon$ for some ε.

- Local maximizer (e.g., $x^{(2)}$)

  $f(x^{(2)}) \ge f(x)$ $\forall\, x \in V$ satisfying $\|x - x^{(2)}\| \le \varepsilon$ for some ε.

- Global maximizer (e.g., $x^{(5)}$)

  $f(x^{(5)}) \ge f(x)$ $\forall\, x \in V$.

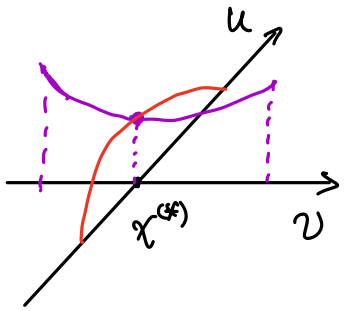
- No information on optimality (e.g., $x^{(4)}$)

- Saddle point (only for V with $\dim(V) \ge 2$)

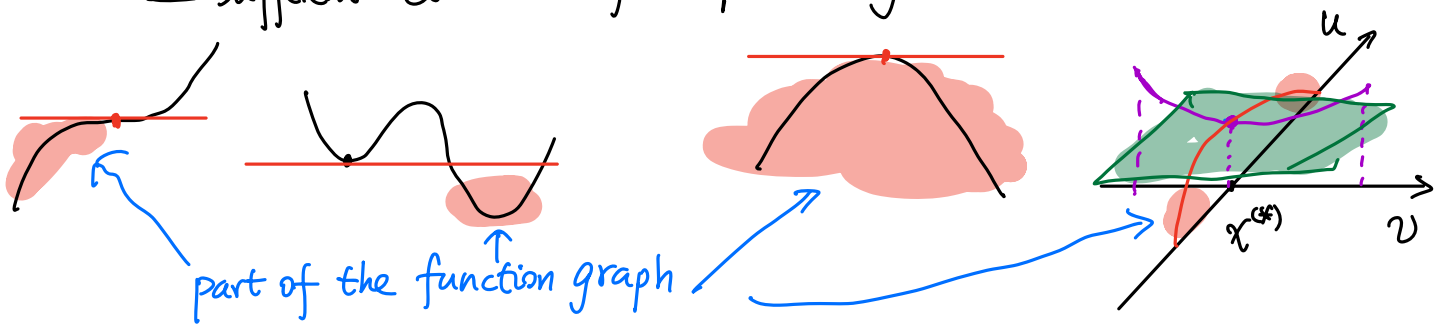
  $\exists\, u, v \in V$

  $$\text{s.t.} \begin{cases} f(x^{(*)}) \ge f(x^{(*)} + tu) \\ f(x^{(*)}) \le f(x^{(*)} + tv) \end{cases}$$

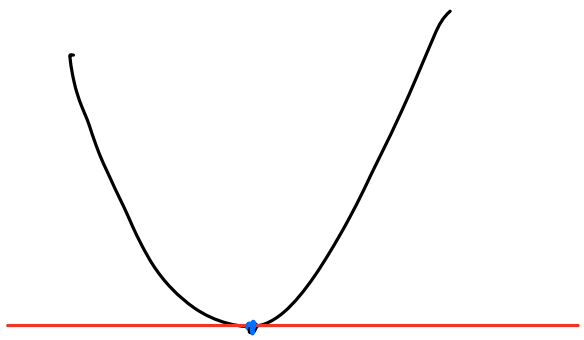
  $\forall\, t \in \mathbb{R} : |t| \le \varepsilon$ for some ε.



- Suffcient condition for optimality.



part of the function graph is below the affine approximation



all and function graph is above affine approximation

$x^{(*)}$ is a global minimizer

Theorem: Let $f: V \to \mathbb{R}$, where V is a Hilbert space.

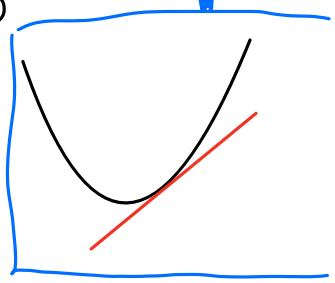
(☆) Assume f is differentiable and

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \qquad \forall\, x, y \in V$$

(Affine approximation always support the function)

Then,

$$x^{(*)} = \operatorname*{argmin}_{x \in V} f(x) \iff \nabla f(x^{(*)}) = 0$$



Proof. "$\Rightarrow$" proved in the previous theorem.

