

# Ch. 3. Inner product, Hilbert space

## § 3.1. Inner Product

Norms give only metrics, which are "scaling" sensitive.

However, in many applications, "scaling" does not matter.

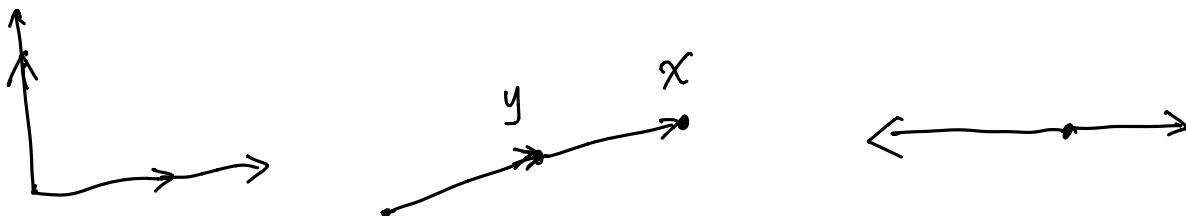
- For example; two images  $x, y$  showing the same scene with different lights.

For simplicity, we may assume, say,  $y = \frac{1}{2}x$  (the first image is with 100% light, and the second 50%)

Then  $\|x-y\| = \frac{1}{2}\|x\|$  — not small.

but  $x, y$  are from the same scene.

- We use inner product for "angle" of two vectors, which is "scaling" insensitive.



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

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

Remark: 1. By ② and ③,  $\langle x, \alpha y_1 + \beta y_2 \rangle = \alpha \langle x, y_1 \rangle + \beta \langle x, y_2 \rangle \quad \forall \alpha, \beta \in \mathbb{R}, \quad x, y_1, y_2 \in V$ .

Therefore,  $\langle \cdot, \cdot \rangle$  is a bi-linear function, i.e., it is linear with respect to one of the variables (with the other fixed).

2. For inner product of vector spaces on  $\mathbb{C}$ , we only need to change ③ to

③'  $\langle x, y \rangle = \overline{\langle y, x \rangle}$ , where  $\overline{\cdot}$  stands for complex conjugate

Example 1:  $\mathbb{R}^n$  is a vector space. We can define an inner product as  
 $\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \dots + x_n y_n$ , where  $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$  and  $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ .  
 $(\equiv x^T y)$

Example 2: Another inner product in  $\mathbb{R}^n$  is as follows.

Let  $A \in \mathbb{R}^{n \times n}$  be a symmetric positive definite (spd)

(Recall spd means:  $A^T = A$  and  $x^T A x > 0 \quad \forall x \neq 0$ )

Then  $\langle x, y \rangle_A = x^T A y$  defines an inner product in  $\mathbb{R}^n$ , because

$\text{if } y=0 \Rightarrow 0$       ①  $\langle x, x \rangle_A = x^T A x \geq 0$  and  $\langle x, x \rangle_A = 0 \Leftrightarrow x^T A x = 0 \Leftrightarrow x = 0$ .

$\text{if } x \neq 0$       ②  $\langle (\alpha x_1 + \beta x_2), y \rangle_A = ((\alpha x_1 + \beta x_2))^T A y = \alpha x_1^T A y + \beta x_2^T A y$   
 $= \alpha \langle x_1, y \rangle_A + \beta \langle x_2, y \rangle_A$ .

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

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

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

- Define

$$\begin{aligned} \langle A, B \rangle &\stackrel{\text{def}}{=} \sum_{i=1}^m \sum_{j=1}^n a_{ij} b_{ij} \quad \forall A, B \in \mathbb{R}^{m \times n} \\ &= \text{trace}(A^T B) \\ &= \text{trace}(B^T A) \end{aligned}$$

This is an inner product on  $\mathbb{R}^{m \times n}$ .

Example 4: For two infinite sequences

$$a = (a_i)_{i=1,2,3,\dots}$$

$$b = (b_i)_{i=1,2,3,\dots}$$

Then  $\langle a, b \rangle = \sum_{i=1}^{\infty} a_i b_i$  defines an inner product.

