Linear functions and Differentiable functions

§2.1 Linear functions.

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

f is a linear function if

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

Example 1: The mean of vectors in $\mathbb{R}^n$. i.e.,

$$f(x) = (x_1 + x_2 + \cdots + x_n)/n, \quad \text{for } x \in \mathbb{R}^n$$

is linear.

Example 2: The maximum element of vectors in $\mathbb{R}^n$, i.e.,

$$f(x) = \max\{x_1, x_2, \cdots, x_n\}, \quad \text{for } x \in \mathbb{R}^n$$

is NOT linear.

Example 3: $f: \mathbb{R}^n \to \mathbb{R}$ with $f(x) = a^T x$ is a Linear function.

Example 4: $F: C[a,b] \to \mathbb{R}$ defined by

$F(f) = f(x)$, where $x \in [a,b]$ is a given number

is linear.

Example 5: $F: L^2(a,b) \to \mathbb{R}$ defined by

$F(f) = \int_a^b f(x)dx$ for $f \in L^2(a,b)$ is linear.

Example 6: $f: V \to \mathbb{R}$ (V is an inner product space) defined by

$f(x) = \langle x, z \rangle$, where $z \in V$ is a given vector in V.

Example 7: A norm function on a vector space V is NOT linear.

To see this,

$$\|-x\| = \|x\|,$$

which contradict with

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

for f being linear on V.

Properties of linear functions

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

  (Because $f(\alpha x) = f(\alpha x + 0x) = \alpha f(x) + 0\, f(x) = \alpha f(x)$)

  It implies $f(0) = 0$, because $f(0) = f(0 \cdot x) = 0 \cdot f(x) = 0 \quad \forall\, x \in V$.

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

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

  To see this, we note that

$$\begin{aligned} f(\alpha_1 x_1 + \cdots + \alpha_k x_k) &= \alpha_1 f(x_1) + f(\alpha_2 x_2 + \cdots + \alpha_k x_k) \\ &= \alpha_1 f(x_1) + \alpha_2 f(x_2) + f(\alpha_3 x_3 + \cdots + \alpha_k x_k) \\ &\ \ \vdots \\ &= \alpha_1 f(x_1) + \cdots + \alpha_k f(x_k). \end{aligned}$$

Inner product representation of a linear function on Hilbert spaces

For simplicity, let's consider a linear function on $\mathbb{R}^n$ equipped with the standard inner product $\langle x, y \rangle = x^T y$ and the induced norm $\|x\|_2 = (\langle x, x \rangle)^{1/2}$.

- From the discussion above,

  For any given $a \in \mathbb{R}^n$, the function $f(x) = \langle a, x \rangle$ is linear.

- The reverse is true, i.e.,

  Any linear function $f : \mathbb{R}^n \to \mathbb{R}$ must be in the form of

  $f(x) = \langle a, x \rangle$ for some $a \in \mathbb{R}^n$.

To see this, let $e_1, e_2, \cdots, e_n$, where $e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow$ ith component, be a basis of $\mathbb{R}^n$,

So that any $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ is written as $x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n$.

Therefore, if f is a linear function, then

$$f(x) = f(x_1e_1 + x_2e_2 + \cdots + x_ne_n)$$
$$= x_1 f(e_1) + x_2 f(e_2) + \cdots + x_n f(e_n) \quad \text{— by property of linear functions}$$
$$= \langle a, x \rangle,$$

where $a = \begin{pmatrix} f(e_1) \\ f(e_2) \\ \vdots \\ f(e_n) \end{pmatrix} \in \mathbb{R}^n$.

- Furthermore, the representation of a linear function $f(x) = \langle a, x \rangle$ is unique, which means there is only one vector $a \in \mathbb{R}$ for which $f(x) = \langle a, x \rangle$ holds for all x.

Indeed, suppose that a is not unique, i.e., we have two vectors a, b such that $f(x) = \langle a, x \rangle$ and $f(x) = \langle b, x \rangle$ for all $x \in \mathbb{R}^n$.

Then, let $x = e_i$: $f(e_i) = \langle a, e_i \rangle = a_i$ and $f(e_i) = \langle b, e_i \rangle = b_i$

So, $a_i = b_i$, $i = 1, 2, \cdots, n$.

Therefore $a = b$.

- Altogeter, we see that

> a linear function $f: \mathbb{R}^n \to \mathbb{R}$ if and only if $f(x) = \langle a, x \rangle$ for some unique $a \in \mathbb{R}^n$

The above holds true for any linear function on Hilbert spaces, widely known as Riesz representation theorem.

Theorem (Riesz representation thereom):

> Let H be a Hilbert space, and f be a function: $H \to \mathbb{R}$. Then f is linear and bounded if and only if $f(x) = \langle a, x \rangle$ for some unique $a \in H$.

Example 1: We know mean(x) is linear on $\mathbb{R}^n$. Since $\mathbb{R}^n$ is a Hilbert space, we can find a unique $a \in \mathbb{R}^n$ s.t.

$$\text{mean}(x) = \langle a, x \rangle.$$

Indeed,

$$\text{mean}(x) = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}x_1 + \frac{1}{n}x_2 + \cdots + \frac{1}{n}x_n = \langle a, x \rangle,$$

where $a = \left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)^T$.

Example 2: Let H be a Hilbert space, and $\|\cdot\|$ is the norm.

It is known that the norm function is NOT linear.

Therefore,

there doesn't exist $a \in H$ such that $\|x\| = \langle a, x\rangle \quad \forall x \in H$.

Hyperplanes

- Again, we consider $\mathbb{R}^n$ as a Hilbert space, where any linear function is written as $\langle a, x\rangle$ for some $a \in \mathbb{R}^n$.

Consider the set

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

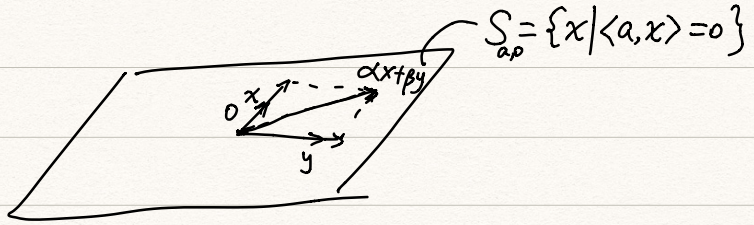
Then, $\forall\ x, y \in S_{a,0}$ and $\alpha, \beta \in \mathbb{R}$,

$$\langle a, \alpha x + \beta y\rangle = \alpha\langle a, x\rangle + \beta\langle a, y\rangle = 0 \Rightarrow \alpha x + \beta y \in S.$$

Therefore, $S_{a,0}$ is a plane.

Since the co-dimension of S_0 is 1 (because it is define by one linear equation)

$S_{a,0}$ is called a hyperplane.



Now let's consider

$$S_{a,b} = \{x \mid \langle a, x\rangle = b\} \quad \text{for } b \in \mathbb{R} \text{ is given.}$$

Let $x_0 \in S_{a,b}$, i.e., $\langle a, x_0\rangle = b$, be fixed.