"$\Leftarrow$": Since $\nabla f(x^{(*)}) = 0$,

$$\begin{aligned} f(x) &\geq f(x^{(*)}) + \langle \nabla f(x^{(*)}), x - x^{(*)} \rangle \\ &= f(x^{(*)}) + \langle 0, x - x^{(*)} \rangle \qquad \forall x \in V \\ &= f(x^{(*)}) \end{aligned}$$

So, $x^{(*)} = \operatorname*{arg\,min}_{x \in V} f(x)$ ⊠

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

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

$$\Rightarrow f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \qquad \forall\, x, y \in V$$

Ex. 2: Let $f_1, f_2, \cdots, f_n : V \to \mathbb{R}$ all satisfy (☆)

Then, $f = \sum_{i=1}^{n} c_i f_i$, where $c_i \geq 0$, $i = 1, \cdots, n$

also satisfy (☆)

proof. f_i satisfies (☆): $i = 1, \cdots, n$

$$f_i(y) \geq f_i(x) + \langle \nabla f_i(x), y - x \rangle \qquad \forall\, x, y \in V$$

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,

$$\sum_{i=1}^{n} c_i f_i(y) \geq \sum_{i=1}^{n} c_i f_i(x) + \left\langle \sum_{i=1}^{n} c_i \cdot \nabla f_i(x), y - x \right\rangle$$

i.e., $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ ☒

§4.4.2. Gradient Descent

Goal: Solve $\min_{x \in V} f(x)$ numerically, where $f: V \to \mathbb{R}$ 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.

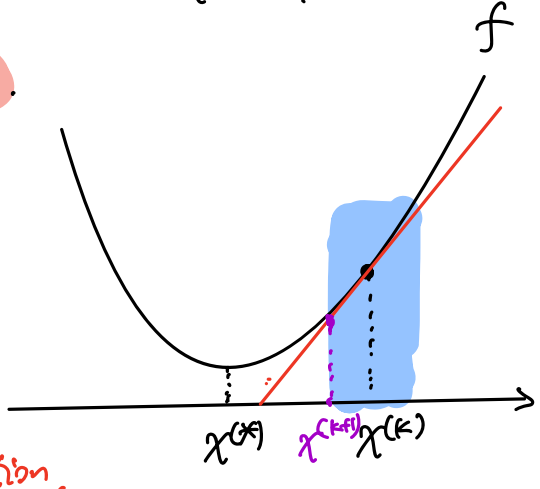
Instead of

$\min_{x \in V} f(x)$

↑ approximate it by $\|x - x^{(k)}\| \leq$ Constant

↑ approximated by affine approximation at $x^{(k)}$

$f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle$



We define

$$x^{(k+1)} = \arg\min_{\text{s.t. } \|x-x^{(k)}\| \le \alpha_k \cdot \|\nabla f(x^{(k)})\|} f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle \quad (*)$$

Notice that (*) where $\alpha_k > 0$ is a constant.

$$(*) \iff \begin{cases} \min \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle \\ \text{s.t} \quad \|x - x^{(k)}\| \le \alpha_k \cdot \|\nabla f(x^{(k)})\| \end{cases}$$

By Cauchy-Schwartz,

$$(**) \quad \begin{aligned} \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle &\overset{①}{\ge} -\|\nabla f(x^{(k)})\| \cdot \|x - x^{(k)}\| \\ &\overset{②}{\ge} -\|\nabla f(x^{(k)})\| \cdot \alpha_k \cdot \|\nabla f(x^{(k)})\| \\ &= -\alpha_k \|\nabla f(x^{(k)})\|^2 \end{aligned}$$

Both ① and ② can be "=" by choosing some x.

