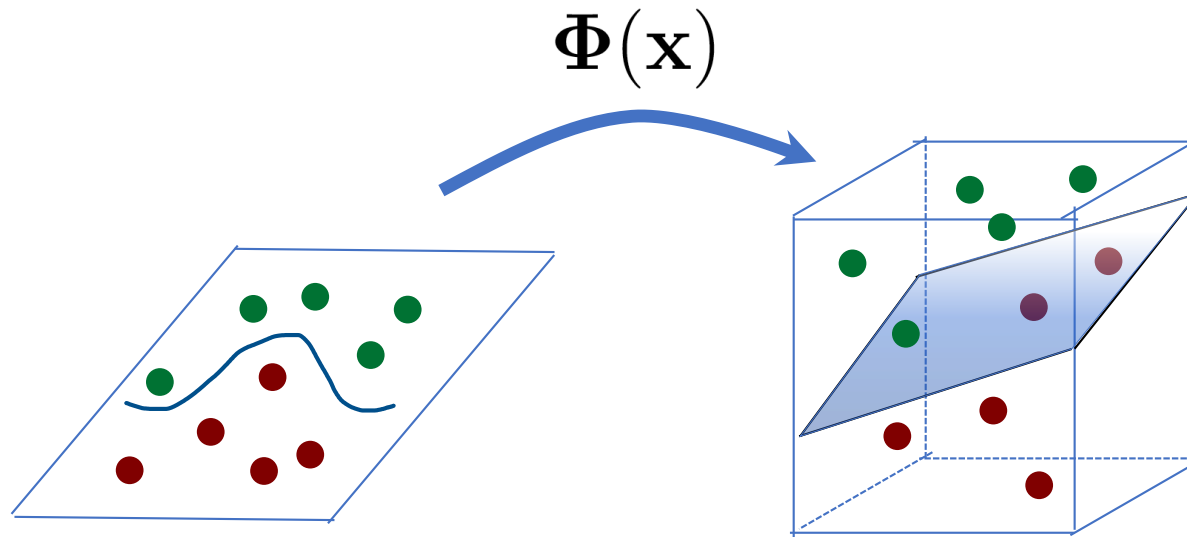


# Kernels

## Part 1

Cynthia Rudin  
Duke University

The “kernel trick” allows the SVM to map all points to a high dimensional space where points are more easily separated.



Credits: Bartlett, Scholkopf and Smola, Cristianini and Shawe-Taylor

The “kernel trick” allows the SVM to map all points to a high dimensional space where points are more easily separated.

Applies much more broadly than SVMs.

Applies to any problem where the  $x_i$ 's appear only within inner products.

Credits: Bartlett, Scholkopf and Smola, Cristianini and Shawe-Taylor

$$\mathbf{x} \longrightarrow \Phi(\mathbf{x})$$

*SVM*

Replace with  $k(\mathbf{x}_i, \mathbf{x}_l)$

Replace with  $\Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_l)$

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,l=1}^n \alpha_i \alpha_l y_i y_l \mathbf{x}_i^T \mathbf{x}_l \leftarrow \text{inner product}$$

$$\text{s.t. } 0 \leq \alpha_i \leq C, i = 1, \dots, n \text{ and } \sum_{i=1}^n \alpha_i y_i = 0$$

*SVM*

Replace with  $k(\mathbf{x}_i, \mathbf{x}_l)$

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,l=1}^n \alpha_i \alpha_l y_i y_l \mathbf{x}_i^T \mathbf{x}_l \leftarrow \text{inner product}$$

$$\text{s.t. } 0 \leq \alpha_i \leq C, i = 1, \dots, n \text{ and } \sum_{i=1}^n \alpha_i y_i = 0$$

The trick:

- You don't need to know  $\Phi$ .
- There could even be multiple  $\Phi$  corresponding to the same  $k$  (and you don't care which one you use!)

The catch: You must use a  $k(\mathbf{x}_i, \mathbf{x}_l)$  that is a valid inner product.

$SVM$

Warning:  $k$  is not just any similarity metric!

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,l=1}^n \alpha_i \alpha_l y_i y_l k(\mathbf{x}_i, \mathbf{x}_l) \leftarrow \text{inner product}$$

$$\text{s.t. } 0 \leq \alpha_i \leq C, i = 1, \dots, n \text{ and } \sum_{i=1}^n \alpha_i y_i = 0$$

## Example 1

2D to 3D

$$[x^{(1)}, x^{(2)}] \rightarrow \Phi([x^{(1)}, x^{(2)}]) = [x^{(1)2}, x^{(2)2}, x^{(1)}x^{(2)}]$$

$$\Phi(\mathbf{x})^T \Phi(\mathbf{z}) = x^{(1)2}z^{(1)2} + x^{(2)2}z^{(2)2} + x^{(1)}x^{(2)}z^{(1)}z^{(2)} = k(\mathbf{x}, \mathbf{z})$$

3D to 9D

$$\begin{aligned} [x^{(1)}, x^{(2)}, x^{(3)}] &\rightarrow \Phi([x^{(1)}, x^{(2)}, x^{(3)}]) \\ &= [x^{(1)2}, x^{(1)}x^{(2)}, x^{(1)}x^{(3)}, x^{(2)}x^{(1)}, x^{(2)2}, x^{(2)}x^{(3)}, x^{(3)}x^{(1)}, x^{(3)}x^{(2)}, x^{(3)2}] \end{aligned}$$

$$\Phi(\mathbf{x})^T \Phi(\mathbf{z}) = \text{Standard inner product in 9D} = k(\mathbf{x}, \mathbf{z})$$

### Example 3

$p$  to ?

$$k(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle_{\mathbf{R}^p}^2 = \left( \sum_{j=1}^p x^{(j)} z^{(j)} \right)^2 = \sum_{j=1}^p \sum_{\ell=1}^p x^{(j)} x^{(\ell)} z^{(j)} z^{(\ell)}.$$

2 to ?

2 to 4

$$\Phi([x^{(1)}, x^{(2)}]) = [x^{(1)2}, x^{(2)2}, x^{(1)}x^{(2)}, x^{(2)}x^{(1)}]$$

$p = 2$  is ok.

$$\Phi(\mathbf{x})^T \Phi(\mathbf{z}) = x^{(1)2} z^{(1)2} + x^{(2)2} z^{(2)2} + 2x^{(1)} x^{(2)} z^{(1)} z^{(2)} = \langle \mathbf{x}, \mathbf{z} \rangle_{\mathbf{R}^2}^2.$$

$p = 3$  is ok too! (See Example 2)... and so are the other  $p$ 's.



### Example 3

$p$  to ?

$$k(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle_{\mathbf{R}^p}^2 = \left( \sum_{j=1}^p x^{(j)} z^{(j)} \right)^2 = \sum_{j=1}^p \sum_{\ell=1}^p x^{(j)} x^{(\ell)} z^{(j)} z^{(\ell)}.$$

3D to 9D

$$\begin{aligned} [x^{(1)}, x^{(2)}, x^{(3)}] &\rightarrow \Phi([x^{(1)}, x^{(2)}, x^{(3)}]) \\ &= [x^{(1)2}, x^{(1)}x^{(2)}, x^{(1)}x^{(3)}, x^{(2)}x^{(1)}, x^{(2)2}, x^{(2)}x^{(3)}, x^{(3)}x^{(1)}, x^{(3)}x^{(2)}, x^{(3)2}] \end{aligned}$$

$$\Phi(\mathbf{x})^T \Phi(\mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle_{\mathbf{R}^3}^2$$

$p = 3$  is ok too! (See Example 2)... and so are the other  $p$ 's.

### Example 4

$$\begin{aligned} k(\mathbf{x}, \mathbf{z}) &= (\mathbf{x}^T \mathbf{z} + c)^2 = \left( \sum_{j=1}^p x^{(j)} z^{(j)} + c \right) \left( \sum_{\ell=1}^p x^{(\ell)} z^{(\ell)} + c \right) \\ &= \sum_{j=1}^p \sum_{\ell=1}^p x^{(j)} x^{(\ell)} z^{(j)} z^{(\ell)} + 2c \sum_{j=1}^p x^{(j)} z^{(j)} + c^2 \\ &= \sum_{j,\ell=1}^p (x^{(j)} x^{(\ell)}) (z^{(j)} z^{(\ell)}) + \sum_{j=1}^p (\sqrt{2c} x^{(j)}) (\sqrt{2c} z^{(j)}) + c^2, \end{aligned}$$

A possible feature map for  $p = 3$ :

$$\Phi(\mathbf{x}) = [x^{(1)2}, x^{(1)} x^{(2)}, \dots, x^{(3)2}, \sqrt{2c} x^{(1)}, \sqrt{2c} x^{(2)}, \sqrt{2c} x^{(3)}, c]$$