Then $S_{a,b} = S_{a,0} + x_0$ because:

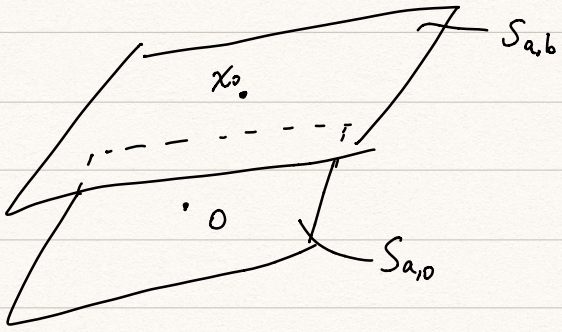
(1) $\forall\ x \in S_{a,b} \quad \langle a, x - x_0\rangle = \langle a, x\rangle - \langle a, x_0\rangle = 0, \Rightarrow x - x_0 \in S_{a,0}.$

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

② $\forall x \in S_{a,0}$ $\quad \langle a, x+x_0 \rangle = \langle a, x \rangle + \langle a, x_0 \rangle = b \implies x+x_0 \in S_{a,b}$

In other words,

$S_{a,b}$ is a shift of a hyperplane, still called a hyperplane.



- This concept can be generalized to any inner product space V.

> The set $\{x \in V \mid \langle a, x \rangle = b\}$, where $a \in V$ and $b \in \mathbb{R}$ are given, is called a Hyperplane in V.

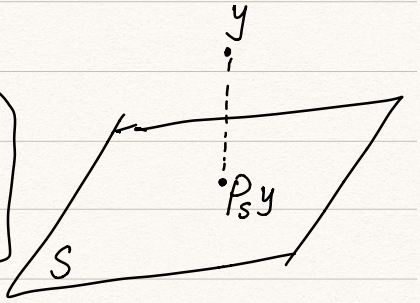
Projection onto hyperplanes

- Consider a Hilbert space V and a hyperplane S

$$S = \{x \in V \mid \langle a, x \rangle = b\}$$

Let $y \in V$ be a given vector.

> The vector on S that is the closest to y is called the projection of y on S, denoted by $P_S y$, i.e., $P_S y = \arg\min_{x \in S} \|x - y\|$.



- Let us find an explicit expression of $P_S y$ in terms of a, b, and y.

> Theorem: z is a solution of $\min_{x \in S} \|x - y\|$ if and only if $z \in S$ and $\langle z - y, z - x \rangle = 0 \quad \forall x \in S$.

Proof: ① We first prove that: If $z \in S$ is a solution of $\min_{x \in S} \|x - y\|$, then $\langle z - y, z - x \rangle = 0 \quad \forall x \in S$.

Since z is a solution, $z \in S$, i.e., $\langle a, z\rangle = b$.

$\forall x \in S$ and $t \in \mathbb{R}$, it is easy to see that

$\langle a, (1+t)z - tx\rangle = (1+t)\langle a, z\rangle - t\langle a, x\rangle = b$.

Therefore, $(1+t)z - tx \in S$.

Since z is closest to y on S, we have

$$\|z-y\|^2 \le \|(1+t)z - tx - y\|^2$$
$$= \|(z-y) + t(z-x)\|^2$$
$$= \|z-y\|^2 + t^2\|z-x\|^2 + 2t\langle z-y, z-x\rangle.$$

i.e., $t\langle z-y, z-x\rangle \ge -\frac{t^2}{2}\|z-x\|^2$.

- If we choose $t > 0$,

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

Letting $t \to 0_+$ gives $\langle z-y, z-x\rangle \ge 0$.

- If we choose $t < 0$,

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

Letting $t \to 0_-$ gives $\langle z-y, z-x\rangle \le 0$

Altogether, z satisfies $\langle z-y, z-x\rangle = 0 \quad \forall x \in S$.

② We then show that $z \in S$ satisfies $\langle z-y, z-x\rangle = 0$, then z is a solution of $\min_{x\in S}\|x-y\|$, by direct calculation.

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

$$\|x-y\|^2 = \|(z-x)-(z-y)\|^2$$
$$= \|z-x\|^2 + \|z-y\|^2 - 2\langle z-x, z-y\rangle$$
$$= \|z-y\|^2 + \|z-x\|^2 \ge \|z-y\|^2 \quad \forall x \in S.$$

This, together with $z \in S$, implies z minimizes $\|x-y\|^2$ in $x \in S$.

⊗

Theorem: The solution of $\min_{x \in S} \|x-y\|^2$ exists and unique, which is given by $y - \left(\frac{\langle a, y\rangle - b}{\|a\|^2}\right) a$

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

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

$= \langle a, y\rangle - (\langle a, y\rangle - b) = b$, so $z \in S$.

② $\forall\, x \in S$,

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

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

(because $\langle a, z\rangle = \langle a, x\rangle = b$).

By the previous theorem, z is a solution of $\min_{x \in S} \|x-y\|$.

It remains to show the uniqueness.

Suppose we have two solutions z_1 and z_2. Then,

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

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

Taking difference leads to $\langle z_1 - z_2, z_1 - z_2\rangle = 0$

$\Rightarrow \|z_1 - z_2\|^2 = 0 \Rightarrow z_1 = z_2$ ⊠.

- In summary, the projection $P_S y$ of $y \in V$ onto the hyperplane

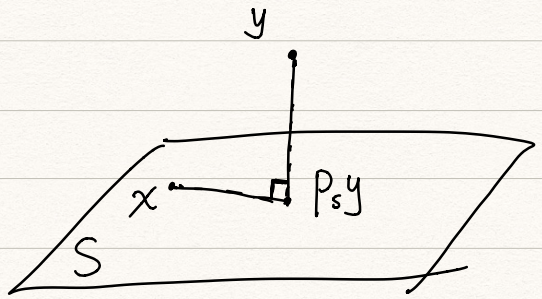
$$S = \{x \in V \mid \langle a, x\rangle = b\}$$

exists and is unique. Furthermore,

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

and it satisfies

$$\langle P_S y - y, P_S y - x\rangle = 0.$$



Affine functions

A linear function plus a constant is called an affine function.
That is, a function $f: V \to \mathbb{R}$ is affine if

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

where $g: V \to \mathbb{R}$ is linear and $b \in \mathbb{R}$ is a constant.

Properties:

- If $f: V \to \mathbb{R}$ is affine, then
  $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$ $\forall\, x, y \in V$ and $\alpha, \beta \in \mathbb{R}$ s.t. $\alpha + \beta = 1$.

  To see this, (g linear; $\alpha+\beta=1$)

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

- If $f: V \to \mathbb{R}$, where V is a Hilbert space, then
  f must be in the form of

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

§2.2 Case Studies: Regression and Classification.

§2.2.1 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 an input feature vector
$y_i \in \mathbb{R}$ is the corresponding response to x_i.
, $i = 1, 2, \cdots, N$.

Given a new input feature vector $x \in \mathbb{R}^n$, how to predict the corresponding response $y \in \mathbb{R}$?

For example, $x_i \in \mathbb{R}^n$ represents n attributes of a house, and $y_i \in \mathbb{R}$ is the selling price.

We want to predict the selling price of a house with feature $x \in \mathbb{R}^n$.

- Mathematically, we need to find a function $f: \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x_i) \simeq y_i \quad , \quad i = 1, 2, \cdots, N$$

This is called regression. In this context,

x_i, are called regressor / independent varibles.

y_i are called dependent variables / outcome / label.