- ① becomes "=" when $x - x^{(k)} = -c \cdot \nabla f(x^{(k)})$ for some $c \ge 0$
- ② becomes "=" when $\|x - x^{(k)}\| = \alpha_k \cdot \|\nabla f(x^{(k)})\|$

i.e., $\|x - x^{(k)}\| = c \cdot \|\nabla f(x^{(k)})\| = \alpha_k \|\nabla f(x^{(k)})\|$

$\Rightarrow c = \alpha_k$

Altogether, when x satisfies $x - x^{(k)} = -\alpha_k \nabla f(x^{(k)})$,

we have $\begin{cases} \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle = -\alpha_k \|\nabla f(x^{(k)})\|^2 \\ \text{and } \|x - x^{(k)}\| \le \alpha_k \|\nabla f(x^{(k)})\| \end{cases}$

By (**):

$$\begin{cases} \min \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle \\ \text{s.t } \|x - x^{(k)}\| \le \alpha_k \cdot \|\nabla f(x^{(k)})\| \end{cases} \ge -\alpha_k \|\nabla f(x^{(k)})\|$$

So, $x^{(k+1)} - x^{(k)} = -\alpha_k \nabla f(x^{(k)})$

i.e., $$\boxed{x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})}$$

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
    $$\lim_{k\to+\infty} \nabla f(x^{(k)}) = 0 \quad \left(\text{Roughly, } \nabla f(x^{(\infty)}) = 0\right)$$
    - If f satisfies (☆), $x^{(\infty)} = \arg\min_{x\in V} f(x)$
    - Otherwise, $x^{(\infty)}$ 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: $\min_{\beta=\begin{bmatrix}a\\b\end{bmatrix}\in\mathbb{R}^{n+1}} \|X\beta - y\|_2^2 + \lambda\|a\|_2^2$

$$\|X\beta - y\|_2^2 + \lambda\|a\|_2^2 = \left\| \begin{matrix} X\beta - y \\ \sqrt{\lambda}\, a \end{matrix} \right\|_2^2$$

$$= \left\| \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} y \\ 0 \end{bmatrix} \right\|_2^2$$

$A \quad x \quad b$

– Kernel Ridge Regression: Can be reformulated as LS.

Let $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ (we want to. $\min_{x\in\mathbb{R}^n} f(x)$)

– $f: \mathbb{R}^n \to \mathbb{R}$ is differentiable. Let's find ∇f.

Use definition of gradient

$y = (x + (y-x))$

$$f(y) = \frac{1}{2}\|Ay-b\|_2^2 = \frac{1}{2}\|(Ax-b)+(Ay-Ax)\|_2^2$$

$$= \frac{1}{2}\|Ax-b\|_2^2 + \langle Ax-b, Ay-Ax\rangle + \frac{1}{2}\|Ay-Ax\|_2^2$$

$$= f(x) + \langle Ax-b, A(y-x)\rangle + \frac{1}{2}\|A(y-x)\|_2^2$$

$$= f(x) + \langle A^T(Ax-b), y-x\rangle + \frac{1}{2}\|A(y-x)\|_2^2 \quad (*)$$

$\langle u, Av\rangle = u^T A v = (A^T u)^T v = \langle A^T u, v\rangle$ ← 证明1

$$\lim_{\|y-x\|_2 \to 0} \frac{|f(y) - (f(x) + \langle A^T(Ax-b), y-x\rangle)|}{\|y-x\|_2}$$

$$\overset{(*)}{=} \lim_{\|y-x\|_2 \to 0} \frac{\frac{1}{2}\|A(y-x)\|_2^2}{\|y-x\|_2}$$

$\|Au\|_2 \le \|A\|_2 \|u\|_2$ ← C-S

$$\le \lim_{\|y-x\|_2 \to 0} \frac{\frac{1}{2}\|A\|_2^2 \cdot \|y-x\|_2^2}{\|y-x\|_2} = 0$$

其导数

$$\Rightarrow \nabla f(x) = A^T(Ax-b)$$

至此，已证明了f可导且求导了

L9

L10 – $\forall\, y, x \in \mathbb{R}^n$,

$$f(y) - \left(f(x) + \langle \nabla f(x), y-x\rangle\right)$$

证明f is convex

$$= f(y) - \left(f(x) + \langle A^T(Ax-b), y-x\rangle\right)$$

(*) 见上文

$$= \frac{1}{2}\|A(y-x)\|_2^2 \geq 0$$, so f is convex

Therefore,

$$x^{(*)} = \arg\min_{x\in\mathbb{R}^n}\left(\frac{1}{2}\|Ax-b\|_2^2\right) \iff A^T(Ax^{(*)}-b)=0$$

$$\iff A^TA\,x^{(*)} = A^Tb$$

A^TA: $n\times n$

正规方程 normal equation

Solve LS ⟺ Solve the normal equation of LS

↳ why? There is a very simple Geometry interpretation as below.