### Example 5

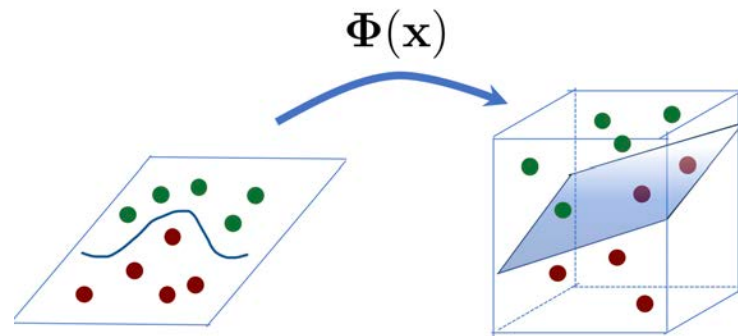
## Polynomial kernels

For any integer  $d \geq 2$

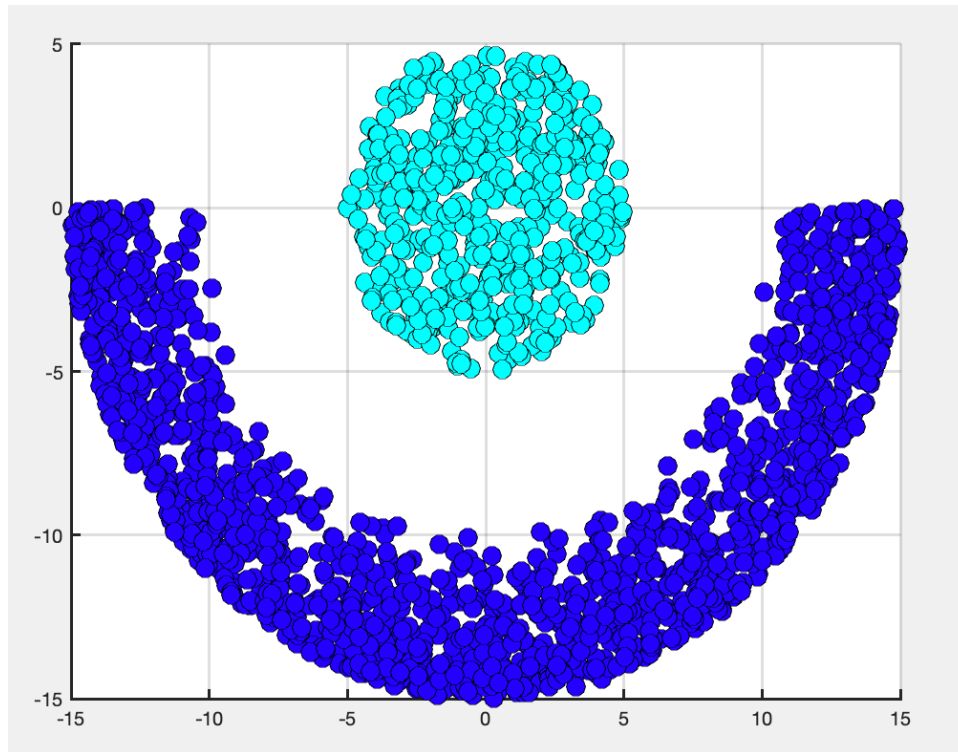
$$k(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z} + c)^d$$

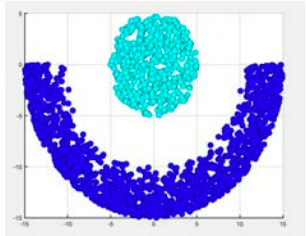
$\Phi$  includes all monomials up to and including degree  $d$ .

The decision boundary in the feature space (of course) is a hyperplane, whereas in the input space it's a polynomial of degree  $d$ .

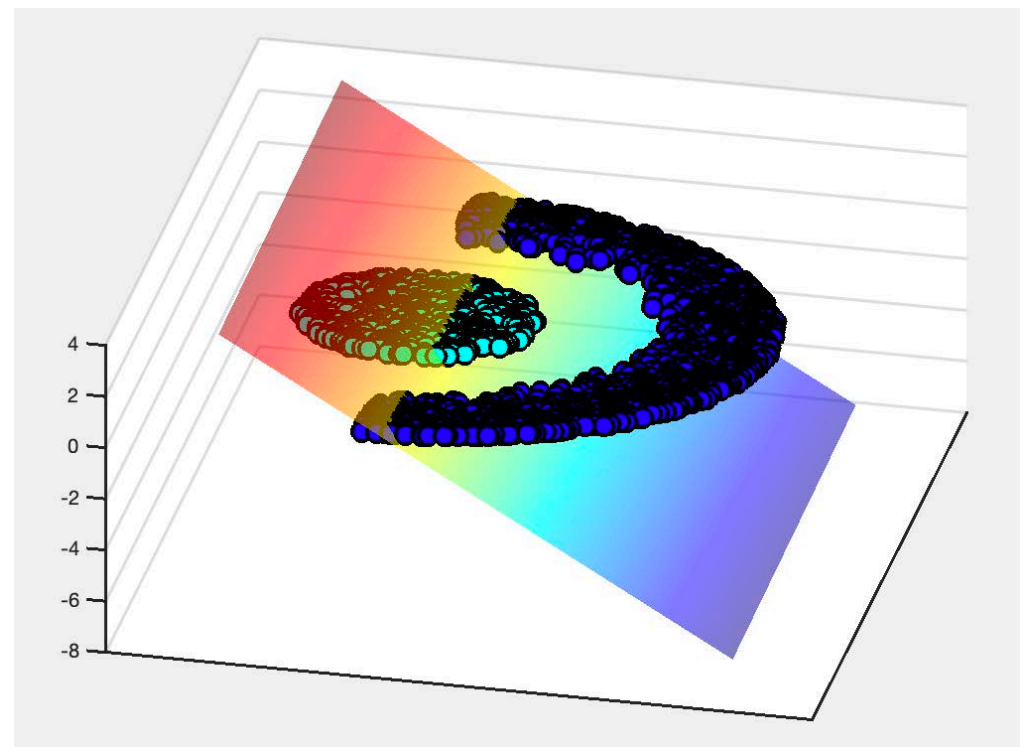
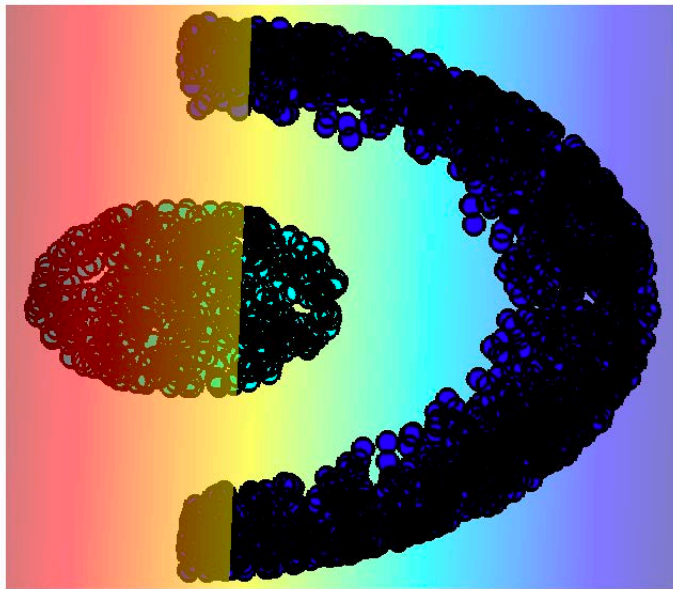


