MATH3332 Data Analytic Tools

Ye Moe

HKUST Fall 2022

Introduction

The purpose of this course is to introduce some crucial mathematical analysis tools for data analysis/machine learning.

According to *Pedro Domingos*,

$$Learning = Representation + Evaluation + Optimization$$

1. Representation
   - How do we represent a learner? Which set should a learner be in? This set is called the hypothesis space of the learner. Some related tools are "space of functions".
   - How do we represent the input? Potential tools include vectors, graphs, manifolds, ...
2. Evaluation
   - How to pick the best learner from the hypothesis space? Needs calculus of "functions of functions" also known as functionals.
   - How to represent the input effectively? Needs Linear Algebra, Graph Theory, Manifolds Calculus, Harmonic Analysis, ...
3. Optimization
   - Numerical optimization solver - how to get the optimal solution numerically by a computer? Many of the resulting optimization is convex optimization and it is related to Convex Analysis.

So this course consists of some

- Basic functional analysis (calculus of functionals)
- Basic convex analysis
- Fourier analysis and Wavelet analysis (if time allowed)

Normed and Inner Product Space

2.1 Vector Spaces

Definition: A vector space over $\mathbb{R}$ is a set $\mathbb{V}$ together with two functions.

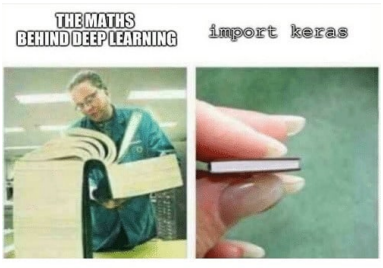
1. Vector addition: $+ : (\mathbb{V}, \mathbb{V}) \to \mathbb{V}$
   i.e. $\forall x, y \in \mathbb{V}, x + y \in \mathbb{V}$
2. Scalar multiplication: $. : (\mathbb{R}, \mathbb{V}) \to \mathbb{V}$
   i.e. $\forall \alpha \in \mathbb{R}, x \in \mathbb{V}, \alpha x \in \mathbb{V}$

These two functions should satisfy the following eight properties:

1. Associativity of addition: $x + (y + z) = (x + y) + z, \forall x, y, z \in \mathbb{V}$
2. Commutativity of addition: $x + y = y + x, \forall x, y \in \mathbb{V}$
3. Zero vector: $\exists$ an element, denoted by 0 in $\mathbb{V}$ s.t. $x+0 = 0+x = x, \forall x \in \mathbb{V}$
4. Negative vector: $\forall x \in \mathbb{V}, \exists$ an elements, denoted by $-x \in \mathbb{V}$ s.t. $x+(-x) = (-x) + x = 0$
5. $\forall x \in \mathbb{V}, 1 \cdot x = x$
6. $\forall x \in \mathbb{V}, \alpha, \beta \in \mathbb{R}, \alpha(\beta x) = (\alpha\beta)x$
7. $\forall x \in \mathbb{V}$ and $\alpha, \beta \in \mathbb{R}, (\alpha + \beta)x = \alpha x + \beta x$
8. $\forall x, y \in \mathbb{V}, \alpha(x + y) = \alpha x + \alpha y$

Remarks: We can define vector space over the complex domain $\mathbb{C}$, but since vector space over complex domain $\mathbb{C}$ is used very rarely, we will only consider vector space in the real domain $\mathbb{R}$.

Some examples of vector space include $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^{m\times n}$, $\mathbb{R}^{m\times n\times l}$, $C[a, b]$ and L_∞.



Machine learning be like

Example: Prove that $\mathbb{R}^n$ is a vector space.
$\forall x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$,

$$x + y = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix} \in \mathbb{R}^n$$

$$\alpha x = \alpha \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{bmatrix} \in \mathbb{R}^n$$

Since it is closed under both vector addition and scalar multiplication, $\mathbb{R}^n$ is a vector space.

Example: Prove that $C[a, b]$ is a vector space.
$\forall f, g \in C[a, b]$ and $\alpha \in \mathbb{R}$,

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

Since it is closed under both vector addition and scalar multiplication, $C[a, b]$ is a vector space.

Remarks: $C[a, b]$ is referred to as a function space, since any vector in this vector space is a function. It might be a hypothesis space of a learner with one input and one output, i.e. Find a $f \in C[a, b]$ s.t. $f(x_i) \approx y_i$ for all i.

Example: Prove that L_∞ is a vector space.

$$L_\infty = \{ \begin{bmatrix} a_1 \\ a_2 \\ \vdots \end{bmatrix} \mid \exists \text{ a finite number } c \text{ s.t. } \mid a_i \mid \leq c \text{ for any } i\}$$

$\forall a, b \in L_\infty$ and $\alpha \in \mathbb{R}$,

$$a + b = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \end{bmatrix} \in L_\infty$$

$$\alpha a = \alpha \begin{bmatrix} a_1 \\ a_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \alpha a_1 \\ \alpha a_2 \\ \vdots \end{bmatrix} \in L_\infty$$

Since it is closed under both vector addition and scalar multiplication, L_∞ is a vector space.

Remarks: This vector space can be used to model stock prices with a very fine time resolution.

Example: Consider the set of all strings.

$$\text{I' + 'am'} \neq \text{'am' + 'I'}$$

The set of all strings violates the commutative properties of a vector space, therefore it isn't a vector space. Hence, we cannot use vector space to model text data in this naïve way.

How do we "vectorize" the text data?
This is a fundamental question in text data analysis.

2.2 Normed and Banach Space

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

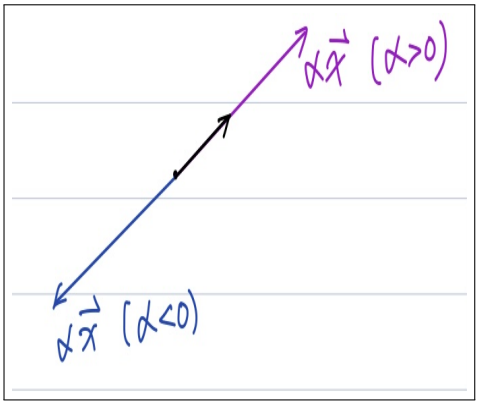
Let $\mathbb{V}$ be a vector space. Let $x, y \in \mathbb{V}$. Then,

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

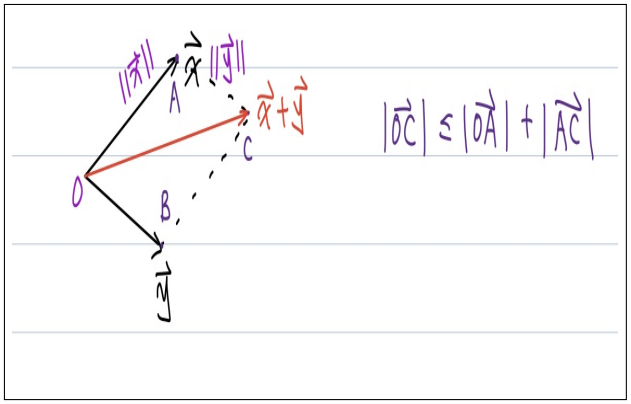
Remarks: Distance should be shift invariant.
To define distance, we only need to define the length of vectors. Let x $\in \mathbb{V}$.
Denote $\|x\|$ be the length of x. Then $\|x\|$ should satisfy:

1. $\|x\| \geq 0$ (the length should be non-negative)
   Moreover, $\|x\| = 0 \iff x = 0$(only zero vector has a zero length)
2. $\|\alpha x\| =\mid \alpha \mid \|x\|$

(length of a scaling of a vector is a scaling of the length of the vector)



3. $\|x + y\| \leq \|x\| + \|y\|$ (also known as triangle inequality)
   (length of direct path should be smaller than the length of indirect path)



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

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

Example: $\mathbb{R}$ is a vector space over $\mathbb{R}$.
Let $\|x\| = \mid x \mid \ \forall x \in \mathbb{R}$. Then it is a norm on $\mathbb{R}$.

Example: $\mathbb{R}^n$ is a vector space over $\mathbb{R}$.
There are many norms on $\mathbb{R}^n$.

- 2-norm: (Euclidean Norm)
  $\|x\|_2 = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$

  Question: Prove that $\|\cdot\|_2$ is indeed a norm for $\mathbb{R}^n$.
  $\forall x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$,

$$\|x\|_2 = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}} \geq 0$$

$$\|x\|_2 = 0 \iff \sum_{i=1}^n x_i^2 = 0 \iff x_i^2 = 0, \ i = 1, ..., n$$

$$\iff x_i = 0, \ i = 1, ..., n \iff x = 0$$

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

$$\|x + y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 + 2\langle x, y \rangle$$

$$\leq \|x\|_2^2 + \|y\|_2^2 + 2\|x\|_2\|y\|_2 \text{ (By Cauchy-Schwartz inequality)}$$

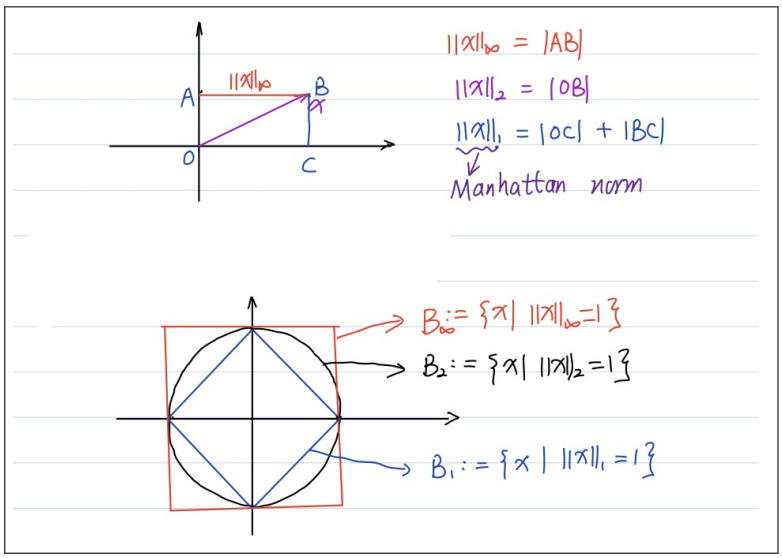
$$= (\|x\|_2 + \|y\|_2)^2$$

$$\|x + y\|_2 \leq \|x\|_2 + \|y\|_2$$

- 1-norm:
  $\|x\|_1 = \sum_{i=1}^n \mid x_i \mid$
- ∞-norm:
  $\|x\|_\infty = \max\limits_{1 \leq i \leq n} \mid x_i \mid$
- p-norm:
  $\|x\|_p = (\sum_{i=1}^n \mid x_i \mid^p)^{\frac{1}{p}}$

Fact: $\|x\|_p$ is a norm on $\mathbb{R}^n \iff$ p ≥ 1

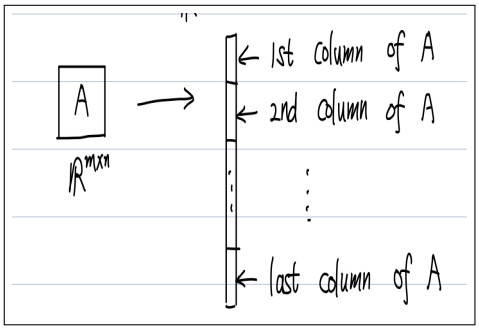
Geometric definition of different norms in $\mathbb{R}^n$



Note that $(\mathbb{R}^n, \|\cdot\|_1), (\mathbb{R}^n, \|\cdot\|_2), (\mathbb{R}^n, \|\cdot\|_\infty), \ldots$ are all different normed spaces. So for a given vector space, we can obtain various normed space by choosing different norms. Also, $\|x\|_p \leq \|x\|_q$ if $p \geq q$.

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

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



We can define vector p-norm for $\mathbb{R}^{m\times n}$.

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

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

  This norm is widely known as the Frobenius norm denoted as $\|A\|_F$.
- p = ∞
  $\|A\|_{\infty,vec} = \underset{i=1,\ldots,m}{max}\ \underset{j=1,\ldots,n}{max} \mid a_{ij} \mid$

2. $\mathbb{R}^{m\times n}$ can be viewed as linear transformation from $\mathbb{R}^n \to \mathbb{R}^m$.
   We can define matrix p-norm for $\mathbb{R}^{m\times n}$.

$$\|A\|_p = \underset{x\neq 0, x\in\mathbb{R}^n}{max} \frac{\|Ax\|_p}{\|x\|_p} = \underset{\|x\|_p=1}{max} \|Ax\|_p$$

   - p = 1
     $\|A\|_1 = \underset{1\leq j\leq n}{max} \sum_{i=1}^{m} \mid a_{ij} \mid$ = maximum absolute column sum
   - p = ∞
     $\|A\|_\infty = \underset{1\leq i\leq m}{max} \sum_{j=1}^{n} \mid a_{ij} \mid$ = maximum absolute row sum
   - p = 2
     $\|A\|_2$ = maximum singular value of A

3. We can also define other matrix norms.

   (a) We can use different norms in $\mathbb{R}^n$ and $\mathbb{R}^m$.

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

   (b) The nuclear norm $\|\cdot\|_*$

Example: $C[a,b]$ is a vector space over $\mathbb{R}$.
$\forall f \in C[a,b]$, define

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

We can check that $\|\cdot\|_\infty$ is indeed a norm on $C[a,b]$.
The distance of two function $f, g \in C[a,b]$ is given by

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

Some other norms on $C[a,b]$.

1. $\|f\|_1 = \int_b^a \mid f(t) \mid dt$
2. $\|f\|_2 = (\int_b^a \mid f(t) \mid^2 dt)^{\frac{1}{2}}$
3. $\|f\|_p = (\int_b^a \mid f(t) \mid^p dt)^{\frac{1}{p}}$

Example: $L_\infty = \{a | a \text{ is a infinite sequence and } \exists\ c > 0 \text{ s.t. } \mid a_i \mid \leq\ c, \forall i\}$

1. $\forall a \in L_\infty$, define

$$\|a\|_\infty = \sup_i \mid a_i \mid$$

Remarks: You cannot replace sup here with max.

2. Define $\|a\|_p = (\sum_{i=1}^\infty \mid a_i \mid^p)^{\frac{1}{p}}$, $\forall a \in L_\infty$ but this is not a norm on L_∞.

e.g. $a = \begin{bmatrix} 1 \\ \frac{1}{2} \\ \vdots \\ \frac{1}{i} \\ \vdots \end{bmatrix} \in L_\infty,$ but $\|a\|_1 = \sum_{i=1}^\infty \mid a_i \mid = \sum_{i=1}^\infty \frac{1}{i} = \infty$

So, $\|\cdot\|_1$ is not a norm on L_∞.
Instead, we consider

$$L_p = \{a \in L_\infty | \|a\|_p < \infty\} \subset L_\infty$$
$$\|\cdot\|_p \text{ is a norm on } L_p.$$

e.g. $a = \begin{bmatrix} 1 \\ \frac{1}{2} \\ \vdots \\ \frac{1}{i} \\ \vdots \end{bmatrix} \in L_\infty$

$\|a\|_\infty = 1$, $\|a\|_2 = (\sum_{i=1}^\infty \frac{1}{i^2})^{\frac{1}{2}} = (\frac{\pi^2}{6})^{\frac{1}{2}} = \frac{\pi}{\sqrt{6}}$, $\|a\|_1 = \infty$
So, $a \in L_\infty$, $a \in L_2$ but $a \notin L_1$. Indeed, $a \in L_p$, $\forall p > 1$.

Limit and Convergence on Normed Vector Space
To define calculus, we first need to define convergent sequence.
Let $\mathbb{V}$ be a normed vector space. Let $\{x^{(k)}\}_{k\in\mathbb{N}}$ be a sequence in $\mathbb{V}$. Let $x \in \mathbb{V}$. We say $\{x^{(k)}\}_{k\in\mathbb{N}}$ converges to x, denoted by $x^{(k)} \to x$, if

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

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

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

$$\|x^{(k)} - x\|_2 = \|x^{(k)}\|_2 = (\textstyle\sum_{i=1}^n (\frac{i}{k})^2)^{\frac{1}{2}} = \frac{1}{k}(\sum_{i=1}^n i^2)^{\frac{1}{2}}$$
$$\lim_{k\to\infty} \|x^{(k)} - x\|_2 = \lim_{k\to\infty} \tfrac{1}{k}(\textstyle\sum_{i=1}^n i^2)^{\frac{1}{2}} = 0$$
$$x^{(k)} \to x$$

Unfortunately, the limit of a sequence may not always be in the same vector space as the original sequence. If this happen, we call this incomplete normed vector space. Otherwise, it is a complete vector space also known as the Banach space.

Example of Banach space:

1. $\mathbb{R}^n$ with any norm
2. $\mathbb{R}^{m\times n}$ with any norm
3. Tensor space $\mathbb{R}^{m\times n\times l}$ with any norm
4. $C[a, b]$ with $\|\cdot\|_\infty$
5. L_p with p-norm, for $p \geq 1$ and $p = \infty$.

Cauchy Sequence
Definition: $\{x^{(k)}\}$ is a Cauchy sequence, if for any $\epsilon > 0$, there exists K such that for any $k, l > K$, $\|x^{(k)} - x^{(l)}\| < \epsilon$.
Facts:

1. If $x^{(k)} \to x$ in $(\mathbb{V}, \|\cdot\|)$, then $\{x^{(k)}\}$ must also be a Cauchy sequence.
   Proof.
   $x^{(k)} \to x$ implies that $\forall\epsilon > 0$, $\exists k$, s.t. $k > K$ $\|x^{(k)} - x\| \leq \frac{\epsilon}{2}$. Therefore, $\|x^{(k)} - x^{(l)}\| \leq \|x^{(k)} - x\| + \|x^{(l)} - x\| \leq \epsilon$, $\forall k, l > K$

2. The reverse is **NOT** necessarily true.

Definition: A vector space $(\mathbb{V}, \|\cdot\|)$ is complete if the limit of all Cauchy sequences in $\mathbb{V}$ is in $\mathbb{V}$.

Remarks: We can always complete an incomplete normed vector space by including all limits of its Cauchy sequence.

Finite Dimensional Vector Space
In most cases, we are dealing with finite dimensional vector space such as $\mathbb{R}^n$, $\mathbb{R}^{m\times n}$ and $\mathbb{R}^{m\times n\times l}$.

Properties related to Finite Dimensional Vector Space:

- Any finite dimensional vector space with any norm is complete. That is, any finite dimensional vector space is Banach space.
- For a finite dimensional vector space $\mathbb{V}$, all norms are equivalent.
  Theorem: For any norms $\|\cdot\|_A$ and $\|\cdot\|_B$, $\exists c_1, c_2 > 0$ s.t.
  $c_1\|a\|_A \le \|a\|_B \le c_2\|a\|_A$, $\forall a \in \mathbb{V}$. (finite dimensional)

Example: Prove that $x^{(k)} \to x$ in $\|\cdot\|_A \iff x^{(k)} \to x$ in $\|\cdot\|_B$.
Since $x^{(k)} \to x$ in $\|\cdot\|_A$,

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

Because of equivalence,

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

$$0 \le \lim_{k\to\infty} \|x^{(k)} - x\|_B \le c_2 \lim_{k\to\infty} \|x^{(k)} - x\|_A = 0$$

$$\lim_{k\to\infty} \|x^{(k)} - x\|_B = 0 \ (\textit{by squeeze theorem})$$

$$x^{(k)} \to x \text{ under } \|\cdot\|_B$$

Similarly for the $\leftarrow$ direction.

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

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

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

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

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

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

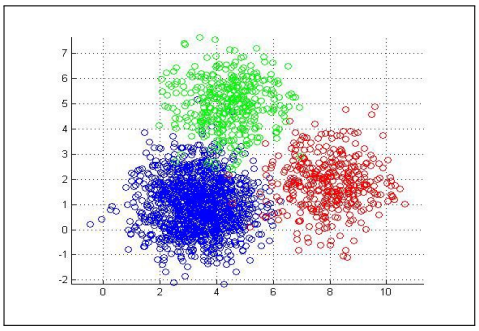
$$\|a\|_\infty \le \|a\|_1 \le n\|a\|_\infty, \forall a \in \mathbb{R}^n$$

Remarks: Though they are equivalent, the speed at which they converge are different. In other words, the convergence speed depends on norms.

Case Study: Clustering, k-means, k-medians

Clustering

Suppose we are given N vectors $x_1, x_2, \cdots, x_N \in \mathbb{R}^n$, the goal of clustering is to group or partition the vectors into k groups or clusters, with the vectors in each group close to each other.



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

- Recommendation system
- Image clustering
- Text data clustering
- Many other applications.

Mathematical formulation:

- **Representation:**
  Let $c_i \in \{1, 2, \cdots, k\}$ be the group that x_i belongs to. $i = 1, 2, \cdots, N$
  Then, group G_j denoted by G_j, is $G_j = \{i | c_i = j\}$. $j = 1, 2, \cdots, k$.
  We assign each group a representative vector, denoted by $z_1, z_2, \cdots, z_k$.
  The representative vectors are not necessarily one of the given vectors.

- **Evaluation:**
  First of all, within one specific group G_j, all vectors should be close to the representative vector z_j. More precisely, let

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

  Then, J_j should be small.
  Secondly, consider all groups, since each J_j is small,

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

  should be small.

Altogether, we solve the following

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

- **Optimization:**
  We may use an alternating minimization to solve this minimization problem.

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

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

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

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

  The two steps are repeated until convergence.

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

For ①:
Finding the partition $G_1, \cdots, G_k$ is equivalent to finding $c_1, \cdots, c_N$. So ① becomes

$$\begin{aligned} &\min_{c_1,\cdots,c_N} (\|x_1 - z_{c_1}\|_2^2 + \cdots + \|x_N - z_{c_N}\|_2^2) \\ \iff &\min_{c_i\in\{1,2,\cdots,k\}} \|x_i - z_{c_i}\|_2^2 \\ \iff &\ c_i = \operatorname*{argmin}_{j\in\{1,2,\cdots,k\}} \|x_i - z_j\|_2^2 \end{aligned}$$

In other words, x_i is assigned to the group whose representative vector is the closest to x_i.

For ②:
It is rewritten as

$$\min_{z_1,\cdots,z_k} (\sum_{i\in G_i} \|x_i - z_1\|_2^2 + \cdots + \sum_{i\in G_k} \|x_i - z_k\|_2^2)$$

Obviously, it is equivalent to minimize each term independently, i.e., solve k independent problems.

$$\begin{aligned} &\min_{z_j} (\sum_{i\in G_j} \|x_i - z_j\|_2^2),\ j = 1, 2, \cdots, k \\ \iff &\ z_j = \tfrac{1}{|G_j|} \sum_{i\in G_j} x_i,\ j = 1, 2, \cdots, k \\ &\text{where } \mid G_j \mid \text{ is the number of elements in } G_j. \end{aligned}$$

In other words, z_j is the mean of all vector in G_j. Note that in the above derivation, when we consider n = 1,

$$\begin{aligned}
&\min_{z_j \in \mathbb{R}} \sum_{i \in G_j} (x_i - z_j)^2 \\
\iff &\sum_{i \in G_j} 2(x_i - z_j) = 0 \\
\iff &\sum_{i \in G_j} z_j = \sum_{i \in G_j} x_i \\
\iff &z_j = \frac{1}{|G_j|} \sum_{i \in G_j} x_i
\end{aligned}$$

Altogether, we get the following clustering algorithm.

k-means Clustering
Initialization: Initialize $z_1, z_2, \cdots, z_k$.
Step 1: Given $z_1, z_2, \cdots, z_k$, compute

$$c_i = \underset{j \in \{1,2,\cdots,k\}}{\operatorname{argmin}} \|x_i - z_j\|_2^2, \; i = 1, 2, \cdots, N$$

and define

$$G_j = \{i | c_i = j\}, \; j = 1, 2, \cdots, k$$

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

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

Go back to step 1.

In k-means, the Euclidean norm is used. We can replace it by 1-norm. We solve

$$\min_{\substack{G_1,\cdots,G_k \\ z_1,\cdots,z_k}} \sum_{j=1}^{k} \|x_i - z_j\|_1$$

k-medians Clustering
Initialization: Initialize $z_1, z_2, \cdots, z_k$.
Step 1: Given $z_1, z_2, \cdots, z_k$, compute

$$c_i = \underset{j \in \{1,2,\cdots,k\}}{\operatorname{argmin}} \|x_i - z_j\|_1, \; i = 1, 2, \cdots, N$$

and define

$$G_j = \{i | c_i = j\}, \; j = 1, 2, \cdots, k$$

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

$$z_j = median\{x_i | i \in G_j\}$$

Go back to step 1.

2.3 Inner Product and Hilbert Space

Question: How do we describe the correlation/centerment between two vectors? Norms are not able to describe it as they are 'scaling sensitive'.

A good answer would be to use angle. A good candidate would be to use inner product since it is 'scaling insensitive'.

Inner Product
Definition: A function $\langle \cdot, \cdot \rangle : (\mathbb{V}, \mathbb{V}) \to \mathbb{R}$ on a vector space $\mathbb{V}$ is called an inner product over $\mathbb{R}$, if:

1. $\forall x \in \mathbb{V}, \langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \iff x = 0$
2. $\langle \alpha x_1 + \beta x_2, y \rangle = \alpha \langle x_1, y \rangle + \beta \langle x_2, y \rangle, \ \forall \alpha, \beta \in \mathbb{R}, x_1, x_2, y \in \mathbb{V}$
3. $\langle x, y \rangle = \langle y, x \rangle, \ \forall x, y \in \mathbb{V}$

Remarks:

1. By 2 and 3, $\langle x, \alpha y_1 + \beta y_2 \rangle = \alpha \langle x, y_1 \rangle + \beta \langle x, y_2 \rangle$, $\forall \alpha\beta \in \mathbb{R}$, $x_1, y_1, y_2 \in \mathbb{V}$. Therefore, $\langle \cdot, \cdot \rangle$ is a bi-linear function, i.e., it is linear with respect to one of the variable with the other fixed.
2. For inner product of vector spaces on $\mathbb{C}$, we only need to change ③ to $\langle x, y \rangle = \overline{\langle y, x \rangle}$, where $\overline{\langle \cdot, \cdot \rangle}$ stands for complex conjugate.

Example: $\mathbb{R}^n$ is a vector space. We can define an inner product as

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

Example: Another inner product in $\mathbb{R}^n$ is as follows. We can define a "weighted" inner product as $\langle x, y \rangle_A = x^T A y$, where $A \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix.
Remarks: A is SPD $\iff$ $A = A^T$ and $x^T A x > 0 \ \forall x \in \mathbb{R}^n$ and $x \neq 0$.