(至此, HM3结束...)

- Geometry (几何解释)

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|\overset{m\times n}{A}x-b\|_2^2$$

$\Updownarrow$ $y = \overset{m\times n}{A}x$ (把 Ax 记作 y), 改写成

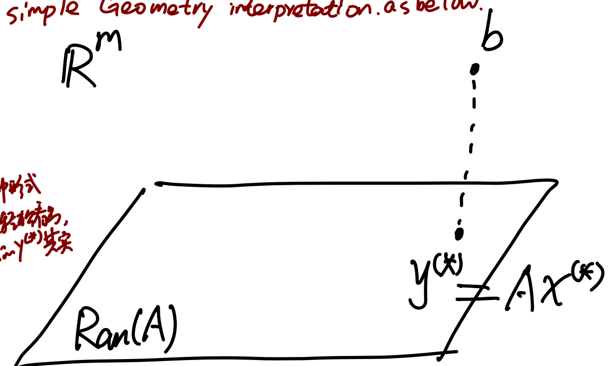
↳ 为什么要写成这种形式? 因为通过这种形式我们可以轻松看出, 我们要求的 $y^{(*)}$ 其实就是:

$$\begin{cases} \min_{y\in\mathbb{R}^m} \frac{1}{2}\|y-b\|_2^2 \\ \text{s.t. } y\in \operatorname{Ran}(A) \end{cases}$$

s.t.: Subject to (表达一个 constraint)

$\operatorname{Ran}(A)$: $\operatorname{col}(A)$; Range of (A) i.e. A 的列空间

√ Projection of b onto $\operatorname{Ran}(A)$, 如右图



Because $y^{(*)}$ is projection,

$b - Ax^{(*)} \perp \operatorname{Ran}(A)$

i.e., $\forall z\in\mathbb{R}^n$,

$$\langle b-Ax^{(*)}, Az\rangle = 0$$

$\Updownarrow$ 由上定义

$$\langle A^T(b-Ax^{(*)}), z\rangle = 0 \quad \forall z\in\mathbb{R}^n$$

As the equation is correct for $\forall z\in\mathbb{R}^n$, so we choose a special z:

$$z = A^T(b-Ax^{(*)})$$

Then

$$\langle A^T(b-Ax^{(*)}), A^T(b-Ax^{(*)})\rangle = 0$$

$$\|A^T(b-Ax^{(*)})\|_2^2 = 0$$

$$A^T(b-Ax^{(*)}) = 0$$

$$A^TA\,x^{(*)} = A^Tb.$$

in fact, the reversal is also true, since $A^T(b-Ax^{(*)})=0$ 代入

- Algorithms for LS ($f(x) = \frac{1}{2}\|Ax-b\|_2^2$, $A\in\mathbb{R}^{m\times n}$, $x\in\mathbb{R}^n$, $b\in\mathbb{R}^m$.)

— Solve the normal equation

$$A^TA\,x^{(*)} = A^Tb$$ by Cholesky decomposition

(LAPACK) linear algebra package

— Gradient Descent:

$$x^{(k+1)} = x^{(k)} - \alpha_k \cdot A^T(Ax^{(k)} - b)$$

- How to choose the step size α_k?

  We can choose an optimal (最优的) α_k by exact line search (精确线搜索).

  在每一步中都找一个使目标函数下降最快的 α_k (step)

  Define $x(\alpha) = x^{(k)} - \alpha A^T(Ax^{(k)} - b)$, where $\alpha \in \mathbb{R}$

  (从 $x^{(k)}$ 处依 α 下降后的 x; $A^T(Ax^{(k)} - b) = \nabla f(x^{(k)})$)

  We solve
  $$\alpha_k = \arg\min_{\alpha \in \mathbb{R}} f(x(\alpha)) \overset{\text{def}}{\equiv} g(\alpha).$$

  which is very intuitive.

  Then $g: \mathbb{R} \to \mathbb{R}$ is differentiable and satisfies
  $$g(\beta) \geq g(\alpha) + g'(\alpha) \cdot (\beta - \alpha) \quad \forall \alpha, \beta \in \mathbb{R}$$
  为什么? 不用管.