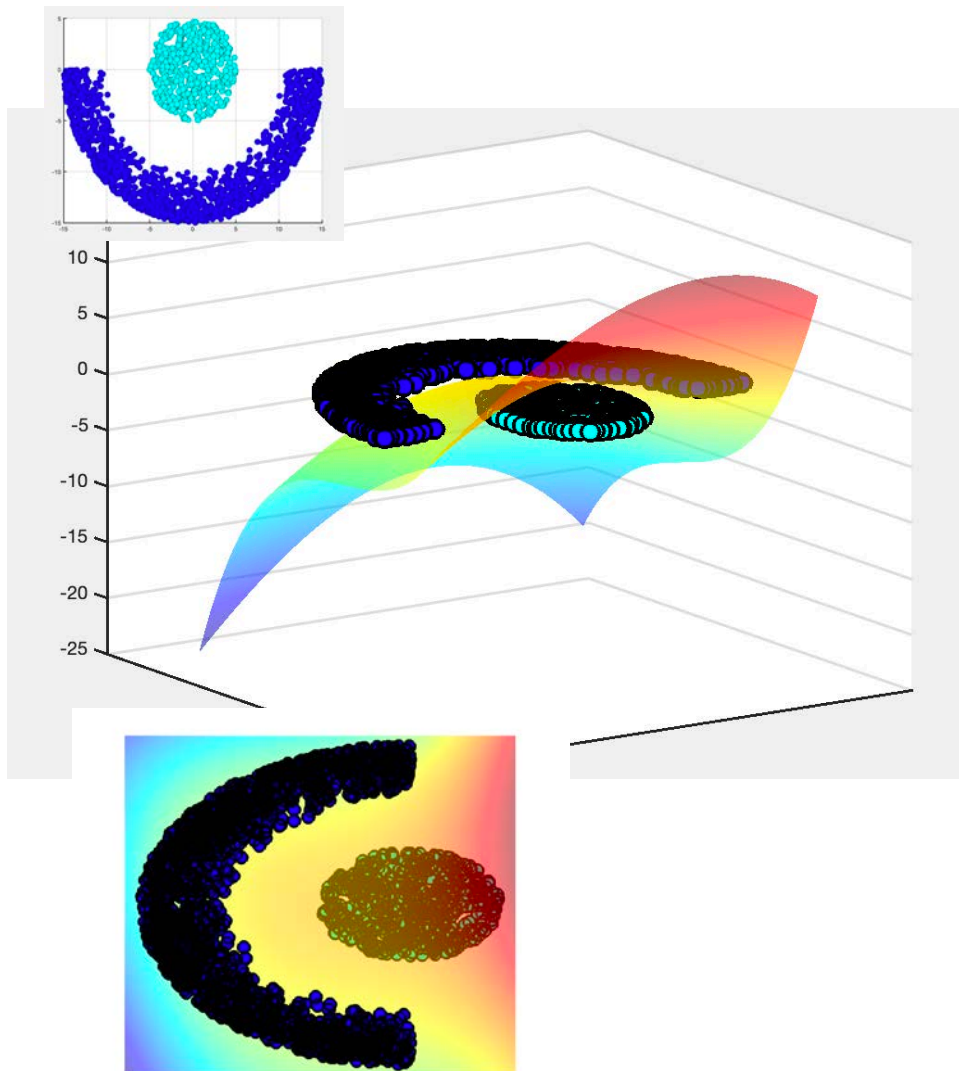
## Crescent-full-moon dataset



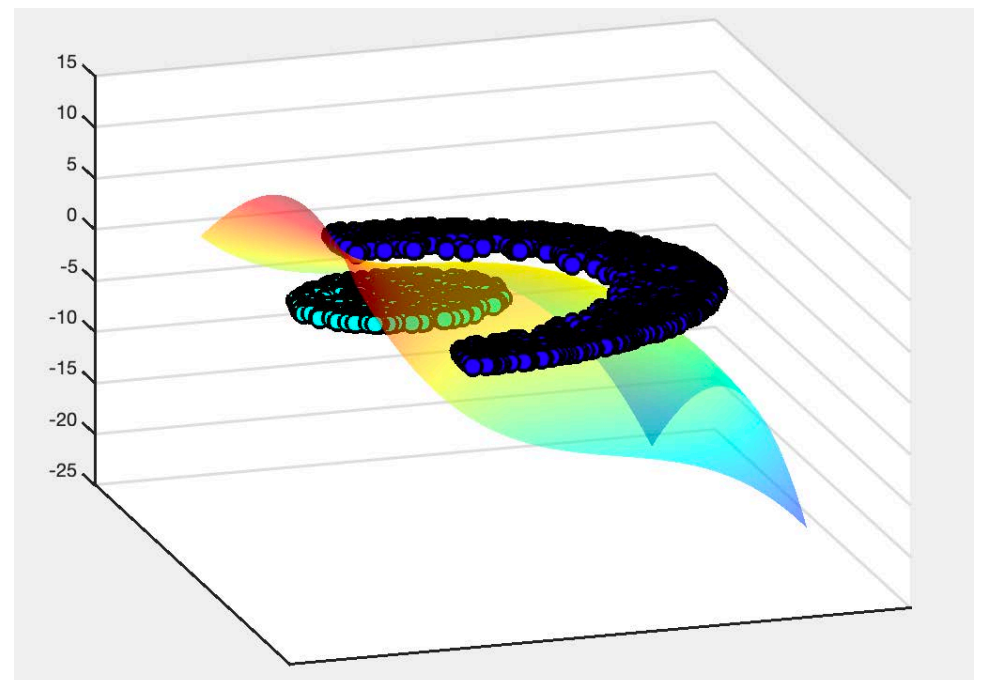


Linear Kernel (plain inner product)