Example: $\mathbb{R}^{m \times n}$ is a vector space. We can define an inner product as

$$\langle A, B \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij}, \ \forall A, B \in \mathbb{R}^{m \times n}$$

Similarly, these are equal to $trace(A^T B)$, $trace(B^T A)$, $trace(AB^T)$ and $trace(BA^T)$, where $trace(A)$ is defined as the sum of the diagonal of matrix A.

Example: In L_2, we can define an inner product as

$$\langle a, b \rangle = \sum_{i=1}^{\infty} a_i b_i, \ \forall a, b \in L_2$$

Example: In $C[a, b]$, we can define an inner product as

$$\langle f, g \rangle = \int_a^b f(t) g(t) dt, \ \forall f, g \in C[a, b]$$

Cauchy-Schwartz Inequality
If $\langle \cdot, \cdot \rangle$ is an inner product on $\mathbb{V}$, then, for any $x, y \in \mathbb{V}$,

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

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

Proof.
Let $\lambda \in \mathbb{R}$ be an arbitrary number,

$$0 \leq \langle x + \lambda y, x + \lambda y \rangle$$

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

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

$$\text{Thus, } \lambda^2 \langle y, y \rangle + 2\lambda \langle x, y \rangle + \langle x, x \rangle \geq 0, \forall \lambda \in \mathbb{R}$$

The left is a quadratic function of λ and is always non-negative. There is at most one root of the quadratic function, hence, the discriminant $b^2 - 4ac \leq 0$.

$$\text{So, } (2\langle x, y \rangle)^2 - 4\langle x, x \rangle \langle y, y \rangle \leq 0$$

$$\implies \langle x, y \rangle^2 \leq \langle x, x \rangle \langle y, y \rangle$$

Finally, when $\langle x, y \rangle^2 = \langle x, x \rangle \langle y, y \rangle$, there is a root, i.e., $\exists$ a unique $\lambda \in \mathbb{R}$, $\lambda^2 \langle y, y \rangle + 2\lambda \langle x, y \rangle + \langle x, x \rangle = 0$.

$$\iff$$

$$\exists \text{ a unique } \lambda \in \mathbb{R}, \langle x + \lambda y, x + \lambda y \rangle = 0.$$

$$\iff$$

$$\exists \text{ a unique } \lambda \in \mathbb{R}, x + \lambda y = 0.$$

$$\iff$$

$$\exists \text{ a unique } \lambda \in \mathbb{R}, x = -\lambda y.$$

With the Cauchy-Schwartz inequality, we can show that

$$\|x\| = (\langle x, x\rangle)^{\frac{1}{2}} \text{ defines a norm.}$$

This is also called "norm induced by the inner product". This one above is for $\mathbb{R}^n$.

Proof.

$$\|x\| = (\langle x, x\rangle)^{\frac{1}{2}} \geq 0 \text{ and } \|x\| = (\langle x, x\rangle)^{\frac{1}{2}} = 0 \iff x = 0$$

$$\|\alpha x\| = (\langle \alpha x, \alpha x\rangle)^{\frac{1}{2}} = (\alpha^2 \langle x, x\rangle)^{\frac{1}{2}} = \mid \alpha \mid \|x\|$$

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

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

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

$$\leq \|x\|^2 + \|y\|^2 + 2\|x\|\|y\|$$

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

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

Remarks: In the proof above, we have used an alternative version of the Cauchy-Schwartz inequality.

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

All kinds of induced norm

1. $\mathbb{R}^n$ with inner product $\langle \cdot, \cdot \rangle : \ \langle x, y\rangle = x^T y$
   The induced norm is
   $\|x\| = (\langle x, x\rangle)^{\frac{1}{2}} = (x^T x)^{\frac{1}{2}} = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}} = \|x\|_2$
2. $\mathbb{R}^n$ with weighted inner product $\langle \cdot, \cdot \rangle_A : \ \langle x, y\rangle_A = x^T A y$
   The induced norm is
   $\|x\|_A = (x^T A x)^{\frac{1}{2}} = (\sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j)^{\frac{1}{2}}$
3. The p-norm of $\mathbb{R}^n$
   $\|x\|_p = (\sum_{i=1}^n \mid x_i \mid^p)^{\frac{1}{p}}$
   When $p = 2$, $\|\cdot\|_2$ is induced by $\langle \cdot, \cdot \rangle$. It is not induced by inner product for all p except for 2.
4. $\mathbb{R}^{m \times n}$ with inner product $\langle \cdot, \cdot \rangle : \ \langle A, B\rangle = \sum_{ij} a_{ij} b_{ij}$
   The induced norm is
   $\|A\| = (\langle A, A\rangle)^{\frac{1}{2}} = (\sum_{ij} a_{ij}^2)^{\frac{1}{2}} = \|A\|_F = \|A\|_{vec,2}$
5. Infinite sequence with inner product $\langle \cdot, \cdot \rangle : \ \langle a, b\rangle = \sum_{i=1}^{\infty} a_i b_i$
   $\|a\| = (\sum_{i=1}^{\infty} a_i^2)^{\frac{1}{2}} = \|a\|_2$
6. $C[a, b]$ with inner product $\langle \cdot, \cdot \rangle : \ \langle f, g\rangle = \int_a^b f(t) g(t) dt$
   $\|f\| = (\int_a^b (f(t))^2 dt)^{\frac{1}{2}} = \|f\|_2$

Angle in inner product spaces
By Cauchy-Schwartz inequality,

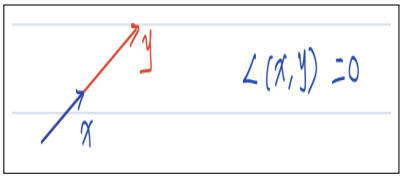
$$|\langle x, y\rangle| \leq \|x\|\|y\| \ \forall x, y \in \mathbb{V}$$

Then,

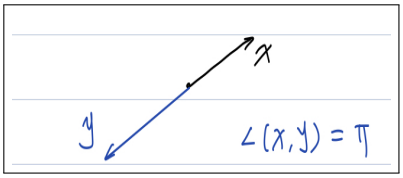
$$-\|x\|\|y\| \ \leq \ \langle x, y\rangle \ \leq \ \|x\|\|y\|$$

$$-1 \ \leq \ \frac{\langle x,y\rangle}{\|x\|\|y\|} \ \leq \ 1 \text{ if } x, y \neq 0$$

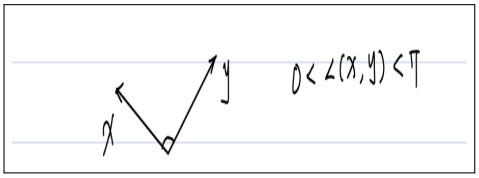
If $\frac{\langle x,y\rangle}{\|x\|\|y\|} = 1$, then $x = \alpha y$ with $\alpha > 0$. Otherwise, if $\alpha \leq 0$, then $\langle x, y\rangle = \alpha\langle y, y\rangle = \alpha\|y\|^2 \leq 0$. (*Contradiction*).



If $\frac{\langle x,y\rangle}{\|x\|\|y\|} = -1$, then $x = \alpha y$ with $\alpha < 0$.



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



Then we define

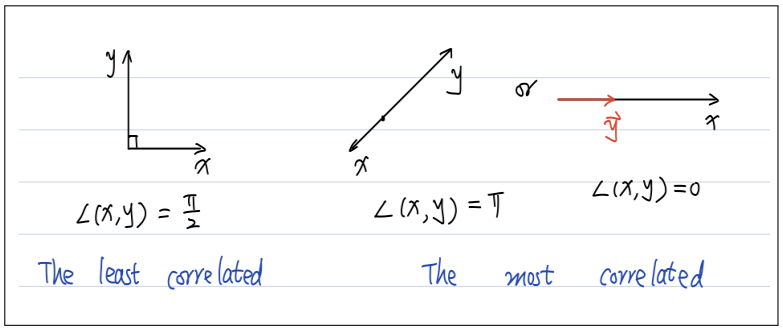
$$\boxed{L(x, y) = \arccos \frac{\langle x,y\rangle}{\|x\|\|y\|}}$$

This definition is consistent with the observation above and the angles of vectors in $\mathbb{R}^2$ and $\mathbb{R}^3$.

Orthogonality
Let $\mathbb{V}$ be a vector space and $\langle \cdot, \cdot \rangle$ be the inner product.

- If $\frac{\langle x,y \rangle}{\|x\|\|y\|} = 1$ or -1, then x and y are the most correlated.
- If $\frac{\langle x,y \rangle}{\|x\|\|y\|} = 0$, then x and y are the least correlated.

If $\langle x, y \rangle = 0$, then we say x and y are orthogonal.



Pythagorean theorem
Definition: Let x, y be two vectors in an inner product space $\mathbb{V}$.
Then $x \perp y \iff \|x+y\|^2 = \|x\|^2 + \|y\|^2$.
Proof.

$$\|x+y\|_2^2 = \langle x+y, x+y \rangle$$

$$= \|x\|^2 + \|y\|^2 + 2\langle x, y \rangle \textbf{(1)}$$

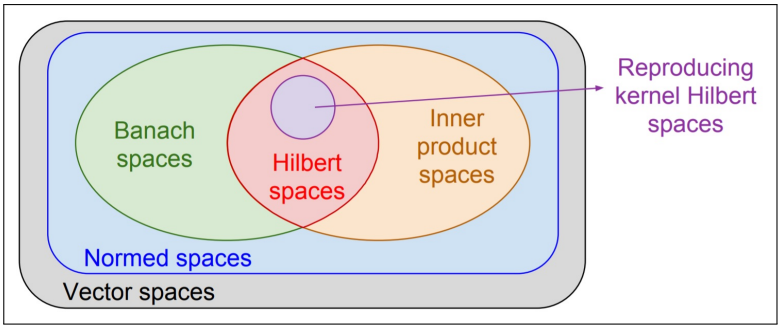
If $x \perp y$, then $\langle x, y \rangle = 0$.

$$\implies \|x+y\|^2 = \|x\|^2 + \|y\|^2$$

If $\|x+y\|^2 = \|x\|^2 + \|y\|^2$, together with **(1)**, we have $\langle x, y \rangle = 0$.

Hilbert Space

Definition: A Hilbert space is a Banach space in which the norm is induced by an inner product.



Examples of Hilbert Space

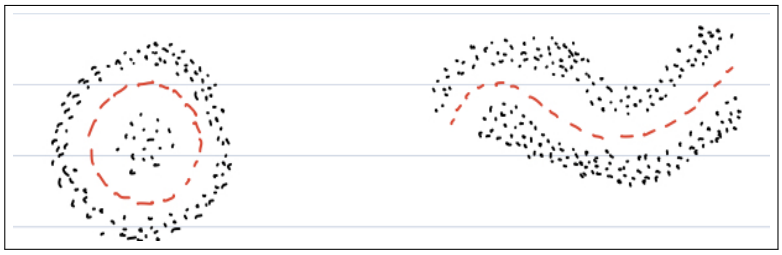
1. $\mathbb{R}^n$ with $\langle \cdot, \cdot \rangle$ is a Hilbert space.
2. $\mathbb{R}^n$ with $\langle \cdot, \cdot \rangle_A$ is a Hilbert space.
3. $\mathbb{R}^{m \times n}$ with $\langle \cdot, \cdot \rangle$ is a Hilbert space.
4. $L_2 = \{a \mid \|a\|_2 < \infty \text{ and } a \text{ is an infinite sequence}\}$ with $\langle \cdot, \cdot \rangle$ is a Hilbert space.
5. $C[a, b]$ with $\langle \cdot, \cdot \rangle$ is **NOT** a Hilbert space, because it is not a Banach space. In other words, the limit of a convergent sequence in $C[a, b]$ may not be in $C[a, b]$. To complete $C[a, b]$ under the norm $\| \cdot \| = (\langle \cdot, \cdot \rangle)^{\frac{1}{2}}$, we need to extend the Riemann integral to the so-called Lebesgue integral, and the resulting Hilbert space is $L^2(a, b)$.

In the following chapters, we will consider calculus on Hilbert/Banach spaces.

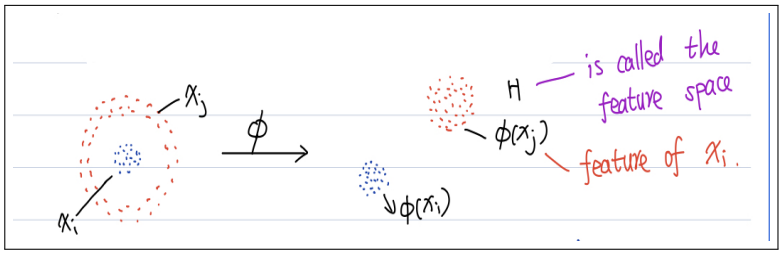
Case Study: Kernel trick, Kernel k-means

Recall that in k-means, we want to group $x_1, x_2, \cdots, x_N \in \mathbb{R}^n$ into k groups. k-means work well only if the data are linearly separable. It will fail on "curved" data sets in $\mathbb{R}^n$.

k-means will fail for the following examples.



To modify k-means to these "curved" data sets in $\mathbb{R}^n$, we transform the "curved" data sets to "uncurved" data set in a Hilbert space $\mathbb{H}$.
Let $\phi : \mathbb{R}^n \to \mathbb{H}$,



Then we apply k-means to $\phi(x_1), \phi(x_2), \cdots, \phi(x_N)$ in $\mathbb{H}$ and let $z_1, z_2, \cdots, z_k$ be the representative vectors in $\mathbb{H}$.

k-means Clustering
Step 0: Initialize $z_1, z_2, \cdots, z_k$
Step 1: Given $z_1, z_2, \cdots, z_k$, compute

$$c_i = \underset{j \in \{1,2,\cdots,k\}}{\operatorname{argmin}} \|\phi(x_i) - z_j\|^2, \; i = 1, 2, \cdots, N$$

and define

$$G_j = \{i | c_i = j\}, \; j = 1, 2, \cdots, k$$

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

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

Go back to step 1.

However, finding the feature map ϕ is not easy, because ϕ depends on the shape of $x_1, x_2, \cdots, x_N$, which generally is very complicated.
The good news is that:

There is no need to know ϕ explicitly in k-means algorithm.

Why?

- First of all, since we care only about the groups of $x_1, \cdots, x_N$, we only need to know $G_1, \cdots, G_k$. The representatives $z_1, \cdots, z_k$ are only intermediate. Therefore, we can eliminate $z_1, \cdots, z_k$ in the k-means algorithm.

  Modified k-means Clustering
  Step 0: Initialize $G_1, \cdots, G_k$
  Step 1: Given $G_1, \cdots, G_k$, compute

  $$c_i = \underset{j \in \{1,2,\cdots,k\}}{\operatorname{argmin}} \|\phi(x_i) - \tfrac{1}{|G_j|} \textstyle\sum_{l \in G_j} \phi(x_l)\|^2,\ i = 1, 2, \cdots, N$$

  and define

  $$G_j = \{i | c_i = j\},\ j = 1, 2, \cdots, k$$

  Go back to step 1.

- Now, since we are talking about Hilbert space $\mathbb{H}$, we can expand the norm by

  $$\begin{gathered} \|\phi(x_i) - \tfrac{1}{|G_j|} \textstyle\sum_{l \in G_j} \phi(x_l)\|^2 \\ = \langle \phi(x_i) - \tfrac{1}{|G_j|} \textstyle\sum_{l \in G_j} \phi(x_l), \phi(x_i) - \tfrac{1}{|G_j|} \textstyle\sum_{l \in G_j} \phi(x_l) \rangle \\ = \langle \phi(x_i), \phi(x_i) \rangle - \tfrac{2}{|G_j|} \textstyle\sum_{l \in G_j} \langle \phi(x_i), \phi(x_l) \rangle + \\ \tfrac{1}{|G_j|^2} \textstyle\sum_{l_1 \in G_j} \sum_{l_2 \in G_j} \langle \phi(x_{l_1}), \phi(x_{l_2}) \rangle \end{gathered}$$

  All terms involved are in the form of

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

  Instead of defining ϕ explicitly, we define a function

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

  s.t. $k(x, y) = \langle \phi(x), \phi(y) \rangle$.

  Therefore, an explicit expression of ϕ is **NOT** necessary. This process is also known as kernel trick.

Kernel function
$k(x, y)$ is called a kernel function. Which kernel function?
Necessary conditions:

1. $k(x, y) = \langle \phi(x), \phi(y) \rangle = \langle \phi(y), \phi(x) \rangle = k(y, x)$
   We say k is a symmetric kernel is $k(x, y) = k(y, x)$, $\forall x, y \in \mathbb{R}^n$.

2. Let $y_1, \cdots, y_m \in \mathbb{R}^n$. Then for any $c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} \in \mathbb{R}^n$,

$$0 \leq \langle \sum_{i=1}^m c_i \phi(y_i), \sum_{j=1}^m c_j \phi(y_j) \rangle = \sum_{i=1}^m \sum_{j=1}^m c_i c_j \langle \phi(y_i), \phi(y_j) \rangle = \sum_{i=1}^m \sum_{j=1}^m c_i c_j k(y_i, y_j) = c^T K c \text{ where } K = [k(y_i, y_j)]_{i,j}$$

That is, $\forall c \in \mathbb{R}^m$, $c^T K c \geq 0$ and $K^T = K$.

Definition: We say a kernel function $k : (\mathbb{R}^n, \mathbb{R}^n) \to \mathbb{R}$ is symmetric positive semi-definite if:

1. $k(x, y) = k(y, x)$, $\forall x, y \in \mathbb{R}^n$
2. For any m and $y_1, y_2, \cdots, y_m \in \mathbb{R}^n$, the matrix

$$K = [k(y_i, y_j)]_{i,j}$$

   is symmetric positive semi-definite.

Mercer's Theorem: If $k : (\mathbb{R}^n, \mathbb{R}^n) \to \mathbb{R}$ is continuous and symmetric positive semi-definite, then there exists a Hilbert space $\mathbb{H}$ and a mapping such that $k(x, y) = \langle \phi(x), \phi(y) \rangle$.

Some popular kernels:

- $k(x, y) = x^T y$ ($\phi(x) = x$ (No transform kernel))
- $k(x, y) = (x^T y + 1)^\alpha$ (α is an integer (Polynomial kernel))
- $k(x, y) = e^{-\frac{\|x-y\|_2^2}{\sigma^2}}$ ($\sigma > 0$ is a parameter (Gaussian kernel))

Kernel k-means Clustering
Step 0: Initialize $G_1, \cdots, G_k$
Step 1: Given $G_1, \cdots, G_k$, compute

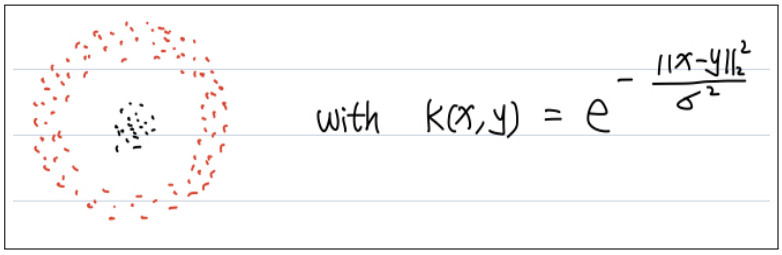
$$c_i = k(x_i, x_i) - \frac{2}{|G_j|} \sum_{l \in G_j} k(x_i, x_l) + \frac{1}{|G_j|^2} \sum_{l_1 \in G_j} \sum_{l_2 \in G_j} k(x_{l_1}, x_{l_2}), \quad i = 1, 2, \cdots, N$$

and define

$$G_j = \{i | c_i = j\},\ j = 1, 2, \cdots, k$$

Go back to step 1.

Why kernel k-means work?

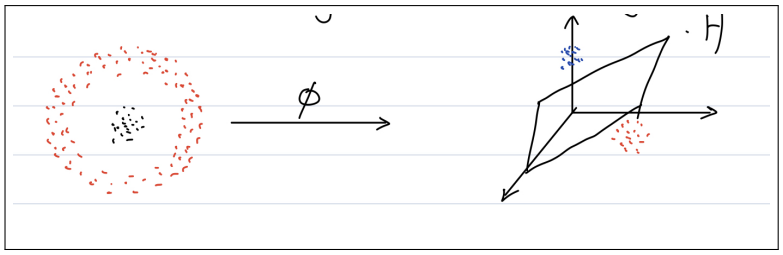


Suppose we use Gaussian kernel,

- $k(x_i, x_i) = e^{-\frac{\|x_i - x_i\|_2^2}{\sigma^2}} = e^{-0} = 1, \forall i$
  so all $\phi(x_1), \cdots, \phi(x_N)$ are on unit sphere in $\mathbb{H}$.
- $k(x_i, x_j)$ $\begin{cases} \approx 1 \text{ if } x_i \approx x_j \\ \approx 0 \text{ if } \|x_i - x_j\|_2 \text{ is large} \end{cases}$

  Since $\langle \phi(x_i), \phi(x_j) \rangle = k(x, y)$,

  - If $x_i \approx x_j$, then
    $\|\phi(x_i) - \phi(x_j)\|^2 = \|\phi(x_i)\|^2 - 2\langle \phi(x_i), \phi(x_j) \rangle + \|\phi(x_j)\|^2 \approx 0$
    $\implies \phi(x_i) = \phi(x_j)$.
  - If $\|x_i - x_j\|_2$ is large, then $\phi(x_i) \perp \phi(x_j)$.

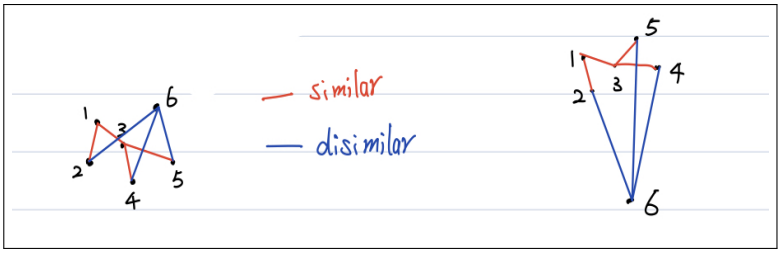


Thus, kernel k-means work for "curved" data sets which k-means fail.

Case Study: Metric Learning

Given a set of data $x_1, x_2, ..., x_n \in \mathbb{R}^n$ and $S : (x_i, x_j) \in S$ if x_i and x_j are similar and $D : (x_i, x_j) \in D$ if x_i and x_j are dissimilar.

Our goal is to find a "new" metric such that for similar pair, it is close and for dissimilar pair, it is far away. In other words, in metric learning, given a set of data in two groups, we need to find the metric that would differentiate between the two groups.



Representation:
Norm induced by weighted inner product
Given $A \in \mathbb{R}^{n \times n}$ is SPD,

$$\langle x, y \rangle = x^T A y \text{ and } \|x\|_A = (x^T A x)^{\frac{1}{2}}$$

Then finding a metric is the same as finding an SPD matrix A.

Remarks:

1. The set of all SPD is **NOT** closed.
2. The closure of the set of all SPD matrices is the set of all SPSD matrices.
3. If A is SPSD, then $\|x\|_A$ is not a norm because $\|x\|_A^2 = 0 \iff x^T A x = 0$ cannot implies $x = 0$.
4. $\|\cdot\|_A$ is still a semi-norm:
   - $\|x\|_A \geq 0$
   - $\|\alpha x\|_A = \mid \alpha \mid \|x\|_A$
   - $\|x + y\|_A \leq \|x\|_A + \|y\|_A$

Evaluation:
Which A is the best?

1. For $(x_i, x_j) \in S$, $dist(x_i, x_j)$ should be small

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

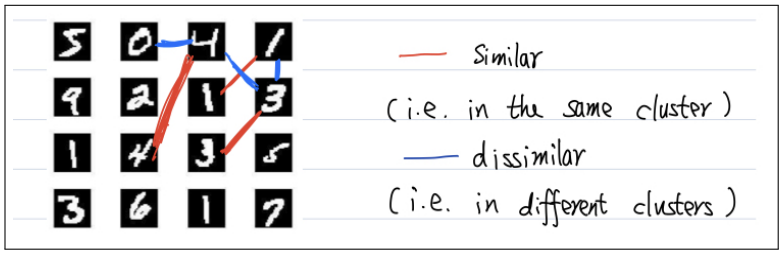
2. For $(x_i, x_j) \in D$, $dist(x_i, x_j)$ should not be small
   Altogether:

$$\begin{aligned} &\min_{A\in\mathbb{R}^{n\times n}} \sum_{(x_i,x_j)\in S} \|x_i - x_j\|_A^2 \\ &\text{s.t.} \sum_{(x_i,x_j)\in D} \|x_i - x_j\|_A^2 \geq 1 \end{aligned}$$

Optimization:
It is too complicated.

One popular application: "Supervised" clustering

- Given a data set with partial clustering information,



- Apply metric learning (i.e. find a distance)
- Cluster the points under the newly learned distance metric.

Linear and Differentiable Functions

3.1 Linear Function

Definition: Let $\mathbb{V}$ be a vector space and $f : \mathbb{V} \to \mathbb{R}$ be a function. f is a linear function if

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

Example: The mean of a vector in $\mathbb{R}^n$.

$\forall x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$, $f(x) = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$ is a linear function because

$$f(\alpha x + \beta y) = \frac{\sum_{i=1}^n (\alpha x_i + \beta y_i)}{n} = \alpha \frac{\sum_{i=1}^n x_i}{n} + \beta \frac{\sum_{i=1}^n y_i}{n} = \alpha f(x) + \beta f(y)$$

Example: The maximum entry of a vector in $\mathbb{R}^n$.

$\forall x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$, $f(x) = \underset{i=1,\cdots,n}{max}\ x_i$ is not a linear function.

One counter example:

$$x = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, y = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \alpha = 1, \beta = 1$$

$$f(\alpha x + \beta y) = f(\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}) = 1 \text{ but } \alpha f(x) = 1 \text{ and } \beta f(y) = 1$$

Hence, $f(\alpha x + \beta y) \neq \alpha f(x) + \beta f(y)$.

Example: $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \langle a, x \rangle$, where $a \in \mathbb{R}^n$ is a fixed vector in $\mathbb{R}^n$ is linear.

Example: $F : C[-1, 1] \to \mathbb{R}$ defined by $F(f) = f(0)$ is linear because

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

Example: $F : C[a, b] \to \mathbb{R}$ defined by $F(f) = \int_a^b f(t)dt$ is linear because

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

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

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

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

Example: Let $\mathbb{V}$ be an inner product space with inner product $\langle \cdot, \cdot \rangle$. Let $a \in \mathbb{V}$ and $f : \mathbb{V} \to \mathbb{R}$ defined by $f(x) = \langle a, x \rangle$ is linear.