  细节不用管

  So, α_k satisfies $\boxed{g'(\alpha_k) = 0}$

  : 易得, 不用管. 细节不用管.

  $$\boxed{\alpha_k = \frac{\|A^T(Ax^{(k)} - b)\|_2^2}{\|AA^T(Ax^{(k)} - b)\|_2^2}}$$

Full alg:

```
Initialize x^(0) ∈ R^n.
for k = 0, 1, 2, ...
    g^(k) = A^T(A x^(k) − b)          (→ ∇f(x^(k)))
    α_k = ||g^(k)||_2^2 / ||A g^(k)||_2^2
    x^(k+1) = x^(k) − α_k g^(k)
endfor
```

此处 g 与上面 $g(\alpha)$ 无关, 为避免混淆, 换为 h

It is also known as the Steepest Descent algorithm.

→ 因为每步都找 ↓ 最快的 α.

Ch 5. Linear Transformations / Operators and Differentiation

§ 5.1. Linear transformations / operators

Let V_1, V_2 be two vector spaces

A mapping $L: V_1 \to V_2$ is a linear operator (or linear transformation)

if $L(\alpha x+\beta y) = \alpha \cdot L(x) + \beta \cdot L(y)$ $\forall\ x, y \in V,\ \alpha, \beta \in \mathbb{R}$

It can easily shown that: if L is linear then $L(\sum_{i=1}^{k} \alpha_i v_i) = \sum_{i=1}^{k} \alpha_i \cdot L(v_i)$

Example 1: Let $A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}$

- Define a mapping $L: \mathbb{R}^n \to \mathbb{R}^m$ by
$$L(x) = Ax \qquad \forall x \in \mathbb{R}^n$$
Then L is a linear transformation because
$$L(\alpha x+\beta y) = A(\alpha x+\beta y) = \alpha \cdot Ax + \beta \cdot Ay \quad \forall\ x, y \in \mathbb{R}^n,\ \alpha, \beta \in \mathbb{R}$$
$$= \alpha \cdot L(x) + \beta \cdot L(y)$$

- Reversly, any linear transformation $L: \mathbb{R}^n \to \mathbb{R}^m$ must be in the form of $L(x) = Ax$ $\forall x \in \mathbb{R}^n$ for some $A \in \mathbb{R}^{m \times n}$

proof. $\forall x \in \mathbb{R}^n$, $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n$

Then $L(x) = L(x_1 e_1 + x_2 e_2 + \cdots + x_n e_n)$
$$= x_1 \cdot L(e_1) + x_2 \cdot L(e_2) + \cdots + x_n \cdot L(e_n)$$
$$= \underbrace{[L(e_1)\ L(e_2)\ \cdots\ L(e_n)]}_{\text{def } A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{x}$$
$$= Ax$$ ⊠

Example 2: Discrete Fourier Transform (DFT) is a linear transformation $\mathbb{C}^n \to \mathbb{C}^n$ defined as follows:

Let $w_n = e^{-\frac{2\pi}{n} i}$, where $i = \sqrt{-1}$ $\quad \left(w_n^n = 1 \right)$
$$\left(= \cos\left(-\tfrac{2\pi}{n}\right) + i \sin\left(-\tfrac{2\pi}{n}\right) \right)$$

and define
$$F_n = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & w_n & w_n^2 & \cdots & w_n^{n-1} \\ 1 & w_n^2 & w_n^4 & \cdots & w_n^{2(n-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w_n^{n-1} & w_n^{2(n-1)} & \cdots & w_n^{(n-1)^2} \end{bmatrix} \in \mathbb{C}^{n \times n}$$

Then for any $x \in \mathbb{C}^n$, its DFT is

$$\hat{x} = F_n x$$

We can plot the j-th column of F_n

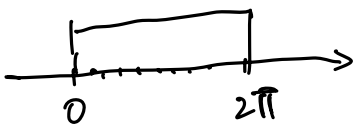
Real part:

$$\mathrm{Re}\begin{pmatrix} 1 \\ w_n^j \\ w_n^{2j} \\ \vdots \\ w_n^{(n-1)j} \end{pmatrix} = \begin{pmatrix} \cos 0 \\ \cos\left(j\frac{2\pi}{n}\right) \\ \cos\left(2j\cdot\frac{2\pi}{n}\right) \\ \vdots \\ \cos\left((n-1)j\cdot\frac{2\pi}{n}\right) \end{pmatrix}$$

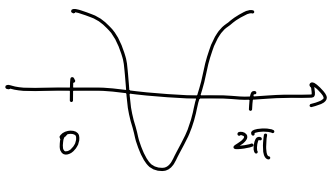
← discretization of $\cos(j\theta)$,

$\theta = 0, \frac{2\pi}{n}, 2\cdot\frac{2\pi}{n}, \cdots, (n-1)\frac{2\pi}{n}$.