Example 5: In  $C[a,b]$ , we can define an inner product as

$$\langle f, g \rangle = \int_a^b f(x) g(x) dx, \quad \forall f, g \in C[a, b].$$

### § 3.2. Properties of inner product, orthogonality

Cauchy-Schwartz Inequality:

If  $\langle \cdot, \cdot \rangle$  is an inner product on  $V$ , then, for any  $x, y \in V$ ,

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

$[\lvert \langle x, y \rangle \rvert \leq \sqrt{\langle x, x \rangle} \sqrt{\langle y, y \rangle}]$

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

Proof • if  $y=0$ . Then obviously

$$|\langle x, y \rangle|^2 = |\langle x, 0 \rangle|^2 = 0 \leq 0 = \langle x, x \rangle \langle y, y \rangle$$

• It remains to prove the inequality with  $y \neq 0$ .

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

$$\begin{aligned} 0 \leq \langle x + \lambda y, x + \lambda y \rangle &= \langle x, x \rangle + \lambda \langle y, x \rangle + \lambda \langle x, y \rangle + \lambda^2 \langle y, y \rangle \\ &= \langle x, x \rangle + 2\lambda \langle x, y \rangle + \lambda^2 \langle y, y \rangle \end{aligned}$$

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

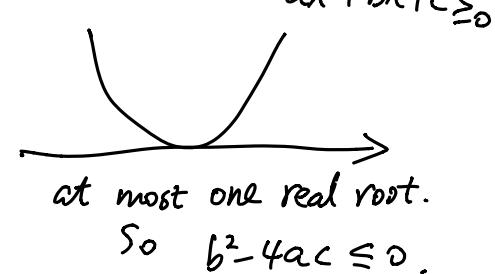
Since  $y \neq 0$ ,  $\langle y, y \rangle > 0$ .  $\Rightarrow f(\lambda) \geq 0$

So,  $f(\lambda)$  a quadratic function of  $\lambda$  that takes non-negative values only.

There is at most one root of the quadratic function.

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

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



Remark:  $|\langle x, y \rangle|^2 = \langle x, x \rangle \langle y, y \rangle \Leftrightarrow \exists \alpha \in \mathbb{R} \text{ s.t. } x = \alpha y$

Part 1:  $\Rightarrow$  • Next when the equality holds true, i.e.,  $\langle x, y \rangle^2 = \langle x, x \rangle \langle y, y \rangle$ ,

1° if  $y=0$ , then obviously  $y=\alpha x$ , where  $\alpha=0$ .

2° if  $y \neq 0$ , and there is exactly one real root of  $f(\lambda)$ .

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

$$\Rightarrow \langle x + \lambda y, x + \lambda y \rangle = 0 \Rightarrow x + \lambda y = 0 \Rightarrow x = \alpha y \text{ with } \alpha = -\lambda.$$

Part 2:  $\Leftarrow$  Finally, we show that if  $x = \alpha y$  or  $y = \alpha x$ , then  $\langle x, y \rangle^2 = \langle x, x \rangle \langle y, y \rangle$ .

By direct calculation:

$$x = \alpha y \Rightarrow \langle x, y \rangle^2 = \langle \alpha y, y \rangle^2 = \alpha^2 \langle y, y \rangle = \underbrace{\langle \alpha y, \alpha y \rangle}_{\alpha^2} \underbrace{\langle y, y \rangle}_{1}$$

$$y \Rightarrow \alpha x \Rightarrow \langle x, y \rangle^2 = \langle x, \alpha x \rangle^2 = \alpha^2 \langle x, x \rangle = \underbrace{\langle \alpha x, \alpha x \rangle}_{\alpha^2} \underbrace{\langle x, x \rangle}_{1}$$



With the Cauchy-Schwartz inequality, we can show that:

$$\|x\| = \sqrt{\langle x, x \rangle}$$
 defines a norm.

Called "norm induced by the inner product".

Proof. ①  $\|x\| = \sqrt{\langle x, x \rangle} \geq 0$  and  $\|x\| = \sqrt{\langle x, x \rangle} = 0 \Leftrightarrow x = 0$ .

②  $\|\alpha x\| = \sqrt{\langle \alpha x, \alpha x \rangle} = \sqrt{\alpha^2 \langle x, x \rangle} = |\alpha| \|x\|$

③  $\|x+y\|_2^2 = \langle x+y, x+y \rangle = \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle$   
 $= \|x\|^2 + \|y\|^2 + 2\langle x, y \rangle$   
 $\stackrel{C-S}{\leq} \|x\|^2 + \|y\|^2 + 2\|x\| \|y\| \quad \left. \begin{array}{l} \text{Note that Cauchy-Schwartz becomes} \\ |\langle x, y \rangle| \leq \|x\| \|y\| \end{array} \right)$   
 $= (\|x\| + \|y\|)^2$



We call the norm  $\|x\| = \sqrt{\langle x, x \rangle}$

the norm induced by inner product  $\langle \cdot, \cdot \rangle$ .