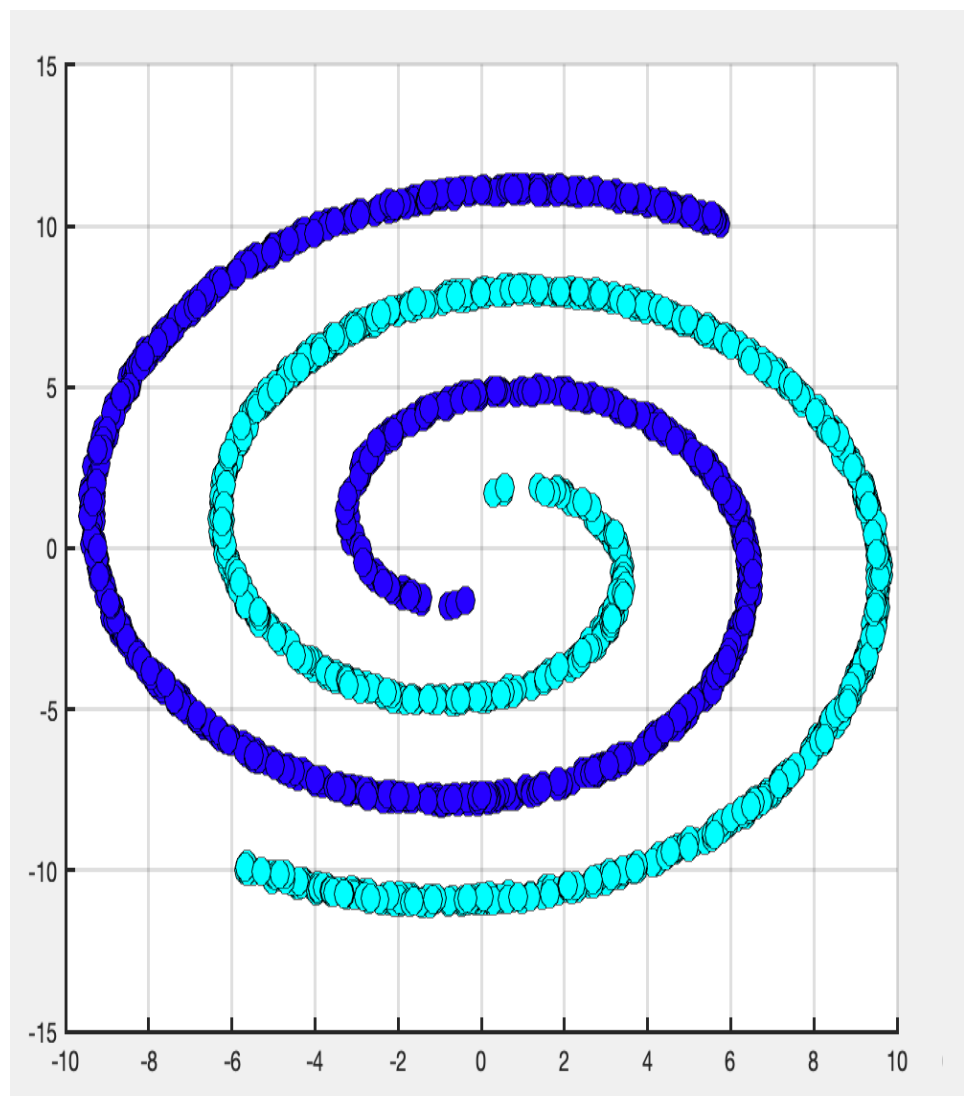


## Polynomial kernels

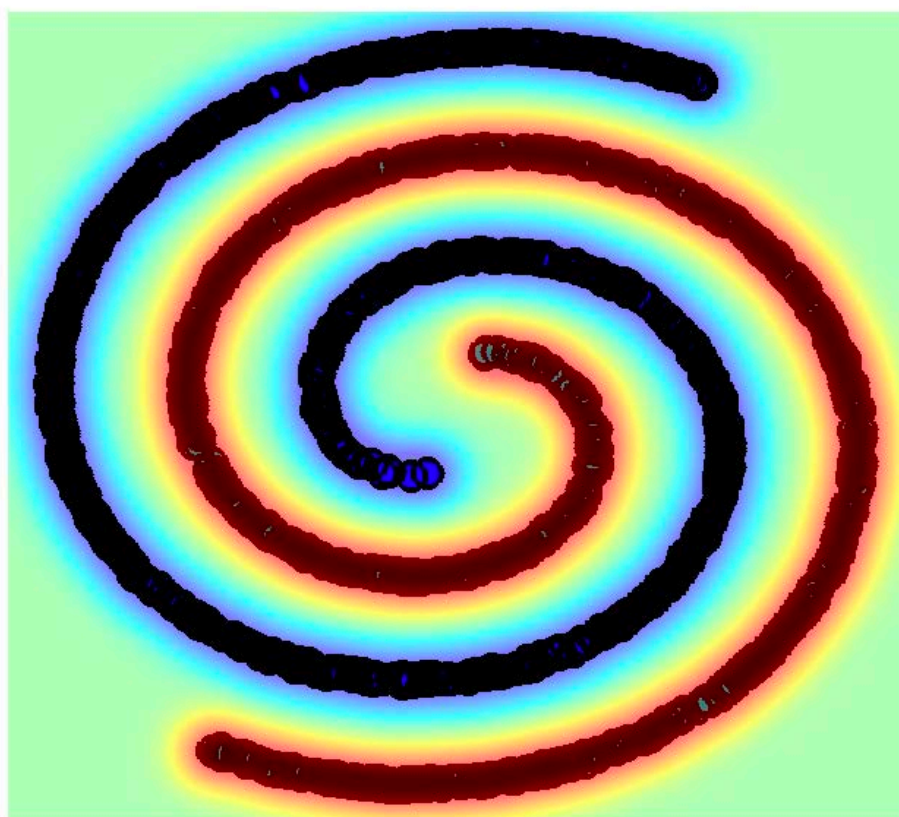


# Kernels

- Linear kernels
- Polynomial kernels
- Later: Gaussian kernels









# Kernels

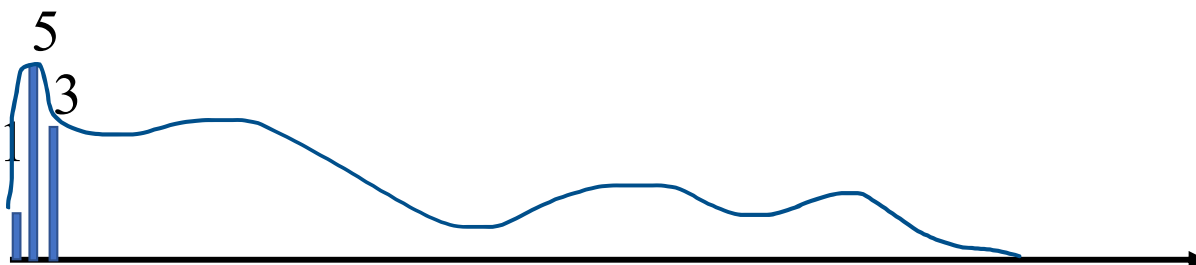
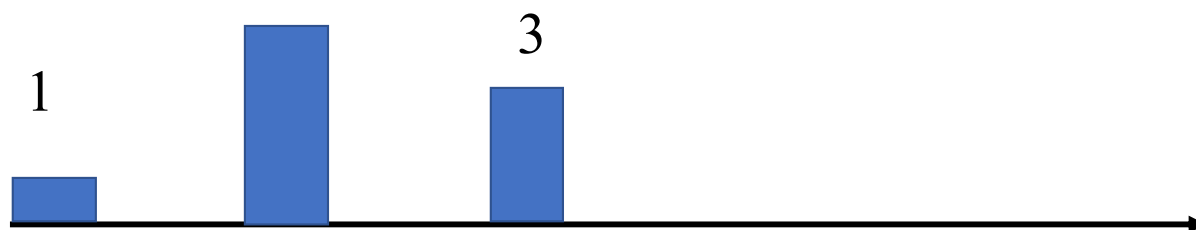
## Part 2

### Evaluating $f(x)$

Cynthia Rudin  
Duke University

Functions as infinite dimensional vectors

$[1, 5, 3]$



Even if the feature space is infinite-dimensional, the solution to the optimization problem is still easy to work with.

How do I make predictions  $f(x)$  for a test sample  $x$ ?

*SVM*

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,l=1}^n \alpha_i \alpha_l y_i y_l \text{ } k(\mathbf{x}_i, \mathbf{x}_l) \leftarrow \text{inner product}$$

$$\text{s.t. } 0 \leq \alpha_i \leq C, i = 1, \dots, n \text{ and } \sum_{i=1}^n \alpha_i y_i = 0$$

Solve the dual, get  $\alpha_i^*$ .

If using ordinary linear kernel, get the primal solution:

$$\boldsymbol{\lambda}^* = \sum_i \alpha_i^* y_i \mathbf{x}_i \quad \lambda_0^* = 1 - \boldsymbol{\lambda}^{*T} \mathbf{x}_{i_{sv}} \\ \text{(for a positive support vector)}$$

$$\begin{aligned} f(\mathbf{x}^{\text{new}}) &= \sum_j \lambda_j^* x^{\text{new}(j)} + \lambda_0^* \\ &= \sum_j \underbrace{\sum_i \alpha_i^* y_i x_i^{(j)}}_{\text{dot product}} x^{\text{new}(j)} + \lambda_0^* \quad \text{(If using kernels, do this instead.)} \\ &= \sum_i \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x}^{\text{new}} + \lambda_0^*, \quad \rightarrow \quad = \sum_i \alpha_i^* y_i k(\mathbf{x}_i, \mathbf{x}^{\text{new}}) + \lambda_0^*. \end{aligned}$$

Solve the dual, get  $\alpha_i^*$ .

If using ordinary linear kernel, get the primal solution:

$$\boldsymbol{\lambda}^* = \sum_i \alpha_i^* y_i \mathbf{x}_i \quad \lambda_0^* = 1 - \boldsymbol{\lambda}^{*T} \mathbf{x}_{i_{sv}}$$

(for a positive support vector)

$$f(\mathbf{x}^{\text{new}}) = \sum_i \alpha_i^* y_i k(\mathbf{x}_i, \mathbf{x}^{\text{new}}) + \lambda_0^*.$$

Solve the dual, get  $\alpha_i^*$ .

If using ordinary linear kernel, get the primal solution:

$$\boldsymbol{\lambda}^* = \sum_i \alpha_i^* y_i \mathbf{x}_i \quad \lambda_0^* = 1 - \boldsymbol{\lambda}^{*T} \mathbf{x}_{i_{sv}}$$

(for a positive support vector)

$$f(\mathbf{x}^{\text{new}}) = \sum_i \alpha_i^* y_i k(\mathbf{x}_i, \mathbf{x}^{\text{new}}) + \lambda_0^*.$$

$$\lambda_0^* = 1 - \boldsymbol{\lambda}^{*T} \mathbf{x}_{i_{sv}} = 1 - \left( \sum_i \alpha_i^* y_i \mathbf{x}_i \right)^T \mathbf{x}_{i_{sv}} = 1 - \sum_i \alpha_i^* y_i \mathbf{x}_i^T \cdot \mathbf{x}_{i_{sv}}$$

Evaluate  $f(\mathbf{x})$  without knowing  $\Phi$

$$\rightarrow \lambda_0^* = 1 - \sum_i \alpha_i^* y_i k(\mathbf{x}_i, \mathbf{x}_{i_{sv}})$$





# Kernels

## Part 3

### Definition of Hilbert Space & RKHS

Cynthia Rudin  
Duke University

A Hilbert Space is a complete inner product space.

- allows you to think about taking inner products on functions and infinite sequences.

A Hilbert Space is a complete inner product space.

An inner product takes two elements of a vector space  $\mathcal{X}$  and outputs a number.  
It must satisfy:

### Symmetry

$$\langle u, v \rangle = \langle v, u \rangle \quad \forall u, v \in \mathcal{X}$$

### Bilinearity

$$\langle \alpha u + \beta v, w \rangle = \alpha \langle u, w \rangle + \beta \langle v, w \rangle \quad \forall u, v, w \in \mathcal{X}, \forall \alpha, \beta \in \mathbf{R}$$

### Strict Positive Definiteness

$$\langle u, u \rangle \geq 0 \quad \forall x \in \mathcal{X}$$

$$\langle u, u \rangle = 0 \iff u = 0.$$

A Hilbert Space is a complete inner product space.

### Example 1

The vector space  $\mathbf{R}^p$  with  $\langle u, v \rangle_{\mathbf{R}^p} = u^T v$

### Example 2

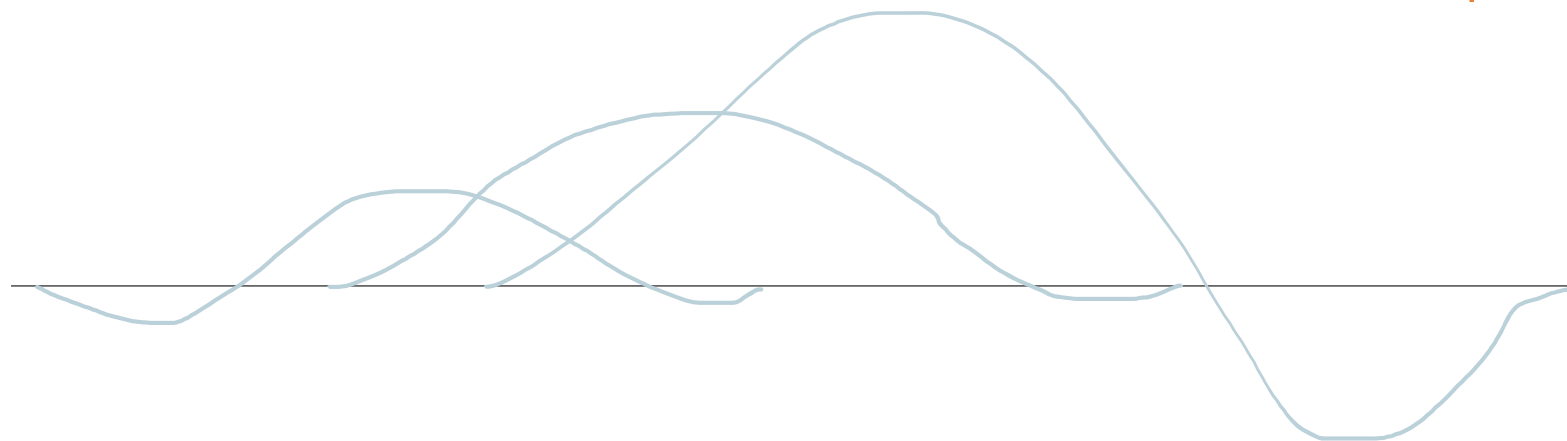
The space  $\ell_2$  of square summable sequences,

with inner product  $\langle u, v \rangle_{\ell_2} = \sum_{i=1}^{\infty} u_i v_i$

A Hilbert Space is a complete inner product space.