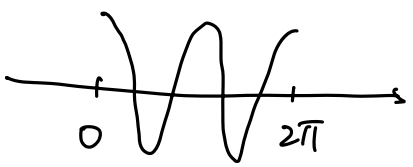
$j=0$



$j=1$



$j=2$



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 $a \circledast : \mathbb{R}^n \to \mathbb{R}^n$ by

$$(a \circledast x)_k = \sum_{(i+j) \bmod n = k} a_i x_j \qquad k = 0, 1, \cdots, n-1$$

$$= \sum_{i=0}^{n-1} a_i x_{(k-i) \bmod n}$$

Then

$$a \circledast x = \begin{bmatrix} a_0 & a_{n-1} & \cdots & a_2 & a_1 \\ a_1 & a_0 & a_{n-1} & \cdots & a_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & a_{n-1} \\ a_{n-1} & & \cdots & a_1 & a_0 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-2} \\ x_{n-1} \end{bmatrix}$$

So, $a \circledast$ is a linear transformation $\mathbb{R}^n \to \mathbb{R}^n$.

Example 4: Let $f : V \to \mathbb{R}$ be a linear function on V.

Then f is a linear transformation $V \to \mathbb{R}$

Example 5: Let $a \in V$. Define $L: \mathbb{R} \to V$ by

$$L(x) = x \cdot a \qquad \forall\, x \in \mathbb{R}$$

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 $D: V_1 \to V_2$ define by

$$Df = f', \qquad \forall\, f \in V_1$$

It is a linear transformation because

$$D(\alpha f + \beta g) = (\alpha f + \beta g)' = \alpha \cdot f' + \beta \cdot g' = \alpha \cdot Df + \beta \cdot Dg$$

Example 7: Fourier Transform $\mathcal{F}: L^2(\mathbb{R}) \to L^2(\mathbb{R})$ defined by

$$(\mathcal{F}f)(\xi) = \int_{-\infty}^{+\infty} f(x)\, e^{-2\pi i \xi x}\, dx, \qquad i = \sqrt{-1}$$

is a linear transformation

Example 8: Given a function $g: \mathbb{R} \to \mathbb{R}$ (smooth enough), define convolution with kernel g by:

$$(g * f)(x) = \int_{-\infty}^{+\infty} g(x-y)\, f(y)\, dy \qquad \forall\, f \in L^2(\mathbb{R})$$

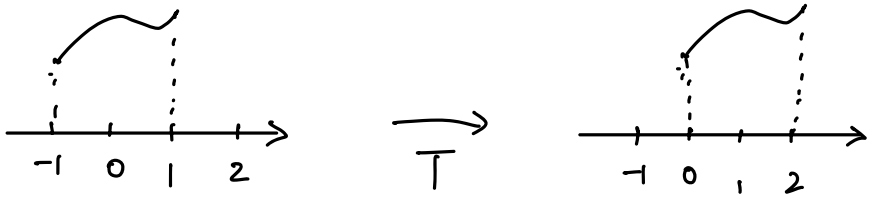
Then

$g*: L^2(\mathbb{R}) \to L^2(\mathbb{R})$ is a linear transformation

Example 9: Consider $C[-1,1]$ and $C[0,2]$

Define translation $T: C[-1,1] \to C[0,2]$ by

$$\forall\, f \in C[-1,1], \qquad (Tf)(t) = f(t-1) \qquad \forall\, t \in [0,2]$$



Example 10: Let $a_1, a_2, \cdots, a_K \in H$, where H is a Hilbert space.

Define $L: H \to \mathbb{R}^k$ 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 is a linear transformation.

- Bounded linear operators

Let $A: V_1 \to V_2$ 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.