This will be considered as a default norm in an inner product space.

Cauchy-Schwartz is restated as:

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

Example 1:  $\mathbb{R}^n$  with inner product  $\langle x, y \rangle = x^T y$ .

The induced norm is

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{x^T x} = \sqrt{\sum_{i=1}^n x_i^2} = \|x\|_2$$

Example 2:  $\mathbb{R}^n$  with inner product  $\langle x, y \rangle_A = x^T A y$ , where  $A$  is SPD.

The induced norm is

$$\|x\|_A = \sqrt{\langle x, x \rangle_A} = \sqrt{x^T A x} = \sqrt{\sum_{i,j} a_{ij} x_i x_j}$$

Example 3: The p-norm in  $\mathbb{R}^n$ ,

$\|x\|_p$ ,  $p \neq 2$ , are not induced by inner products.

Example 4:  $\mathbb{R}^{m \times n}$  with inner product

$$\langle A, B \rangle = \sum_{i,j} a_{ij} b_{ij}$$

The induced norm is

$$\|A\| = \sqrt{\langle A, A \rangle} = \left( \sum_{i,j} a_{ij}^2 \right)^{1/2} \equiv \|A\|_F - \text{the Frobenius norm.}$$

"Angles" in inner product spaces.

The equality " $=$ " is attained if and only if  $x$  and  $y$  are aligned exactly, which should have the least angle. Therefore, we use the ratio of the two sides of Cauchy-Schwartz

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

to quantize the closeness to exact alignment. Hence to define the angles between  $x$  and  $y$ .

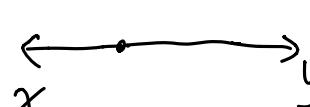
- When  $x = \alpha y$  with  $\alpha > 0$ ,

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


Since  $x$  and  $y$  are in the same direction,

It is naturally to define  $\langle x, y \rangle = 0$

- When  $x = \alpha y$  with  $\alpha < 0$

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


Since  $x$  and  $y$  are in the opposite direction,  
it is naturally to define  $\langle x, y \rangle = \pi$ .

- We define

$$\cos \angle(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|}$$

or, equivalently,  $\angle(x, y) = \arccos \frac{\langle x, y \rangle}{\|x\| \|y\|}$

This definition coincides with the above two cases and the vectors in  $\mathbb{R}^2$  or  $\mathbb{R}^3$  equipped with the standard inner product

**Orthogonality:** Let  $V$  be an inner product space.

- When the angle of  $x$  and  $y$  is  $\frac{\pi}{2}$ , we call they are orthogonal, denoted by  $x \perp y$ , i.e.,

$$x \perp y \text{ if } \langle x, y \rangle = 0$$

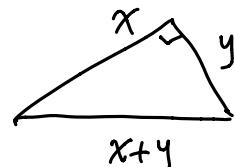
- When  $x \perp y$ , they are least relevant.

- Pythagoras' theorem: Let  $x, y$  be two vectors in an inner product space.

$$\text{If } x \perp y, \text{ then } \|x+y\|^2 = \|x\|^2 + \|y\|^2$$

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

$$\begin{aligned} &= \langle x, x \rangle + \underbrace{\langle x, y \rangle + \langle y, x \rangle}_{=0} + \langle y, y \rangle \\ &= \|x\|^2 + \|y\|^2 \quad \text{⊗} \end{aligned}$$