- The class of all functions $\mathbb{R}^n \to \mathbb{R}$ is too large, and the given data set $\{(x_i, y_i)\}_{i=1}^N$ is not enough to determine a function uniquely.

So, we need to find a function class Φ where we search f.

Intuitively, larger N, larger function class Φ.

- Linear model:

We search f in the class of all affine functions,

i.e., $f(x) = \langle a, x \rangle + b$ for some $a \in \mathbb{R}^n$, $b \in \mathbb{R}$.

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

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

by minimizing the error of the linear equations.

While there are many possible definitions of error, it is popular to consider the square error as follows:

$$\left(\langle a, x_i \rangle + b - y_i\right)^2, \quad i = 1, 2, \cdots, N.$$

Therefore, we find $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ by solving

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

This problem is called the Least squares (LS) problem.

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

and $y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N$.

Then LS problem becomes

$$\boxed{\min_{\beta \in \mathbb{R}^{n+1}} \|X\beta - y\|_2^2.}$$

Since we have N linear equations to fit and $n+1$ unknowns, $N \geq n+1$ is required to have a unique solution.

- Regularization.

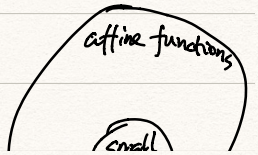
  In many applications, we don't have enough data (e.g., $N < n+1$), so that the class of affine functions is too large to search f.

  We search f is a sub class of all affine functions by regularizing f.

  Since $f(x) = \langle a, x \rangle + b$, regularizing f is equivalent to regularizing $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$.

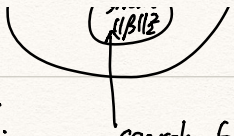
  — Ridge regression: We solve

$$\boxed{\min_{\beta \in \mathbb{R}^{n+1}} \frac{1}{2}\|X\beta - y\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2.}$$



Here $\|\beta\|_2^2$ is the regularization term

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



search f here in ridge regression

In other words, we find β such that

the error of data fitting and the 2-norm of β

are minimized simultaneously.

Therefore, ridge regression gives a β such that

$X\beta \simeq y$ and $\|\beta\|_2$ is small.

— LASSO regression: we solve

$$\min_{\beta \in \mathbb{R}^{n+1}} \frac{1}{2}\|X\beta - y\|_2^2 + \frac{\lambda}{2}\|\beta\|_1$$

Here: $\|\beta\|_1$ is regularization term

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

So, we find β such that

the error of data fitting and the 1-norm of β

are minimized simultaneously.

Small $\|\beta\|_1$ tends to give a sparse β

(i.e., many entries of β are zeros)

Consequently, LASSO gives a solution β that

$X\beta \simeq y$ and $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$ is sparse

Thus, given $x \in \mathbb{R}^n$, the prediction is given by

$$\langle x, a \rangle + b = \sum_{i=1}^{n} a_i x_i + b = \sum_{i \in S} a_i x_i + b$$

Let $S = \{i \mid a_i \neq 0\}$

Since S is a small set, the prediction depends only on

a small portion entries of x.

This is preferred because the prediction is interpretable.

- Kernel regression

Feature map $\phi: \mathbb{R}^n \to H$

Then do regression in H.

However, since H is very large, the set of all linear functions is also two large. We need regularization.

We solve

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

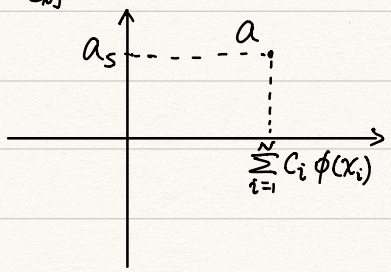
Representer Theorem:

The solution must be in the form of $a = \sum_{i=1}^{N} c_i \phi(x_i)$ for some $c = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N$.

Proof. For any $a \in H$, we claim 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 $\langle a_S, \phi(x_i) \rangle = 0$ for $i = 1, 2, \cdots, N$.



Indeed, Let $c = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N$ be satisfying

$$\langle a - \sum_{i=1}^{N} c_i \phi(x_i), \phi(x_j) \rangle = 0, \quad j = 1, 2, \cdots, N$$

(This is N linear equation with N unknows, and it can be checked there exists at least one solution)

Denote $a_S = a - \sum_{i=1}^{N} c_i \phi(x_i)$. It can be checked

$$\langle a_S, \phi(x_j) \rangle = 0, \quad j = 1, 2, \cdots, N$$

due to the construction of c and a_S.

Therefore,

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

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

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

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

$$= \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} c_j\, K(x_i,x_j) - y_i\right)^2 + \lambda\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j\, K(x_i,x_j) + \lambda\|a_s\|_H^2$$

$$= \frac{1}{2}\|Kc - y\|_2^2 + \lambda c^T K c + \lambda\|a_s\|_H^2$$

where $K = \begin{bmatrix} K(x_1,x_1) & K(x_1,x_2) & \cdots & K(x_1,x_N) \\ \vdots & & & \\ K(x_N,x_1) & K(x_N,x_2) & \cdots & K(x_N,x_N) \end{bmatrix} \in \mathbb{R}^{N\times N}$ $\quad c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N$.

Because $\|a_s\|_H^2 \geq 0$, the objective function is minimized when $\|a_s\|_H^2 = 0$, i.e., $a_s = 0$.

Thus, $a \in H$ is a solution only if $a = \sum_{i=1}^{N} c_i\,\phi(x_i)$ ⊠

From the above proof, we see that

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

$$\Updownarrow$$

$$\boxed{\min_{c\in\mathbb{R}^N}\ \frac{1}{2}\|Kc - y\|_2^2 + \lambda c^T K c}$$

Let the solution be $c \in \mathbb{R}^N$.

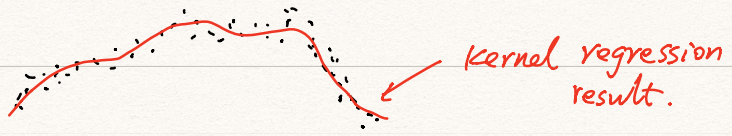
Then the predicted output y for input $x \in \mathbb{R}^n$ is

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

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

All the computation involves only the kernel function $K(\cdot,\cdot)$,

No explicit feature map $\phi(\cdot)$ is needed.



§ 2.2.2 Classification

- Classification: Giving training data

$$(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N), \quad x_i \in \mathbb{R}^n, \quad y_i \in \{-1, +1\}, \quad i = 1, \cdots, N,$$

find a classifier (a function) f such that

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

We use hyperplanes to separate the points

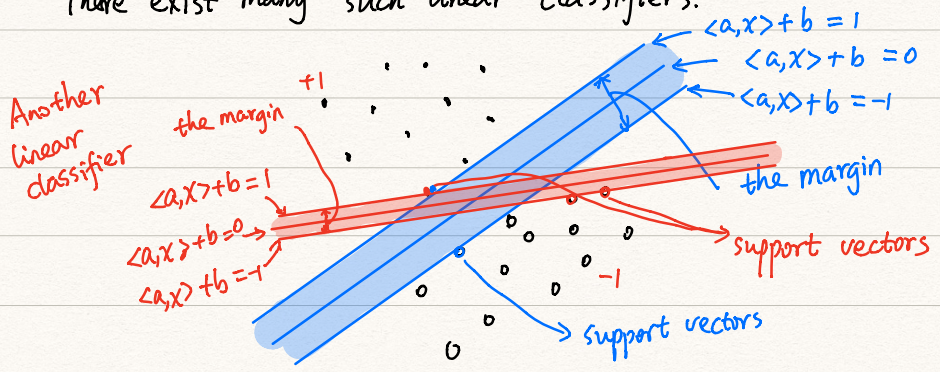
$$f(x) = \langle a, x \rangle + b, \quad \text{where } a \in \mathbb{R}^n, \quad b \in \mathbb{R}.$$

The weights $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are noralized such that

$$\langle a, x_i \rangle + b_i \begin{cases} \geq 1 & \text{if } y_i = +1 \\ \leq -1 & \text{if } y_i = -1 \end{cases}$$

- Support Vector Machine (SVM)

There exist many such linear classifiers.