### Example 3

The space  $L_2(\mathcal{X}, \mu)$  of square integrable functions, that is, functions  $f$  such that  $\int f(x)^2 d\mu(x) < \infty$ , with inner product  $\langle f, g \rangle_{L_2(\mathcal{X}, \mu)} = \int \underbrace{f(x)g(x)}_{\text{product}} d\mu(x)$ .



A Reproducing Kernel Hilbert Space (RKHS) has a special function  $k$  that obeys the *reproducing property*:

$$f(x) = \underbrace{\langle k(x, \cdot), f(\cdot) \rangle}_{\text{reproducing property}}_{\mathcal{H}}$$

$k$  evaluates  $f$  at the point  $x$ .





# Kernels

## Part 4

### A finite world

Cynthia Rudin  
Duke University

Given that  $k$  is going to be an inner product, what properties should it have?

Start simple. Live in a finite-sized world.

Feature space is of size  $m$ .

$$\{x_1, \dots, x_m\}$$

This is not too unrealistic.

Predict stroke from:

age 120 values

gender 2 values

past history of strokes 2 values

blood thinner 2 values

congestive heart failure 2 values

hypertension 2 values

$$m = 120 \times 2 \times 2 \times 2 \times 2 \times 2$$

The Gram matrix of all inner products:

$$\mathbf{K} = \begin{matrix} & \begin{matrix} 1 & l & m \end{matrix} \\ \begin{matrix} m \\ i \\ 1 \end{matrix} & \left[ \begin{array}{ccc} & & \\ & k(x_i, x_l) & \\ & & \end{array} \right] \end{matrix} \quad \begin{matrix} \text{Every possible inner} \\ \text{product in the space.} \\ \\ \end{matrix} \\ = \mathbf{V} \mathbf{\Lambda} \mathbf{V}'$$

Inner products are symmetric

$$k(x_i, x_l) = k(x_l, x_i)$$

$\mathbf{K}$  must be symmetric. This means it can be diagonalized.

The Gram matrix of all inner products:

$$\mathbf{K} = \begin{matrix} & \begin{matrix} 1 & l & m \end{matrix} \\ \begin{matrix} 1 \\ i \\ m \end{matrix} & \begin{bmatrix} k(x_1, x_1) & k(x_1, x_l) & k(x_1, x_m) \\ \vdots & \ddots & \vdots \\ k(x_m, x_1) & k(x_m, x_l) & k(x_m, x_m) \end{bmatrix} \end{matrix} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}'$$

$\mathbf{V}$  is a matrix of **eigenvectors**  $\mathbf{V}_t$  (represented by vertical lines in the diagram).  
 $\mathbf{\Lambda}$  is a diagonal matrix of **eigenvalues**  $\lambda_t$  (represented by a diagonal line in the diagram).

Consider this feature map:

$$\Phi(x_i) = [\sqrt{\lambda_1}v_1^{(i)}, \dots, \sqrt{\lambda_t}v_t^{(i)}, \dots, \sqrt{\lambda_m}v_m^{(i)}]$$

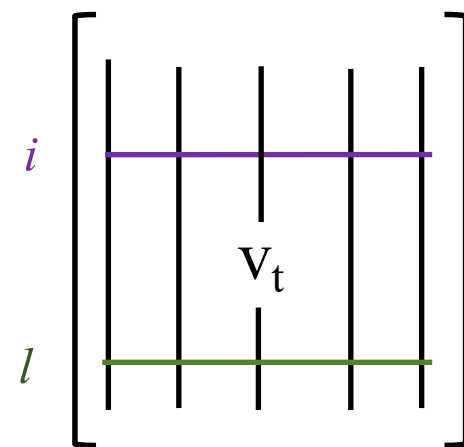
(assume nonnegative)

Write also for  $x_l$ :

$$\Phi(x_l) = [\sqrt{\lambda_1}v_1^{(l)}, \dots, \sqrt{\lambda_t}v_t^{(l)}, \dots, \sqrt{\lambda_m}v_m^{(l)}]$$

Take regular dot product in  $\mathbf{R}^m$ :

$$\langle \Phi(x_i), \Phi(x_l) \rangle_{\mathbf{R}^m} = \sum_{t=1}^m \lambda_t v_t^{(i)} v_t^{(l)}$$



Consider this feature map:

$$\Phi(x_i) = [\sqrt{\lambda_1}v_1^{(i)}, \dots, \sqrt{\lambda_t}v_t^{(i)}, \dots, \sqrt{\lambda_m}v_m^{(i)}]$$

(assume nonnegative)

Why do we assume the eigenvalues are nonnegative?

Say  $\lambda_s < 0$ . Coefficients are elements of eigenvector  $\mathbf{v}_s$

Take this special point:  $\mathbf{z} = \sum_{i=1}^m v_s^{(i)} \Phi(x_i)$

So, if  $k$  is an inner product, its Gram matrix  $\mathbf{K}$  had better be positive semidefinite!  
(nonnegative eigenvalues)

$$\begin{aligned} \|\mathbf{z}\|_2^2 &= \langle \mathbf{z}, \mathbf{z} \rangle_{\mathbf{R}^m} = \sum_i \sum_l v_s^{(i)} \Phi(x_i)^T \Phi(x_l) v_s^{(l)} = \sum_i \sum_l v_s^{(i)} K_{il} v_s^{(l)} \\ &= \mathbf{v}_s^T \mathbf{K} \mathbf{v}_s = \lambda_s < 0 \quad \text{bad} \end{aligned}$$

So far...

If  $k$  is going to be an inner product:

It must be symmetric.

Its Gram matrix  $\mathbf{K}$  must be positive semidefinite.





# Kernels

## Part 5

### Defining Kernels via Gram Matrices

Cynthia Rudin  
Duke University

So far...

If  $k$  is going to be an inner product:

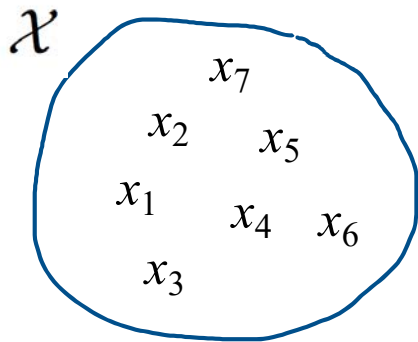
It must be symmetric.

Its Gram matrix  $\mathbf{K}$  must be positive semidefinite.

Let's officially define a kernel. We will give it properties we want.

A function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbf{R}$  is a kernel if

- $k$  is symmetric:  $k(x, z) = k(z, x)$ .
- $k$  gives rise to a positive semi-definite "Gram matrix," i.e., for any number of states  $m \in \mathbf{N}$  and any set of states  $x_1, \dots, x_m$  chosen from  $\mathcal{X}$ , the Gram matrix  $\mathbf{K}$  defined by  $K_{il} = k(x_i, x_l)$  is positive semidefinite.



$$\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{bmatrix}$$

$k(x_i, x_l)$

A convenient way to show that a matrix is positive semidefinite:

$$\forall \mathbf{c} \in \mathbf{R}^m, \mathbf{c}^T \mathbf{K} \mathbf{c} \geq 0$$

(equivalent to showing that all the eigenvalues are nonnegative)

This is useful! It allows us to prove:

☀  $k(u, u) \geq 0$  Gram Matrix of  $m = 1$ .  $\mathbf{K}$  is just  $k(u, u)$ .

$$\mathbf{c}^T \mathbf{K} \mathbf{c} \geq 0 \quad \longrightarrow \quad \mathbf{c}^2 \mathbf{K} \geq 0 \quad \longrightarrow \quad \mathbf{K} \geq 0 \quad \longrightarrow \quad k(u, u) \geq 0$$

☀  $k(u, v) \leq \sqrt{k(u, u)k(v, v)}$  (This is the Cauchy-Schwarz inequality.)

Let's show it for  $m = 2$ .