Example: A norm function on $\mathbb{V}$ is **NOT** linear.

Proof.
Let $\| \cdot \| : \mathbb{V} \to \mathbb{R}$. Then $\| - x\| = \|x\|$ by norm property.
If $\| \cdot \|$ is linear, then

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

Properties of Linear Function:

1. *Homogeneity*:

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

$$\text{because } f(\alpha x) = f(\alpha x + 0 \cdot y) = \alpha f(x) + 0 \cdot f(y) = \alpha f(x)$$

$$\text{Choosing } \alpha = 0, \text{ then we obtain } f(0) = 0.$$

2. *Additivity*:

$$f(x + y) = f(x) + f(y)$$

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

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

$$= \alpha_1 f(x_1) + \alpha_2 f(x_2) + f(\alpha_3 x_3 + \cdots + \alpha_k x_k)$$

$$= \cdots$$

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

Linear Function on Hilbert Space
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 $\|x\|_2 = (\langle x, x \rangle)^{\frac{1}{2}}$.

- From one of the examples 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 $a \in \mathbb{R}^n$.

We are assuming that $\mathbb{H} = \mathbb{R}^n$ for simplicity, but this theorem actually holds for any forms of Hilbert Space $\mathbb{H}$.

Theorem: For any linear function $f : \mathbb{R}^n \to \mathbb{R}$, there exists a unique $a \in \mathbb{R}^n$ s.t. $f(x) = \langle a, x \rangle$, $\forall x \in \mathbb{R}^n$.

Proof.
Let e_1, e_2, .. e_n be the natural basis of $\mathbb{R}^n$ where e_i is a vector where the i-th entry is 1 and 0 elsewhere.

$$\forall x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n,\ x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n$$

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

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

$$= \langle \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \begin{bmatrix} f(e_1) \\ f(e_2) \\ \vdots \\ f(e_n) \end{bmatrix} \rangle$$

$$= \langle a, x \rangle \text{ where } a = \begin{bmatrix} f(e_1) \\ f(e_2) \\ \vdots \\ f(e_n) \end{bmatrix}$$

Now we prove the uniqueness of this theorem.
Suppose a is **NOT** unique, $\exists a, b \in \mathbb{R}^n$ s.t.

$$f(x) = \langle a, x \rangle = \langle b, x \rangle,\ \forall x \in \mathbb{R}^n$$

Then choose $x = e_i$, $i = 1, \cdots, n$

$$f(e_i) = \langle a, e_i \rangle = \langle b, e_i \rangle \implies a_i = b_i,\ i = 1, \cdots, n$$

$$\implies a = b$$

(Contradiction)

Therefore, a must be unique.

Riesz Representation Theorem

Extending previous theorem to the entirety of Hilbert space $\mathbb{H}$.

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

Example: We know that mean of a vector on $\mathbb{R}^n$ is linear.

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

Example: Let $\mathbb{H}$ be a Hilbert space and $\|\cdot\|$ is **NOT** linear. So there is no such $a \in \mathbb{H}$ s.t. $\|x\| = \langle a, x \rangle$, $\forall x \in \mathbb{H}$.

Example: $\mathbb{R}^{n \times n}$ with inner product

$$\langle A, B \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} b_{ij},\ \forall A, B \in \mathbb{R}^{n \times n}$$

Define $trace(A) = \sum_{i=1}^{n} a_{ii},\ \forall A \in \mathbb{R}^{n \times n}$ is linear. We have

$$trace(A) = \langle A, I \rangle$$

Remarks:

1. In finite dimensional Hilbert space, linear $\iff$ linear and bounded.
2. In infinite dimensional Hilbert space, there exists linear but unbounded function.

Example: $L^2(-1,1)$ - the completion of $C[-1,1]$ under the inner product $\langle f,g\rangle = \int_{-1}^{1} f(t)g(t)dt$ and $\|f\|_2 = (\langle f,f\rangle)^{\frac{1}{2}} = (\int_{-1}^{1} \mid f(t) \mid^2 dt)^{\frac{1}{2}}$. Consider $F(f) = f(0)$, $\forall f \in L^2(-1,1)$
But $F(f)$ is unbounded since

$$\exists f \in L^2(-1,1) \text{ s.t. } \mathrm{F(f)} = \infty$$

$$\text{e.g. } f(t) = \begin{cases} 1 & t \neq 0 \text{ and } t \in (-1,1) \\ \infty & t = 0 \end{cases}$$

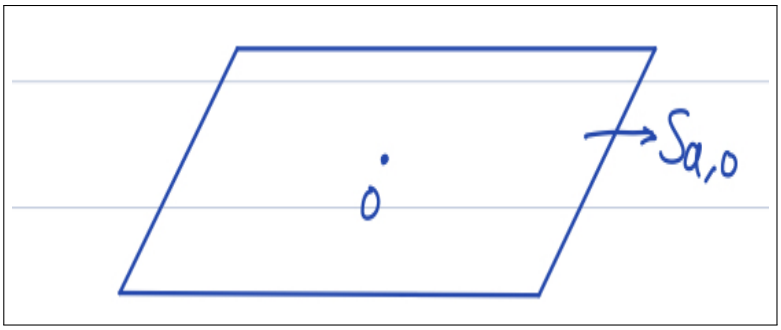
There exists no inner product representation for $F(f) = f(0)$ even though $F(f)$ is linear.

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

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

G is linear.
G is bounded because for any $f \in L^2(-1,1)$,
$G(f) = \int_{-1}^{1} f(t)dt = \int_{-1}^{1} f(t) \cdot 1dt = \langle f,1\rangle \leq \|f\|_2(\int_{-1}^{1} 1^2 dt)^{\frac{1}{2}} \leq 2\|f\|_2$
Riesz $\implies$ $g \in L^2(-1,1)$ s.t. $G(f) = \langle f,g\rangle$. Indeed, $g(t) = 1$, $\forall t \in (-1,1)$.

Hyperplane
Let $\mathbb{H}$ be a Hilbert space and $a \in \mathbb{H}$.
Consider $S_{a,0} = \{x \in \mathbb{H} | \langle a,x\rangle = 0\} \subset \mathbb{H}$,
Then $\forall \alpha, \beta \in \mathbb{R}$ and $\forall x, y \in S_{a,0}$
$\langle a, \alpha x + \beta y\rangle = \alpha\langle a,x\rangle + \beta\langle a,y\rangle = 0$. That is, $\alpha x + \beta y \in S_{a,0} \implies S_{a,0}$ is a linear space (subspace of $\mathbb{H}$).

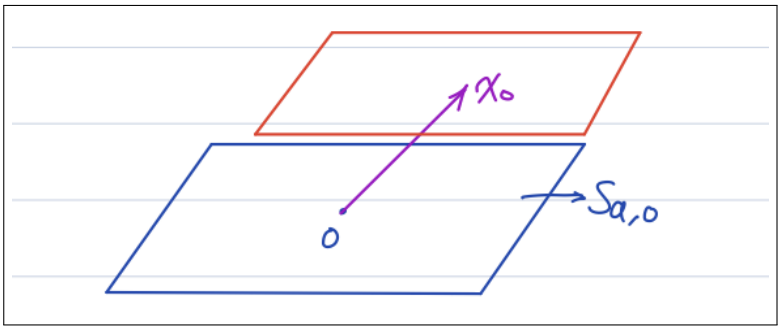


$S_{a,b} = \{x \in \mathbb{H} | \langle a, x \rangle = b\} \subset \mathbb{H}$
Let $x_0 \in S_{a,b}$, then $\langle a, x_0 \rangle = b$.

1. $\forall x \in S_{a,b}$
   $\langle a, x - x_0 \rangle = \langle a, x \rangle - \langle a, x_0 \rangle = b - b = 0$
   $\implies x - x_0 \in S_{a,0} \implies x \in x_0 + S_{a,0} \implies S_{a,b} \subset x_0 + S_{a,0}$

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

(1) and (2) $\implies S_{a,b} = S_{a,0} + x_0$
$S_{a,b}$ is the shift of subspace $S_{a,0}$.



Thus, $S_{a,b}$ is a plane on $\mathbb{H}$. So we call $S_{a,b}$ a hyperplane. Also, its co-dimension is 1 because it is defined by one linear equation.

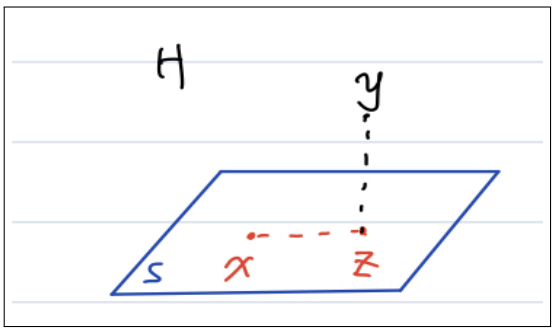
Projection Onto Hyperplane
Consider a hyperplane in $\mathbb{H}$

$$S = \{x \in \mathbb{H} | \langle a, x \rangle = b\}$$

Given $y \in \mathbb{H}$, the vector on S that is closest to y is called the projection of y onto S, denoted by $P_S y$.

$$P_S y = \underset{x \in S}{\operatorname{argmin}} \|y - x\|$$

Our goal here is to find the explicit form of $P_s y$ in terms of a, b and y.



Theorem: z is a solution of $\underset{x\in S}{\operatorname{argmin}}\|y-x\| \iff z \in S$ and $\langle z-y, x-z\rangle = 0$, $\forall x \in S$.

Remarks: Since $\forall x \in S$, $x - z \in S - z$ and $z \in S \implies x - z \in S_{a,0}$, so $\langle z-y, x-z\rangle = 0$ implies $z - y \perp S_{a,0}$.

Proof.
($\Rightarrow$) Assume z is a solution of $\underset{x\in S}{\operatorname{argmin}}\|y-x\|$, then $z \in S$.
$\forall x \in S$ and $\forall t \in \mathbb{R}$,

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

Hence, $z + t(x-z) \in S$.
Since z is a minimizer,

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

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

$$\implies 2t\langle z-y, x-z\rangle \ \ge\ -t^2\|x-z\|^2$$

- If $t > 0$, then

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

  Let $t \to 0_+$, then

$$\langle z-y, x-z\rangle \ \ge\ \lim_{t\to 0_+}(-\tfrac{t}{2}\|x-z\|^2) = 0$$

- If $t < 0$, then

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

  Let $t \to 0_-$, then

$$\langle z-y, x-z\rangle \ \le\ \lim_{t\to 0_-}(-\tfrac{t}{2}\|x-z\|^2) = 0$$

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

($\Leftarrow$) Assume $z \in S$ and $\langle z-y, x-z\rangle = 0$, $\forall x \in S$,

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

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

$$= \|x-z\|^2 + \|z-y\|^2 \ge \|z-y\|^2$$

$$\implies z = \underset{x\in S}{\operatorname{argmin}}\|x-y\|$$

Theorem: Let $\mathbb{H}$ be a Hilbert space and $a \in \mathbb{H}$. Let $b \in \mathbb{R}$ and $S = \{x \in \mathbb{H} | \langle a, x \rangle = b\}$. Given $y \in \mathbb{H}$, the solution of

$$\underset{x \in S}{\operatorname{argmin}} \|y - x\|$$

exists and is unique, which is given by

$$y - (\tfrac{\langle a,y \rangle - b}{\|a\|^2})a$$

Proof.
$z = y - (\tfrac{\langle a,y \rangle - b}{\|a\|^2})a$

Then

1. $\langle a, z \rangle = \langle a, y \rangle - \langle a, \tfrac{\langle a,y \rangle - b}{\|a\|^2} a \rangle$

   $= \langle a, y \rangle - \tfrac{\langle a,y \rangle - b}{\|a\|^2} \langle a, a \rangle$

   $= \langle a, y \rangle - (\langle a, y \rangle - b)$

   $= b \implies z \in S$

2. For any $x \in S$,
   $\langle z - y, x - z \rangle = \langle (-\tfrac{\langle a,y \rangle - b}{\|a\|^2})a, x - z \rangle$

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

   $= -\tfrac{\langle a,y \rangle - b}{\|a\|^2} (b - b)$

   $= 0$

Hence, z is a solution of $\underset{x \in S}{\operatorname{argmin}} \|y - x\|$. It remains to check the uniqueness.
Suppose it has two solutions z_1 and z_2. Then z_1,$z_2 \in S$.

$$\begin{gathered} z_1 \text{ is a solution } \implies \langle z_1 - y, z_2 - z_1 \rangle = 0 \\ z_2 \text{ is a solution } \implies \langle z_2 - y, z_1 - z_2 \rangle = 0 \implies \langle y - z_2, z_2 - z_1 \rangle = 0 \end{gathered}$$

Adding the two identities, we have

$$\begin{gathered} \langle z_1 - z_2, z_2 - z_1 \rangle = 0 \\ \iff -\|z_1 - z_2\|^2 = 0 \\ \iff z_1 = z_2 \textit{ (Contradiction)} \end{gathered}$$

In summary,
Let $\mathbb{H}$ be a Hilbert space and

$$S = \{x \in \mathbb{H} | \langle a, x \rangle = b\}$$

Let $y \in \mathbb{H}$. Then the projection of y onto S is

$$P_S y = \underset{x \in S}{\operatorname{argmin}} \|y - x\|$$

and can be given by

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

Affine Function
Definition: A linear function plus a constant is an affine function. i.e. f is affine if $f(x) = g(x) + b$ where $g : \mathbb{H} \to \mathbb{R}$ is linear and $b \in \mathbb{R}$.

Properties:

1. If $f : \mathbb{H} \to \mathbb{R}$ is affine, then for any $\alpha, \beta \in \mathbb{R}$ and $\alpha + \beta = 1$, we have

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

   To see this,

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

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

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

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

2. If $\mathbb{H}$ is a Hilbert space and f is bounded, then f is affine if and only if

   $$f(x) = \langle a, x \rangle + b \text{ for some } a \in \mathbb{H} \text{ and } b \in \mathbb{R}$$

Case Study: Regression, Classification

Regression

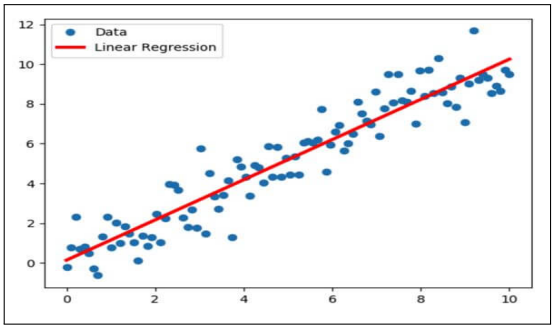
Given a set of data $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$, where $x_i \in \mathbb{R}^n$ is the input vector/independent variable and y_i is the corresponding output/dependent variable.

Given a new $x \in \mathbb{R}^n$, how do we predict the corresponding response $y \in \mathbb{R}$?

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

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

The process of finding such a function is called regression.



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 f uniquely. So, we need to find a function class ϕ (a subset of all functions) where we can search for f. Intuitively, the larger the N, the larger the function class ϕ.

How do we choose ϕ?

We choose

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

We obtain linear models. 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.

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

by minimizing the error of the linear equations.

While there are many possible definitions of error, here we are considering the square error as follows:

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

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

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

This problem is known as the least squares (LS) problem.

Write

$$X = \begin{bmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \vdots \\ x_N^T & 1 \end{bmatrix} \in \mathbb{R}^{N\times(n+1)}, \beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1} \text{ and } y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

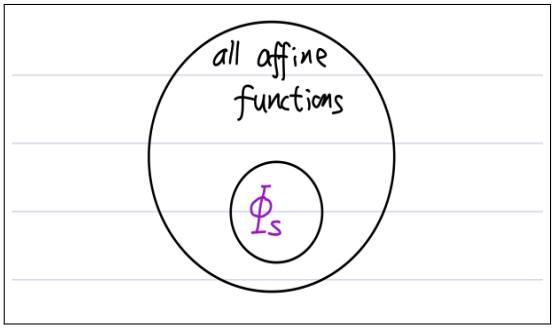
$$X\beta - y = \begin{bmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \vdots \\ x_N^T & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_1^T a + b - y_1 \\ x_2^T a + b - y_2 \\ \vdots \\ x_N^T a + b - y_N \end{bmatrix}$$

Then LS problem becomes

$$\underset{\beta\in\mathbb{R}^{n+1}}{min} \|X\beta - y\|_2^2$$

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

Regularization
In many applications, we generally don't have enough data (i.e, more unknowns than equation ($N < n+1$)) so that the class of affine functions is too large to search for f. Then we search for f in a subclass of all affine functions by regularizing f.



$\phi_s = \{f | f(x) = \langle a, x\rangle + b, \text{and } \beta = \begin{bmatrix} a \\ b \end{bmatrix} \in S\}$

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

Ridge Regression: We choose

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

So, we solve

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

$$\Longleftrightarrow$$

$$\min_{\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}} \|X\beta - y\|_2^2 + \lambda \|a\|_2^2$$

where $\lambda > 0$ is a constant depending on c and others.
Here:

$$\|X\beta - y\|_2^2 \text{ - data fitting term}$$
$$\|a\|_2^2 \text{ - regularization term}$$
$$\lambda \text{ - regularization parameter}$$

The larger the λ, the smaller ϕ_s, the looser fitting to data.
The smaller the λ, the larger ϕ_s, the better fitting to data.

LASSO Regression: We choose

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

So, we solve

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

$$\Longleftrightarrow$$

$$\min_{\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}} \|X\beta - y\|_2^2 + \lambda \|a\|_1$$

Here:

$$\|X\beta - y\|_2^2 \text{ - data fitting term}$$
$$\|a\|_1 \text{ - regularization term}$$
$$\lambda \text{ - regularization parameter}$$

Small $\|a\|_1$ tends to give a sparse vector a, i.e., many entries of β are zeros.

$$\begin{aligned} f(x) &= \langle a, x \rangle + b \\ &= \sum_{i=1}^{n} a_i x_i + b \\ &= \sum_{i \in supp(a)} a_i x_i + b \end{aligned}$$

Only x_i, $i \in supp(a)$ contribute to the prediction, where $supp(a)$ is the index of all non-zero entries of a. Since this set is small, the prediction depends only on a small portion of entries of x. The prediction is more interpretable.

Kernel Ridge Regression
To improve the linear model, we use kernel method.

1. Transform:

$$\begin{gathered}\phi : \mathbb{R}^n \to \mathbb{H} \text{ (feature mapping)} \\ x_i \to \phi(x_i) \in \mathbb{H}\end{gathered}$$

2. Linear regression is applied in $\mathbb{H}$:

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

We need to find an affine function $f : \mathbb{H} \to \mathbb{R}$ s.t.

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

Then, how do we find? In fact, we do not need an affine functions.
Let $f : \mathbb{H} \to \mathbb{R}$ be affine and bounded,

$$f(\phi(x)) = g(\phi(x)) + b = \langle \phi(x), a \rangle + b, \text{ where } a \in \mathbb{H}, b \in \mathbb{R} \ (by\ Riesz)$$

Then we define a new $\widetilde{\phi}$ and $\widetilde{\mathbb{H}} = (\mathbb{H}, \mathbb{R})$.

$$\begin{gathered}\widetilde{\phi}(x) = \begin{bmatrix}\phi(x) \\ 1\end{bmatrix} \in \widetilde{\mathbb{H}} \\ f(\phi(x)) = \langle \phi(x), a \rangle + 1 \cdot b = \langle \begin{bmatrix}\phi(x) \\ 1\end{bmatrix}, \begin{bmatrix}a \\ b\end{bmatrix} \rangle \\ = \langle \widetilde{\phi}(x), \widetilde{a} \rangle,\ \widetilde{a} = \begin{bmatrix}a \\ b\end{bmatrix} \in \widetilde{\mathbb{H}} \\ \implies f(\phi(x)) \text{ is a linear and bounded function on } \widetilde{\mathbb{H}}.\end{gathered}$$

Hence, a linear and bounded function on $\mathbb{H}$ is enough.
So, the linear regression in $\mathbb{H}$ becomes:
Find $a \in \mathbb{H}$ s.t.

$$\langle a, \phi(x_i) \rangle \approx y_i,\ i = 1, 2, \cdots, N$$

So, we solve

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

However, since $\mathbb{H}$ is very large, the set of all linear functions is also too large.
Thus, we need regularization.
So, we solve

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

This is still an infinitely dimensional problem with the need to look for explicit ϕ and a.

Representer Theorem
Definition: 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 $a \in \mathbb{H}$, 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 $i = 1, 2, \cdots, N$.

Proof of the claim:
We prove only the case when $N = 1$. Let $S = \{a \in \mathbb{H} | \langle a, \phi(x_1) \rangle = 0\}$. Then S is a hyperplane/subspace. So, $a = P_s a + (a - P_s a)$.

- For $P_s a$, since $P_s a$ is the projection,

$$\langle a - P_s a, P_s a - v \rangle = 0, \forall v \in S$$

  Since $0 \in S$, we can choose $v = 0$ and get

$$\langle a - P_s a, P_s a \rangle = 0$$

- For $a - P_s a$, by the explicit formula of $P_s a$

$$a - P_s a = \frac{\langle \phi(x_1), a \rangle}{\|\phi(x_1)\|_{\mathbb{H}}^2} \phi(x_1) \stackrel{\delta}{=} c_1 \phi(x_1)$$

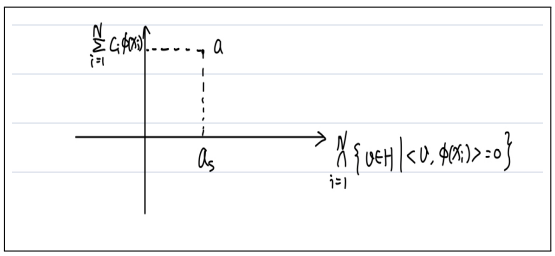
So we obtain (by setting $P_s a \stackrel{\delta}{=} a_s$)

$$a = a_s + c_1 \phi(x_1)$$

where $\langle a_s, \phi(x_1) \rangle = 0$.

For general N, we can use projection onto the intersection of

$$\{a \in \mathbb{H} | \langle a, \phi(x_i) \rangle = 0\}, \ i = 1, 2, \cdots, N$$



Therefore, $\sum_{i=1}^{N}(\langle a, \phi(x_i)\rangle - y_i)^2 + \lambda\|a\|_{\mathbb{H}}^2$
$= \sum_{i=1}^{N}(\langle a_s + \sum_{j=1}^{N} c_j\phi(x_j), \phi(x_i)\rangle - y_i)^2 + \lambda\|a_s + \sum_{j=1}^{N} c_j\phi(x_j)\|_{\mathbb{H}}^2$
$= \sum_{i=1}^{N}(\sum_{j=1}^{N} c_j\langle\phi(x_j), \phi(x_i)\rangle - y_i)^2 + \lambda(\|a_s\|_{\mathbb{H}}^2 + \|\sum_{j=1}^{N} c_j\phi(x_j)\|_{\mathbb{H}}^2)$
$= \sum_{i=1}^{N}(\sum_{j=1}^{N} c_j\langle\phi(x_j), \phi(x_i)\rangle - y_i)^2 + \lambda\|a_s\|_{\mathbb{H}}^2 + \lambda\sum_{i=1}^{N}\sum_{j=1}^{N} c_ic_j\langle\phi(x_i), \phi(x_j)\rangle$

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

$$\sum_{i=1}^{N}(\sum_{j=1}^{N} c_j\langle\phi(x_j), \phi(x_i)\rangle - y_i)^2 + \lambda\|a_s\|_{\mathbb{H}}^2 + \lambda\sum_{i=1}^{N}\sum_{j=1}^{N} c_ic_j\langle\phi(x_i), \phi(x_j)\rangle$$
$$= \|K^Tc - y\|_2^2 + \lambda c^TKc + \lambda\|a_s\|_{\mathbb{H}}^2$$

So, we solve

$$\begin{aligned} &\min_{\substack{a_s\in\mathbb{H}\\ c\in\mathbb{R}^N}} \|K^Tc - y\|_2^2 + \lambda c^TKc + \lambda\|a_s\|_{\mathbb{H}}^2 \\ &\text{s.t. } \langle a_x, \phi(x_i)\rangle = 0,\ i = 1, \cdots, N \end{aligned}$$

Because $\|a_s\|_{\mathbb{H}}^2 \geq 0$, the objective function is minimized when $\|a_s\|_H^2 = 0$, i.e., $a_s = 0$. Thus, $a \in \mathbb{H}$ is a solution only if $a = \sum_{i=1}^{N} c_i\phi(x_i)$.

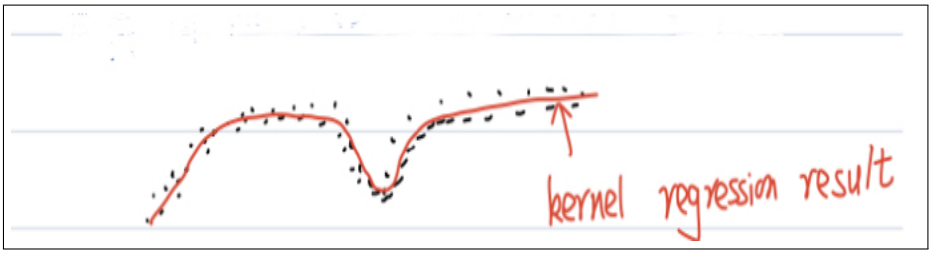
From the above proof, we see that

$$\begin{aligned} &\min_{a\in\mathbb{H}} \sum_{i=1}^{N}(\langle a, \phi(x_i)\rangle - y_i)^2 + \lambda\|a\|_{\mathbb{H}}^2 \\ &\qquad\Longleftrightarrow \\ &\underset{c\in\mathbb{R}^N}{argmin}\|K^Tc - y\|_2^2 + \lambda c^TKc \end{aligned}$$

Let the solution be $c \in \mathbb{R}^n$. Then the predicted output y for the input $x \in \mathbb{R}^N$ is

$$\begin{aligned} y &= \langle a, \phi(x)\rangle \\ &= \langle\sum_{i=1}^{N} c_i\phi(x_i), \phi(x)\rangle \\ &= \sum_{i=1}^{N} c_ik(x_i, x) \end{aligned}$$

All the computations involve only the kernel function $k(\cdot, \cdot)$. No explicit feature map $\phi(\cdot)$ is needed.