- $\forall\ A, B: V_1 \to V_2$ linear, define $A+B$ by

$$(A+B)(x) = A(x) + B(x) \qquad \forall\ x \in V_1$$

We can check $A+B: V_1 \to V_2$ is a linear operator

$$\begin{aligned}(A+B)(\alpha x + \beta y) &= A(\alpha x + \beta y) + B(\alpha x + \beta y) \\ &= \alpha \cdot A(x) + \beta \cdot A(y) + \alpha \cdot B(x) + \beta \cdot B(y) \\ &= \alpha \cdot (A(x) + B(x)) + \beta \cdot (A(y) + B(y)) \\ &= \alpha \cdot (A+B)(x) + \beta \cdot (A+B)(y)\end{aligned}$$

- $\forall\ \alpha \in \mathbb{R}$ and $\forall\ A: V_1 \to V_2$ linear, define $\alpha \cdot A$ by

$$(\alpha \cdot A)(x) = \alpha \cdot A(x) \qquad \forall\ x \in V_1$$

Then $\alpha \cdot A: V_1 \to V_2$ (we can $\alpha \cdot A$ is a linear operator)

We define norm of linear operators $V_1 \to V_2$

Let $A: V_1 \to V_2$ be a linear operator, and V_1, V_2 are normed vector spaces.

(V_1 with norm $\|\cdot\|_{V_1}$, V_2 with norm $\|\cdot\|_{V_2}$)

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.

① $\|A\| \geq 0$ obvious.

$$\|A\| = 0 \iff \sup_{\|x\|_{V_1}=1} \|Ax\|_{V_2} = 0 \iff Ax = 0 \quad \forall\, x : \|x\|_{V_1} = 1$$

$$\iff A \cdot \frac{y}{\|y\|_{V_1}} = 0 \quad \forall\, y \neq 0,\ y \in V_1$$

$$\iff Ay = 0 \quad \forall\, y \neq 0,\ y \in V_1$$

$$\iff Ay = 0 \quad \forall\, y \in V_1 \iff A = 0$$

② $$\|\alpha A\| = \sup_{\|x\|_{V_1}=1} \|(\alpha A)(x)\|_{V_2} = \sup_{\|x\|_{V_1}=1} \left(|\alpha|\, \|A(x)\|_{V_2}\right)$$

$$= |\alpha| \cdot \sup_{\|x\|_{V_1}=1} \|A(x)\|_{V_2} = |\alpha| \cdot \|A\|$$

③ $$\|A+B\| = \sup_{\|x\|_{V_1}=1} \|(A+B)(x)\|_{V_2} = \sup_{\|x\|_{V_1}=1} \|A(x)+B(x)\|_{V_2}$$

$$\leq \sup_{\|x\|_{V_1}=1} \left(\|A(x)\|_{V_2} + \|B(x)\|_{V_2}\right) \leq \sup_{\|x\|_{V_1}=1} \|A(x)\|_{V_2} + \sup_{\|x\|_{V_1}=1} \|B(x)\|_{V_2}$$

$$= \|A\| + \|B\|$$

Additionally, $\|\cdot\|$ satisfies: consistency.

④ $$\boxed{\|Ax\|_{V_2} \leq \|A\| \cdot \|x\|_{V_1} \quad \forall\, x \in V_1}$$

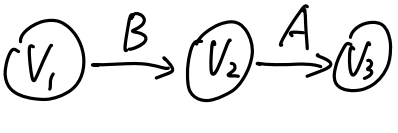
proof. $$\|A\| = \sup_{\substack{y\in V_1 \\ y\neq 0}} \frac{\|Ay\|_{V_2}}{\|y\|_{V_1}} \geq \frac{\|Ax\|_{V_2}}{\|x\|_{V_1}} \quad \text{if } x \neq 0$$

$$\Rightarrow \|Ax\|_{V_2} \leq \|A\| \cdot \|x\|_{V_1}$$

· If $x = 0$, obvious. ⊠

⑤ Let $B : V_1 \to V_2$ linear

$A : V_2 \to V_3$



Then we define AB by $(AB)(x) = A(B(x)) \quad \forall\, x \in V_1$

(i.e., $AB = A \circ B$)

Then $AB: V_1 \to V_3$ is a linear operator. (prove it!)