A convenient way to show that a matrix is positive semidefinite:

$$\forall \mathbf{c} \in \mathbf{R}^m, \mathbf{c}^T \mathbf{K} \mathbf{c} \geq 0$$

(equivalent to showing that all the eigenvalues are nonnegative)

$$k(u, v) \leq \sqrt{k(u, u)k(v, v)} \quad (\text{This is the Cauchy-Schwarz inequality.})$$

Let's show it for  $m = 2$ .

$$\mathbf{K} = \begin{pmatrix} k(u, u) & k(u, v) \\ k(v, u) & k(v, v) \end{pmatrix} \quad \text{Choose } \mathbf{c} = \begin{bmatrix} k(v, v) \\ -k(u, v) \end{bmatrix}$$

Because  $\mathbf{K}$  is positive semidefinite:

$$0 \leq \mathbf{c}^T \mathbf{K} \mathbf{c} = [k(v, v)k(u, u) - k(u, v)^2]k(v, v) \geq 0$$
$$k(v, v)k(u, u) \geq k(u, v)^2$$

So far...

We have a definition of kernel!

It is symmetric and gives rise to positive semidefinite Gram matrices.

Now we can use them to define a RKHS.



# Kernels

## Part 6

### Defining Reproducing Kernel Hilbert Space

Cynthia Rudin

Duke University

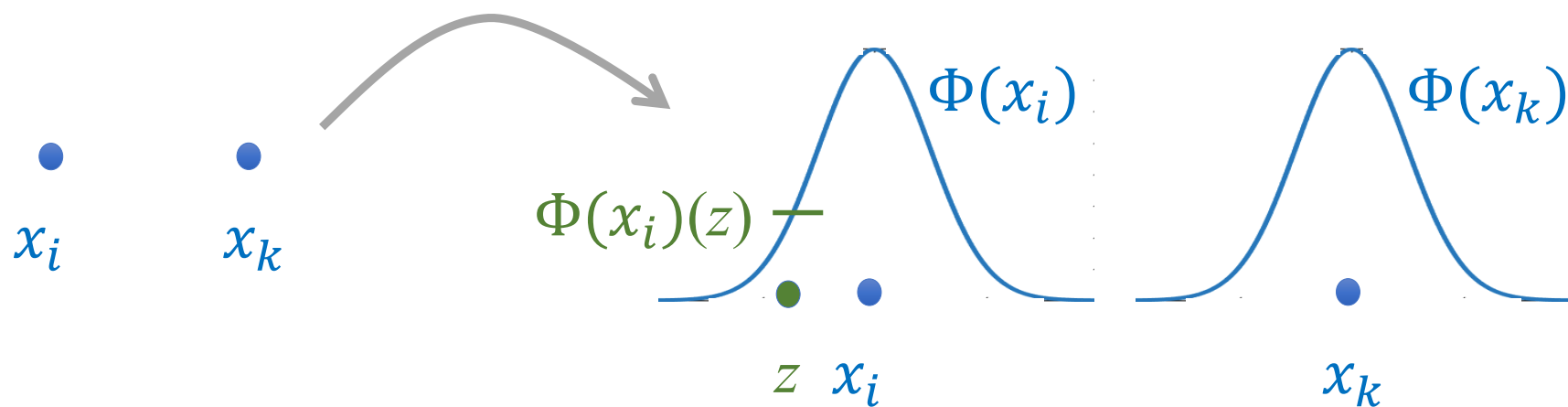


Need to do four things:

- Define a feature map  $\Phi$
- Use that to define elements of our Hilbert space
- Define inner product of the space
- Show that  $k$  is the special function needed for the reproducing property.

Define the feature map  $\Phi : \mathcal{X} \rightarrow (\text{functions from } \mathcal{X} \text{ to } \mathbf{R})$

$$\Phi : x \longmapsto k(\cdot, x) \quad (k \text{ is your choice})$$



$\Phi(x_i)(z)$  is a number. It is  $k(z, x_i)$ .

Define the feature map

Construct the **vectors** for our **vector space**,  $\Phi: x \mapsto k(\cdot, x)$

$$f(\cdot) = \sum_{i=1}^m \alpha_i k(\cdot, x_i) \leftarrow \text{“vectors”}$$

where  $m, \alpha_i$  and  $x_1 \dots x_m \in \mathcal{X}$  can be anything.

The **vector space** is:

$$\text{span}(\{\Phi(x) : x \in \mathcal{X}\}) = \left\{ f(\cdot) = \sum_{i=1}^m \alpha_i k(\cdot, x_i) : m \in \mathbf{N}, x_i \in \mathcal{X}, \alpha_i \in \mathbf{R} \right\}$$

The **inner product** is:

$$\begin{aligned} g(\cdot) &= \sum_{j=1}^{m'} \beta_j k(\cdot, x'_j) \\ f(\cdot) &= \sum_{i=1}^m \alpha_i k(\cdot, x_i) \end{aligned} \quad \begin{array}{c} \text{ } \\ \text{ } \end{array} \rightarrow \langle f, g \rangle_{H_k} = \sum_{i=1}^m \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)$$

$$\begin{aligned}
 g(\cdot) &= \sum_{j=1}^{m'} \beta_j k(\cdot, x'_j) \\
 f(\cdot) &= \sum_{i=1}^m \alpha_i k(\cdot, x_i)
 \end{aligned}
 \quad \xrightarrow{\quad} \quad
 \langle f, g \rangle_{H_k} = \sum_{i=1}^m \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)$$

Is it well-defined?

It's symmetric, since  $k$  is symmetric:  $\langle g, f \rangle_{H_k} = \sum_{j=1}^{m'} \sum_{i=1}^m \beta_j \alpha_i k(x'_j, x_i) = \langle f, g \rangle_{H_k}$ .

It's bilinear:

$$\begin{aligned}
 \langle f_1 + f_2, g \rangle_{H_k} &\stackrel{\star}{=} \sum_{j=1}^{m'} \beta_j (f_1(x'_j) + f_2(x'_j)) \\
 &= \sum_{j=1}^{m'} \beta_j f_1(x'_j) + \sum_{j=1}^{m'} \beta_j f_2(x'_j) \\
 &\stackrel{\star}{=} \langle f_1, g \rangle_{H_k} + \langle f_2, g \rangle_{H_k}
 \end{aligned}$$

$$\langle f, g \rangle_{H_k} \stackrel{\star}{=} \sum_{j=1}^{m'} \beta_j \underbrace{\sum_{i=1}^m \alpha_i k(x_i, x'_j)}_{f(x'_j)} = \sum_{j=1}^{m'} \beta_j f(x'_j)$$

Can do same for other side:

$$\langle f, g_1 + g_2 \rangle_{H_k} = \langle f, g_1 \rangle_{H_k} + \langle f, g_2 \rangle_{H_k}$$

$$\begin{aligned}
 g(\cdot) &= \sum_{j=1}^{m'} \beta_j k(\cdot, x'_j) \\
 f(\cdot) &= \sum_{i=1}^m \alpha_i k(\cdot, x_i)
 \end{aligned}
 \quad \begin{array}{c} \bullet \\ \nearrow \\ \bullet \end{array} \quad \langle f, g \rangle_{H_k} = \sum_{i=1}^m \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)$$

Is it well-defined?

It's strictly positive definite:

$$\langle f, f \rangle_{H_k} = \sum_{i,j=1}^m \alpha_i \alpha_j k(x_i, x_j) = \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \geq 0$$

Because  $k$  is a kernel,  $\mathbf{K}$  is a positive semidefinite Gram matrix

So we got positive semidefinite...

$$\begin{aligned}
 g(\cdot) &= \sum_{j=1}^{m'} \beta_j k(\cdot, x'_j) \\
 f(\cdot) &= \sum_{i=1}^m \alpha_i k(\cdot, x_i)
 \end{aligned}
 \quad \cdot \quad \rightarrow \quad \langle f, g \rangle_{H_k} = \sum_{i=1}^m \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)$$

Interlude

RKHS

$$\begin{aligned}
 &k(\cdot, x) \\
 f(\cdot) &= \sum_{i=1}^m \alpha_i k(\cdot, x_i)
 \end{aligned}
 \quad \cdot \quad \rightarrow \quad \langle k(\cdot, x), f \rangle_{H_k}$$

Reproducing property!

$$\begin{aligned}
 &k(\cdot, x) \\
 &k(\cdot, x')
 \end{aligned}
 \quad \cdot \quad \rightarrow \quad \langle k(\cdot, x), k(\cdot, x') \rangle_{H_k}$$

A reproducing kernel!

$$\begin{aligned}
 g(\cdot) &= \sum_{j=1}^{m'} \beta_j k(\cdot, x'_j) \\
 f(\cdot) &= \sum_{i=1}^m \alpha_i k(\cdot, x_i)
 \end{aligned}
 \quad \Rightarrow \quad
 \langle f, g \rangle_{H_k} = \sum_{i=1}^m \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)$$

Is it well-defined?