Classification
Given training data

$$(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N),\ x_i \in \mathbb{R}^n, y_i \in \{-1, +1\},\ i = 1, 2, \cdots, N,$$

find a classifier (a function) f such that

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

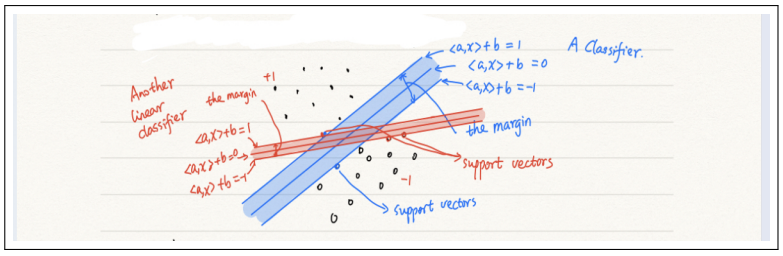
We use hyperplanes to separate the points

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

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

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

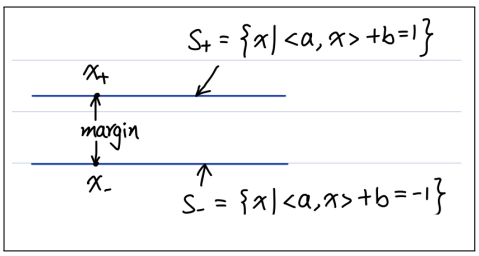
Support Vector Machine (SVM)
There exists many such linear classifiers.



Which one is better?
From the picture illustrated above, the blue one 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

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



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

$$\|x_+ - x_-\|_2 = dist(S_+, S_-)$$

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

$$x_+ = x_- - \frac{\langle a, x_-\rangle + b - 1}{\|a\|_2^2} a$$

$$= x_- - \frac{-1 - b + b - 1}{\|a\|_2^2} a \ (since\ x_- \in S_-)$$

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

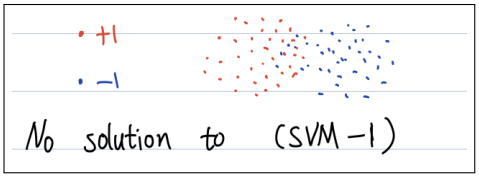
Thus, $\|x_+ - x_-\|_2 = \|\frac{2}{\|a\|_2^2} a\|_2 = \frac{2}{\|a\|_2}$, i.e., the margin is $\frac{2}{\|a\|_2}$.
So in SVM, we solve

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

Equivalently,

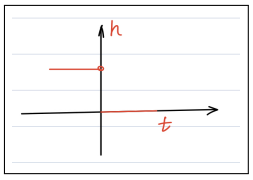
$$\begin{gathered} \underset{\substack{a \in \mathbb{R}^n \\ b \in \mathbb{R}}}{min} \|a\|_2 \\ \text{s.t. } y_i(\langle a, x_i\rangle + b) \geq 1 \end{gathered}$$

We call this SVM-1. However, it is **NOT** robust to noise.



We reformulate SVM-1 as follows.
Define a function $h : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$

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

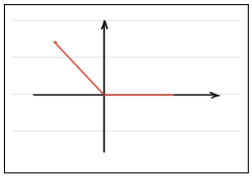


Then SVM-1 becomes

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

Now we consider a "soft" version of h.

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



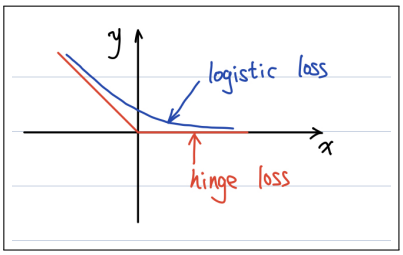
Then we obtain SVM-2

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

The error is 0 if $y_i(\langle a, x_i \rangle + b) \geq 1$, and the error is the absolute value of $y_i(\langle a, x_i \rangle + b) - 1$ if $y_i(\langle a, x_i \rangle + b) < 1$.

But the h in SVM-2 is non-smooth (**NOT** good for optimization), we approximate h by some smooth functions.
Notice that the h in SVM-2 is equivalent to

$$h(t) = log(1 + e^{-t}) \ \textit{(logistic loss)}$$



Then we obtain logistic regression.

Kernel SVM
The linear SVMs don't work for the so called "curved" data. We need to use kernel method.
Let $\phi : \mathbb{R}^n \to \mathbb{H}$ be a feature map. So, we use linear functions on $\mathbb{H}$ to classify the points. By the Riesz representation theorem, any linear function comes in the form of $\langle a, x\rangle$ for some $a \in \mathbb{H}$.
Therefore, SVM-2 becomes

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

Again, one can easily 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 $a_s \in \mathbb{H}$ and $\langle a_s, \phi(x_i)\rangle = 0$, $\forall i$. The rest is the same as the proof given for the linear regression case previously.
Thus, K-SVM becomes

$$\min_{c\in\mathbb{R}^N} \sum_{i=1}^N h(y_i(\sum_{j=1}^N k(x_i, x_j)c_j) - 1) + \lambda c^T Kc$$

where $K = [k(x_i, x_j)]_{i=1,j=1}^{N,N} \in \mathbb{R}^{N\times N}$.
The prediction of the input $x \in \mathbb{R}^n$ is given by

$$sgn(\sum_{i=1}^N k(x_i, x)c_j)$$

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

3.2 First-Order Derivative

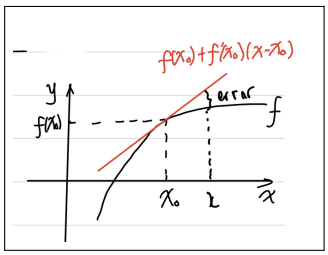
Recall that for a function $f : \mathbb{R} \to \mathbb{R}$, the derivative at x_0 is

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

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, for differentiation at x_0,

1. f is approximated by an affine function that passes through $(x_0, f(x_0))$.
2. the error of the approximation is $o(| x - x_0 |)$. (little o, i.e., $\lim_{x \to x_0} \frac{|error|}{|x-x_0|} = 0$)

This idea can be used to define differentiation of functions on the Hilbert space. (We used Hilbert space for simplicity. It can be easily adapted to Banach space as well.)

Let $f : \mathbb{H} \to \mathbb{R}$. Consider the differentiation of f at $x^{(0)} \in \mathbb{H}$.

1. By Riesz representation theorem, any affine function is in the form of $\langle v, x \rangle + a$ for some $v \in \mathbb{H}$ and $a \in \mathbb{R}$. Since it passes through $(x^{(0)}, f(x^{(0)}))$, $\langle v, x^{(0)} \rangle + a = f(x^{(0)})$. 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 error of the approximation is

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

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

$$\frac{|f(x)-f(x^{(0)})-\langle v, x-x^{(0)} \rangle|}{\|x-x_0\|} \to 0 \text{ as } x \to x^{(0)}$$

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

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

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

Example: $f(x) = \|x\|^2$, where $\|x\|$ is the norm on $\mathbb{H}$ induced by the inner product.
At $x^{(0)} \in \mathbb{H}$,
$f(x) = \|x\|^2 = \|x^{(0)} + (x - x^{(0)})\|^2 = \|x^{(0)}\|^2 + 2\langle x^{(0)}, x - x^{(0)}\rangle + \|x - x^{(0)}\|^2$
Therefore,

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

So

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

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

Example: $f(x) = \langle a, x\rangle$ for some $a \in \mathbb{H}$.
At $x^{(0)} \in \mathbb{H}$,
$f(x) = \langle a, x\rangle = \langle a, x^{(0)}\rangle + \langle a, x - x^{(0)}\rangle = f(x^{(0)}) + \langle a, x - x^{(0)}\rangle + 0$
Therefore,

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

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

Example: $f(x) = \|x - a\|^2$, where $\|\cdot\|$ is the norm on $\mathbb{H}$ and $a \in \mathbb{H}$.
At $x^{(0)} \in \mathbb{H}$,
$f(x) = \|x - a\|^2 = \|x^{(0)} - a + x - x^{(0)}\|^2$
$= \|x^{(0)} - a\|^2 + 2\langle x^{(0)} - a, x - x^{(0)}\rangle + \|x - x^{(0)}\|^2$
Therefore,

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

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

Properties of Fréchet Differentiation

1. Fréchet Differentiation is linear.
   $\forall \alpha, \beta \in \mathbb{R}, f, g : \mathbb{H} \to \mathbb{R}, \boldsymbol{\nabla}(\alpha f + \beta g)(x) = \alpha \boldsymbol{\nabla} f(x) + \beta \boldsymbol{\nabla} g(x)$, provided that $\boldsymbol{\nabla} f(x)$ and $\boldsymbol{\nabla} g(x)$ exist.

   Proof.

By definition,

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

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

$$0 \leq \lim_{\|y-x\| \to 0} \frac{|(\alpha f + \beta g)(y) - ((\alpha f + \beta g)(x) + \langle \alpha \nabla f(x) + \beta \nabla g(x), y-x \rangle)|}{\|y-x\|}$$

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

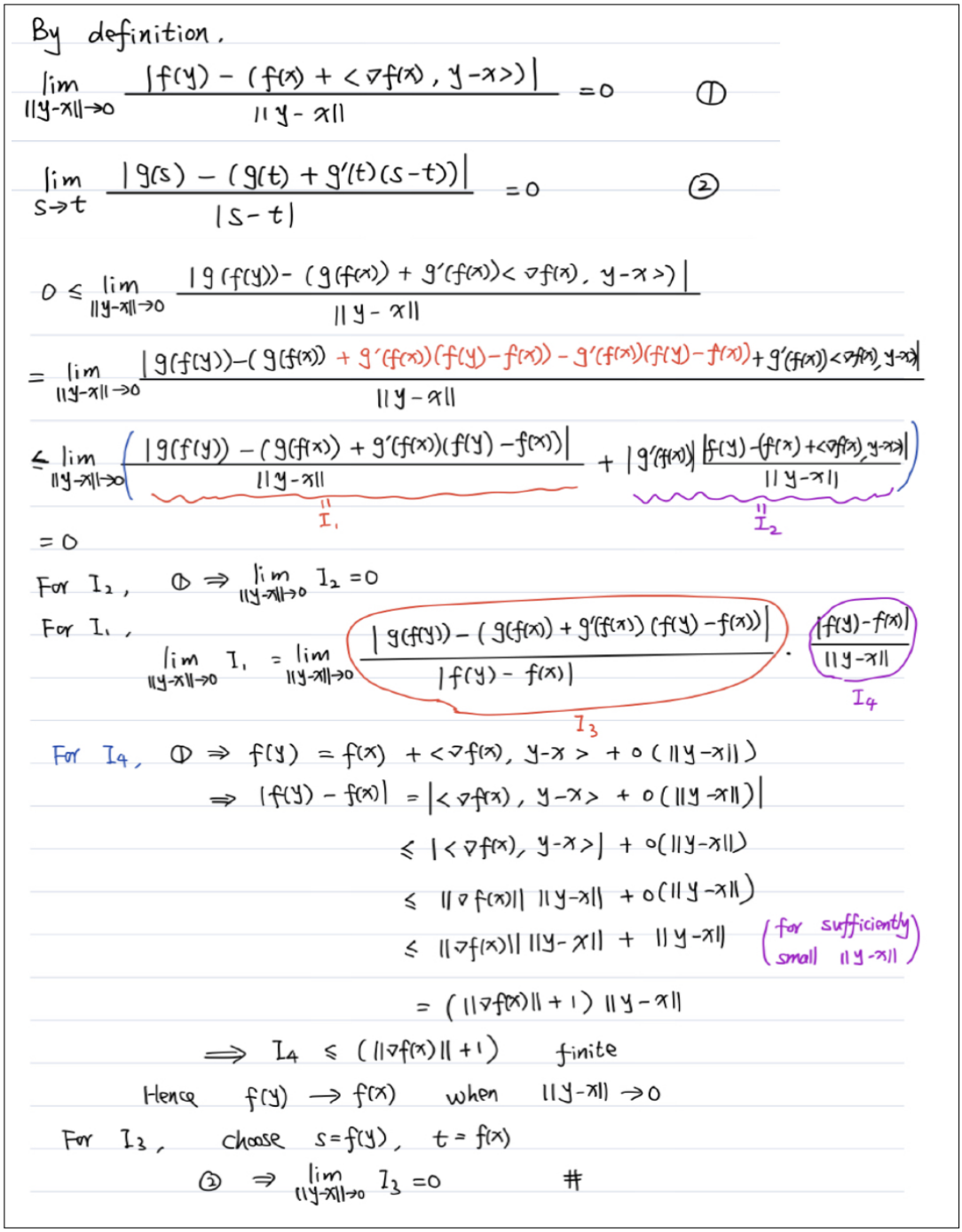
$$\leq \lim_{\|y-x\| \to 0} \frac{|\alpha| \, |f(y) - (f(x) + \langle \nabla f(x), y-x \rangle)| + |\beta| \, |g(y) - (g(x) + \langle \nabla g(x), y-x \rangle)|}{\|y-x\|}$$

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

$$= 0 \quad \#$$

2. Chain Rule
Let $f : \mathbb{H} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$. If $g \circ f : \mathbb{H} \to \mathbb{R}$, then
$\nabla(g \circ f)(x) = g'(f(x))\nabla f(x)$.

Proof.

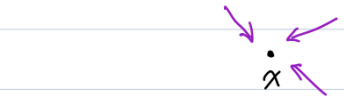


3. For functions $f : \mathbb{R}^n \to \mathbb{R}$, where $\mathbb{R}^n$ is with the standard inner product, if f is differentiable at $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$, then $\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial x}{\partial x_n}(x) \end{bmatrix}$

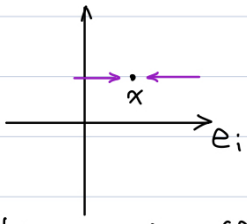
Proof.

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

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



Choose $y = x + te_i$, where $e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ ← ith entry , $t \in \mathbb{R}$



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

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

$$\Rightarrow \quad g'(0) = \langle \nabla f(x), e_i \rangle$$

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

$$\Rightarrow \quad \langle \nabla f(x), e_i \rangle = \frac{\partial f}{\partial x_i}(x)$$

$$\Rightarrow \quad [\nabla f(x)]_i = \frac{\partial f}{\partial x_i}(x) \qquad \#$$

4. Let $f : \mathbb{H} \to \mathbb{R}$ and $u \in \mathbb{R}^n$ with $\|u\| = 1$. Assume f is differentiable at $x \in \mathbb{R}^n$, then $\langle \nabla f(x), u \rangle = \frac{d}{dt} f(x + tu)|_{t=0}$. i.e., $\langle \nabla f(x), u \rangle$ is the directional derivative along the u direction.

Proof.

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

Choose $y = x + tu$, $t \in \mathbb{R}$

$$\lim_{t \to 0} \frac{|\underbrace{f(x+tu)}_{\triangleq g(t)} - (f(x) + t\langle \nabla f(x), u \rangle)|}{|t|} = 0$$

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

$$\Rightarrow g'(0) = \langle \nabla f(x), u \rangle$$

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

↑ directional derivative

Example: Let $f(x) = \|x\|$. Find its derivative for any $x \in \mathbb{H}$.
Let $f_1(x) = \|x\|^2$ for $x \in \mathbb{H}$ and $f_2(t) = \sqrt{t}$ for $t \in \mathbb{R}$. Then $f(x) = \|x\| = f_2(f_1(x))$.
When $x \neq 0$, both f_1 and f_2 are differentiable. By the chain rule,

$$\nabla f(x) = f_2'(f_1(x)) \nabla f_1(x) = \frac{x}{\|x\|}$$

When $x = 0$, f_1 is differentiable at x = 0 and f_2 is **NOT** differentiable at $f_1(x) = 0$. We cannot apply the chain rule. By using other means, we can prove f is **NOT** differentiable when x = 0.

Taylor Expansion
For functions $f : \mathbb{R} \to \mathbb{R}$

$$f(x) \approx f(x^{(0)}) + f'(x^{(0)})(x - x^{(0)})$$

Assume $f : \mathbb{H} \to \mathbb{R}$ is differentiable at $x^{(0)} \in \mathbb{H}$,

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

$$\iff$$

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

Differentiation of Functions on Banach Space
Let $\mathbb{V}$ be a Banach space. Let $f : \mathbb{V} \to \mathbb{R}$. We use affine approximation for differentiation. Let $x^{(0)} \in \mathbb{V}$. We find a function $g : \mathbb{V} \to \mathbb{R}$ s.t.

- g is affine

$$\implies g(x) = Lx + a, \text{ where } L : \mathbb{V} \to \mathbb{R} \text{ is linear, } a \in \mathbb{R}$$

- $g(x^{(0)}) = f(x^{(0)})$

$$\begin{gathered} g(x^{(0)}) = Lx^{(0)} + a = f(x^{(0)}) \\ \implies g(x) = Lx + a \\ = Lx - Lx^{(0)} + Lx^{(0)} + a \\ = L(x - x^{(0)}) + Lx^{(0)} + a \\ = L(x - x^{(0)}) + f(x^{(0)}) \end{gathered}$$

- $error = o(\|x - x^{(0)}\|)$

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

Definition: Let $\mathbb{V}$ be a Banach space and $x^{(0)} \in \mathbb{V}$. Let $f : \mathbb{V} \to \mathbb{R}$. f is differentiable at $x^{(0)}$ if there exists a linear function $L : \mathbb{V} \to \mathbb{R}$ s.t.

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

The linear function L is called the differentiation of f, denoted by $Df(x)$.

3.3 Linear Operator

Let $\mathbb{V}_1, \mathbb{V}_2$ be two vector spaces. A map $L : \mathbb{V}_1 \to \mathbb{V}_2$ is a linear transformation/linear operator if

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

where $\alpha, \beta \in \mathbb{R}$ and $x, y \in \mathbb{V}_1$.

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

Define a transformation $\mathbb{R}^n \to \mathbb{R}^m$ by $x \in \mathbb{R}^n \to Ax \in \mathbb{R}^m$, where Ax is matrix-vector product. Then it is a linear transformation because $A(\alpha x + \beta y) = \alpha Ax + \beta Ay$. Reversely, 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: Let $f : \mathbb{V} \to \mathbb{R}$ be a linear function on $\mathbb{V}$. Then f is a linear transformation from $\mathbb{V}$ to $\mathbb{R}$, as $\mathbb{R}$ is a vector space as well.

Example: Let $a \in \mathbb{R}$. Then define $L : \mathbb{R} \to \mathbb{V}$ by $L(x) = ax$. Then L is a linear transformation.

Example: Let $a_1, a_2, \cdots, a_k \in \mathbb{H}$. Then define $L : \mathbb{H} \to \mathbb{R}^k$ by

$$L(x) = \begin{bmatrix} \langle a_1, x \rangle \\ \langle a_2, x \rangle \\ \vdots \\ \langle a_k, x \rangle \end{bmatrix} \in \mathbb{R}^k$$

Then L is a linear transformation.

Example: Let $V_1 = \{f | f$ and $f'$are continuous on $[a, b]\}$ and $V_2 = \{f | f$ is continuous on $[a, b]\}$. Then define $D : V_1 \to V_2$ by

$$Df = f', \forall f \in V_1$$

Then D is a linear transformation.

Example: Consider $C[-1, 1]$ and $C[0, 2]$. Define $T : C[-1, 1] \to C[0, 2]$ by

$$T(f(t)) = f(t - 1), \forall f \in C[-1, 1], \forall t \in [0, 2]$$

Then T is a linear transformation.

Operator Norm
Consider the set of all linear operators $\mathbb{V}_1 \to \mathbb{V}_2$, where $\mathbb{V}_1$, $\mathbb{V}_2$ are two normed space. (**NOT** necessarily Banach space)

Definition: Let A, B be linear operators $\mathbb{V}_1 \to \mathbb{V}_2$ and $\alpha \in \mathbb{R}$.

- Define $A + B$ by

$$(A + B)(x) = Ax + Bx, \forall x \in \mathbb{V}_1$$

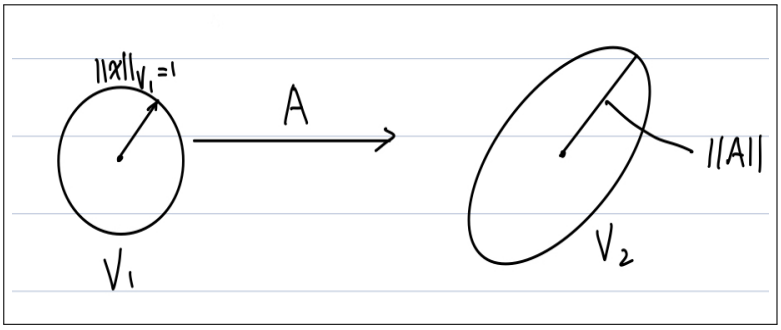
- Define αA by

$$(\alpha A)(x) = \alpha A(x), \forall x \in \mathbb{V}_1$$

Then the set of all linear operators $\mathbb{V}_1 \to \mathbb{V}_2$ is a vector space. So, we can define a norm on it.

For any linear operator $A : \mathbb{V}_1 \to \mathbb{V}_2$,

$$\|A\| = \sup_{\|x\|_{\mathbb{V}_1}=1} \|Ax\|_{\mathbb{V}_2}$$



This is indeed a norm.

Proof.

1. $\|A\| \geq 0$ is obvious

$$\|A\| = 0 \iff \sup_{\|x\|_{\mathbb{V}_1}=1} \|Ax\|_{\mathbb{V}_2} = 0$$

$$\iff \|Ax\|_{\mathbb{V}_2} = 0, \forall x \text{ satisfying } \|x\|_{\mathbb{V}_1} = 1$$

$$\iff \|A\frac{y}{\|y\|_{\mathbb{V}_1}}\|_{\mathbb{V}_2} = 0, \forall y \in \mathbb{V}_1, y \neq 0$$

$$\iff \|Ay\|_{\mathbb{V}_2} = 0, \forall y \in \mathbb{V}_1$$

$\iff A = 0$

2. $\|\alpha A\| = \sup_{\|x\|_{\mathbb{V}_1}=1} \|(\alpha A)x\|_{\mathbb{V}_2}$

   $= \sup_{\|x\|_{\mathbb{V}_1}=1} \|\alpha Ax\|_{\mathbb{V}_2} =\mid \alpha \mid \sup_{\|x\|_{\mathbb{V}_1}=1} \|Ax\|_{\mathbb{V}_2} = \alpha\|A\|$

3. $\|A + B\| = \sup_{\|x\|_{\mathbb{V}_1}=1} \|Ax + Bx\|_{\mathbb{V}_2}$

   $\leq \sup_{\|x\|_{\mathbb{V}_1}=1} \|Ax\|_{\mathbb{V}_2} + \|Bx\|_{\mathbb{V}_2} = \sup_{\|x\|_{\mathbb{V}_1}=1} \|Ax\|_{\mathbb{V}_2} + \sup_{\|x\|_{\mathbb{V}_1}=1} \|Bx\|_{\mathbb{V}_2}$

   $= \|A\| + \|B\|$

In addition, it also satisfies the following properties:

1. $\|Ax\|_{\mathbb{V}_2} \leq \|A\|\|x\|_{\mathbb{V}_1}$, $\forall x \in \mathbb{V}_1$

   Proof.
   If $x = 0$, then $0 = \|Ax\|_{\mathbb{V}_2} \leq 0 = \|A\|\|x\|_{\mathbb{V}_1}$.
   If $x \neq 0$, then

   $$\|\tfrac{x}{\|x\|_{\mathbb{V}_1}}\|_{\mathbb{V}_1} = 1 \text{ and } \|A\| = \sup_{\|z\|_{\mathbb{V}_1}=1} \|Az\|_{\mathbb{V}_2} \geq \|A\tfrac{x}{\|x\|_{\mathbb{V}_1}}\|_{\mathbb{V}_2} = \tfrac{\|Ax\|_{\mathbb{V}_2}}{\|x\|_{\mathbb{V}_1}}$$

   $\implies \|Ax\|_{\mathbb{V}_2} \leq \|A\|\|x\|_{\mathbb{V}_1}$

2. $\|AB\| \leq \|A\|\|B\|$, where $A : \mathbb{V}_2 \to \mathbb{V}_3$ and $B : \mathbb{V}_1 \to \mathbb{V}_2$ are linear

   Proof.
   $\|AB\| = \sup_{\|x\|_{\mathbb{V}_1}=1} \|ABx\|_{\mathbb{V}_3}$

   $\leq \sup_{\|x\|_{\mathbb{V}_1}=1} \|A\|\|Bx\|_{\mathbb{V}_2}$ (*by previous property*)

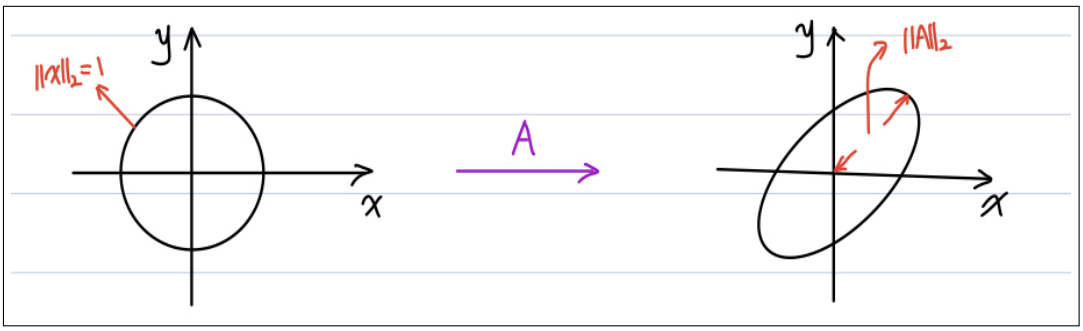
   $= \|A\| \sup_{\|x\|_{\mathbb{V}_1}=1} \|Bx\|_{\mathbb{V}_2}$

   $= \|A\|\|B\|$

Example: 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 $\|A\|_2$.

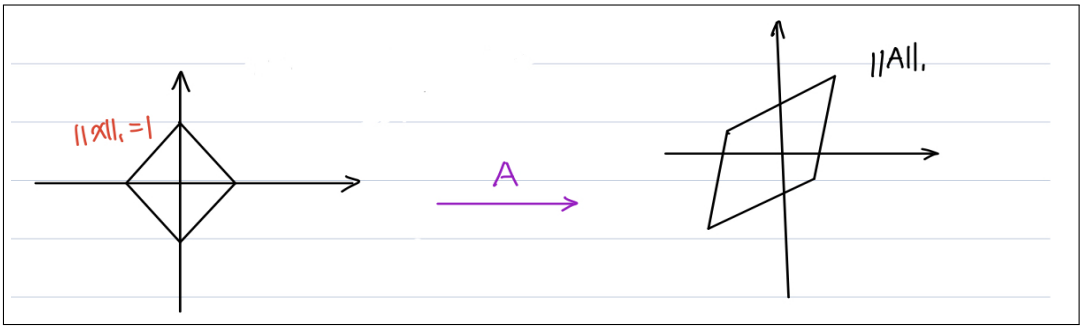
$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2 = (\sup_{\|x\|_2=1} \|Ax\|_2^2)^{\frac{1}{2}}$$

$$= (\sup_{x^Tx=1} x^TA^TAx)^{\frac{1}{2}} = (\text{max eigenvalue of } A^TA)^{\frac{1}{2}}$$



Example: Consider $A \in \mathbb{R}^{m\times n}$ as linear operators from $\mathbb{R}^n$ to $\mathbb{R}^m$. But this time $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with 1-norm. Then, for any $A \in \mathbb{R}^{m\times n}$, we call the corresponding operator norm the 1-norm, denoted by $\|A\|_1$.

$$\|A\|_1 = \sup_{\|x\|_1=1} \|Ax\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |\, a_{ij} \,| = \text{maximum absolute column sum}$$



Theorem: $\|A\|_1$ is the maximum absolute column sum.