- Parallelogram Law: If  $x, y \in H$  — an inner product space,

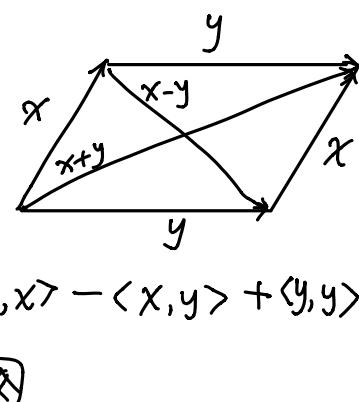
$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$$

proof.  $\|x+y\|^2 + \|x-y\|^2$

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

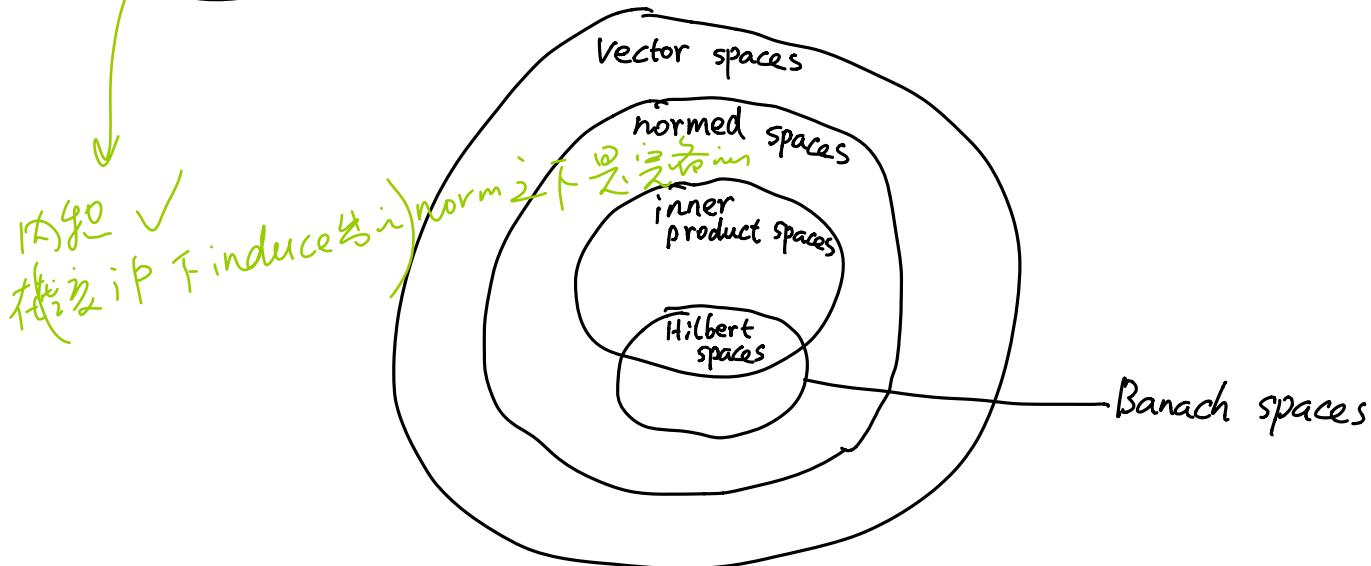
$$= \langle x, x \rangle + \langle y, x \rangle + \langle x, y \rangle + \langle y, y \rangle + \langle x, x \rangle - \langle y, x \rangle - \langle x, y \rangle + \langle y, y \rangle$$

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



Hilbert space:

A Hilbert Space is a Banach space in which an inner product is equipped and the norm is induced by the inner product.



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

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

is a Hilbert space.

Example 2:  $\mathbb{R}^n$  with inner product

$$\langle x, y \rangle_A = x^T A y, \quad \text{where } A \text{ is an spd matrix}$$

is a Hilbert space. The norm on this space is

$$\|x\|_A = (x^T A x)^{\frac{1}{2}}.$$

Example 3:  $\mathbb{R}^{m \times n}$  with inner product  $\langle A, B \rangle = \text{trace}(A^T B)$   
is a Hilbert space.

Example 4:  $l_2 = \{a \mid a \text{ is an infinite sequence and } \|a\|_2 < +\infty\}$   
with inner product  $\langle a, b \rangle = \sum_{i=1}^{\infty} a_i b_i$   
is a Hilbert space.

Example 5:  $C[a, b]$  with inner product

$$\langle f, g \rangle = \int_a^b f(x)g(x)dx$$

is NOT a Hilbert space, because it is not complete.

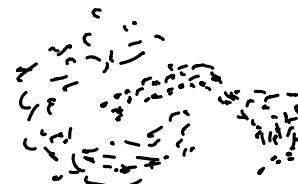
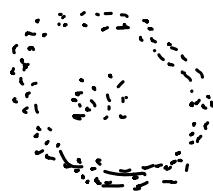
To complete  $C[a, b]$  under the norm  $\| \cdot \| = (\langle \cdot, \cdot \rangle)^{\frac{1}{2}}$ ,

we need to extend the Riemannian integral to the so-called Lebesgue integral, and the resulting Hilbert space is  $L^2(a, b)$ .