We have $\boxed{\|AB\| \le \|A\| \cdot \|B\|}$.

proof. $\|AB\| = \sup_{\|x\|_{V_1}=1} \|(AB)(x)\|_{V_3} = \sup_{\|x\|_{V_1}=1} \|A(Bx)\|_{V_3}$

$\le \sup_{\|x\|_{V_1}=1} \|A\| \cdot \|Bx\|_{V_2} = \|A\| \cdot \left(\sup_{\|x\|_{V_1}=1} \|Bx\|_{V_2}\right)$

$= \|A\| \cdot \|B\|$ ∎

Example 1: $A \in \mathbb{R}^{m\times n}$ is a linear transformation $\mathbb{R}^n \to \mathbb{R}^m$

$$\|A\| = \sup_{\|x\|_{\mathbb{R}^n}=1} \|Ax\|_{\mathbb{R}^m}$$

- If $\|\cdot\|_{\mathbb{R}^n} = \|\cdot\|_2$, $\|\cdot\|_{\mathbb{R}^m} = \|\cdot\|_2$, then $\|A\| = \|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$
  $= \left(\text{max eigenvalue of } A^TA\right)^{1/2}$
- $\|\cdot\|_1$ ... $\|A\|_1$
- $\|\cdot\|_\infty$ ... $\|A\|_\infty$

Example 2: DFT $F_n: \mathbb{C}^n \to \mathbb{C}^n$

Define $F_n^H = \overline{F_n^T}$ — conjugate transpose

Then, it has been proved $F_n^H \cdot F_n = F_n \cdot F_n^H = I$

(i.e, F_n is unitary)

(and F_n^H is the inverse discrete Fourier transform)

So, $\|F_n\|_2 = \sup_{\|x\|_2=1} \|F_n x\|_2 = \left(\sup_{\|x\|_2=1} \|F_n x\|_2^2\right)^{1/2}$

$= \left(\sup_{\|x\|_2=1} (x^H F_n^H F_n x)\right)^{1/2} = \left(\sup_{\|x\|_2=1} (x^H x)\right)^{1/2}$

$= \left(\sup_{\|x\|_2=1} \|x\|_2^2\right)^{1/2} = 1$

For $z \in \mathbb{C}^n$, $\|z\|_2^2 = z^H z$

Example 3: Fourier transform $\mathcal{F}: L^2(\mathbb{R}) \to L^2(\mathbb{R})$

$$(\mathcal{F}f)(\xi) = \int_{-\infty}^{+\infty} f(x)\, e^{-2\pi i \xi x}\, dx$$

Plancherel identity: $\|\mathcal{F}f\|_2 = \|f\|_2$

Then $\|\mathcal{F}\| = \sup_{\|f\|_2=1} \|\mathcal{F}f\|_2 = 1$

Example 4: Let $T: C[-1, 1] \to C[0, 2]$ defined by

$$(Tf)(t) = f(t-1) \quad \forall\, t \in [0, 2]$$

...

$$\|T\| = 1$$

Example 5: Let $V_1 = \{f \mid f \text{ and } f' \text{ are continous on } [0,1]\}$

$V_2 = \{f \mid f \text{ is continuous on } [0,1]\}$

The differentiation operator $D: V_1 \to V_2$ 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 $f_k(t) = \sin(2\pi k t)$, k is an integer

Then $(Df_k)(t) = f_k'(t) = 2\pi k \cdot \cos(2\pi k t)$

So, $\|f_k\|_\infty = 1$, $\|Df_k\|_\infty = 2\pi k$

So, $\|D\| \geq \lim_{k\to+\infty} \|Df_k\|_\infty = \lim_{k\to+\infty} 2\pi k = +\infty$

Bounded linear operators $V_1 \to V_2$, where V_1, V_2 are normed vector spaces.

$$\mathcal{L}(V_1, V_2) = \{A \mid A \text{ is a linear operator } V_1 \to V_2 \text{ with } \|A\| < +\infty\}$$

(All linear and bounded operators $V_1 \to V_2$)

— $\mathcal{L}(V_1, V_2)$ is closed under "+" and "$\cdot$", because

- If $A, B \in \mathcal{L}(V_1, V_2)$, then $\|A\| < +\infty$, $\|B\| < +\infty$, and

$$\|A+B\| \leq \|A\| + \|B\| < +\infty$$

So, $A+B \in \mathcal{L}(V_1, V_2)$

and $\|\alpha A\| = |\alpha| \|A\| < +\infty.$

So, $\alpha \cdot A \in \mathcal{L}(V_1, V_2)$

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:

- $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m) = \mathbb{R}^{m \times n}$
- $\mathcal{L}(V_1, V_2) = \{A \mid A$ is a linear operator $V_1 \to V_2\}$ if V_1, V_2 are finite-dimensional norm vector spaces