Let $A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m\times n}$, where $a_i \in \mathbb{R}^m$ are columns of A. Then

$$\|A\|_1 = \max_{1\le i\le n} \|a_i\|_1$$

Proof.
It suffices to prove that

1. $\|A\|_1 \le \max_{1\le i\le n} \|a_i\|_1$
2. $\|A\|_1 \ge \max_{1\le i\le n} \|a_i\|_1$

For ①: For any $x \in \mathbb{R}^n$,

$$\|Ax\|_1 = \|\sum_{i=1}^n x_i a_i\|_1 \le \sum_{i=1}^n \|x_i a_i\|_1 (\textit{by triangle inequality})$$

$$= \sum_{i=1}^n (|\ x_i\ |\ \|a_i\|_1) \le (\sum_{i=1}^n |\ x_i\ |)(\max_{1\le i\le n} \|a_i\|_1)$$

$$= \|x\|_1 (\max_{1\le i\le n} \|a_i\|_1) = \max_{1\le i\le n} \|a_i\|_1$$

Taking supremum over x, we obtain

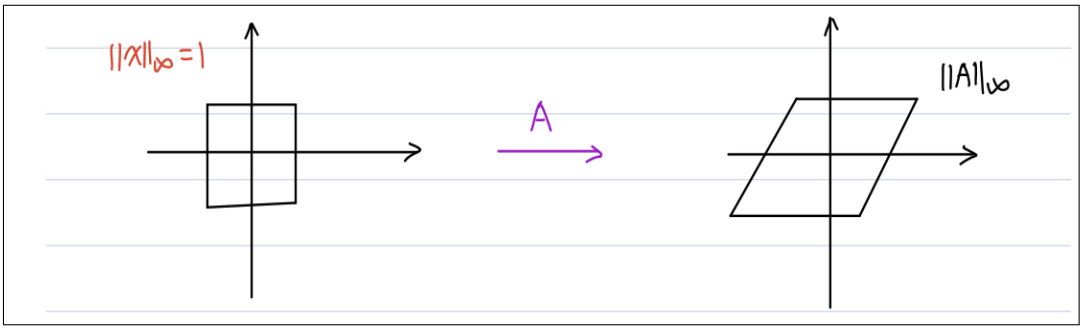
$$\|A\|_1 = \sup_{\|x\|_1=1} \|Ax\|_1 \le \max_{1\le i\le n} \|a_i\|_1$$

For ②: Let $i_o = \underset{1\le i\le n}{argmax} \|a_i\|_1$. Then

$$\max_{1\le i\le n} \|a_i\|_1 = \|a_{i_o}\|_1 = \|Ae_{i_o}\|_1 \le \sup_{\|x\|_1=1} \|Ax\|_1 = \|A\|_1$$

Example: Consider $A \in \mathbb{R}^{m\times n}$ as linear operators from $\mathbb{R}^n$ to $\mathbb{R}^m$. But $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with ∞-norm. Then, for any $A \in \mathbb{R}^{m\times n}$, we call the corresponding operator norm the ∞-norm, denoted by $\|A\|_\infty$.

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{1\le i\le m} \sum_{j=1}^n |\ a_{ij}\ | = \text{maximum absolute row sum}$$



Theorem: $\|a\|_\infty$ is the maximum absolute row sum.

More specifically, let $A = \begin{bmatrix} a_{(1)}^T \\ a_{(2)}^T \\ \vdots \\ a_{(n)}^T \end{bmatrix}$, where $a_{(i)}^T \in \mathbb{R}^{1\times n}$ is the i-th row of A. Then

$$\|A\|_\infty = \max_{1\le i\le m} \|a_{(i)}\|_1.$$

(*Proof is left as an exercise to the reader.*)

Example: Let $T : C[-1,1] \to C[0,2]$ be the translation operator defined by:

$$\forall f \in C[-1,1],\ (Tf)(t) = f(t-1),\ \forall t \in [0,2].$$

It is easy to prove that T is linear. The norm on $C[-1,1]$ is
$\|f\|_\infty = \max_{t\in[a,b]} \mid f(t) \mid$. Then

$$\|T\| = \sup_{\|f\|_\infty=1} \|Tf\|_\infty = \sup_{\max_{t\in[-1,1]}|f(t)|=1} \max_{t\in[0,2]} \mid f(t-1) \mid = 1$$

Example: Consider the differentiation operator D. Let
$V_1 = \{f | f, f' \text{ are continuous on } [0,1]\}$ and $V_2 = \{f | f \text{ is continuous on } [0,1]\}$, both equipped with the norm $\|.\|_\infty$. Then $D : V_1 \to V_2$ defined by $Df = f'$ is linear. Find $\|D\|$.

Consider $f_k(t) = sin(2\pi kt)$, $t \in [0,1]$, where $k \in \mathbb{N}$. Then
$f_k'(t) = 2\pi k cos(2\pi kt)$, $t \in [0,1]$. Therefore, $f_k(t) \in V_1$ with $\|f_k\|_\infty = 1$.

$$\|Df_k\|_\infty = \|2\pi k cos(2\pi kt)\|_\infty = 2\pi k$$

Hence,

$$\|D\| = \sup_{\|f\|_\infty=1} \|Df\|_\infty \ge \lim_{k\to\infty} \|Df_k\|_\infty \ (since\ \|f_k\|_\infty = 1)$$

$$= \lim_{k\to\infty} 2\pi k = \infty$$

Thus, $\|D\| = +\infty$.

Let $\mathbb{V}_1$, $\mathbb{V}_2$ be two normed vector spaces with $\|\cdot\|_{\mathbb{V}_1}$ and $\|\cdot\|_{\mathbb{V}_2}$.

$$\mathcal{L}(\mathbb{V}_1, \mathbb{V}_2) = \{A | A \text{ is linear } \mathbb{V}_1 \to \mathbb{V}_2 \text{ and } \|A\| < +\infty\}$$

Then $\mathcal{L}(\mathbb{V}_1, \mathbb{V}_2)$ is a normed vector space of all linear and bounded operators.

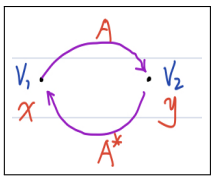
Remarks: If both $\mathbb{V}_1$ and $\mathbb{V}_2$ are Banach space, then $\mathcal{L}(\mathbb{V}_1, \mathbb{V}_2)$ is also a Banach space.
As a special case,

$$\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m) = \mathbb{R}^{m\times n}$$

where $\mathbb{R}^n$ and $\mathbb{R}^m$ can be with any norm and $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ is a Banach space.

Adjoint Operator
Let $\mathbb{V}_1$, $\mathbb{V}_2$ be two Hilbert spaces with $\langle\cdot,\cdot\rangle_{\mathbb{V}_1}$ and $\langle\cdot,\cdot\rangle_{\mathbb{V}_2}$ respectively. Let $A \in \mathcal{L}(\mathbb{V}_1, \mathbb{V}_2)$. The adjoint operator of A, denoted by A^*, is defined by

$$\langle Ax, y\rangle_{\mathbb{V}_2} = \langle x, A^*y\rangle_{\mathbb{V}_1},\ \forall x \in \mathbb{V}_1,\ y \in \mathbb{V}_2$$



Example: Consider $A \in \mathbb{R}^{m\times n} = \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. Then $\forall x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$

$$\langle Ax, y\rangle_{\mathbb{R}^m} = y^TAx = (y^TA)x = (A^Ty)^Tx = \langle x, A^Ty\rangle_{\mathbb{R}^n}$$

Definitely $A^T \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n) \implies A^* = A^T$
Therefore, the adjoint of A is its transpose A^T.

Example: Let $\mathbb{H}$ be a Hilbert space. Let $f : \mathbb{H} \to \mathbb{R}$ be linear and bounded. We have

$$f(x) = \langle a, x\rangle_{\mathbb{H}} \text{ for some } a \in \mathbb{H}$$

Then $\forall x \in \mathbb{H}$, $y \in \mathbb{R}$

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

Therefore, if we set $g(y) = ya$ for all $y \in \mathbb{R}$, then

- g is linear
  $g(\alpha y_1 + \beta y_2) = (\alpha y_1 + \beta y_2)a = \alpha y_1 a + \beta y_2 a = \alpha g(y_1) + \beta g(y_2)$
- g is bounded
  $\|g\| = \sup\limits_{|y|=1} \|g(y)\|_{\mathbb{H}} = \sup\limits_{|y|=1} \|ya\|_{\mathbb{H}} = \|a\|_{\mathbb{H}} < +\infty$

Thus, $f^*(y) = g(y) = ya$.

Let's check that $f^* \in \mathcal{L}(\mathbb{R}, \mathbb{H})$.

1. For any $y, z \in \mathbb{R}$ and $\alpha, \beta \in \mathbb{R}$

   $$f^*(\alpha y + \beta z) = (\alpha y + \beta z)a = \alpha(ya) + \beta(za) = \alpha f^*(y) + \beta f^*(z)$$

   Hence, f^* is linear.

2. $\|f^*\| = \sup\limits_{y\in\mathbb{R}} \frac{\|f^*(y)\|}{|y|} = \sup\limits_{y\in\mathbb{R}} \frac{|y|\|a\|}{|y|} = \|a\| < +\infty$

   Hence, f^* is bounded.

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

Example: Let $a_1, \cdots, a_k \in \mathbb{H}$, where $\mathbb{H}$ is a Hilbert space. Then $L : \mathbb{H} \to \mathbb{R}^k$ defined by

$$L(x) = \begin{bmatrix} \langle a_1, x \rangle \\ \langle a_2, x \rangle \\ \vdots \\ \langle a_k, x \rangle \end{bmatrix}, \ x \in \mathbb{H} \text{ is linear}$$

L's linearity is trivial.
We only show that L is bounded

$$\|L\| = \sup_{\|x\|=1} \|L(x)\| = \sup_{\|x\|=1} (\textstyle\sum_{i=1}^k (\langle a_i, x \rangle)^2)^{\frac{1}{2}}$$

$$\leq \sup_{\|x\|=1} (\textstyle\sum_{i=1}^k \|a_i\|^2 \|x\|^2)^{\frac{1}{2}} = (\textstyle\sum_{i=1}^k \|a_i\|^2)^{\frac{1}{2}} < \infty$$

Therefore, $L \in \mathcal{L}(\mathbb{H}, \mathbb{R}^k)$. Let's find L^*.
$\forall x \in \mathbb{H}, u \in \mathbb{R}^k$

$$\langle L(x), u \rangle_{\mathbb{R}^k} = \textstyle\sum_{i=1}^k \langle a_i, x \rangle_{\mathbb{H}} u_i = \sum_{i=1}^k \langle u_i a_i, x \rangle_{\mathbb{H}} = \langle x, \sum_{i=1}^k u_i a_i \rangle_{\mathbb{H}}$$

Thus $L^* : \mathbb{R}^k \to \mathbb{H}$ is defined by

$$L^*(u) = \textstyle\sum_{i=1}^k u_i a_i$$

It remains to show that $L^* \in \mathcal{L}(\mathbb{R}^k, \mathbb{H})$.

1. For any $u, v \in \mathbb{R}^k$ and $\alpha, \beta \in \mathbb{R}$

$$L^*(\alpha u + \beta v) = \textstyle\sum_{i=1}^k (\alpha u_i + \beta v_i) a_i$$

$$= \alpha \textstyle\sum_{i=1}^k u_i a_i + \beta \sum_{i=1}^k v_i a_i = \alpha L^*(u) + \beta L^*(v)$$

   Hence, L^* is linear.

2. $\|L^*\| = \sup_{\|u\|=1} \|L^*(u)\| = \sup_{\|u\|=1} \|\sum_{i=1}^k u_i a_i\|$

$$\leq \sup_{\|u\|=1} \textstyle\sum_{i=1}^k \mid u_i \mid \|a_i\| = \sup_{\|u\|=1} \langle \begin{bmatrix} \mid u_1 \mid \\ \mid u_2 \mid \\ \vdots \\ \mid u_k \mid \end{bmatrix}, \begin{bmatrix} \|a_1\| \\ \|a_2\| \\ \vdots \\ \|a_k\| \end{bmatrix} \rangle$$

$$= \sup_{\|u\|=1} \langle \widetilde{u}, \widetilde{a} \rangle \leq \sup_{\|u\|=1} \|\widetilde{u}\| \|\widetilde{a}\| = \|\widetilde{a}\| \quad (\textit{because } \|\widetilde{u}\| = \|u\|)$$

$$= (\textstyle\sum_{i=1}^k \|a_i\|^2)^{\frac{1}{2}} < +\infty$$

   Hence, L^* is bounded.

Therefore, $L^* \in \mathcal{L}(\mathbb{R}^k, \mathbb{H})$.

Example: Let $T : L^2(-1,1) \to L^2(0,2)$ be the translation operator defined by

$$\forall f \in L^2(-1,1),\ (Tf)(t) = f(t-1),\ \forall t \in (0,2)$$

Recall that on $L^2(a,b)$, $\langle f, g\rangle = \int_a^b f(t)g(t)dt$ and the norm is $\|f\|_2 = (\int_a^b \mid f(t) \mid^2 dt)^{\frac{1}{2}}$.

Then

1. T is linear
   For any $f, g \in L^2(-1,1)$ and $\alpha, \beta \in \mathbb{R}$

   $$T(\alpha f + \beta g)(t) = (\alpha f + \beta g)(t-1)$$

   $$= \alpha f(t-1) + \beta g(t-1) = \alpha T(f)(t) + \beta T(g)(t)$$

2. T is bounded
   $\|T\| = \sup_{\|f\|_2=1} \|Tf\|_2 = \sup_{\int_{-1}^1 |f(t)|^2 dt = 1} (\int_0^2 \mid f(t-1) \mid^2 dt)^{\frac{1}{2}} = 1$

Therefore, $T \in \mathcal{L}(L^2(-1,1), L^2(0,2))$.

Let's find T^*.
Let $f \in L^2(-1,1), g \in L^2(0,2)$

$$\langle Tf, g\rangle = \int_0^2 (Tf)(t)g(t)dt = \int_0^2 f(t-1)g(t)dt = \int_{-1}^1 f(s)g(s+1)ds = \langle f, \widetilde{T}g\rangle$$

where $(\widetilde{T}g)(t) = g(t+1)$, $\forall t \in (-1,1)$
We can then show that

- $\widetilde{T}$ is linear from $L^2(0,2) \to L^2(-1,1)$.
- $\|\widetilde{T}\| = 1$ (can be done similarly to $\|T\|$)

Therefore, $\widetilde{T} \in \mathcal{L}(L^2(0,2), L^2(-1,1))$ and $\langle Tf, g\rangle = \langle f, \widetilde{T}g\rangle$, $\forall f \in L^2(-1,1)$ and $g \in L^2(0,2)$. Thus, $T^* = \widetilde{T}$. In other words, the adjoint of a translation is the translation backward.

3.4 Higher-Order Derivative

Differentiation of transformation between normed spaces
Let $\mathbb{V}_1$ and $\mathbb{V}_2$ be two normed spaces with inner products $\|\cdot\|_{\mathbb{V}_1}$ and $\|\cdot\|_{\mathbb{V}_2}$. Let $F : \mathbb{V}_1 \to \mathbb{V}_2$ be a map (**NOT** necessarily linear). Then, at any point $x^{(0)} \in \mathbb{V}_1$, then the linear approximation passing through $(x^{(0)}, F(x^{(0)}))$ is

$$F(x) \approx F(x^{(0)}) + L(x - x^{(0)})$$

where $L \in \mathcal{L}(\mathbb{V}_1, \mathbb{V}_2)$.

If this approximation is $o(\|x - x^{(0)}\|_{\mathbb{V}_1})$, then L is called the differentiation of F at $x^{(0)}$.

Definition: $F : \mathbb{V}_1 \to \mathbb{V}_2$ is differentiable at $x^{(0)} \in \mathbb{V}_1$, if there exists $L \in \mathcal{L}(\mathbb{V}_1, \mathbb{V}_2)$ such that

$$\lim_{\|x-x^{(0)}\|_{\mathbb{V}_1} \to 0} \frac{\|F(x)-F(x^{(0)})-L(x-x^{(0)})\|_{\mathbb{V}_2}}{\|x-x^{(0)}\|_{\mathbb{V}_1}} = 0$$

In this case, L is called the differentiation of F at $x^{(0)}$, denoted by

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

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

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

Example: Let $A \in \mathcal{L}(\mathbb{V}_1, \mathbb{V}_2)$, then, for any $x^{(0)} \in \mathbb{V}_1$,

$$\lim_{\|x-x^{(0)}\|_{\mathbb{V}_1} \to 0} \frac{\|Ax-(Ax^{(0)}+A(x-x^{(0)}))\|_{\mathbb{V}_2}}{\|x-x^{(0)}\|_{\mathbb{V}_1}} = 0$$

Thus, $DA(x^{(0)}) = A \in \mathcal{L}(\mathbb{V}_1, \mathbb{V}_2)$.

Chain Rule
Definition: Let $F : \mathbb{V}_1 \to \mathbb{V}_2$ and $G : \mathbb{V}_2 \to \mathbb{V}_3$. Then $G \circ F : \mathbb{V}_1 \to \mathbb{V}_3$.

$$D(G \circ F)(x) = DG(F(x)) \circ DF(x)$$

where $D(G \circ F)(x) \in \mathcal{L}(\mathbb{V}_1, \mathbb{V}_3)$, $DG(F(x)) \in \mathcal{L}(\mathbb{V}_2, \mathbb{V}_3)$ and $DF(x) \in \mathcal{L}(\mathbb{V}_1, \mathbb{V}_2)$.

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

Define $F : \mathbb{V} \to \mathbb{R}^2$ by $F(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix}$, $\forall x \in \mathbb{V}$ and $G : \mathbb{R}^2 \to \mathbb{R}$ by $G(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}) = \alpha\beta$, $\forall \alpha, \beta \in \mathbb{R}$. Then $f(x) = G(F(x)) = (G \circ F)(x)$.

By the Chain Rule,

$$Df(x) = D(G \circ F)(x) = DG(F(x)) \circ DF(x) = DG(F(x))(DF(x))$$

For DG:

$$DG(\begin{bmatrix} \alpha \\ \beta \end{bmatrix})(y) = \langle \nabla G(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}), y \rangle = \langle \begin{bmatrix} \beta \\ \alpha \end{bmatrix}, y \rangle, \ \forall y \in \mathbb{R}^2$$

For DF: Since f_1, f_2 are differentiable at $x \in \mathbb{V}$,

1. $f_1(z) = f_1(x) + \langle \nabla f_1(x), z - x \rangle + o(\|z - x\|) = \epsilon_1$
2. $f_2(z) = f_2(x) + \langle \nabla f_2(x), z - x \rangle + o(\|z - x\|) = \epsilon_2$

So

$$F(z) = F(x) + (\begin{bmatrix} \langle \nabla f_1(x), z - x \rangle \\ \langle \nabla f_2(x), z - x \rangle \end{bmatrix}) + o(\|z - x\|)$$

For the blue term,

$$\| \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} \|_2 = \sqrt{\epsilon_1^2 + \epsilon_2^2} \leq \sqrt{\epsilon_1^2 + \epsilon_2^2 + 2 \mid \epsilon_1 \mid\mid \epsilon_2 \mid}$$

$$= \sqrt{(\mid \epsilon_1 \mid + \mid \epsilon_2 \mid)^2} = \mid \epsilon_1 \mid + \mid \epsilon_2 \mid \sim o(\|z - x\|)$$

Thus, $DF(x)(y) = \begin{bmatrix} \langle \nabla f_1(x), y \rangle \\ \langle \nabla f_2(x), y \rangle \end{bmatrix}$, $\forall y \in \mathbb{V}$.

Therefore,

$$Df(x)(y) = \langle \begin{bmatrix} f_2(x) \\ f_1(x) \end{bmatrix}, \begin{bmatrix} \langle \nabla f_1(x), y \rangle \\ \langle \nabla f_2(x), y \rangle \end{bmatrix} \rangle$$

$$= f_2(x)\langle \nabla f_1(x), y \rangle + f_1(x)\langle \nabla f_2(x), y \rangle$$

$$= \langle f_2(x)\nabla f_1(x) + f_1(x)\nabla f_2(x), y \rangle, \ \forall y \in \mathbb{V}$$

Then,

$$\nabla f(x) = f_2(x) \cdot \nabla f_1(x) + f_1(x) \cdot \nabla f_2(x)$$

Example: Let $A \in \mathcal{L}(\mathbb{V}, \mathbb{W})$, where $\mathbb{V}$, $\mathbb{W}$ are Hilbert spaces. Let $f : \mathbb{V} \to \mathbb{R}$ be defined by

$$f(x) = \tfrac{1}{2}\|Ax - b\|_{\mathbb{W}}^2$$

where $b \in \mathbb{W}$. Find $\boldsymbol{\nabla} f(x)$.

Note that $Df(x)(y) = \langle \boldsymbol{\nabla} f(x), y \rangle$.

$$f(y) = \tfrac{1}{2}\|Ay - b\|_{\mathbb{W}}^2 = \tfrac{1}{2}\|(Ax - b) + (Ay - Ax)\|_{\mathbb{W}}^2$$

$$= \tfrac{1}{2}\|Ax - b\|_{\mathbb{W}}^2 + \langle Ax - b, A(y - x)\rangle_{\mathbb{W}} + \tfrac{1}{2}\|A(y - x)\|_{\mathbb{W}}^2$$

$$= f(x) + \langle A^*(Ax - b), y - x\rangle_{\mathbb{V}} + \tfrac{1}{2}\|A(y - x)\|_{\mathbb{W}}^2$$

$$0 \leq \lim_{\|y-x\|_{\mathbb{V}} \to 0} \frac{\frac{1}{2}\|A(y-x)\|_{\mathbb{W}}^2}{\|y-x\|_{\mathbb{V}}} \leq \lim_{\|y-x\|_{\mathbb{V}} \to 0} \frac{\frac{1}{2}(\|A\|\|y-x\|_{\mathbb{V}})^2}{\|y-x\|_{\mathbb{V}}} = 0$$

$$\implies \boldsymbol{\nabla} f(x) = A^*(Ax - b) \text{ or } Df(x)(y) = \langle A^*(Ax - b), y\rangle$$

In particular, if $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, then $\mathbb{V} = \mathbb{R}^n, \mathbb{W} = \mathbb{R}^m, A^* = A^T$,

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

$$\boldsymbol{\nabla} f(x) = A^T(Ax - b)$$

More rules on differentiation

1. **Linearity**
   Let $F, G : \mathbb{V}_1 \to \mathbb{V}_2$

   $$D(\alpha F + \beta G)(x) = \alpha DF(x) + \beta DG(x)$$

   where $x \in \mathbb{V}_1, \alpha, \beta \in \mathbb{R}$.

2. **Chain Rule**
   Let $F : \mathbb{V}_1 \to \mathbb{V}_2$, $G : \mathbb{V}_2 \to \mathbb{V}_3$

   $$D(G \circ F) = DG \circ DF$$

   $\forall x, y \in \mathbb{V}_1$

   $$D(G \circ F)(x)(y) = DG(F(x))(DF(x)(y))$$

   where $D(G \circ F)(x) \in \mathcal{L}(\mathbb{V}_1, \mathbb{V}_3)$, $DG(F(x)) \in \mathcal{L}(\mathbb{V}_2, \mathbb{V}_3)$ and $DF(x) \in \mathcal{L}(\mathbb{V}_1, \mathbb{V}_2)$.

   In particular:

   - If $\mathbb{V}_2 = \mathbb{V}_3 = \mathbb{R}$, then $F : \mathbb{V}_1 \to \mathbb{R}$, $G : \mathbb{R} \to \mathbb{R}$, then $G \circ F : \mathbb{V}_1 \to \mathbb{R}$

     $$D(G \circ F)(x)(y) = DG(F(x))(DF(x)(y))$$

$$DG(F(x))(\alpha) = G'(F(x))\alpha$$

$$DF(x)(y) = \langle \boldsymbol{\nabla} F(x), y \rangle, \forall x, y \in \mathbb{V}_1$$

Hence
$D(G \circ F)(x)(y) = G'(F(x))\langle \boldsymbol{\nabla} F(x), y \rangle = \langle G'(F(x))\boldsymbol{\nabla} F(x), y \rangle$
$\implies \boldsymbol{\nabla}(G \circ F)(x) = G'(F(x))\boldsymbol{\nabla} F(x)$

- If $\mathbb{V}_3 = \mathbb{R}$, then $F : \mathbb{V}_1 \to \mathbb{V}_2$, $G : \mathbb{V}_2 \to \mathbb{R}$, where $\mathbb{V}_1$, $\mathbb{V}_2$ are Hilbert spaces, then $G \circ F : \mathbb{V}_1 \to \mathbb{R}$

$$D(G \circ F)(x)(y) = DG(F(x))(DF(x)(y))$$

$$DG(F(x))(z) = \langle \boldsymbol{\nabla} G(F(x)), z \rangle$$

$$DG(F(x))(DF(x)(y)) = \langle \boldsymbol{\nabla} G(F(x)), DF(x)(y) \rangle$$

$$= \langle (DF(x))^* \boldsymbol{\nabla} G(F(x)), y \rangle, \forall x, y \in \mathbb{V}_1$$

$\implies \boldsymbol{\nabla}(G \circ F)(x) = (DF(x))^* \boldsymbol{\nabla} G(F(x))$

-If G is linear and bounded, then $\exists g \in \mathbb{V}_2$, s.t. $G(z) = \langle g, z \rangle$, $z \in \mathbb{V}_2$, and $\boldsymbol{\nabla} G(x) = g$, so $\forall x \in \mathbb{V}_2$, $\boldsymbol{\nabla}(G \circ F)(x) = (DF(x))^* g$.