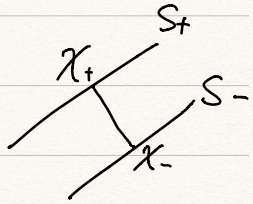


Which one is the better?

From the figure, the blue is better, because it has a larger margin, and hence a larger buffer zone of mis-classification.

Therefore, we want to maximize the margin among all candidates.

Let us calculate the margin in terms of a and b.



The margin is the distance between the two hyperplanes

$$S_+ = \{x \mid \langle a, x \rangle + b = 1\} \quad \text{and} \quad S_- = \{x \mid \langle a, x \rangle + b = -1\}$$

Let $x_+ \in S_+$ and $x_- \in S_-$ such that

$$\|x_+ - x_-\|_2 = \text{dist}(S_+, S_-).$$

Since x_+ is a projection of x_- onto $S_+ = \{x \mid \langle a, x\rangle = 1-b\}$

$$\begin{aligned} x_+ &= x_- - \frac{\langle a, x_-\rangle + b - 1}{\|a\|_2^2} a \\ &= x_- - \frac{-1-b+b-1}{\|a\|^2} a \quad (\text{since } x_- \in S_-) \\ &= x_- + \frac{2}{\|a\|_2^2} a \end{aligned}$$

Thus, $\|x_+ - x_-\|_2 = \left\|\frac{2}{\|a\|_2^2} a\right\|_2 = \frac{2}{\|a\|_2}$

i.e., the margin $= \frac{2}{\|a\|_2}$

Support Vector Machine (SVM)

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

which is equivalent to

$$\begin{aligned} \min_{a\in\mathbb{R}^n,\, b\in\mathbb{R}} \quad & \frac{1}{2}\|a\|_2^2 \\ \text{s.t} \quad & y_i(\langle a, x_i\rangle + b) \geq 1 \end{aligned} \qquad \text{(SVM-1)}$$

- The above SVM is NOT robust to noise.

For example shown on the right, even we have only two noisy points, there is no solution to (SVM-1).

We consider a "soft" version of (SVM-1) as follows.

We minimize the error to the separation

$$\sum_{i=1}^{N} h\left(\left(y_i \langle a, x_i\rangle + b\right) - 1\right),$$

where the function $h(t) = \begin{cases} 0 & \text{if } t \geq 0 \\ |t| & \text{if } t < 0 \end{cases}$

In other words, if $y_i\langle a, x_i\rangle + b \geq 1$, the error is 0.

if $y_i\langle a, x_i\rangle + b < 1$, the error is the absdute value of $y_i\langle a, x_i\rangle + b_i - 1$.

Also, the margin $\frac{1}{2}\|a\|_2^2$ can be viewed as a regularization.

Altogether, we solve

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

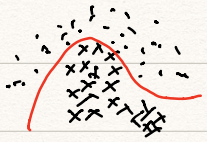
We call this SVM with a soft margin.

— h can be approximated by some smooth function. If h is chosen the so-called logistic function, we call it logistic regression.

- Kernel SVM

The linear SVMs don't work for curved data.

The kernel method can be used.



Let $\phi : \mathbb{R}^n \to H$ be a feature map.

So, we use linear functions on H to classify the points

By Riesz representation theorem, any linear function in the form of $\langle a, x\rangle$ for some $a \in H$.

Therefore, (SVM-2) becomes

$$\min_{a\in H} h\left(y_i\langle a, \phi(x_i)\rangle - 1\right) + \frac{1}{2}\|a\|_H^2 \qquad (K\text{-}SVM)$$

Aagain, one can prove the following representer theorem.

Theorem: Any solution of (K-SVM) is in the form of

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

proof. Write $a = \sum_{i=1}^{N} c_i \phi(x_i) + a_s$ for some $a_s \in H$ and $\langle a_s, \phi(x_i) \rangle = 0 \ \forall i$.

The rest is the same as the linear regression case. ⊠.

Thus, (K-SVM) becomes

$$\min_{c \in \mathbb{R}^N} h\left(y_i \left(\sum_{j=1}^{N} K(x_i, x_j) c_j \right) - 1 \right) + \frac{1}{2} c^T K c,$$

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

The prediction of the input x is given by

$$\operatorname{sgn}\left(\sum_{j=1}^{N} K(x, x_j) c_j \right)$$

Again, only $K(\cdot, \cdot)$ is needed in the kernel SVM, and no explicit feature map $\phi(\cdot)$ is required.

§ 2.3 First-order derivative

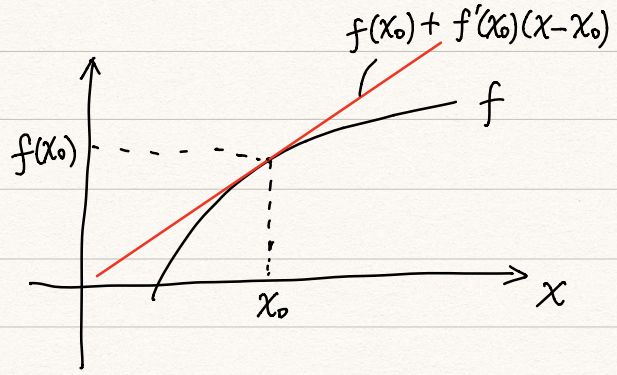
Recall that for a function $f: \mathbb{R} \to \mathbb{R}$, the derivative at x_0 is

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

which is the same as

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

Notice that $f(x_0) + f'(x_0)(x - x_0)$ is an affine function in $\mathbb{R}$ that passes through $(x_0, f(x_0))$.



In other words, in differentiation at x_0,

(1) f is approximated by an affine function passes thru $(x_0, f(x_0))$.

(2) the error of the approximation is $o(|x - x_0|)$.

(little o, i.e., $\lim_{x \to x_0} \frac{|\text{error}|}{|x - x_0|} \to 0$)

This can be used to define differentiation of functions on Hilbert spaces.
(We use Hilbert space for simplicity. It can be easily adapted to Banach spaces.)

Let $f: V \to \mathbb{R}$ with V a Hilbert space.

Consider the differentiation of f at $x^{(0)} \in V$.