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

### § 3.3 Case Study: Kernel trick, Kernel k-means

The k-means will not work for the following examples



Recall in a clustering, we want to group  $x_1, x_2, \dots, x_n \in \mathbb{R}^n$  into  $K$  groups.

The k-means algorithm works like:

Initialize  $z_1, z_2, \dots, z_K$

Step 1: Given  $z_1, z_2, \dots, z_K$ , update the groups  $G_1, \dots, G_K$  by

① for each  $x_i$ , assign  $c_i$ , the group that  $x_i$  belongs to, by

$$c_i = \arg \min_{j \in \{1, \dots, K\}} \|x_i - z_j\|_2^2$$

② Then  $G_j = \{i | c_i = j\}$ , for  $j = 1, 2, \dots, K$ .

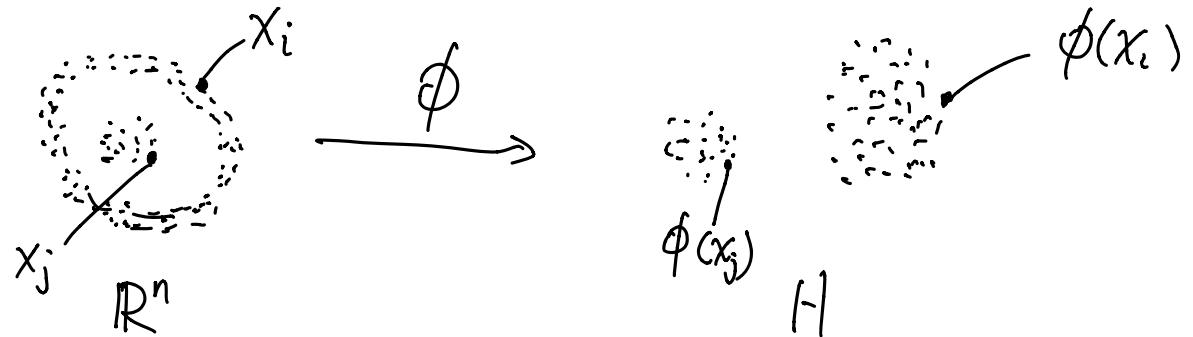
Step 2: Given  $G_1, \dots, G_K$ , update their representatives by

$$z_j = \frac{1}{|G_j|} \left( \sum_{i \in G_j} x_i \right), \text{ for } j = 1, 2, \dots, K.$$

To modify k-means to those "curved" data sets in  $\mathbb{R}^n$

we use a transform to "uncurve" the data sets in a Hilbert space.

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



$\phi(x_i)$  is called the feature of  $x_i$ .

$\phi$  is called the feature map.

$H$  is called the feature space.

Then we apply k-means to  $\phi(x_1), \phi(x_2), \dots, \phi(x_N)$  in  $H$ .

Let  $z_1, \dots, z_k$  be the representative vectors in  $H$ .

Step 1: Given  $z_1, \dots, z_k$ ,

$$c_i = \arg \min_{j \in \{1, 2, \dots, k\}} \|\phi(x_i) - z_j\|^2, \quad i=1, 2, \dots, N.$$

and

$$G_j = \{i \mid c_i = j\}, \quad j=1, 2, \dots, k.$$

Step 2: Given  $G_1, \dots, G_k$ ,

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

Repeat.

However, finding the feature map  $\phi$  is not easy, because  $\phi$  depends on the shape of  $x_1, x_2, \dots, x_N$ , which generally is very complicated.

The good news is that:

[There is no need to know  $\phi$  explicitly in the k-means algorithm.]

This is seen as in below:

- First of all, since we care only the groups of  $x_1, \dots, x_N$ , we only need to know  $G_1, \dots, G_k$ . The representatives  $z_1, \dots, z_k$  are only intermediate.