-If F is linear and bounded,

$$F \in \mathcal{L}(\mathbb{V}_1, \mathbb{V}_2)$$

$$DF(x) = F, \forall x \in \mathbb{V}_1$$

so $\boldsymbol{\nabla}(G \circ F)(x) = F^* \boldsymbol{\nabla} G(F(x))$.

3. **Jacobian Matrix**
Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a vector-valued function. Assume that $\mathbb{R}^m$, $\mathbb{R}^n$ are endowed with the standard inner product. Then

$$DF(x) = (\frac{\partial (F(x))_i}{\partial x_j})_{i=1,j=1}^{m,n} \in \mathbb{R}^{m \times n}$$

where $(F(x))_i$ is the i-th component of $F(x) \in \mathbb{R}^m$.

Proof. Denote $(F(x))_i = f_i(x) \in \mathbb{R}$, i.e., $F(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix} \in \mathbb{R}^m$. By differentiability,

$f_i(y) = f_i(x) + \langle \nabla f_i(x), y-x \rangle + \varepsilon_i$ — the error

where $\varepsilon_i \sim o(\|y-x\|_2)$, i.e.

$$\lim_{\|y-x\|_2 \to 0} \frac{|\varepsilon_i|}{\|y-x\|_2} = 0 \quad \text{for} \quad i = 1, \cdots, m$$

Therefore,

$$F(y) = F(x) + \begin{pmatrix} \langle \nabla f_1(x), y-x \rangle \\ \langle \nabla f_2(x), y-x \rangle \\ \vdots \\ \langle \nabla f_m(x), y-x \rangle \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix} \equiv \mathcal{E} \in \mathbb{R}^m$$

Since

$$\|\varepsilon\|_2 \le \|\varepsilon\|_1 = |\varepsilon_1| + |\varepsilon_2| + |\varepsilon_3| + |\varepsilon_4| + \cdots + |\varepsilon_m|$$

$$0 \le \lim_{\|y-x\|_2 \to 0} \frac{\|\varepsilon\|_2}{\|y-x\|_2} \le \lim_{\|y-x\|_2 \to 0} \frac{|\varepsilon_1| + |\varepsilon_2| + \cdots + |\varepsilon_m|}{\|y-x\|_2} = 0$$

Therefore,

$$DF(x)(y) = \begin{pmatrix} \langle \nabla f_1(x), y \rangle \\ \langle \nabla f_2(x), y \rangle \\ \vdots \\ \langle \nabla f_m(x), y \rangle \end{pmatrix} = \begin{pmatrix} (\nabla f_1(x))^T y \\ (\nabla f_2(x))^T y \\ \vdots \\ (\nabla f_m(x))^T y \end{pmatrix}$$

$$= \begin{pmatrix} (\nabla f_1(x))^T \\ (\nabla f_2(x))^T \\ \vdots \\ (\nabla f_m(x))^T \end{pmatrix} y \equiv DF(x) \in \mathbb{R}^{m \times n}$$

Remarks: $DF(x)$ is known as the Jacobian matrix. From the proof, we can see that differentiation is an extension of the Jacobian matrix.

4. **Multiplication Rule**

- Number multiplication rule (from the example)
  Let $f_1, f_2 : \mathbb{V} \to \mathbb{R}$, where $\mathbb{V}$ is a Banach space, then
  $$D(f_1 f_2)(x)(y) = Df_1(x)(y) \cdot f_2(x) + f_1(x) \cdot Df_2(x)(y)$$
  In particular, if $\mathbb{V}$ is a Hilbert space, then
  $$\boldsymbol{\nabla}(f_1 f_2)(x) = f_2(x) \cdot \boldsymbol{\nabla} f_1(x) + f_1(x) \cdot \boldsymbol{\nabla} f_2(x)$$
- Scalar multiplication rule
  Let $f : \mathbb{V} \to \mathbb{R}$, $F : \mathbb{V} \to \mathbb{V}$. Define $fF : \mathbb{V} \to \mathbb{V}$ by
  $(fF)(x) = f(x) \cdot F(x)$, $\forall x \in \mathbb{V}$.
  $$D(fF)(x)(y) = f(x) \cdot DF(x)(y) + f'(x) \cdot F(x)(y)$$

Proof. This formula can be obtained by a similar argument in the example.

Both number and scalar multiplications are bilinear, i.e. $x \cdot y$ is linear in x (resp. y) if y (resp. x) is fixed.

Let $\widetilde{F}(x,y)$ be bilinear, where x, y are two vectors (not necessarily in the same vector space.)

For any x_0, y_0,

$$\widetilde{F}(x,y) = \widetilde{F}(x_0, y_0) + \widetilde{F}(x_0, y-y_0) + \widetilde{F}(x-x_0, y_0) + \widetilde{F}(x-x_0, y-y_0)$$

(This can be checked by bilinearity.)

Since $\widetilde{F}(x_0, y-y_0) + \widetilde{F}(x-x_0, y_0)$ is linear in $(x-x_0, y-y_0)$, and if we assume,

$$\widetilde{F}(x-x_0, y-y_0) = o\left(\sqrt{\|x-x_0\|^2 + \|y-y_0\|^2}\right)$$

(This is true for $F(x,y)=xy$)

We have $D\widetilde{F}(x_0,y_0)(u,v) = \widetilde{F}(x_0, v) + \widetilde{F}(u, y_0)$

Let $G(x) = (f(x), F(x))$ Then $\widetilde{F} \circ G(x) = f(x) \cdot F(x) = (fF)(x)$

Therefore, by the chain rule,

$$\begin{aligned} D(fF)(x)(y) &= D\widetilde{F}(G(x))\left(DG(x)(y)\right) \\ &= D\widetilde{F}(G(x))\left(Df(x)(y), DF(x)(y)\right) \\ &= D\widetilde{F}\left(f(x), F(x)\right)\left(Df(x)(y), DF(x)(y)\right) \\ &= \widetilde{F}\left(f(x), DF(x)(y)\right) + \widetilde{F}\left(Df(x)(y), F(x)\right) \\ &= f(x) \cdot DF(x)(y) + Df(x)(y) \cdot F(x) \end{aligned}$$

5. **Matrix Multiplication Rule**
   Let $F : \mathbb{V} \to \mathbb{R}^{m\times n}$ and $G : \mathbb{V} \to \mathbb{R}^{n\times p}$. Then we consider $F \cdot G : \mathbb{V} \to \mathbb{R}^{m\times p}$ defined by

$$(F \cdot G)(x) = F(x)G(x),\ \forall x \in \mathbb{V}$$

   Since $F \cdot G$ is bilinear in F, G and $\|(x - x_0)(y - y_0)\| \leq \|x - x_0\|\|y - y_0\| \sim o(\sqrt{\|x - x_0\|^2 + \|y - y_0\|^2})$, we have

$$D(F \cdot G)(x)(y) = DF(x)(y)G(x) + F(x)DG(x)(y)$$

   where $D(F \cdot G)(x)(y) \in \mathbb{R}^{m\times p}$, $DF(x)(y) \in \mathbb{R}^{m\times n}$, $DG(x)(y) \in \mathbb{R}^{n\times p}$.

6. **Inner Product**
   Let $F : \mathbb{V} \to \mathbb{W}$ and $G : \mathbb{V} \to \mathbb{W}$. Then we consider $\langle F, G\rangle : \mathbb{V} \to \mathbb{R}$ ($\mathbb{V}$, $\mathbb{W}$ are Hilbert spaces) defined by

$$\langle F, G\rangle(x) = \langle F(x), G(x)\rangle_{\mathbb{W}},\ \forall x \in \mathbb{V}$$

   Again, since $\langle F, G\rangle$ is bilinear in F, G and $|\ \langle x - x_0, y - y_0\rangle\ | \leq \|x - x_0\|\|y - y_0\| \sim o(\sqrt{\|x - x_0\|^2 + \|y - y_0\|^2})$, we have

$$D(\langle F, G\rangle)(x)(y) = \langle DF(x)(y), G(x)\rangle_{\mathbb{W}} + \langle F(x), DG(x)(y)\rangle_{\mathbb{W}}$$

$$= \langle (DF(x))^* G(x), y\rangle_{\mathbb{V}} + \langle (DG(x))^* F(x), y\rangle_{\mathbb{V}}$$

$$= \langle (DF(x))^* G(x) + (DG(x))^* F(x), y\rangle_{\mathbb{V}}$$

   Therefore, $D(\langle F, G\rangle)(x) = (DF(x))^* G(x) + (DG(x))^* F(x)$.

   Remarks: Besides number, scalar, matrix multiplication and inner product, there are other bilinear mappings satisfying a similar differentiation rule. e.g. convolution.

Hessian of Functions on Hilbert Space
Let $f : \mathbb{V} \to \mathbb{R}$, where $\mathbb{V}$ is a Hilbert space. For first-order derivative, it is $\boldsymbol{\nabla} f(x) \in \mathbb{V}$. But for second-order derivative, we can view $x \to \boldsymbol{\nabla} f(x)$ as a mapping from $\mathbb{V} \to \mathbb{V}$. Then $D(\boldsymbol{\nabla} f)(x)$ is a second-order derivative.

Definition: The Hessian of f is

$$\boldsymbol{\nabla}^2 f(x) := D(\boldsymbol{\nabla} f)(x)$$

In particular, $f : \mathbb{R}^n \to \mathbb{R}$, then

$$\boldsymbol{\nabla}^2 f(x) = D(\boldsymbol{\nabla} f)(x) = [\tfrac{\partial^2 f}{\partial x_i \partial x_j}]_{i=1,j=1}^{n,n} \in \mathbb{R}^{n \times n}$$

$$(\boldsymbol{\nabla} f(x))_{x_i} = \tfrac{\partial f}{\partial x_i}(x)$$

Example: $f(x) = \frac{1}{2}\|Ax - b\|_{\mathbb{W}}^2$, where $A \in \mathcal{L}(\mathbb{V}, \mathbb{W})$, $b \in \mathbb{W}$, $\|\cdot\|_{\mathbb{W}}$, $x \in \mathbb{V}$.

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

1. $A \in \mathcal{L}(\mathbb{V}, \mathbb{W})$
2. $A^* \in \mathcal{L}(\mathbb{W}, \mathbb{V})$

① and ② $\implies$ $A^*A \in \mathcal{L}(\mathbb{V}, \mathbb{V})$

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

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

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

Then $\boldsymbol{\nabla} f(x) = A^T(Ax - b)$ and $\boldsymbol{\nabla}^2 f(x) = A^TA$.

Example: $f(x) = F(Ax)$, where $A \in \mathcal{L}(\mathbb{V}, \mathbb{W})$ and $F : \mathbb{W} \to \mathbb{R}$.

By the chain rule,

$$\boldsymbol{\nabla} f(x) = A^* \boldsymbol{\nabla} F(Ax)$$

For the Hessian,

$$D(\boldsymbol{\nabla} f)(x)(y) = D(A^* \boldsymbol{\nabla} F(Ax))(y),\ \forall y \in \mathbb{V}$$

$$= DA^*(\boldsymbol{\nabla} F(Ax))(D(\boldsymbol{\nabla} F)(Ax))(D(Ax)(y))$$

$A^* \in \mathcal{L}(\mathbb{W}, \mathbb{V})$ $\implies$ $DA^*(\boldsymbol{\nabla} F(Ax)) = A^*$

$A \in \mathcal{L}(\mathbb{V}, \mathbb{W})$ $\implies$ $D(Ax) = A$

$D(\boldsymbol{\nabla} F)(Ax) = \boldsymbol{\nabla}^2 F(Ax)$

Hence, $D(\boldsymbol{\nabla} f)(x)(y) = A^* \boldsymbol{\nabla}^2 F(Ax) A$.

Example: $f(x) = \sum_{i=1}^{m} f_i(\langle a_i, x\rangle)$ where $a_i \in \mathbb{R}^n$, $x \in \mathbb{R}^n$, $f_i : \mathbb{R} \to \mathbb{R}$.

Let $A = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_m^T \end{bmatrix} \in \mathbb{R}^{m\times n}$, $F(y) = \sum_{i=1}^{m} f_i(y_i)$, $\forall y \in \mathbb{R}^m$, then $f(x) = F(Ax)$.

Since $F : \mathbb{R}^n \to \mathbb{R}$,

$$\boldsymbol{\nabla} F(y) = \begin{bmatrix} \frac{\partial F}{\partial y_1}(y) \\ \frac{\partial F}{\partial y_2}(y) \\ \vdots \\ \frac{\partial F}{\partial y_m}(y) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial y_1}(\sum_{i=1}^{m} f_i(y_i)) \\ \frac{\partial}{\partial y_2}(\sum_{i=1}^{m} f_i(y_i)) \\ \vdots \\ \frac{\partial}{\partial y_m}(\sum_{i=1}^{m} f_i(y_i)) \end{bmatrix} = \begin{bmatrix} f_1^{'}(y_1) \\ f_2^{'}(y_2) \\ \vdots \\ f_m^{'}(y_m) \end{bmatrix}$$

Therefore,

$$\boldsymbol{\nabla} f(x) = A^T \begin{bmatrix} f_1^{'}(a_1^T x) \\ f_2^{'}(a_2^T x) \\ \vdots \\ f_m^{'}(a_m^T x) \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \cdots & a_m \end{bmatrix} \begin{bmatrix} f_1^{'}(a_1^T x) \\ f_2^{'}(a_2^T x) \\ \vdots \\ f_m^{'}(a_m^T x) \end{bmatrix} = \sum_{i=1}^{m} f_i^{'}(a_i^T x) a_i$$

For the Hessian, we need

$$\boldsymbol{\nabla}^2 F(y) = [\tfrac{\partial^2 F}{\partial y_i \partial y_j}]_{i=1,j=1}^{m} = \begin{bmatrix} f_1^{''}(y_1) & & & \\ & f_2^{''}(y_2) & & \\ & & \ddots & \\ & & & f_m^{''}(y_m) \end{bmatrix}$$

Therefore,

$$\boldsymbol{\nabla}^2 f(x) = A^T \begin{bmatrix} f_1^{''}(a_1^T x) & & & \\ & f_2^{''}(a_2^T x) & & \\ & & \ddots & \\ & & & f_m^{''}(a_m^T x) \end{bmatrix} A$$

$$= \begin{bmatrix} a_1 & a_2 & \cdots & a_m \end{bmatrix} \begin{bmatrix} f_1^{''}(a_1^T x) & & & \\ & f_2^{''}(a_2^T x) & & \\ & & \ddots & \\ & & & f_m^{''}(a_m^T x) \end{bmatrix} \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_m^T \end{bmatrix}$$

$$= \sum_{i=1}^{m} f_i^{''}(a_i^T x) a_i a_i^T$$

Example: $f(x) = \|x\|$, where $x \in \mathbb{V}$ with $\mathbb{V}$ being a Hilbert space.

As we know that when $x \neq 0$,

$$\boldsymbol{\nabla} f(x) = \frac{x}{\|x\|} = \frac{1}{\|x\|}x$$

Let $g : \mathbb{V} \to \mathbb{R}$ by $g(x) = \frac{1}{\|x\|}$ and $I : \mathbb{V} \to \mathbb{V}$ by $I(x) = x$. Then $\boldsymbol{\nabla} f(x) = gI$. Since I is linear, $DI(x) = I$.

For $\boldsymbol{\nabla} g$:

Let $g_1(t) = \frac{1}{\sqrt{t}} : \mathbb{R} \to \mathbb{R}$ and $g_2(t) = \|x\|^2 : \mathbb{V} \to \mathbb{R}$. Then $g(x) = g_1(g_2(x))$, i.e. $g = g_1 \circ g_2$.

$$\boldsymbol{\nabla} g(x) = g_1^{'}(g_2(x))\boldsymbol{\nabla} g_2(x) = -\frac{1}{2}(\|x\|^2)^{\frac{3}{2}} \cdot 2x = -\frac{x}{\|x\|^3}$$

$$D(\boldsymbol{\nabla} f)(x)(y) = D(gI)(x)(y)$$

$$= Dg(x)(y) \cdot I(x) + g(x) \cdot DI(x)(y) = \langle -\frac{x}{\|x\|^3}, y\rangle x + \frac{1}{\|x\|}y$$

So $\boldsymbol{\nabla}^2 f(x)(y) = \langle -\frac{x}{\|x\|^3}, y\rangle x + \frac{y}{\|x\|}$.

In particular, when $\mathbb{V} = \mathbb{R}^n$,

$$\boldsymbol{\nabla}^2 f(x)(y) = -\frac{xx^T y}{\|x\|^3} + \frac{Iy}{\|x\|} = (\frac{1}{\|x\|}I - \frac{1}{\|x\|^3}xx^T)y$$

and hence

$$\boldsymbol{\nabla}^2 f(x) = \frac{1}{\|x\|}I - \frac{1}{\|x\|^3}xx^T \in \mathbb{R}^{n\times n}$$

Function Expansion
Let $f : \mathbb{V} \to \mathbb{R}$ be a differentiable function, where $\mathbb{V}$ is a Hilbert space.

From the definition of the gradient,

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

Now we derive the expansion up to second-order derivative.

For this purpose, we consider

$$f(x + tu), \text{ where } x, u \in \mathbb{V} \text{ are given and } t \in \mathbb{R}$$

Then $g(t) \equiv f(x + tu)$ is a function $\mathbb{R} \to \mathbb{R}$.

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

This is derived by the chain rule:

$$G(t) = x + tu,\ G : \mathbb{R} \to \mathbb{V} \text{ and } f(G(t)) = f(x + tu)$$

Since $\nabla G(t)(s) = u$ and $(DG(t))^* v = \langle v, u \rangle$,

$$\frac{d}{dt} f(G(t)) = (DG(t))^* \nabla f(G(t)) = \langle \nabla f(G(t)), u \rangle$$

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

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

Similarly,

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

i.e. $\langle \nabla^2 f(x) u, u \rangle$ is the second-order derivative of $f(x)$ along u.

We can similarly show that

$$\frac{\partial^2}{\partial t_1 \partial t_2} f(x + t_1 u + t_2 v)|_{t_1 = t_2 = 0} = \frac{\partial^2}{\partial t_2 \partial t_1} f(x + t_1 u + t_2 v)|_{t_1 = t_2 = 0}$$

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

Hence, $\nabla^2 f(x) = (\nabla^2 f(x))^*$. (Hessian is self-adjoint.)

Now we present the second-order expansion of $f : \mathbb{V} \to \mathbb{R}$.
Let $g(t) = f(x^{(0)} + tu), x^{(0)}, u \in \mathbb{V}, t \in \mathbb{R}$, then by Taylor's expansion on $g(t)$,

$$g(t) = g(t_0) + g'(t_0)(t - t_0) + \tfrac{1}{2}g''(t_0)(t - t_0)^2 + o(\mid t - t_0 \mid^2)$$

which is equivalent to

$$f(x^{(0)} + tu) = f(x^{(0)} + t_0 u) + \tfrac{d}{dt} f(x^{(0)} + tu)|_{t=t_0}(t - t_0$$

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

From previous derivation, we have $\frac{d}{dt} f(x^{(0)} + tu)|_{t=t_0} = \langle \boldsymbol{\nabla} f(x^{(0)} + t_0 u), u \rangle$ and $\frac{d^2}{dt^2} f(x^{(0)} + tu)|_{t=t_0} = \langle \boldsymbol{\nabla}^2 f(x^{(0)} + t_0 u)u, u \rangle$.

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

$$x^{(0)} + tu = x^{(0)} + \|x - x^{(0)}\| \tfrac{x - x^{(0)}}{\|x - x^{(0)}\|} = x$$

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

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

This is the second-order expansion of $f : \mathbb{V} \to \mathbb{R}$. Many algorithms and theories are based on this expansion.

Newton–Raphson method (aka Newton's method)
Consider

$$\underset{x\in\mathbb{V}}{min} f(x)$$

Given $x^{(k)}$, we find a better $x^{(k+1)}$ by

- expanding $f(x)$ around $x^{(k)}$ as

  $$f(x) \approx f(x^{(k)}) + \langle \boldsymbol{\nabla} f(x^{(k)}), x - x^{(k)} \rangle + \tfrac{1}{2} \langle \boldsymbol{\nabla}^2 f(x^{(k)})(x - x^{(k)}), x - x^{(k)} \rangle$$

  We denote this approximation by $F_k(x)$.

- Instead of minimizing $f(x)$, we minimize its approximation $F_k(x)$ to obtain $x^{(k+1)}$.

  $$x^{(k+1)} = \underset{x\in\mathbb{V}}{argmin} F_k(x)$$

  By 1-st order optimality,

  $$\boldsymbol{\nabla} F_k(x^{(k+1)}) = 0$$

  Then by direct calculation,

  $$\boldsymbol{\nabla} F_k(x) = \boldsymbol{\nabla} f(x^{(k)}) + \boldsymbol{\nabla}^2 f(x^{(k)})(x - x^{(k)})$$

  So $\boldsymbol{\nabla} f(x^{(k)}) + \boldsymbol{\nabla}^2 f(x^{(k)})(x^{(k+1)} - x^{(k)}) = 0$.

Then we just derive the Newton's method

$$x^{(k+1)} = x^{(k)} - (\boldsymbol{\nabla}^2 f(x^{(k)}))^{-1} \boldsymbol{\nabla} f(x^{(k)})$$

Facts on Newton's method

1. It is convergent if $\|x^{(0)} - x^{(*)}\|$ is small enough, and might be divergent if $\|x^{(0)} - x^{(*)}\|$ is large.

2. It is more expensive per iteration than gradient descent, because the inverse of the Hessian is required in the Newton's method.

3. Needs fewer iterations than gradient descent if it converges, because the second-order approximation $F_k(x)$ has a much higher accuracy than the first-order approximation in gradient descent.

Improving upon Newton's method

1. One main concern is the local convergence of Newton's method, as it converges to x^* only when $\|x^{(0)} - x^{(*)}\|$ is sufficiently small.

   Since $F_k(x) \approx F(x)$ only when $\|x^{(k)} - x\|$ is small, similar to gradient descent, we may modify the Newton's method by

   $$\begin{aligned} x^{(k+1)} &= \underset{x \in \mathbb{V}}{argmin} F_k(x) \\ \text{s.t. } \|x - x^{(k)}\| &\leq \alpha_k \|\boldsymbol{\nabla} f(x^{(k)})\| \end{aligned}$$

   which has an explicit form as

   $$x^{(k+1)} = x^{(k)} - (\boldsymbol{\nabla}^2 f(x^{(k)}) + \lambda_k I)^{-1} \boldsymbol{\nabla} f(x^{(k)})$$

   This algorithm is a form of trust region algorithm, which converges globally.

2. Another concern of Newton's method is that it is too expensive to compute the inversion of Hessian.

   We may approximate $(\boldsymbol{\nabla}^2 f(x^{(k)}))^{-1}$ by some linear operator B_k:

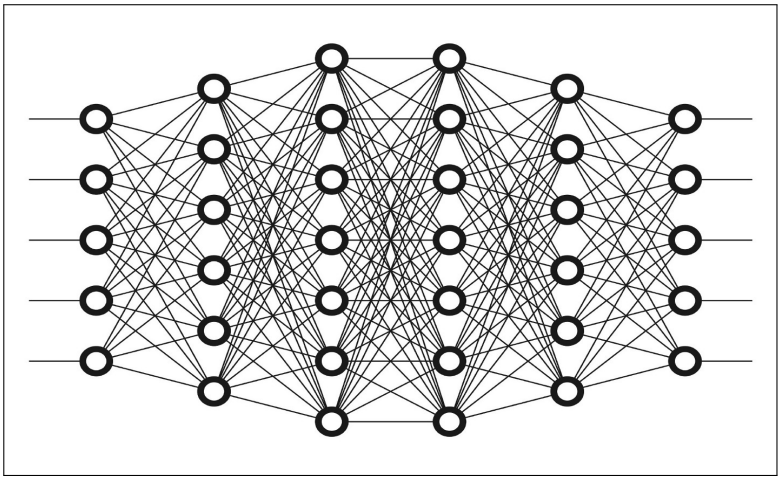
   $$x^{(k+1)} = x^{(k)} - B_k \boldsymbol{\nabla} f(x^{(k)})$$

   This is known as Quasi-Newton method.

   Some well-known B_k's:

   - If B_k is diagonal, (in the special case when $\mathbb{V} = \mathbb{R}^n$, $B_k u = b_k \circ u$ for some $b_k \in \mathbb{R}^n$ with $\circ$ being component multiplication), then we obtain a scaled gradient descent.
   - BFGS (Broyden, Fletcher, Goldfarb and Shanno)
   - Other structured approximation to $(\boldsymbol{\nabla}^2 f(x^{(k)}))^{-1}$

Case Study: Deep Learning

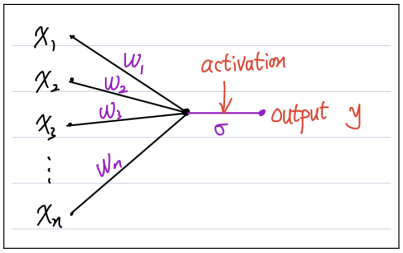


Given $(x^{(i)}, y_i)_{i=1}^m$, where $x^{(i)} \in \mathbb{R}^n, y_i \in \mathbb{R}$, we want to find a function $f : \mathbb{R}^n \to \mathbb{R}$, s.t.

$$f(x^{(i)}) \approx y_i,\ i = 1, \cdots, m$$

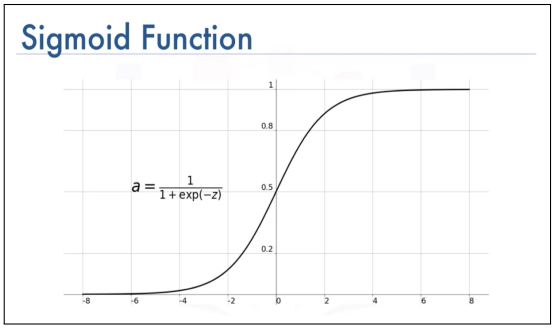
In linear regression, we choose f to be an affine function. In kernel regression, we choose f to be a linear function in the feature domain. In deep learning, we choose f to be a function generated by deep neural networks.

For simplicity, we consider the fully-connected neural network.