(1) By Riesz representation theorem, any affine function is in the form

of $\langle v, x\rangle + a$ for some $v \in V$ and $a \in \mathbb{R}$. Since it passes thru $(x^{(0)}, f(x^{(0)}))$, $\langle v, x^{(0)}\rangle + a = f(x^{(0)})$ Therefore, the affine function is in the form of

$$\begin{aligned}\langle v, x\rangle + a &= \langle v, x - x^{(0)}\rangle + (\langle v, x^{(0)}\rangle + a) \\ &= f(x^{(0)}) + \langle v, x - x^{(0)}\rangle.\end{aligned}$$

(2). The approxition error is

$$\text{error} = \left| f(x) - f(x^{(0)}) - \langle v, x - x^{(0)}\rangle \right|,$$

The error should be in the order of $o(\|x - x^{(0)}\|)$, i.e.,

$$\frac{\text{error}}{\|x - x^{(0)}\|} \to 0 \quad \text{as} \quad x \to x^{(0)}.$$

Definition: Let V be a Hilbert space. Let $f: V \to \mathbb{R}$. Then f is said Frechet differentiable if there exists a $v \in V$ such that

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

If f is differentiable at $x^{(0)}$, v is called the gradient of f at $x^{(0)}$, denoted by $\nabla f(x^{(0)})$.

Example 1: $f(x) = \|x\|^2$, where $\|x\|$ is the norm on V.

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

$$\begin{aligned}\|x\|^2 &= \|(x - x^{(0)}) + x^{(0)}\|^2 = \langle (x - x^{(0)}) + x^{(0)}, (x - x^{(0)}) + x^{(0)}\rangle \\ &= \|x - x^{(0)}\|^2 + \|x^{(0)}\|^2 + 2\langle x^{(0)}, x - x^{(0)}\rangle\end{aligned}$$

Therefore, $\|x\|^2 - \underbrace{\left(\|x^{(0)}\|^2 + 2\langle x^{(0)}, x - x^{(0)}\rangle\right)}_{\text{affine approximation}} = \|x - x^{(0)}\|^2$

So $$\lim_{x \to x^{(0)}} \frac{\left|\|x\|^2 - \|x^{(0)}\|^2 - \langle 2x^{(0)}, x - x^{(0)}\rangle\right|}{\|x - x^{(0)}\|} = \lim_{x \to x_0} \frac{\|x - x^{(0)}\|^k}{\|x - x^{(0)}\|} = 0.$$

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

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

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

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

Therefore, $\lim_{x \to x^{(0)}} \frac{|\langle a, x \rangle - \langle a, x^{(0)} \rangle - \langle a, x - x^{(0)} \rangle|}{\|x - x^{(0)}\|} = \lim_{x_0 \to x} \frac{0}{\|x - x^{(0)}\|} = 0.$

Thus, $\nabla f(x^{(0)}) = a$.

Example 3: $f(x) = \|x - a\|^2$ for some $a \in V$.

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

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

So, $\lim_{x \to x^{(0)}} \frac{|f(x) - f(x^{(0)}) - \langle 2(x^{(0)} - a), x - x^{(0)} \rangle|}{\|x - x^{(0)}\|} = \lim_{x \to x^{(0)}} \|x - x^{(0)}\| = 0.$

Therefore, $\nabla f(x^{(0)}) = 2(x^{(0)} - a)$

Properties:

① Frechet differentiation is linear (in f), i.e.,

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

② Chain rule: Let $f: V \to \mathbb{R}$ and $g: \mathbb{R} \to \mathbb{R}$. Then $g \circ f: V \to \mathbb{R}$ and

$$\nabla(g \circ f)(x) = g'(f(x)) \cdot \nabla f(x)$$

if f and g are differentiable at x and $f(x)$ respectively.

Example 4: $f(x) = \|x\| \quad \forall\, x \in V$.

This is a composition of $f_1(x) = \|x\|^2$ from $V \to \mathbb{R}$

and $f_2(t) = \sqrt{t}$ from $\mathbb{R} \to \mathbb{R}$.

When $\|x\| \neq 0$, both f_1 and f_2 are differentiable.

Also, $\nabla f_1(x) = 2x$, $f_2'(t) = \frac{1}{2\sqrt{t}}$ if $t \neq 0$.

So, $\nabla f(x) = \nabla (f_2 \circ f_1)(x) = f_2'(f_1(x)) \cdot \nabla f_1(x)$

$$= \frac{1}{2\sqrt{\|x\|^2}} \cdot 2x = \frac{x}{\|x\|}.$$

When $\|x\| = 0$, (i.e., $x = 0$), $f_2(t)$ is NOT differentiable at $f_1(x) = 0$.

It can be shown that $f(x) = \|x\|$ is NOT differentiable at $x = 0$.

③ For functions on $\mathbb{R}^n$: $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}, \quad \text{where} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

i.e., Frechet differentiation is the same as the standard differentiation in multivariate calculus.

Taylor's expansion

From the definition, we see that

$$f(x) \approx f(x^{(0)}) + \langle \nabla f(x^{(0)}), x - x^{(0)} \rangle$$

This is a generalization of Taylor's expansion.

In paticular, if $f : \mathbb{R}^n \to \mathbb{R}$

$$f(x) \approx f(x^{(0)}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^{(0)}) \, (x_i - x_i^{(0)}).$$

§ 2.4. Linear operators and Higher-order derivatives

- **Linear operator / Linear transformations**

Now we consider functions between vector spaces.

Let V_1, V_2 be two vector spaces

A map $L: V_1 \to V_2$ is a linear operator (a.k.a. linear transformation) if: $L(\alpha x + \beta y) = \alpha L(x) + \beta L(y)$ $\forall\ \alpha, \beta \in \mathbb{R}$, $x, y \in V_1$.

Example 1: Let $A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m\times n}$

Then the transformation: $\mathbb{R}^n \to \mathbb{R}^m$ defined by

$x \to Ax$, where Ax is the matrix-vector product.

is a linear transformation, because

$$A(\alpha x + \beta y) = \alpha Ax + \beta Ay.$$

Reversly, any linear transformation $L: \mathbb{R}^n \to \mathbb{R}^m$ must be in the form of $L(x) = Ax$ for some matrix $A \in \mathbb{R}^{m\times n}$.

Example 2: Let $f: V \to \mathbb{R}$ be a linear function on V.

Then f is a linear transformation from V to $\mathbb{R}$, as $\mathbb{R}$ is a vector space.

Example 3: Let $a \in \mathbb{R}$. Then define $L: \mathbb{R} \to V$ by

$L(x) = ax$. Then L is a linear operator.

Example 4: Let $V_1 = \{f \mid f \text{ and } f' \text{ is continous on } [a,b]\}$

and $V_2 = \{f \mid f \text{ is continuous on } [a,b]\}$.

Then the differentiation operator $D: V_1 \to V_2$ defined by

$Df = f'$ is an linear operator.

- Operator norm

Consider the set of all linear operators $V_1 \to V_2$, where V_1, V_2 are two normed space.

- $\forall$ A, B linear operators $V_1 \to V_2$, define $A+B$ by
$$(A+B)(x) = A(x) + B(x) \qquad \forall\, x \in V_1$$
- $\forall\, \alpha \in \mathbb{R}$ and A linear operator $V_1 \to V_2$, define αA by
$$(\alpha A)(x) = \alpha\, A(x) \qquad \forall\, x \in V_1.$$

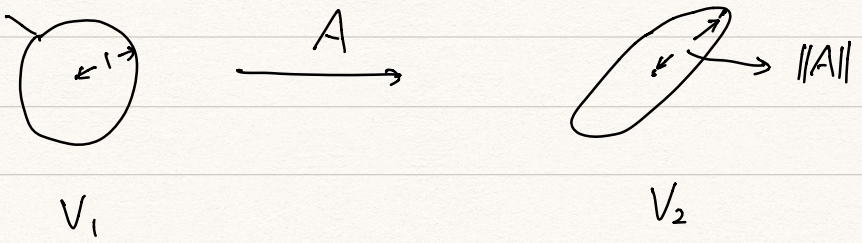
Then, the set of all linear operator $V_1 \to V_2$ is a vector space.