**Last thing:**  $\langle f, f \rangle_{H_k} = 0 \Rightarrow f = 0$  for all  $x$

$$\begin{aligned}
 |f(x)|^2 &= |\langle k(\cdot, x), f \rangle_{H_k}|^2 \\
 &\leq \langle k(\cdot, x), k(\cdot, x) \rangle_{H_k} \cdot \langle f, f \rangle_{H_k} = k(x, x) \langle f, f \rangle_{H_k} = 0
 \end{aligned}$$

reproducing property

Cauchy-Schwarz  
(must have it to be an inner product)

reproducing property

For completeness, define a norm  $\|f\|_{H_k} = \sqrt{\langle f, f \rangle_{H_k}}$

And include its completion:  $H_k = \overline{\{f : f = \sum_i \alpha_i k(\cdot, x_i)\}}$

We say  $\mathcal{H}$  is a *Reproducing Kernel Hilbert Space* if there exists a  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbf{R}$ , such that

1.  $k$  has the reproducing property, i.e.,  $f(x) = \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}}$
2.  $k$  spans  $\mathcal{H}$ , that is,  $\mathcal{H} = \overline{\text{span}\{k(\cdot, x) : x \in \mathcal{X}\}}$



The RKHS I described is from the Moore-Aronszajn Theorem (1950) that states that for every positive definite function  $k(\cdot, \cdot)$  there exists a unique RKHS.

There is another way to construct an RKHS that is closer to what we did in the finite case, based on Mercer's theorem. (Think eigenvalues and eigenvectors.)

$$\langle k(\cdot, x), f \rangle_{H_k} = \sum_i \alpha_i k(x_i, x) = f(x) \quad \Rightarrow \text{Reproducing property!} \quad \Leftarrow$$

$$\langle k(\cdot, x), k(\cdot, x') \rangle_{H_k} = k(x, x') \quad \Rightarrow \text{A reproducing kernel!} \quad \Leftarrow$$



# Kernels

## Part 7

What is not a Reproducing Kernel Hilbert Space?

Cynthia Rudin  
Duke University

$L_2$ , the space of square integrable functions.

The kernel would need to be the Dirac delta function. But it is not in  $L_2$ .

$$f(x) = \int_z \underbrace{\delta(x-z)}_{\downarrow} f(z) dz.$$

$$\int_z \delta(z)^2 dz \leftarrow \text{not finite.}$$



# Kernels

## Part 8

### Representer Theorem

Cynthia Rudin  
Duke University

Start with SVM – we want to find solutions to this problem:

$$f^* = \operatorname{argmin}_{f \in H_k} R^{\text{train}}(f)$$

$$R^{\text{train}}(f) := \sum_{i=1}^n \text{hinge loss}(f(x_i), y_i) + C \|f\|_{H_k}^2$$

What about any loss function?

$$R^{\text{train}}(f) := \sum_{i=1}^n \ell(f(x_i), y_i) + C \|f\|_{H_k}^2$$

Let's even use a generic regularization term

$$R^{\text{train}}(f) := \sum_{i=1}^n \ell(f(x_i), y_i) + \Omega(\|f\|_{H_k}^2)$$

where  $\Omega$  is nondecreasing

## Representer Theorem (Kimeldorf and Wahba, 1971)

Fix a set  $\mathcal{X}$ , kernel  $k$ , and let  $H_k$  be the corresponding RKHS.

Let  $\Omega : \mathbb{R} \rightarrow \mathbb{R}$  be a nondecreasing function.

For any loss function  $\ell : \mathbb{R}^2 \rightarrow \mathbb{R}$ , the solutions of

$$f^* \in \operatorname{argmin}_{f \in H_k} \sum_{i=1}^n \ell(f(x_i), y_i) + \Omega(\|f\|_{H_k}^2)$$

can be expressed in the following form:

$$f^* = \sum_{i=1}^n \alpha_i k(x_i, \cdot).$$

To solve SVM, all we  
need are the  $\alpha_i$ 's.  
(We knew that!)

Even if we're trying to solve an optimization problem in an *infinite dimensional space*  $H_k$ ,  
where an *arbitrary* loss depends on *arbitrary*  $x_i$ 's,  
then the solution lies in the span of the  $n$  kernels centered on these  $x_i$ 's.



*Proof:* Project  $f$  onto the subspace  $\text{span}\{k(x_i, \cdot) : 1 \leq i \leq n\}$


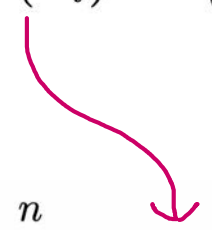

$$f = f_s + f_\perp$$

(perpendicular)  $\|f\|_{H_k}^2 = \|f_s\|_{H_k}^2 + \|f_\perp\|_{H_k}^2 \geq \|f_s\|_{H_k}^2$

(monotonicity)  $\Omega(\|f\|_{H_k}^2) \geq \Omega(\|f_s\|_{H_k}^2)$

---

$$\begin{aligned} f(x_i) &= \langle f, k(x_i, \cdot) \rangle_{H_k} = \langle f_s, k(x_i, \cdot) \rangle_{H_k} + \langle f_\perp, k(x_i, \cdot) \rangle_{H_k} \\ &= \langle f_s, k(x_i, \cdot) \rangle_{H_k} = f_s(x_i) \end{aligned}$$


$$\sum_{i=1}^n \ell(f(x_i), y_i) = \sum_{i=1}^n \ell(f_s(x_i), y_i)$$

$$\begin{array}{ccc} \text{minimize}_f & \sum_{i=1}^n \ell(f(x_i), y_i) + \Omega(\|f\|_{H_k}^2) & \\ & \parallel & \vee \\ & \sum_{i=1}^n \ell(f_s(x_i), y_i) & \Omega(\|f_s\|_{H_k}^2) \end{array}$$

Thus, to minimize, set  $f_{\perp}$  to 0.

So, the minimizer is in  $\text{span}\{k(x_i, \cdot) : 1 \leq i \leq n\}$ .



## Representer Theorem (Kimeldorf and Wahba, 1971)


Fix a set  $\mathcal{X}$ , kernel  $k$ , and let  $H_k$  be the corresponding RKHS.

Let  $\Omega : \mathbb{R} \rightarrow \mathbb{R}$  be a nondecreasing function.

For any loss function  $\ell : \mathbb{R}^2 \rightarrow \mathbb{R}$ , the solutions of

$$f^* \in \operatorname{argmin}_{f \in H_k} \sum_{i=1}^n \ell(f(x_i), y_i) + \Omega(\|f\|_{H_k}^2)$$

can be expressed in the following form:

$$f^* = \sum_{i=1}^n \alpha_i k(x_i, \cdot).$$




# Kernels

## Part 9

### Constructing Kernels

Cynthia Rudin  
Duke University

Let's construct kernels from other kernels. Say  $k_1$  and  $k_2$  are kernels.

$$k(x, z) = \alpha k_1(x, z) + \beta k_2(x, z) \text{ for } \alpha, \beta \geq 0$$

1

$k_1$  has  $\Phi_1$  and  $\langle, \rangle_{H_{k_1}}$

$k_2$  has  $\Phi_2$  and  $\langle, \rangle_{H_{k_2}}$

$$\alpha k_1(x, z) = \langle \sqrt{\alpha} \Phi_1(x), \sqrt{\alpha} \Phi_1(z) \rangle_{H_{k_1}}$$

$$\beta k_2(x, z) = \langle \sqrt{\beta} \Phi_2(x), \sqrt{\beta} \Phi_2(z) \rangle_{H_{k_2}}$$

$$k(x, z) = \alpha k_1(x, z) + \beta k_2(x, z)$$

$$= \langle \sqrt{\alpha} \Phi_1(x), \sqrt{\alpha} \Phi_1(z) \rangle_{H_{k_1}} + \langle \sqrt{\beta} \Phi_2(x), \sqrt{\beta} \Phi_2(z) \rangle_{H_{k_2}}$$

$$=: \langle [\sqrt{\alpha} \Phi_1(x), \sqrt{\beta} \Phi_2(x)], [\sqrt{\alpha} \Phi_1(z), \sqrt{\beta} \Phi_2(z)] \rangle_{H_{\text{new}}}$$

so  $k$  is an inner product

2

$$k(x, z) = k_1(x, z)k_2(x, z)$$

3

$$k(x, z) = k_1(h(x), h(z)), \text{ where } h : \mathcal{X} \rightarrow \mathcal{X}$$

$$k_1(h(x), h(z)) = \langle \Phi(h(x)), \Phi(h(z)) \rangle_{H_{k_1}}$$

$$=: \langle \Phi_h(x), \Phi_h(z) \rangle_{H_{\text{new}}}$$

4

$$k(x, z) = g(x)g(z) \text{ for } g : \mathcal{X} \rightarrow \mathbb{R}.$$

5

$$k(x, z) = h(k_1(x, z)) \text{ where } h \text{ is a polynomial with positive coefficients}$$

$$k(x, z) = k_1(x, z)k_2(x, z)$$

2

$$k(x, z) = \alpha k_1(x, z) + \beta k_2(x, z) \text{ for } \alpha, \beta \geq 0$$

1

$$k(x, z) = h(k_1(x, z)) \text{ where } h \text{ is a polynomial with positive coefficients}$$