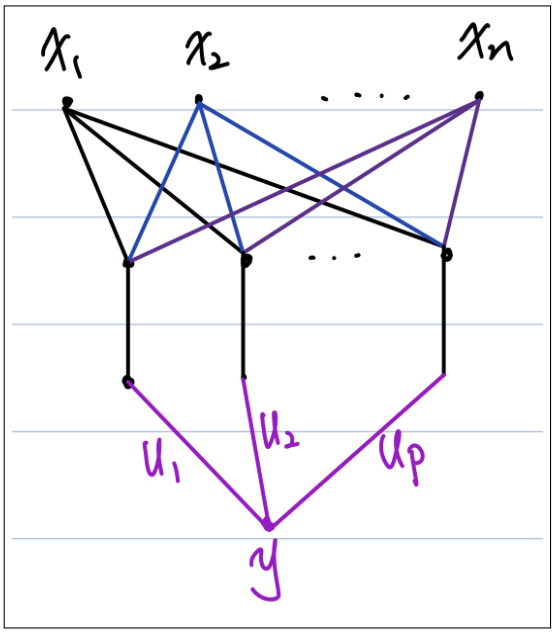


$$y = \sigma(\textstyle\sum_{i=1}^n w_i x_i) = \sigma(\langle w, x \rangle) = \sigma(w^T x)$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function known as the sigmoid function.



Neural Network



1-layer neural network formulation

$$y = \sum_{i=1}^{p} u_i \sigma(\langle w_i, x \rangle)$$

Let $W = \begin{bmatrix} w_1^T \\ w_2^T \\ \vdots \\ w_p^T \end{bmatrix} \in \mathbb{R}^{p \times n}$, $u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{bmatrix} \in \mathbb{R}^p$ and $\sigma : \mathbb{R}^p \to \mathbb{R}^p$ by $\sigma(y) = \begin{bmatrix} \sigma(y_1) \\ \sigma(y_2) \\ \vdots \\ \sigma(y_p) \end{bmatrix}$.

Then

$$y = \langle u, \sigma(Wx) \rangle \equiv f_{W,u}(x)$$

The neural network for regression becomes

Find $W \in \mathbb{R}^{p \times n}$, $u \in \mathbb{R}^p$ s.t.

$$f_{W,u}(x^{(i)}) \approx y_i,\ i = 1, \cdots, m$$

Therefore,

$$\min_{\substack{W \in \mathbb{R}^{p \times n} \\ u \in \mathbb{R}^p}} \sum_{i=1}^{m} (f_{W,u}(x^{(i)}) - y_i)^2$$

Training of Neural Network

Define $F_i(W,u) = (f_{W,u}(x^{(i)}) - y_i)^2$ and $F = \sum_{i=1}^m F_i(W,u)$.

We solve $\min\limits_{W,u} F(W,u) \iff \min\limits_{W,u} \sum_{i=1}^m F_i(W,u)$.
Remarks:

1. $F(W,u)$ is differentiable if $\sigma : \mathbb{R} \to \mathbb{R}$ is differentiable.
2. $F(W,u)$ is **NOT** convex in general.
3. Gradient descent is **NOT** guaranteed to find the global minimizer for non-convex function.
4. Stochastic gradient descent will give a good solution of the minimizer in the training of the neural network.

To apply gradient descent, we need to find $\nabla F(W,u)$, which is a combination of $\nabla_W F(W,u)$ and $\nabla_u F(W,u)$.

Further, $F = \sum_{i=1}^m F_i \implies \nabla_W F = \sum_{i=1}^m \nabla_W F_i$ and $\nabla_u F = \sum_{i=1}^m \nabla_u F_i$.

For $\nabla_W F_i(W,u)$,

Recall $F_i(W,u) = (f_{W,u}(x^{(i)}) - y_i)^2$. Define $H(W) = f_{W,u}(x^{(i)}) \in \mathbb{R}$, $H : \mathbb{R}^{p\times n} \to \mathbb{R}$ and $G(t) = (t - y_i)^2$, $G : \mathbb{R} \to \mathbb{R}$.

Then $F_i(W,u) = G(H(W)) = G \circ H(W)$.

By the chain rule,

$$\nabla_W F_i(W,u) = G'(H(W))\nabla(W) = 2(f_{W,u}(x^{(i)}) - y_i)\nabla_W f_{W,u}(x^{(i)})$$

It remains to find

$$\nabla_W f_{W,u}(x^{(i)}) = \nabla_W(\langle u, \sigma(Wx^{(i)})\rangle)$$

Define $H(W) = \sigma(Wx^{(i)}) \in \mathbb{R}^p$, $H : \mathbb{R}^{p\times n} \to \mathbb{R}^p$ and $G(z) = \langle u, z\rangle$, $\forall z \in \mathbb{R}^p$. ($G$ is linear)

Then $\langle u, \sigma(Wx^{(i)})\rangle = G(H(W)) = (G \circ H)(W)$.

By the chain rule,

$$\nabla_W(\langle u, \sigma(Wx^{(i)}))\rangle = \nabla_W(G \circ H)(W)$$

$$= (DH(W))^* u = (D_W \sigma(Wx^{(i)}))^* u$$

Now, it remains to fine

$$D_W \sigma(Wx^{(i)})$$

Let $G(W) = Wx^{(i)}$, $G : \mathbb{R}^{p\times n} \to \mathbb{R}^p$ is linear
$\implies DG(W)(V) = G(V) = Vx^{(i)}$.

Then $\sigma(Wx^{(i)}) = (\sigma \circ G)(W)$, $\sigma : \mathbb{R}^p \to \mathbb{R}^p$ is non-linear.

$\forall V \in \mathbb{R}^{p \times n}$,

$$D(\sigma \circ G)(W)(V) = D\sigma(G(W))(DG(W)(V)) = D\sigma(G(W))(Vx^{(i)})$$

For $D\sigma$, since $\sigma : \mathbb{R}^p \to \mathbb{R}^p$,

$$D\sigma(z) = \begin{bmatrix} \sigma'(z_1) & & & \\ & \sigma'(z_2) & & \\ & & \ddots & \\ & & & \sigma'(z_p) \end{bmatrix} = diag(\sigma'(z)),\ \forall z \in \mathbb{R}^p$$

Then $D\sigma(G(W))(Vx^{(i)}) = diag(\sigma'(Wx^{(i)}))Vx^{(i)}$.

For $(D(\sigma \circ G)(W))^*$,

$$\langle D(\sigma \circ G)(W)(V), z\rangle = \langle diag(\sigma'(Wx^{(i)}))Vx^{(i)}, z\rangle$$

$$= z^T diag(\sigma'(Wx^{(i)}))Vx^{(i)} = trace(x^{(i)}z^T diag(\sigma'(Wx^{(i)}))V)$$

$$= \langle V, diag(\sigma'(Wx^{(i)}))z(x^{(i)})^T\rangle$$

$$\implies (D(\sigma \circ G)(W))^* z = diag(\sigma'(Wx^{(i)}))z(x^{(i)})^T$$

Finally,

$$\nabla_W F_i(W, u) = 2(f_{W,u}(x^{(i)}) - y_i)diag(\sigma'(Wx^{(i)}))u(x^{(i)})^T$$

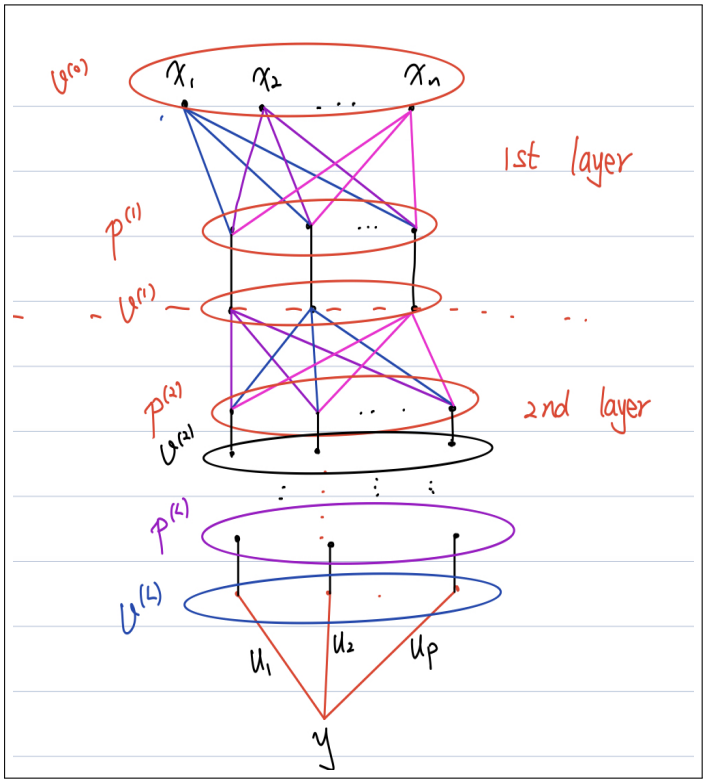
For $\nabla_W F_i(W, u)$,

$\nabla_u F_i(W, u) = 2(f_{W,u}(x^{(i)}) - y_i)\nabla_u(f_{W,u}(x^{(i)}))$

$= 2(f_{W,u}(x^{(i)}) - y_i)\nabla_u(\langle u, \sigma(Wx^{(i)})\rangle)$

$= 2(f_{W,u}(x^{(i)}) - y_i)\sigma(Wx^{(i)})$

Multi-layer neural network formulation



For 1-layer, we define $f_{W,u}(x) = \langle u, \sigma(Wx)\rangle$. Assume that we have L-layer, then we redefined $f_{W,u}(x)$ as

$$f_{W,u}(x) = \langle u, \sigma(W^{(L)} \cdots \sigma(W^{(2)}\sigma(W^{(1)}x)))\rangle$$

$$= \langle u, \sigma \circ W^{(L)} \circ \cdots \circ \sigma \circ W^{(2)} \circ \sigma \circ W^{(1)}(x)\rangle$$

where $W = [W^{(1)}; W^{(2)}; \cdots; W^{(L)}]$.

When L is large, we call the neural network training process as deep learning.

The following derivation is copied and pasted from Professor.YE's own writing. (Typesetting is painful sometimes.)

Find W, u s.t.

$$f_{W,u}(x^{(i)}) \approx y_i, \quad i = 1, \cdots, m$$

Solve

$$\min_{W,u} \sum_{i=1}^{m} (f_{W,u}(x^{(i)}) - y_i)^2$$

$F_i(W, u)$

$\Leftrightarrow$ Find W, u

$$\min_{W,u} \sum_{i=1}^{m} F_i(W,u) \quad , \quad F_i(W,u) = (f_{W,u}(x^{(i)}) - y_i)^2$$

For algorithms, we need

$$\nabla_{W^{(l)}} F_i(W,u) \quad \text{and} \quad \nabla_u F_i(W,u)$$

$l = 1, 2, \cdots, L$

Define

$$\left.\begin{aligned} v^{(0)} &= x^{(i)} \\ v^{(1)} &= \sigma(W^{(1)} v^{(0)}) \\ v^{(2)} &= \sigma(W^{(2)} v^{(1)}) \\ v^{(3)} &= \sigma(W^{(3)} v^{(2)}) \\ &\vdots \\ v^{(L)} &= \sigma(W^{(L)} v^{(L-1)}) \end{aligned}\right\} \quad \begin{aligned} v^{(0)} &= x^{(i)} \\ v^{(l)} &= \sigma(W^{(l)} v^{(l-1)}) \\ l &= 1, \cdots, L \end{aligned}$$

$v^{(l)}$ is the output of the l-th layer

Define

$$s^{(L)} = \sigma \quad (\sigma: \text{activation in } L\text{-th layer}), \quad s^{(L-1)} = \sigma \circ W^{(L)} \circ \sigma, \quad \cdots$$

Equivalently,

$$f_{W,u}(x^{(i)}) = \langle u, \underbrace{\sigma \circ W^{(L)} \circ \sigma \circ \cdots \circ \sigma \circ W^{(l)}}_{s^{(l)}} \circ \underbrace{\sigma \cdots \sigma \circ W^{(1)}(x^{(i)})}_{v^{(l-1)}} \rangle$$

$$f_{W,u}(x^{(i)}) = \langle u, v^{(L)} \rangle$$

$$= \langle u, S^{(L)}(W^{(L)} v^{(L-1)}) \rangle$$
$$\vdots$$
$$= \langle u, S^{(L)}(W^{(L)} v^{(L-1)}) \rangle, \quad l = 1, 2, \cdots, L$$

For $\nabla_u F_i(W, u)$:

$$\nabla_u F_i(W,u) = \nabla_u ((f_{W,u}(x^{(i)}) - y_i)^2)$$
$$= 2(f_{W,u}(x^{(i)}) - y_i) \nabla_u (f_{W,u}(x^{(i)}))$$
$$= 2(f_{W,u}(x^{(i)}) - y_i) \nabla_u (\langle u, v^{(L)} \rangle)$$
$$= 2(f_{W,u}(x^{(i)}) - y_i) v^{(L)}$$

For $\nabla_{W^{(l)}} F_i(W, u)$

$$\nabla_{W^{(l)}} F_i(W,u) = \nabla_{W^{(l)}} (f_{W,u}(x^{(i)}) - y_i)^2)$$
$$= 2(f_{W,u}(x^{(i)}) - y_i) \nabla_{W^{(l)}} (f_{W,u}(x^{(i)}))$$
$$= 2(f_{W,u}(x^{(i)}) - y_i) \nabla_{W^{(l)}} (\langle u, S^{(L)}(W^{(L)} v^{(L-1)}) \rangle)$$

$$\nabla_{W^{(l)}} (\langle u, S^{(L)}(W^{(L)} v^{(L-1)}) \rangle)$$
$$= (DS^{(L)}(W^{(L)} v^{(L-1)}))^T u (v^{(L-1)})^T$$

denoted by $M^{(L)}$

$S^{(L)}: \mathbb{R}^{n_L} \mapsto \mathbb{R}^{n_L}$, n_L = # of neurons in layer L

$\nabla_W (\langle u, \sigma(Wx^{(i)}) \rangle)$ 1-layer case

$= \text{diag}(\sigma'(Wx^{(i)})) u (x^{(i)})^T$ $\sigma: \mathbb{R}^p \mapsto \mathbb{R}^p$

We use recursion to find $M^{(L)}$, $L = 1, \cdots, L$

$$M^{(L)} = DS^{(L)}(W^{(L)} v^{(L-1)})$$
$$= D\sigma(W^{(L)} v^{(L-1)})$$
$$= \text{diag}(\sigma'(W^{(L)} v^{(L-1)}))$$

Since $S^{(L)} = S^{(L+1)} \circ W^{(L+1)} \circ \sigma$

By the chain rule,

$$DS^{(l)} = DS^{(l+1)} \circ DW^{(l+1)} \circ D\sigma$$

$$DS^{(l)}(W^{(l)}v^{(l-1)}) = DS^{(l+1)}(W^{(l+1)}v^{(l)}) \cdot DW^{(l+1)}(v^{(l)}) \cdot D\sigma(W^{(l)}v^{(l-1)})$$

in $\mathbb{R}^{n_L \times n_l}$ in $\mathbb{R}^{n_L \times n_{l+1}}$ in $\mathbb{R}^{n_{l+1} \times n_l}$ in $\mathbb{R}^{n_l \times n_l}$

$$\Rightarrow M^{(l)} = M^{(l+1)} W^{(l+1)} \operatorname{diag}(\sigma'(W^{(l)}v^{(l-1)}))$$

$$l = L, L-1, \cdots, 1$$

So,

$$\nabla_{W^{(l)}} F_i(W, u) = 2(f_{W,u}(x^{(i)}) - y_i)(M^{(l)})^T u (v^{(l-1)})^T$$

$(M^{(l)})^T u = z^{(l)}$

Denote $p^{(l)} = W^{(l)} v^{(l-1)}$

Recall $v^{(l)} = \sigma(W^{(l)} v^{(l-1)}) = \sigma(p^{(l)})$

$v^{(l)}$: output of l-th layer; $p^{(l)}$: preactivation output of l-layer

$$z^{(l)} = (M^{(l)})^T u$$

$$= \operatorname{diag}(\sigma'(p^{(l)}))(W^{(l+1)})^T (M^{(l+1)})^T u$$

$$(\Sigma = \sigma'(\cdot) \odot) = \operatorname{diag}(\sigma'(p^{(l)}))(W^{(l+1)})^T z^{(l+1)}$$

$$= (\Sigma \circ (W^{(l+1)})^T z^{(l+1)}$$

$$\vdots$$

$$= (\Sigma \circ (W^{(l+1)})^T \circ \Sigma \circ \cdots \circ \Sigma \circ (W^{(L)})^T) z^{(L)}$$

$$= (\Sigma \circ (W^{(l+1)})^T \circ \Sigma \circ \cdots \circ \Sigma \circ (W^{(L)})^T \circ \Sigma) u$$

$$\Rightarrow \nabla_{W^{(l)}} F_i(W, u) = 2(f_{W,u}(x^{(i)}) - y_i) z^{(l)} (v^{(l-1)})^T$$

$z^{(l)} = (\Sigma \circ (W^{(l+1)})^T \circ \cdots \circ (W^{(L)})^T \circ \Sigma) u$

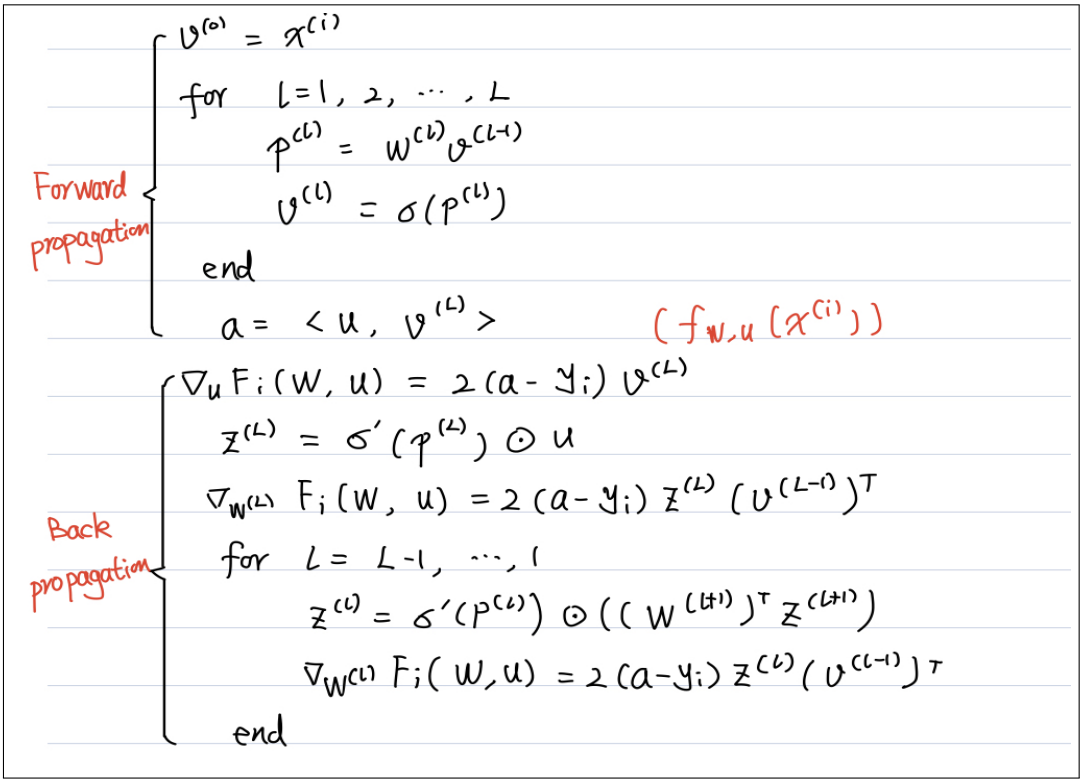
$v^{(l-1)} = (\sigma \circ W^{(l-1)} \circ \cdots \circ \sigma \circ W^{(1)}) x^{(i)}$

$a, b \in \mathbb{R}^s$

$\operatorname{diag}(a) \cdot b = \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_s \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_s \end{pmatrix} = \begin{pmatrix} a_1 b_1 \\ a_2 b_2 \\ \vdots \\ a_s b_s \end{pmatrix} \equiv a \odot b$

entry wise multiplication

Final Algorithm



Case Study: Matrix Differentiation

In many machine learning tasks, we need to find the differentiation of the function mappings of matrix inputs. To find the differentiation, we consider the Hilbert space $\mathbb{R}^{m\times n}$ with inner product

$$\langle X, Y\rangle = \sum_i \sum_j x_{ij}y_{ij} = trace(X^TY) = trace(Y^TX)$$

for $X, Y \in \mathbb{R}^{m\times n}$ and the induced norm is

$$\|X\|_F = (\langle X, X\rangle)^{\frac{1}{2}} = (\sum_{i,j} x_{ij}^2)^{\frac{1}{2}}$$

Matrix Multiplication
Let $A \in \mathbb{R}^{p\times m}$ and $B \in \mathbb{R}^{n\times q}$ be given
Consider the mapping $F : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$

$$F(X) = AXB$$

Find $DF(X)$.

Since $F(X) = AXB$ is a linear mapping, then

$$DF(X)(Y) = AYB, \forall Y \in \mathbb{R}^{m\times n}$$

Quadratic Matrix Multiplication
Let $A \in \mathbb{R}^{n\times n}$ be given.
Consider the mapping $F : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$

$$F(X) = XAX$$

Find $DF(X)$.

For any $X_0 \in \mathbb{R}^{n\times n}$

$$XAX = (X - X_0 + X_0)A(X - X_0 + X_0)$$

$$= \underbrace{X_0AX_0}_{F(X_0)} + (X - X_0)AX_0 + X_0A(X - X_0) + (X - X_0)A(X - X_0)$$

Therefore, $$0 \le \lim_{\|X - X_0\|_F \to 0} \frac{\|XAX - (X_0AX_0 + (X - X_0)AX_0 + X_0A(X - X_0))\|_F}{\|X - X_0\|_F}$$

$$\leq \lim_{\|X-X_0\|_F \to 0} \frac{\|(X-X_0)A(X-X_0)\|_F}{\|X-X_0\|_F} \quad \cdots \quad (1)$$

It can be shown that $\|BC\|_F \leq \|B\|_2 \|C\|_F$ and $\|BC\|_F \leq \|B\|_F \|C\|_2$

So, $\|(X-X_0)A(X-X_0)\|_F \leq \|(X-X_0)A\|_2 \|X-X_0\|_F$

$$\leq \|X-X_0\|_2 \|A\|_2 \|X-X_0\|_F$$

$$\leq \|A\|_2 \|X-X_0\|_F^2$$

(The last inequality is obtained by the fact
$\|X-X_0\|_2 =$ maximum singular value of $X-X_0$
and $\|X-X_0\|_F = (\sum \sigma_i^2)^{\frac{1}{2}}$ where σ_i, $i=1,\cdots,n$
are singular values of $X-X_0$.)

Together with (1)

$$0 \leq \lim_{\|X-X_0\|_F \to 0} \frac{\|XAX - (X_0AX_0 + (X-X_0)AX_0 + X_0A(X-X_0))\|_F}{\|X-X_0\|_F}$$

$$\leq \lim_{\|X-X_0\|_F \to 0} \frac{\|(X-X_0)A(X-X_0)\|_F}{\|X-X_0\|_F}$$

$$\leq \lim_{\|X-X_0\|_F \to 0} \|A\|_2 \|X-X_0\|_F = 0$$

Hence $D(X_0AX_0)(Y) = YAX_0 + X_0AY$, $\forall Y \in \mathbb{R}^{n\times n}$.

That is $D(XAX)(Y) = YAX + XAY$, $\forall Y \in \mathbb{R}^{n\times n}$.

Matrix Inversion
Consider the mapping $F : X \to X^{-1}$, where $X \in \mathbb{R}^{n\times n}$ is invertible. Find $DF(X)(Y)$ for $Y \in \mathbb{R}^{n\times n}$.

Notice that $XX^{-1} = I$, where I is an identity matrix.

$$D(XX^{-1})(Y) = D(I)(Y) = 0$$

By matrix multiplication rule of differentiation,

$$D(XX^{-1})(Y) = D(X)(Y)X^{-1} + XD(X^{-1})(Y) = 0$$

Since the mapping $X \to X$ is linear, $D(X)(Y) = Y$. Hence
$YX^{-1} + XD(X^{-1})(Y) = 0$. That is

$$D(X^{-1})(Y) = -X^{-1}YX^{-1}$$

Matrix Trace
Let $X \in \mathbb{R}^{n\times n}$. Consider the mapping
$F : X \to trace(X) = \sum_{i=1}^{n} x_{ii} = \sum_{i=1}^{n} \lambda_i(x)$ where $\lambda_i(x)$ are eigenvalues of X.
Find $DF(X)(Y)$.

Since $F(X) = \sum_{i=1}^{n} x_{ii} = \langle X, I\rangle$, $F(X)$ is a linear transformation. Hence
$DF(X)(Y) = \langle Y, I\rangle$. That is, $\boldsymbol{\nabla} F(X) = I$ or $\boldsymbol{\nabla} trace(X) = I$.

There are more, but I think we have covered enough of the topic, but if you are interested you can refer to other reference materials.

Optimality and Convexity

4.1 Smooth Unconstrained Optimization

We have seen that many data analysis tasks can be formulated as an optimization problem in the form of

$$\min_{x \in \mathbb{R}^n} f(x)$$

We can extend most optimization problems to Hilbert space $\mathbb{H}$.

Consider the following unconstrained optimization

$$\min_{x \in \mathbb{H}} f(x) \text{ (OPT)}$$

We assume that $f(x)$ is differentiable.

Solvability of (OPT)
We say x^* is a solution of (OPT) if

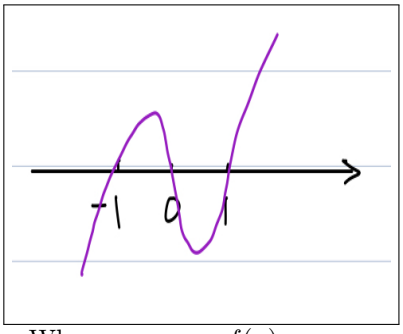
$$f(x^*) \leq f(x), \forall x \in \mathbb{H} \ (0^{th} \text{ order optimality condition})$$

We call x^* a global minimizer of f, denoted by

$$x^* = \underset{x \in \mathbb{H}}{argmin} f(x)$$

Remarks: The existence of a solution of (OPT) is **NOT** guaranteed automatically.

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



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

We assume (OPT) has at least one solution.