Therefore, we can eliminate  $z_1, \dots, z_k$  in the k-means algorithm.

$$\begin{aligned} \textcircled{1} \quad c_i &= \arg \min_{j \in \{1, 2, \dots, k\}} \left\| \phi(x_i) - \frac{1}{|G_j|} \sum_{l \in G_j} \phi(x_l) \right\|^2, \quad i=1,2,\dots,N. \\ \textcircled{2} \quad G_j &\leftarrow \{i \mid c_i = j\}, \quad j=1,2,\dots,k. \end{aligned}$$

- Now, only ① involves the feature mapping  $\phi$ . Since  $H$  is a Hilbert space, we can expand the norm in ① by

$$\begin{aligned} &\left\| \phi(x_i) - \frac{1}{|G_j|} \sum_{l \in G_j} \phi(x_l) \right\|^2 \\ &= \left\langle \phi(x_i) - \frac{1}{|G_j|} \sum_{l \in G_j} \phi(x_l), \phi(x_i) - \frac{1}{|G_j|} \sum_{l \in G_j} \phi(x_l) \right\rangle \\ &= \langle \phi(x_i), \phi(x_i) \rangle - \frac{2}{|G_j|} \sum_{l \in G_j} \langle \phi(x_i), \phi(x_l) \rangle \\ &\quad + \frac{1}{|G_j|^2} \sum_{l_1 \in G_j} \sum_{l_2 \in G_j} \langle \phi(x_{l_1}), \phi(x_{l_2}) \rangle \end{aligned}$$

We see that

Only inner products in the feature space are involved.

Therefore,

An explicit expression of  $\phi$  is NOT necessary.

Kernel trick:

Instead of defining  $\phi(x)$  explicitly, we define a kernel function  $K(x, y)$ , which satisfies  $K(x, y) = \langle \phi(x), \phi(y) \rangle$ .

The kernel function  $K(x, y)$  can be seen as an explicit quantification of similarity of  $x$  and  $y$ .

Not all function  $K(x, y)$  satisfying  $K(x, y) = \langle \phi(x), \phi(y) \rangle$  for some feature map  $\phi$ . (Example:  $K(x, y) = -1$  is not good because it violates with  $\langle \phi(x), \phi(x) \rangle \geq 0$ ) Which function  $K(x, y)$  can be an inner product  $\langle \phi(x), \phi(y) \rangle$  for some feature mapping  $\phi$ ?

- First of all, inner product property

$$K(x, y) = \langle \phi(x), \phi(y) \rangle \stackrel{?}{=} \langle \phi(y), \phi(x) \rangle = K(y, x).$$

(We say  $K(\cdot, \cdot)$  is symmetric if  $K(x, y) = K(y, x)$  for all  $x, y \in \mathbb{R}^n$ )

- Secondly, let  $y_1, y_2, \dots, y_m$  be  $m$  vectors in  $\mathbb{R}^n$ , then,

for any  $C = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{pmatrix} \in \mathbb{R}^m$ ,

$$\left\langle \sum_{i=1}^m c_i \phi(y_i), \sum_{i=1}^m c_i \phi(y_i) \right\rangle \geq 0 \quad (\text{By inner product property})$$

On the other hand,

$$\begin{aligned} & \left\langle \sum_{i=1}^m c_i \phi(y_i), \sum_{j=1}^m c_j \phi(y_j) \right\rangle = \left\langle \sum_{i=1}^m c_i \phi(y_i), \sum_{j=1}^m c_j \phi(y_j) \right\rangle \\ & \quad \xrightarrow{\text{By inner product}} = \sum_{i=1}^m \sum_{j=1}^m c_i c_j \langle \phi(y_i), \phi(y_j) \rangle \\ & \quad = \sum_{i=1}^m \sum_{j=1}^m c_i c_j K(y_i, y_j) \\ & \quad = C^T \begin{bmatrix} K(y_1, y_1) & K(y_1, y_2) & \cdots & K(y_1, y_m) \\ K(y_2, y_1) & K(y_2, y_2) & \cdots & K(y_2, y_m) \\ \vdots & \vdots & \ddots & \vdots \\ K(y_m, y_1) & K(y_m, y_2) & \cdots & K(y_m, y_m) \end{bmatrix} C \geq 0. \quad \forall C \in \mathbb{R}^m \end{aligned}$$

In other words, the matrix  $\begin{bmatrix} K(y_1, y_1) & K(y_1, y_2) & \cdots & K(y_1, y_m) \\ K(y_2, y_1) & K(y_2, y_2) & \cdots & K(y_2, y_m) \\ \vdots & \vdots & \ddots & \vdots \\ K(y_m, y_1) & K(y_m, y_2) & \cdots & K(y_m, y_m) \end{bmatrix}$  is symmetric positive semi-definite.