So, we can define a norm on it. For any liear operator $A: V_1 \to V_2$

$$\|A\| = \sup_{\|x\|_{V_1}=1} \|Ax\|_{V_2}$$

($\|\cdot\|_{V_1}$, $\|\cdot\|_{V_2}$ — the norm on V_1 and V_2 respectively)

the unit ball in $\|\cdot\|_{V_1}$



This is indeed a norm.

① $\|A\| \geq 0$ is obvious.

$$\|A\| = 0 \iff \sup_{\|x\|_{V_1}=1} \|Ax\|_{V_2} = 0 \iff Ax = 0 \ \ \forall\ x : \|x\|_{V_1} = 1$$
$$\iff A\frac{y}{\|y\|_{V_1}} = 0 \qquad \forall\, y \in V_1,\ y \neq 0$$
$$\iff Ay = 0 \ \ \forall\, y \in V_1 \iff A = 0.$$

② $$\|\alpha A\| = \sup_{\|x\|_{V_1}=1} \|\alpha A x\|_{V_2} = \sup_{\|x\|_{V_1}=1} \alpha \|Ax\|_{V_2} = \alpha \|A\|$$

③ $$\|A+B\| = \sup_{\|x\|_{V_1}=1} \|(A+B)x\|_{V_2} = \sup_{\|x\|_{V_1}=1} \|Ax + Bx\|_{V_2}$$

$$\le \sup_{\|x\|_{V_1}=1} \left(\|Ax\|_{V_2} + \|Bx\|_{V_2}\right) \le \sup_{\|x\|_{V_1}=1} \|Ax\|_{V_2} + \sup_{\|x\|_{V_1}=1} \|Bx\|_{V_2}$$
$$= \|A\| + \|B\|.$$

In addition, it sasfies the following consitency properties:

④ $\|Ax\|_{V_2} \le \|A\|\,\|x\|_{V_1}$ $\quad \forall\, x \in V_1$,

Proof. If $x=0$, then $\|Ax\|_{V_2} = 0 \le 0 = \|A\|\|x\|_{V_1}$

If $x \neq 0$, then, $\left\| \frac{x}{\|x\|_{V_1}} \right\|_{V_1} = 1$ and

$$\|A\| = \sup_{\|y\|_{V_1}=1} \|Ay\|_{V_2} \ge \left\| A \frac{x}{\|x\|_{V_1}} \right\|_{V_2} = \frac{\|Ax\|_{V_2}}{\|x\|_{V_1}}$$
$$\Rightarrow \quad \|Ax\|_{V_2} \le \|A\|\,\|x\|_{V_1}. \qquad \otimes$$

⑤ $\|AB\| \le \|A\|\,\|B\|$ $\quad \forall\ B \in \mathcal{L}(V_1, V_2),\ A \in \mathcal{L}(V_2, V_3)$

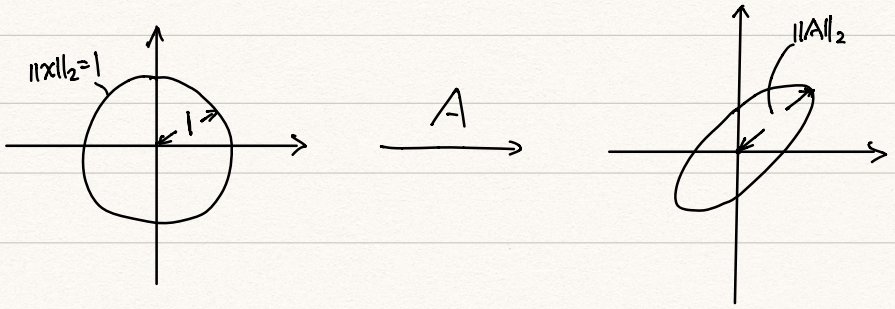
Proof. $\|AB\| = \sup_{\|x\|_{V_1}=1} \|ABx\|_{V_3} \le \sup_{\|x\|_{V_1}=1} \|A\|\,\|Bx\|_{V_2}$

$$= \|A\| \sup_{\|x\|_{V_1}=1} \|Bx\|_{V_2} = \|A\| \cdot \|B\| \qquad \otimes$$

Example 1: As we have seen, $\mathbb{R}^{m\times n}$ is the set of all linear operators from $\mathbb{R}^n$ to $\mathbb{R}^m$. We endow $\mathbb{R}^n$ and $\mathbb{R}^m$ with 2-norm.

Then, for any $A \in \mathbb{R}^{m\times n}$, we call the corresponding operator norm the 2-norm, denoted by $\|A\|_2$. So,

$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2 = \left(\sup_{\|x\|_2=1} \|Ax\|_2^2\right)^{1/2}$$
$$= \left(\sup_{x^Tx=1} x^TA^TAx\right)^{1/2} = \left(\text{max eigenvalue of } A^TA\right)^{1/2}$$

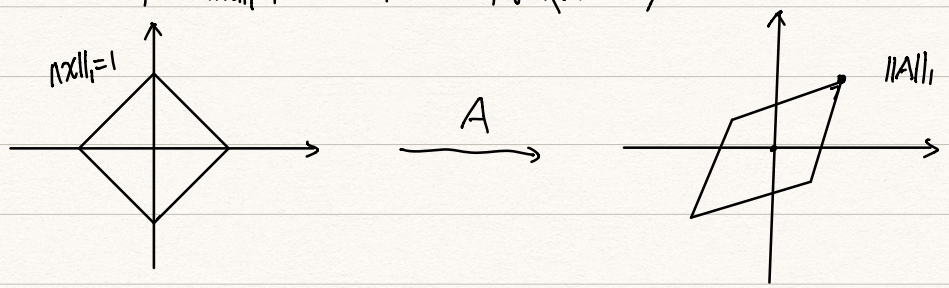


Example 2: Consider $\mathbb{R}^{m\times n}$ as linear operators from $\mathbb{R}^n$ to $\mathbb{R}^m$

But $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with 1-norms.

we call the resulting operator norm 1-norm, denoted by $\|A\|_1$

$$\|A\|_1 = \sup_{\|x\|_1=1} \|Ax\|_1 = \max_{1\le j\le n}\left(\sum_{i=1}^{m}|a_{ij}|\right)$$

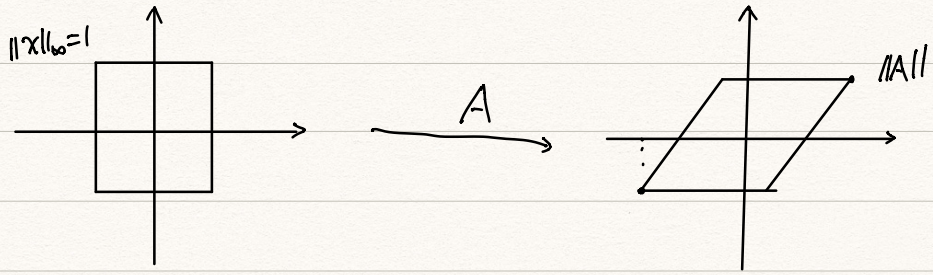


Example 3: $\mathbb{R}^{m\times n}$ as linear operators from $\mathbb{R}^n \to \mathbb{R}^m$.