Characterization of the solution of (OPT)
Necessary condition for optimality

1. 0^{th} order optimality condition

$$x^* = \underset{x\in\mathbb{H}}{argmin} f(x)$$

$$\Longleftrightarrow$$

$$f(x^*) \leq f(x), \forall x \in \mathbb{H}$$

2. 1^{st} order optimality condition
Theorem: Assume $f : \mathbb{H} \to \mathbb{R}$ is differentiable at $x^* \in \mathbb{H}$, then

$$x^* = \underset{x\in\mathbb{H}}{argmin} f(x)$$

$$\Longrightarrow$$

$$\boldsymbol{\nabla} f(x^*) = 0$$

Proof.
By Taylor expansion,

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

Suppose $\boldsymbol{\nabla} f(x^*) \neq 0$, Choose $\widetilde{x} = x^* - t\boldsymbol{\nabla} f(x^*)$ with $t > 0$, we have

$$\begin{aligned} f(\widetilde{x}) &= f(x^*) + \langle \boldsymbol{\nabla} f(x^*), -t\boldsymbol{\nabla} f(x^*) \rangle + o(\mid t \mid \|\boldsymbol{\nabla} f(x^*)\|) \\ &= f(x^*) - t\|\boldsymbol{\nabla} f(x^*)\|^2 + o(\mid t \mid \|\boldsymbol{\nabla} f(x^*)\|) \end{aligned}$$

Because

$$\lim_{t\to 0} \frac{o(|t|\|\boldsymbol{\nabla} f(x^*)\|)}{|t|\|\boldsymbol{\nabla} f(x^*)\|} = 0$$

$\forall c > 0$, $\exists t$ s.t.

$$ct\|\boldsymbol{\nabla} f(x^*)\| > o(\mid t \mid \|\boldsymbol{\nabla} f(x^*)\|)$$

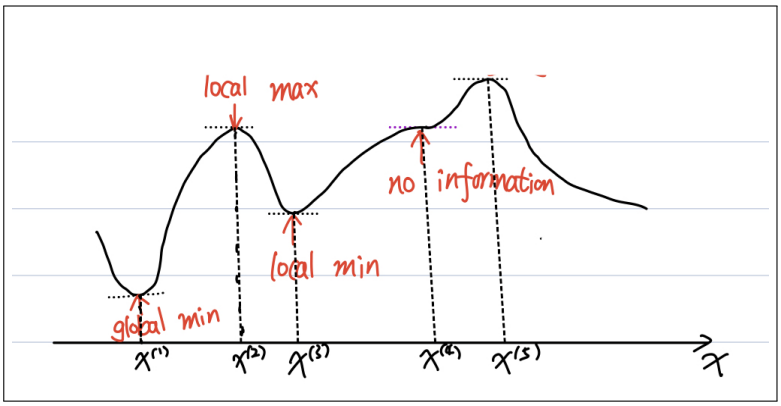
Now we choose $c = \|\boldsymbol{\nabla} f(x^*)\| \neq 0$. Then

$$t\|\boldsymbol{\nabla} f(x^*)\|^2 > o(\mid t \mid \|\boldsymbol{\nabla} f(x^*)\|)$$

$\Longrightarrow$ $f(\widetilde{x}) < f(x^*)$ (*Contradiction*)

Remarks: The reverse is **NOT** true in general.

Example:



All $x^{(i)}$ satisfies $\nabla f(x^{(i)}) = 0$, only $x^{(1)} = \underset{x \in \mathbb{H}}{argmin} f(x)$. Actually, when $\nabla f(x^*) = 0$, x^* can be

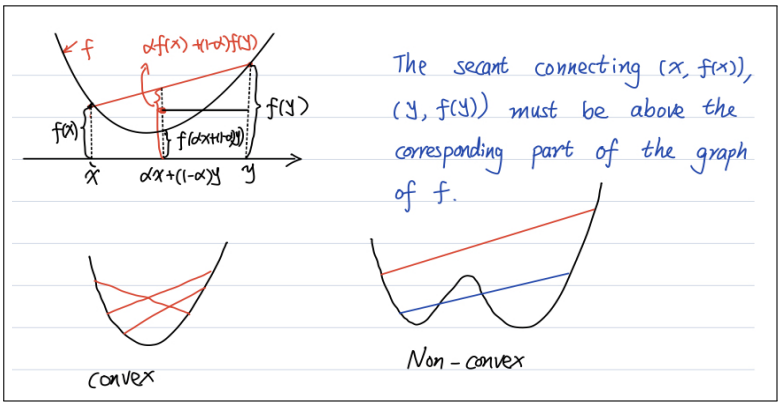
- Global minimizer, i.e., $x^* = \underset{x \in \mathbb{H}}{argmin} f(x)$ (see $x^{(1)}$)
- Local minimizer, i.e., $\exists \epsilon > 0$, s.t. $\forall x : \|x - x^*\| \leq \epsilon$, we have $f(x^*) \leq f(x)$ (see $x^{(3)}$)
- Local maximizer, i.e., $\exists \epsilon > 0$, s.t. $\forall x : \|x - x^*\| \leq \epsilon$, we have $f(x^*) \geq f(x)$ (see $x^{(2)}$)
- Global maximizer, i.e., $\forall x \in \mathbb{H}$, $f(x^*) \geq f(x)$ (see $x^{(5)}$)
- Saddle point (only for $\mathbb{H}$ with $dim(\mathbb{H}) \geq 2$), i.e., $\exists u, v \in \mathbb{H}$, s.t.

  $$f(x^*) \geq f(x^* + tu) \text{ and } f(x^*) \leq f(x^* + tv) \ \forall t :| \ t \ | \leq \epsilon$$

  That is, $f(x^*)$ is a local minimum along v and a local maximum along u.
- None of the above (see $x^{(4)}$)

Sufficient condition for optimality
Convexity
Definition: A function $f : \mathbb{H} \to \mathbb{R}$ is called 'convex' if

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y), \ \forall x, y \in \mathbb{H}, \alpha \in [0, 1]$$



Example: Let $\mathbb{H}$ be a Hilbert space and $f(x) = \|x\|^2$, $x \in \mathbb{H}$. Prove that $f(x)$ is convex.

Proof.
$\forall \alpha \in [0,1], x, y \in \mathbb{H}$

$$
\begin{aligned}
f(\alpha x + (1-\alpha)y) &= \|\alpha x + (1-\alpha)y\|^2 \\
&= \alpha^2 \|x\|^2 + 2\alpha(1-\alpha)\langle x, y\rangle + (1-\alpha)^2 \|y\|^2 \\
&= \alpha \|x\|^2 + (1-\alpha)\|y\|^2 + 2\alpha(1-\alpha)\langle x, y\rangle + (\alpha^2 - \alpha)\|x\|^2 \\
&\quad + (\alpha^2 - \alpha)\|y\|^2 \\
&= \alpha \|x\|^2 + (1-\alpha)\|y\|^2 - \alpha(1-\alpha)\left[-2\langle x, y\rangle + \|x\|^2 + \|y\|^2\right] \\
&= \alpha f(x) + (1-\alpha) f(y) - \alpha(1-\alpha)\|x - y\|^2 \\
&\leq \alpha f(x) + (1-\alpha) f(y)
\end{aligned}
$$

So $f(x)$ is convex.

Example: Let $\mathbb{V}$ be a vector space with norm $\|\cdot\|$. Let $f(x) = \|x\|$ for any x in $\mathbb{V}$. Prove that $f(x)$ is convex.

Proof.
$\forall \alpha \in [0,1], x, y \in \mathbb{V}$

$$
\begin{aligned}
f(\alpha x + (1-\alpha)y) &= \|\alpha x + (1-\alpha)y\| \\
&\overset{\text{triangle inequality}}{\leq} |\alpha| \|x\| + |1-\alpha| \|y\| \\
&= \alpha \|x\| + (1-\alpha)\|y\| \\
&= \alpha f(x) + (1-\alpha) f(y)
\end{aligned}
$$

So $f(x)$ is convex.
Remarks: $f(x) = \|x\|$ is convex for any norm on $\mathbb{V}$.

Example: Prove that any affine function is convex.

Proof.
Let $f : \mathbb{H} \to \mathbb{R}$ be affine. Then

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

where $g(x)$ is linear and b is a constant.
$\forall \alpha \in [0,1], x, y \in \mathbb{H}$

$$\begin{aligned} f(\alpha x + (1-\alpha)y) &= g(\alpha x + (1-\alpha)y) + b \\ &= \alpha g(x) + (1-\alpha) g(y) + (\alpha + 1 - \alpha) b \\ &= \alpha (g(x) + b) + (1-\alpha)(g(y) + b) \\ &= \alpha f(x) + (1-\alpha) f(y) \end{aligned}$$

Example: Let $f_1, f_2, \cdots, f_n$ are convex. Let $f = \sum_{i=1}^{n} c_i f_i$, where $c_i \geq 0$, $i = 1, \cdots, n$. Show that f is convex.

Proof.
$\forall \alpha \in [0,1], x, y \in \mathbb{H}$

$$\begin{aligned} f(\alpha x + (1-\alpha)y) &= \sum_{i=1}^{n} c_i f_i(\alpha x + (1-\alpha)y) \\ &\overset{c_i \geq 0 \ \forall i}{\leq} \sum_{i=1}^{n} c_i \left(\alpha f_i(x) + (1-\alpha) f_i(y)\right) \\ &= \alpha \sum_{i=1}^{n} c_i f_i(x) + (1-\alpha) \sum_{i=1}^{n} c_i f_i(y) \end{aligned}$$

Remarks: We put the constraint $c_i \geq 0$, because suppose there are two convex functions f_1 and f_2, then $f_1 - f_2$ is **NOT** necessarily convex. One example would be $f_1 = x$ and $f_2 = x^2$.

Example: Let f be convex and g be affine. Show that $f \circ g$ is convex.

Proof.
$\forall \alpha \in [0,1], x, y \in \mathbb{H}$

$$\begin{aligned} f \circ g(\alpha x + (1-\alpha)y) &= f(g(\alpha x + (1-\alpha)y)) \\ &= f(\alpha g(x) + (1-\alpha) g(y)) \\ &\leq \alpha f(g(x)) + (1-\alpha) f(g(y)) \\ &= \alpha (f \circ g)(x) + (1-\alpha)(f \circ g)(y) \end{aligned}$$

Remarks: If f and g are both convex, then $f \circ g$ is **NOT** necessarily convex. One example would be $g(x) = \frac{1}{x}$, $h(x) = -log(x)$. Then
$h(g(x)) = -log(\frac{1}{x}) = log(x)$ is concave.

Theorem: If $f : \mathbb{H} \to \mathbb{R}$ is convex and differentiable, then

$$x^* = \underset{x \in \mathbb{H}}{argmin} f(x) \iff \nabla f(x^*) = 0$$

Lemma: If $f : \mathbb{H} \to \mathbb{R}$ is differentiable, then

$$f \text{ is convex } \iff f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \ \forall x, y \in \mathbb{H}$$

This characterization in the above lemma is using tangent line.

Proof of Lemma: We first prove the lemma when $\mathbb{H} = \mathbb{R}$.
($\Leftarrow$) Assume $f : \mathbb{R} \to \mathbb{R}$ is convex, then $\alpha \in [0, 1]$ and $x, y \in \mathbb{R}$ with $x \neq y$.

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y)$$

$$\implies f(y) \geq \frac{f(\alpha x + (1-\alpha)y) - \alpha f(x)}{1 - \alpha}$$

$$= f(x) + \frac{f(\alpha x + (1-\alpha)y) - f(x)}{(1-\alpha)(y-x)}(y - x)$$

$$= f(x) + \frac{f(x + (1-\alpha)(y-x)) - f(x)}{(1-\alpha)(y-x)}(y - x)$$

Let $\alpha \to 1_-$,

$$f(y) \geq f(x) + f'(x)(y - x)$$

($\Rightarrow$) Assume $f(y) \geq f(x) + f'(x)(y - x)$, $\forall x, y \in \mathbb{R}$, then we choose $x, y \in \mathbb{R}$ with $x \neq y$.
Define $z = \alpha x + (1 - \alpha)y$ for some $\alpha \in [0, 1]$.

$$f(x) \geq f(z) + f'(z)(x - z) \ \textbf{(1)}$$
$$f(y) \geq f(z) + f'(z)(y - z) \ \textbf{(2)}$$

Combining the two

$$\alpha f(x) + (1 - \alpha) f(y)$$
$$\geq \alpha f(z) + \alpha f'(z)(x - z) + (1 - \alpha) f(z) + (1 - \alpha) f'(z)(y - z)$$
$$= f(z) + f'(z)[\alpha(x - z) + (1 - \alpha)(y - z)] = f(z)$$

Here $[\alpha(x - z) + (1 - \alpha)(y - z)] = \alpha x + (1 - \alpha)y - z = 0$.
$\implies f(z) \leq \alpha f(x) + (1 - \alpha) f(y)$
$\implies f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y)$
$\implies f$ is convex.

Next, we prove the general case.
($\Leftarrow$) Assume $f : \mathbb{H} \to \mathbb{R}$ is convex. For any $x, y \in \mathbb{H}$,

$$g(t) = f(tx + (1-t)y),\ t \in \mathbb{R}, \text{ we have}$$

1. $g(t) : \mathbb{R} \to \mathbb{R}$ is convex, since $f(x)$ is convex. More specifically, $g(\alpha s + (1-\alpha)t) \leq \alpha g(s) + (1-\alpha)g(t)$.
2. $g'(t) = \langle \nabla f(tx + (1-t)y), x - y \rangle$

Since $g(t) : \mathbb{R} \to \mathbb{R}$, we can use our previous result when $\mathbb{H} = \mathbb{R}$.

$$g(0) \geq g(1) + g'(1)(0-1),\ \forall x, y \in \mathbb{R}$$

which is the same as

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

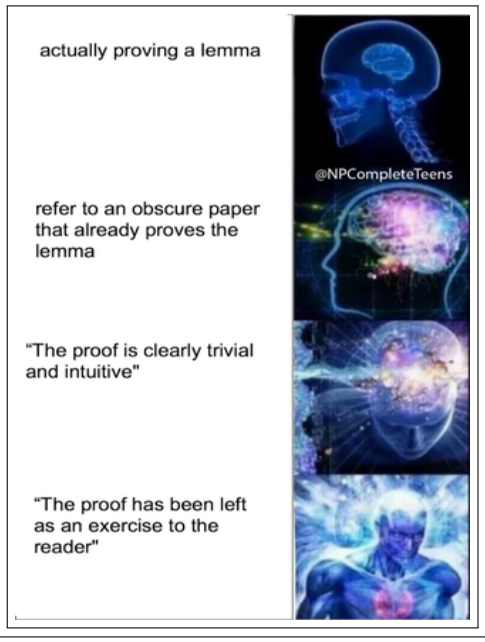
($\Rightarrow$) Assume $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$, $\forall x, y \in \mathbb{H}$, $\alpha \in [0, 1]$.
Define $z = \alpha x + (1-\alpha)y$

$$f(x) \geq f(z) + \langle \nabla f(z), x - z \rangle \ \textbf{(1)}$$
$$f(y) \geq f(z) + \langle \nabla f(z), y - z \rangle \ \textbf{(2)}$$

Combining the two

$$\begin{aligned} \alpha f(x) + (1-\alpha) f(y) &\geq f(z) + \langle \nabla f(z), \alpha(x - z) + (1-\alpha)(y - z) \rangle \\ &= f(z) + \langle \nabla f(z), 0 \rangle = f(z) = f(\alpha x + (1-\alpha)y) \end{aligned}$$

$\implies$ f is convex.



Now we prove the theorem using the proven lemma.

Proof.
Previously we have proven that if $x^* = \underset{x\in\mathbb{H}}{argmin} f(x)$, then $\boldsymbol{\nabla} f(x^*) = 0$. So we have only considered "$\Leftarrow$" direction.

Since f is convex and differentiable, for any $x \in \mathbb{H}$,

$$f(x) \geq f(x^*) + \langle \boldsymbol{\nabla} f(x^*), x - x^* \rangle$$

By assumption, $\boldsymbol{\nabla} f(x^*) = 0$.

$$\implies f(x) \geq f(x^*),\ \forall x \in \mathbb{H}$$

$$\implies x^* = \underset{x\in\mathbb{H}}{argmin} f(x)$$

Uniqueness of Global Minimizer
Definition: A function $f : \mathbb{H} \to \mathbb{R}$ is strictly convex if

$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y),\ \forall x \neq y, \alpha \in (0,1)$$

Theorem: Assume f is strictly convex, then the solution of $\underset{x\in\mathbb{H}}{argmin} f(x)$ is unique if it exists.

Proof.
Suppose that there are at least two solution x^* and y^*, then $f(x^*) = f(y^*)$. Consider $z = \alpha x^* + (1-\alpha)y^*$ with $\alpha \in (0,1)$. Then $z \neq x^*$ and $z \neq y^*$. Moreover,

$$f(z) = f(\alpha x^* + (1-\alpha)y^*)$$

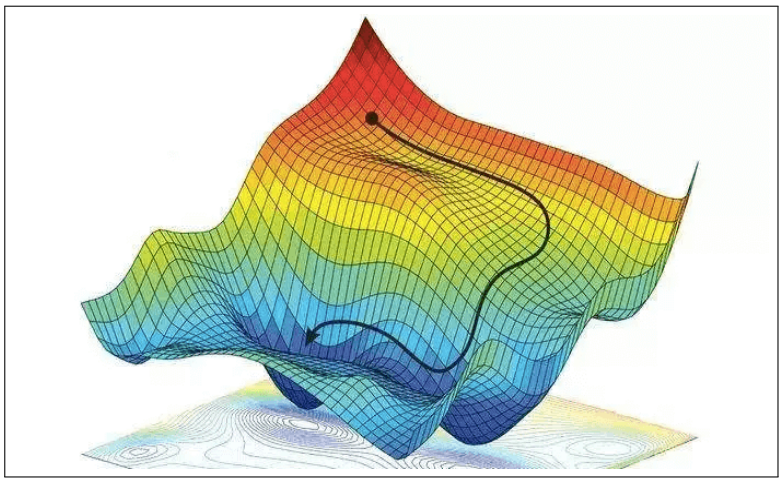
$$< \alpha f(x^*) + (1-\alpha) f(y^*)$$

$$= \alpha f(x^*) + (1-\alpha) f(x^*) = f(x^*)$$

i.e., $f(z) < f(x^*)$, but $x^* = \underset{x\in\mathbb{H}}{argmin} f(x)$ $(Contradiction)$

4.2 Gradient Descent

In the case when f is differentiable, the simplest algorithm for finding a solution of $\underset{x\in\mathbb{H}}{argmin} f(x)$ is gradient descent.



Let $x^{(k)}$ be the current estimation of x^*. We want to find $x^{(k+1)}$ such that $f(x^{(k+1)}) < f(x^{(k)})$. Since f is differentiable, we can expand it at $x^{(k)}$ as

$$f(x^{(k+1)}) \approx f(x^{(k)}) + \langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle$$

If $\|x^{(k+1)} - x^{(k)}\|$ is small, then this implies that

$$f(x^{(k+1)}) - f(x^{(k)}) \approx \langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle$$

So the approximation is only good when $\|x - x^{(k)}\|$ is small. We approximate it by

$$x^{(k+1)} = \underset{x\in\mathbb{H}}{argmin} [f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle \text{] } \textbf{(*)}$$
$$\text{s.t. } \|x - x^{(k)}\| \leq \alpha_k \|\nabla f(x^{(k)})\|$$

where $\alpha_k > 0$ is a small number.

Now we need to find the solution of $(*)$.

- **(*)** $\iff \underset{x\in\mathbb{H}}{min} \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle$
- By Cauchy-Schwartz,

$$\langle \nabla f(x^{(k)}), x - x^{(k)} \rangle$$

$$\geq -\|\nabla f(x^{(k)})\| \|x - x^{(k)}\| \textbf{ (1)}$$

$$\geq -\|\nabla f(x^{(k)})\| \alpha_k \|\nabla f(x^{(k)})\| \textbf{ (2)}$$

$$= -\alpha_k \|\nabla f(x^{(k)})\|^2$$

The minimum value of **(*)** $\geq -\alpha_k \|\nabla f(x^{(k)})\|^2$

- "=" is attained in **(1)** when

$$x - x^{(k)} = -c\nabla f(x^{(k)}), \text{ where } c > 0$$

- "=" is attained in **(2)** when

$$\|x - x^{(k)}\| = \alpha_k \|\nabla f(x^{(k)}\|$$

Together with $\|x - x^{(k)}\| = c\|\nabla f(x^{(k)})\|$, we have $c = \alpha_k$.

$$x - x^{(k)} = -\alpha_k \nabla f(x^{(k)})$$

$\implies x = x^{(k)} - \alpha_k \nabla f(x^{(k)})$ (the solution of **(*)**)

Gradient Descent Algorithm
Given $x^{(0)}$ (initial guess),

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}),\ k = 0, 1, \cdots$$

where α_k is known as the step size.

Choosing the right α_k is crucial. There are many strategies for finding good α_k, two well known method are:

- Backtracking line search
- Exact line search

Convergence of Gradient Descent
Gradient Descent converges with sufficiently small α_k. If it converges, then the limit x^∞ satisfies

$$\nabla f(x^\infty) = 0$$

- If f is convex, then x^∞ is the global minimizer.
- If f is non-convex, then x^∞ is **NOT** guaranteed to be the global minimizer. It only finds a point x satisfying the condition that $\nabla f(x) = 0$.

Application of Gradient Descent
Least Square

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2$$

where $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$ are given

Let $f(x) = \frac{1}{2}\|Ax-b\|_2^2$. In order to utilize gradient descent, we first need to show that f is convex and differentiable.

1. $f : \mathbb{R}^n \to \mathbb{R}$ is convex.

   Proof.
   Let $g(y) = \frac{1}{2}\|y\|_2^2$, where $y \in \mathbb{R}^m$. Then $g : \mathbb{R}^m \to \mathbb{R}$ is convex. (proven in previous example) So, $f(x) = \frac{1}{2}\|Ax-b\|_2^2 = g(Ax-b)$ and $\forall x, y \in \mathbb{R}^n$, $\alpha \in [0,1]$.

$$f(\alpha x + (1-\alpha)y) = g(A(\alpha x + (1-\alpha)y) - b)$$

$$= g(\alpha Ax + (1-\alpha)Ay - \alpha b - (1-\alpha)b)$$

$$= g(\alpha(Ax-b) + (1-\alpha)(Ay-b))$$

$$\leq \alpha g(Ax-b) + (1-\alpha)g(Ay-b)$$

$$= \alpha f(x) + (1-\alpha)f(y)$$

2. $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable.

   Theorem: Let $g : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function. Let $A \in \mathbb{R}^{m\times n}$. Define $h : \mathbb{R}^n \to \mathbb{R}$ by $h(x) = g(Ax)$. Then $\nabla h(x) = A^T \nabla g(Ax)$. (an extension of the chain rule)

   Proof.
   Consider $E(x,y) := h(y) - (h(x) + \langle A^T \nabla g(Ax), y-x\rangle)$
   $= h(y) - (h(x) + \langle \nabla g(Ax), A(y-x)\rangle$

$$0 \leq \lim_{\|y-x\|_2\to 0} \frac{|E(x,y)|}{\|y-x\|_2}$$

$$= \lim_{\|Ay-Ax\|_2\to 0} \frac{|E(x,y)|}{\|y-x\|_2} \leq \|A\|_2 \lim_{\|Ay-Ax\|_2\to 0} \frac{|E(x,y)|}{\|Ay-Ax\|_2} = 0 \text{ (*)}$$

$$\text{(using } \|Ay - Ax\|_2 \leq \|A\|_2 \|y - x\|_2)$$

Since g is differentiable at Ax, **(*)**

$$\lim_{\|Ay - Ax\|_2 \to 0} \frac{|E(x,y)|}{\|Ay - Ax\|_2} = 0$$

This then implies (by squeeze theorem) that

$$\lim_{\|y - x\|_2 \to 0} \frac{|E(x,y)|}{\|y - x\|_2} = 0$$

Since $f(x) = g(Ax - b)$ with $g(y) = \frac{1}{2}\|y\|^2$,

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

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

$$x^* = \underset{x \in \mathbb{R}^n}{argmin} \frac{1}{2}\|Ax - b\|_2^2 \iff A^T(Ax^* - b) = 0 \iff A^T Ax^* = A^T b$$

The equation above is known as the normal equation in linear algebra.

Geometric Interpretation

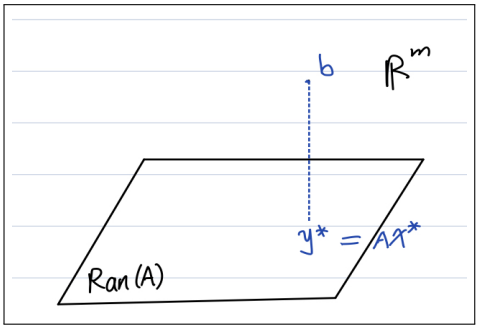
$$\underset{x \in \mathbb{R}^n}{min} \frac{1}{2}\|Ax - b\|_2^2$$

By setting $y = Ax$, the minimization become

$$\underset{y \in \mathbb{R}^m}{min} \frac{1}{2}\|y - b\|_2^2$$

$$\text{s.t. } y \in Ran(A)$$

This is equivalent to projecting b onto $Ran(A)$.



Remarks: b is **NOT** necessarily in $Ran(A)$.

Since $b - Ax^* \perp Ran(A)$,

$$\langle b - Ax^*, Az \rangle = 0, \forall z \in \mathbb{R}^n$$

$$\Updownarrow$$

$$\langle A^T(b - Ax^*), z \rangle = 0, \forall z \in \mathbb{R}^n$$

$$\Updownarrow$$

$$A^T(b - Ax^*) = 0$$

$$\Updownarrow$$

$$A^T Ax^* = A^T b$$

It can be shown that: If $A^T A$ is invertible, then $f(x)$ is strictly convex, and therefore the solution of least square is unique.

To find x^*, we now use gradient descent

$$x^{(k+1)} = x^{(k)} - \alpha_k A^T(Ax^{(k)} - b),\ k = 0, 1, 2, \cdots$$

To choose a good α_k, we may use "line search", i.e., we set

$$\alpha_k = \underset{\alpha \in \mathbb{R}}{argmin} f(x^{(k)} - \alpha A^T(Ax^{(k)} - b))$$

In other words, α_k is the optimal step size.
Let $g(\alpha) = f(x^{(k)} - \alpha A^T(Ax^{(k)} - b))$. It can then be checked that $g(\alpha)$ is convex, and therefore $g'(\alpha_k) = 0$, we get

$$\alpha_k = \frac{\|A^T(Ax^{(k)} - b)\|_2^2}{\|AA^T(Ax^{(k)} - b)\|_2^2}$$

This leads to the steepest descent algorithm for least square.

Algorithm:

Initialize $x^{(0)}$
for $k = 0, 1, 2, \cdots$
$g^{(k)} = A^T(Ax^{(k)} - b)$
$\alpha_k = \frac{\|g^{(k)}\|_2^2}{\|Ag^{(k)}\|_2^2}$
$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$

end