We say a function  $K(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  is symmetric positive semi-definite if:

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

ii) For any  $m$  and any vectors  $y_1, y_2, \dots, y_m \in \mathbb{R}^n$ , the matrix

$$\begin{bmatrix} K(y_1, y_1) & K(y_1, y_2) & \cdots & K(y_1, y_m) \\ K(y_2, y_1) & K(y_2, y_2) & \cdots & K(y_2, y_m) \\ \vdots & \vdots & \ddots & \vdots \\ K(y_m, y_1) & K(y_m, y_2) & \cdots & K(y_m, y_m) \end{bmatrix}$$

is symmetric positive semi-definite.

Mercer's theorem tells us that: If a kernel function  $K(\cdot, \cdot)$  is symmetric positive semi-definite, then there exists a feature map  $\phi$  such that  $K(x, y) = \langle \phi(x), \phi(y) \rangle$

Some popular kernels:

- ①  $K(x, y) = x^T y$  ( $\phi(x) = x$ . No transform)
- ②  $K(x, y) = (x^T y)^\alpha$  polynomial kernels
- ③  $K(x, y) = e^{-\frac{\|x-y\|_2^2}{\sigma^2}}$  Gaussian kernel

Kernel k-means algorithm

- choose a kernel function  $K(\cdot, \cdot)$
- Initialize  $G_1, G_2, \dots, G_K$  by, e.g., one step of k-means.

→ Set

$$c_i = \arg \min_{j \in \{1, 2, \dots, K\}} \left( K(x_i, x_i) - \frac{2}{|G_j|} \sum_{l \in G_j} K(x_i, x_l) + \frac{1}{|G_j|^2} \sum_{l \in G_j} \sum_{l' \in G_j} K(x_l, x_{l'}) \right)$$

for  $i=1, 2, \dots, N$ .

- update  $G_1, G_2, \dots, G_K$  by
- $$G_j = \{i \mid c_i = j\}, \text{ for } j=1, 2, \dots, K.$$

- go back and repeat

The kernel k-means works for some datasets for which k-means fail.

Example:

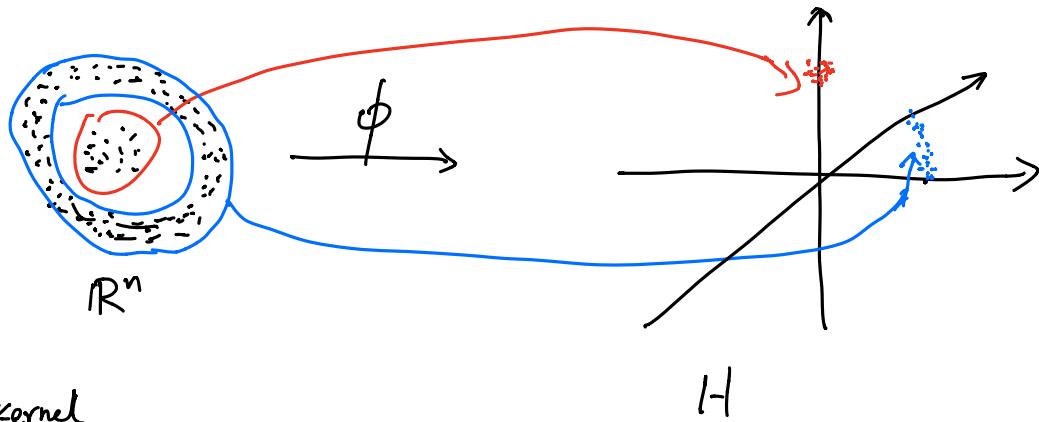


If we use Gaussian kernel  $K(x, y) = e^{-\frac{\|x-y\|^2}{\sigma^2}}$

then

- $K(x_i, x_i) = e^{-\frac{\|x_i-x_i\|^2}{\sigma^2}} = 1$ 
  - so all  $\phi(x_1), \dots, \phi(x_n)$  are on unit sphere in  $H$ .
- $K(x_i, x_j) \begin{cases} \approx 0 & \text{if } \|x_i - x_j\|_2 \text{ is large} \\ \approx 1 & \text{if } \|x_i - x_j\|_2 \text{ is small.} \end{cases}$ 
  - so  $\phi(x_i), \phi(x_j)$  are orthogonal in  $H$  if  $\|x_i - x_j\|_2$  large.
  - $\phi(x_i) \approx \phi(x_j)$  in  $H$  if  $\|x_i - x_j\|$  small.