But $\mathbb{R}^n$ and $\mathbb{R}^m$ are with ∞-norms.

The resulting operator norm is call ∞-norm, denoted by $\|\cdot\|_\infty$

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{1\le i\le m}\left(\sum_{j=1}^{n}|a_{ij}|\right)$$



Now, we consider the set

$$\mathcal{L}(V_1,V_2) = \{A : A \text{ is linear } V_1\to V_2\} \cap \{A : \|A\|<+\infty\}$$

Then $\mathcal{L}(V_1,V_2)$ is a normed space if V_1, V_2 are normed vector space.

and

$\mathcal{L}(V_1,V_2)$ is a Banach space if V_1, V_2 are Banach spaces.

- Adjoint operator.

Let $(V_1, \langle\cdot,\cdot\rangle_{V_1})$ and $(V_2, \langle\cdot,\cdot\rangle_{V_2})$ are two inner product spaces.

Let $A \in \mathcal{L}(V_1,V_2)$. Then, its adjoint operator $A^* \in \mathcal{L}(V_2,V_1)$ is defined by

$$\langle Ax, y\rangle_{V_2} = \langle x, A^*y\rangle_{V_1} \quad \forall\ x\in V_1 \text{ and } y\in V_2.$$

Example 1: Consider $A \in \mathbb{R}^{m\times n} = \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$

Then $\langle Ax, y\rangle = y^TAx = (A^Ty)^Tx = \langle x, A^Ty\rangle \quad \forall x\in\mathbb{R}^n, y\in\mathbb{R}^m$.

Therefore, the adjoint of A is its transpose A^T.

Example 2: Let V be an inner product space.

Consider $\mathcal{L}(V, \mathbb{R})$, i.e, bounded linear functions on V.

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

Therefore, $\forall x\in V$, $y\in\mathbb{R}$,

$$\langle f(x), y\rangle_{\mathbb{R}} = yf(x) = y\langle a, x\rangle_V = y\langle x, a\rangle_V = \langle x, ya\rangle_V$$

Therefore, $f^*(y) = ya$, $\forall\ y\in\mathbb{R}$,

Obviously, $f^*\in\mathcal{L}(\mathbb{R}, V)$.

- Differentiation of transformations between normed spaces

Let V_1, V_2 be two normed spaces with inner products $\|\cdot\|_{V_1}$ and $\|\cdot\|_{V_2}$.

Let $F: V_1\to V_2$ be a map (not necessarily linear)

Then, at any point $x^{(0)}\in V_1$, then linear approximation passing thru $(x^{(0)}, F(x^{(0)}))$ is

$$F(x) \approx F(x^{(0)}) + L(x-x^{(0)}), \quad \text{where } L\in\mathcal{L}(V_1, V_2).$$

If this approximation is $o(\|x-x^{(0)}\|_{V_1})$, then L is called the differentiation.

Definition: $F: V_1\to V_2$ is differentiable at $x^{(0)}\in V_1$, if there exists $L\in\mathcal{L}(V_1, V_2)$ such that

$$\lim_{x\to x^{(0)}} \frac{\|F(x) - F(x^{(0)}) - L(x-x^{(0)})\|_{V_2}}{\|x-x^{(0)}\|_{V_1}} = 0.$$

L is called a derivative of F at $x^{(0)}$, denoted by

$$DF(x^{(0)}) = L.$$

Example 1: If $f: V \to \mathbb{R}$ with V a Hilbert space, then

$$Df(x^{(0)})\, y = \langle \nabla f(x^{(0)}), y \rangle \quad \forall\ y \in V.$$

Example 2: Let $A \in \mathcal{L}(V_1, V_2)$, then, for any $x^{(0)} \in V_1$,

$$\lim_{x \to x^{(0)}} \frac{\| Ax - Ax^{(0)} - A(x - x^{(0)}) \|_{V_2}}{\| x - x^{(0)} \|_{V_1}} = \lim_{x \to x^{(0)}} \frac{0}{\| x - x^{(0)} \|_{V_1}} = 0.$$

Therefore, $D(Ax^{(0)}) = A$.

Chain Rule: Let $F: V_1 \to V_2$, $G: V_2 \to V_3$,

Then $G \circ F : V_1 \to V_3$

Then $\boxed{D(G \circ F)(x) = DG(F(x)) \circ DF(x).}$

Example 3: $f(x) = f_1(x) f_2(x)$, where $f, f_1, f_2 : V \to \mathbb{R}$

Define $F: V \to \mathbb{R}^2$ as $F(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix}$ $\forall x \in V$

$G: \mathbb{R}^2 \to \mathbb{R}$ as $G\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \alpha\beta$ $\forall \alpha, \beta \in \mathbb{R}$.

Then $f(x) = G(F(x))$

So $Df(x) = DG(F(x)) \circ DF(x)$

Let's calculate

① $DG\begin{pmatrix} \alpha \\ \beta \end{pmatrix} z = \nabla G\begin{pmatrix} \alpha \\ \beta \end{pmatrix} z = \left\langle \begin{pmatrix} \beta \\ \alpha \end{pmatrix}, z \right\rangle \quad \forall z \in \mathbb{R}^2$

② For $DF(x)$

$$f_1(x) = f_1(x^{(0)}) + Df_1(x^{(0)})(x - x^{(0)}) + o(\|x - x^{(0)}\|)$$

$$f_2(x) = f_2(x^{(0)}) + Df_2(x^{(0)})(x - x^{(0)}) + o(\|x - x^{(0)}\|)$$

So, $F(x) = F(x^{(0)}) + DF(x^{(0)})(x - x^{(0)}) + o(\|x - x^{(0)}\|)$

where $DF(x^{(0)}): V \to \mathbb{R}$ defined by $DF(x^{(0)})(y) = \begin{pmatrix} Df_1(x^{(0)})(y) \\ Df_2(x^{(0)})(y) \end{pmatrix}$

Therefore,

$$Df(x)(y) = \left\langle \begin{pmatrix} f_2(x) \\ f_1(x) \end{pmatrix}, \begin{matrix} Df_1(x)(y) \\ Df_2(x)(y) \end{matrix} \right\rangle$$

$$\boxed{Df(x)(y) = Df_1(x)(y)\cdot f_2(x) + f_1(x)\cdot Df_2(x)(y)} \quad \forall\, y \in V.$$

In particular, if V is a Hilbert space,

$$\boxed{\nabla f(x) = f_2(x)\cdot \nabla f_1(x) + f_1(x)\cdot \nabla f_2(x)} \quad \text{product rule.}$$

Example 4: Let $A \in \mathcal{L}(V,W)$, where V, W two Hilbert spaces.

Let $f: V \to \mathbb{R}$ be $f(x) = \frac{1}{2}\|Ax-b\|_W^2$, where $b \in W$.