Therefore,



Thus,  $\overset{\text{Kernel}}{\text{k-means}}$  works for this data set.

## § 3.4. Case Study: Metric Learning

- Given a set of vectors  $x_1, x_2, \dots, x_N \in \mathbb{R}^n$ , and given information that certain pairs of them are **similar** / **dissimilar**

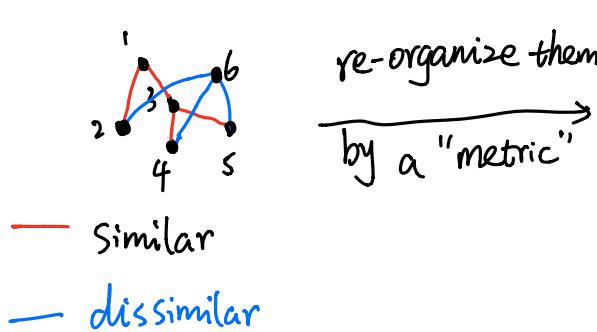
$S$ :  $(x_i, x_j) \in S$  if  $x_i$  and  $x_j$  are similar.

$D$ :  $(x_i, x_j) \in D$  if  $x_i$  and  $x_j$  are dissimilar.

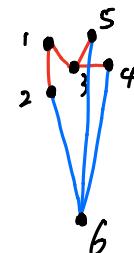
Can we find a distance metric such that

"similar" points are close to each other

and "dissimilar" points are far away from each other ?



blank unknown  
(e.g.,  $(x_2, x_3)$   
 $(x_3, x_6)$ )



So, the similarities are measured by the distance. Then we can infer that, e.g.,  $(x_2, x_3)$  are similar  $(x_3, x_6)$  are dissimilar.

- Potential applications

— "Supervised" clustering.

- Given a data set with partial clustering information

5	0	4	1
9	2	1	3
1	4	3	5
3	6	1	7

— Similar (i.e., in the same cluster)

— dissimilar (i.e., in different clusters)

- Metric learning (i.e., find a distance)
- cluster the points under the learned distance metric.

Representation:

There are many different forms of norms, even in  $\mathbb{R}^n$ .

Candidate forms of norms

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

- This class of norms is too small, because there is only one parameter ( $p$ ) in the definition of these norms.

— norms induced by the  $A$ -inner product

$$\|x\|_A = (\langle x, x \rangle_A)^{\frac{1}{2}} = (x^T A x)^{\frac{1}{2}},$$

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

- This class of norms are parametrized by an  $n \times n$  matrix  $A$  that contains enough parameters ( $n^2$  parameters)
- We use this class of norms as candidate norms for our distance metric learning.

- Then

find a distance metric  $\iff$  Find an SPD matrix  $A \in \mathbb{R}^{n \times n}$  and the distance metric is  $\|\cdot\|_A$

Remarks:

① The SPD-ness of  $A$  is usually relaxed to

$A$  is symmetric positive semi-definite (SPSD).

Correspondingly, the resulting norm will NOT satisfy

$\|x\|_A = 0 \iff x = 0$ , and other conditions in the definition of norms are satisfied; thus,  $\|\cdot\|_A$  with SPSD  $A$  is NOT exactly a norm, but a pseudo-norm.

② Given  $x, y \in \mathbb{R}^n$ , the distance induced by  $\|\cdot\|_A$

$$\|x-y\|_A = \sqrt{(x-y)^T A (x-y)}$$

is also known as *Mahalanobis distance*.

Evaluation: Which  $A$  is the best for metric learning?

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

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

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

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

So, we find the best  $A \in \mathbb{R}^{n \times n}$  is the sense of

$$\min_{\substack{A \in \mathbb{R}^{n \times n} \\ \text{is SPSD}}} \frac{\sum_{(x_i, x_j) \in S} \|x_i - x_j\|_A^2}{\sum_{(x_i, x_j) \in D} \|x_i - x_j\|_A^2}$$

Or, equivalently, we solve

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

Optimization:

The optimization is introduced later.