Let us calculate $Df(x^{(0)})$ and $\nabla f(x^{(0)})$.

$$f(x) = \tfrac{1}{2}\|Ax-b\|_W^2 = \tfrac{1}{2}\|(Ax^{(0)}-b) + A(x-x^{(0)})\|_W^2$$

$$= \tfrac{1}{2}\|Ax^{(0)}-b\|_W^2 + \langle Ax^{(0)}-b,\, A(x-x^{(0)})\rangle_W + \tfrac{1}{2}\|A(x-x^{(0)})\|_W^2$$

$$= f(x^{(0)}) + \underbrace{\langle A^*(Ax^{(0)}-b),\, x-x^{(0)}\rangle_V}_{L(x-x^{(0)}),\ \text{where } L: V\to\mathbb{R}} + \tfrac{1}{2}\|A(x-x^{(0)})\|_W^2.$$

So, $f(x) - f(x^{(0)}) - \langle A^*(Ax^{(0)}-b),\, x-x^{(0)}\rangle_V = \frac{1}{2}\|A(x-x^{(0)})\|_W^2$.

Also, $$\frac{\frac{1}{2}\|A(x-x^{(0)})\|_W^2}{\|x-x^{(0)}\|_V} \le \frac{\frac{1}{2}\|A\|^2\|x-x^{(0)}\|_W^2}{\|x-x^{(0)}\|_V} = \tfrac{1}{2}\|A\|^2\|x-x^{(0)}\|_V$$

$\to 0$ as $\|x-x^{(0)}\|_V \to 0$.

Therefore, $Df(x^{(0)})(y) = \langle A^*(Ax^{(0)}-b),\, y\rangle_V \quad \forall\, y \in V$

and $\nabla f(x^{(0)}) = A^*(Ax^{(0)}-b)$.

In particular, if $x \in \mathbb{R}^n$, $f(x) = \frac{1}{2}\|Ax-b\|_2^2$, $A \in \mathbb{R}^{m\times n}$,

then $\nabla f(x) = A^T(Ax-b)$.

- Hessian of functions on Hilbert spaces

Let $f: V \to \mathbb{R}$ where V is a Hilbert space.

1-st order derivative: $\nabla f(x) \in V$.

2-nd order derivative: view $x \to \nabla f(x)$ as a map $V \to V$, and its derivative $D(\nabla f)$ is the 2nd order derivative of of $f: V \to \mathbb{R}$.

Definition: $\nabla^2 f(x) \equiv D(\nabla f(x))$

Therefore, $\nabla^2 f(x)$ is a bounded linear transformation $V \to V$.

Example: $f(x) = \frac{1}{2}\|Ax-b\|_W^2$, where A bounded linear $V \to W$

$b \in W$, $x \in V$.

Then, $\nabla f(x) = A^*(Ax-b) = A^*Ax - A^*b$

and $D(\nabla f(x)) = A^*A$

So, $\nabla^2 f(x) = A^*A \in \mathcal{L}(V, V)$

In particular, if $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$

$$f(x) = \frac{1}{2}\|Ax-b\|_2^2$$

Then $\nabla f(x) = A^T(Ax-b)$

$$\nabla^2 f(x) = A^TA$$

- Function expansion
  - $\frac{d}{dt} f(x+tu) = \langle \nabla f(x+tu), u \rangle$, where $t \in \mathbb{R}$, $x, u \in V$.

because $f(x+su) = f(x+tu) + \langle \nabla f(x+tu), (s-t)u \rangle + o(|s-t|\|u\|)$

$= f(x+tu) + (s-t)\langle \nabla f(x+tu), u \rangle + o(|s-t|)$

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

i.e., $\boxed{\langle \nabla f(x), u \rangle \text{ is the directional derivative of } f(x) \text{ along } u.}$

- Similarly,

$$\frac{d^2}{dt^2} f(x+tu) = \frac{d}{dt}\langle \nabla f(x+tu), u \rangle = \langle \nabla^2 f(x+tu)\, u, u \rangle$$

because $\langle \nabla f(x+su), u \rangle = \langle \nabla f(x+tu), u \rangle$

$$+ \langle \nabla^2 f(x+tu)(s-t)u, u \rangle + o(|s-t|\|u\|^2)$$

$$= \langle \nabla f(x+tu), u \rangle + (s-t) \langle \nabla^2 f(x+tu)\, u, u \rangle + o(|s-t|)$$

Set $t=0$; $\frac{d^2}{dt^2} f(x+tu)\big|_{t=0} = \langle \nabla^2 f(x)\, u, u \rangle$.

That is, $\boxed{\langle \nabla^2 f(x)\, u, u \rangle \text{ is the 2-nd order derivative of } f(x) \text{ along } u.}$

- We can similarly show that

$$\frac{\partial}{\partial t_1}\frac{\partial}{\partial t_2} f(x+t_1u+t_2v)\big|_{t_1=t_2=0} = \frac{\partial}{\partial t_2}\frac{\partial}{\partial t_1} f(x+t_1u+t_2v)\big|_{t_1=t_2=0}$$

$$\langle \nabla^2 f(x)\, v, u \rangle = \langle \nabla^2 f(x)\, u, v \rangle$$

$$\forall x, u, v \in V \quad t_1, t_2 \in \mathbb{R}.$$

Therefore, $\langle \nabla^2 f(x)\, v, u \rangle = \langle v, \nabla^2 f(x)\, u \rangle \quad \forall\, u, v \in \mathbb{R}$,

which implies $\boxed{\nabla^2 f(x) = (\nabla^2 f(x))^* \text{ (i.e., Hessian is self-adjoint)}}$

- Consider function $g(t) = f(x^{(0)} + tu)$, $\quad x^{(0)}, u \in V$, $\quad t \in \mathbb{R}$.

By Taylor's expansion on $g(t)$,

$$g(t) = g(t_0) + g'(t_0)(t-t_0) + \frac{1}{2} g''(t_0)(t-t_0)^2 + o(|t-t_0|^2),$$

which is the same as

$$f(x^{(0)}+tu) = f(x^{(0)}+t_0u) + \frac{d}{dt} f(x^{(0)}+tu)\big|_{t=t_0} (t-t_0) + \frac{d^2}{dt^2} f(x^{(0)}+tu)\big|_{t=t_0} \frac{1}{2}(t-t_0)^2 + o(|t-t_0|^2)$$

So,

$$f(x^{(0)}+tu) = f(x^{(0)}+t_0u) + (t-t_0)\langle \nabla f(x^{(0)}+t_0u), u \rangle + \frac{1}{2}(t-t_0)^2 \langle \nabla f^2(x^{(0)}+t_0u)\, u, u \rangle + o(|t-t_0|^2)$$

$\forall\, x \in V$, choose $u = \frac{x-x^{(0)}}{\|x-x^{(0)}\|}$, and $t = \|x-x^{(0)}\|$, $t_0 = 0$

Then $x^{(0)} + tu = x^{(0)} + \|x-x^{(0)}\| \frac{x-x^{(0)}}{\|x-x^{(0)}\|} = x$, and

$$\boxed{f(x) = f(x^{(0)}) + \langle \nabla f(x^{(0)}), x-x^{(0)} \rangle + \frac{1}{2} \langle \nabla f^2(x^{(0)})(x-x^{(0)}), (x-x^{(0)}) \rangle + o(\|x-x^{(0)}\|^2)}$$

This is an expansion of $f(x)$ up to 2-nd derivative.

Many algorithms and theory are based on this expansion.