5



6

$$k(x, z) = \exp(k_1(x, z))$$

polynomial with positive coefficients  $\frac{1}{i!}$

$$\exp(z) = \lim_{i \rightarrow \infty} \left( 1 + z + \cdots + \frac{z^i}{i!} \right)$$

$$\begin{matrix} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{matrix} & \left[ \begin{array}{cccccc} & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{array} \right] \end{matrix}$$

 $k(x_a, x_l)$ 

Each element is a limit of polynomials

$$k(x, z) = \exp(k_1(x, z)) \quad (6)$$

$$k(\mathbf{x}, \mathbf{z}) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{z}\|_{\ell_2}^2}{\sigma^2}\right) \quad (7)$$

Gaussian kernel

$$\begin{aligned} k(\mathbf{x}, \mathbf{z}) &= \exp\left(\frac{-\|\mathbf{x} - \mathbf{z}\|_{\ell_2}^2}{\sigma^2}\right) = \exp\left(\frac{-\|\mathbf{x}\|_{\ell_2}^2 - \|\mathbf{z}\|_{\ell_2}^2 + 2\mathbf{x}^T \mathbf{z}}{\sigma^2}\right) \\ &= \left(\exp\left(\frac{-\|\mathbf{x}\|_{\ell_2}^2}{\sigma^2}\right) \exp\left(\frac{-\|\mathbf{z}\|_{\ell_2}^2}{\sigma^2}\right)\right) \exp\left(\frac{2\mathbf{x}^T \mathbf{z}}{\sigma^2}\right) \end{aligned}$$

$$k(x, z) = g(x)g(z) \text{ for } g : \mathcal{X} \rightarrow \mathbb{R} \quad (4)$$

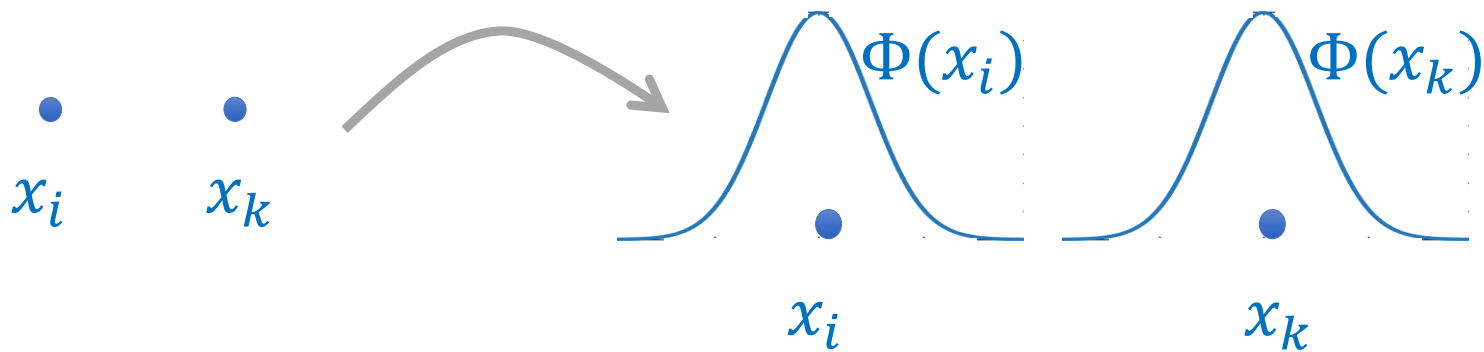
$$k(x, z) = k_1(x, z)k_2(x, z) \quad (2)$$

$$k(\mathbf{x}, \mathbf{z}) = \exp \left( \frac{-\|\mathbf{x} - \mathbf{z}\|_{\ell_2}^2}{\sigma^2} \right)$$


7

Gaussian kernel

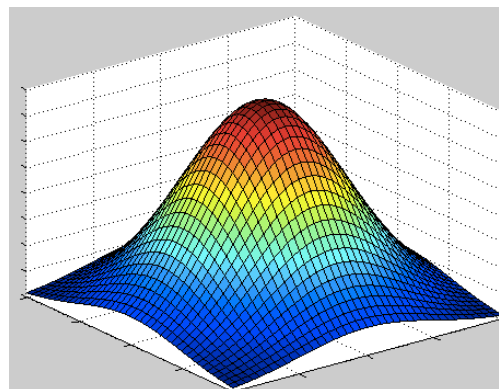
$$\Phi(\mathbf{x}) = k(\mathbf{x}, \cdot) = \exp \left( \frac{-\|\mathbf{x} - \cdot\|_{\ell_2}^2}{\sigma^2} \right)$$



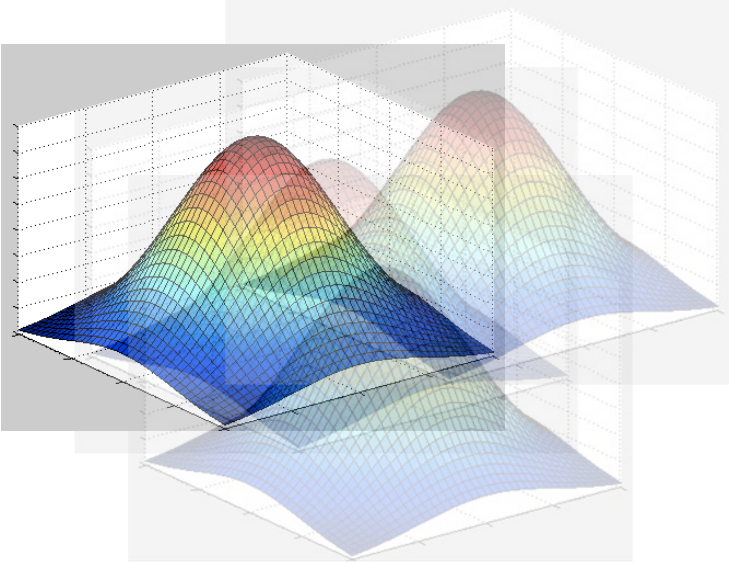
Reminder:

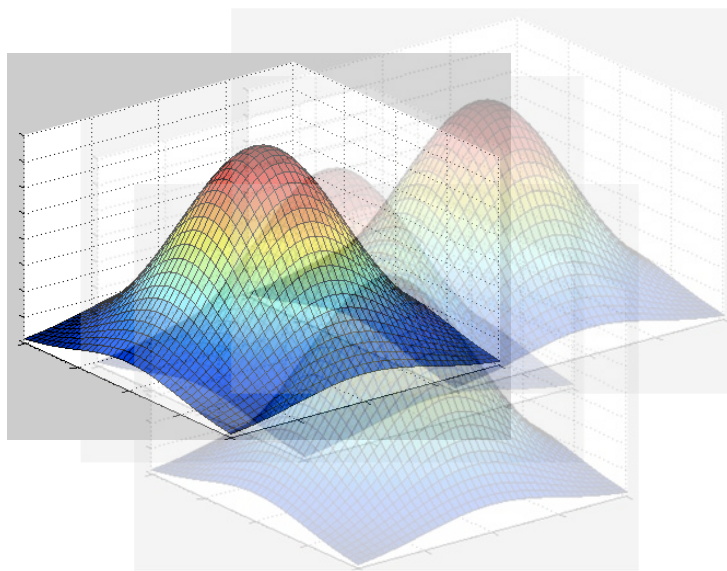
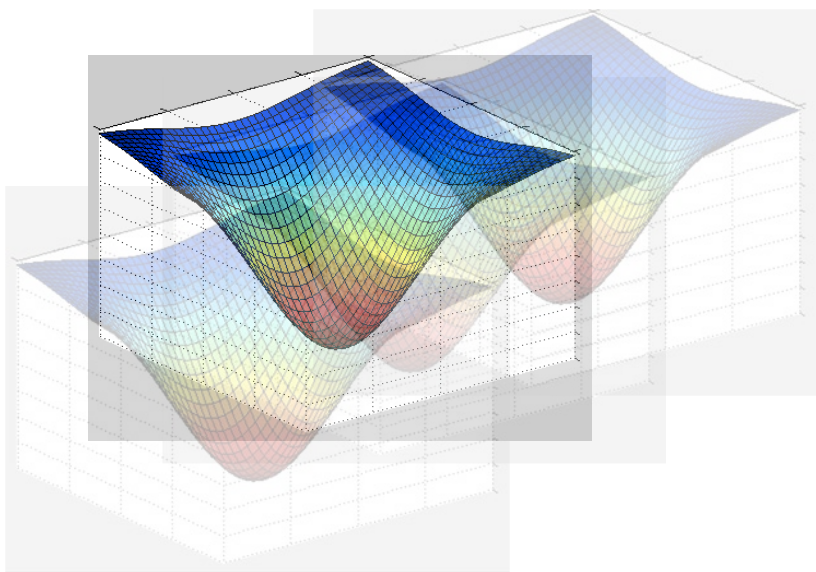
$$f(\mathbf{x}) = \sum_i \alpha_i^* y_i k(\mathbf{x}_i, \mathbf{x}) + \lambda_0^*$$


$$\Phi(\mathbf{x}_i) = k(\mathbf{x}_i, \mathbf{x}) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}\|_{\ell_2}^2}{\sigma^2}\right)$$

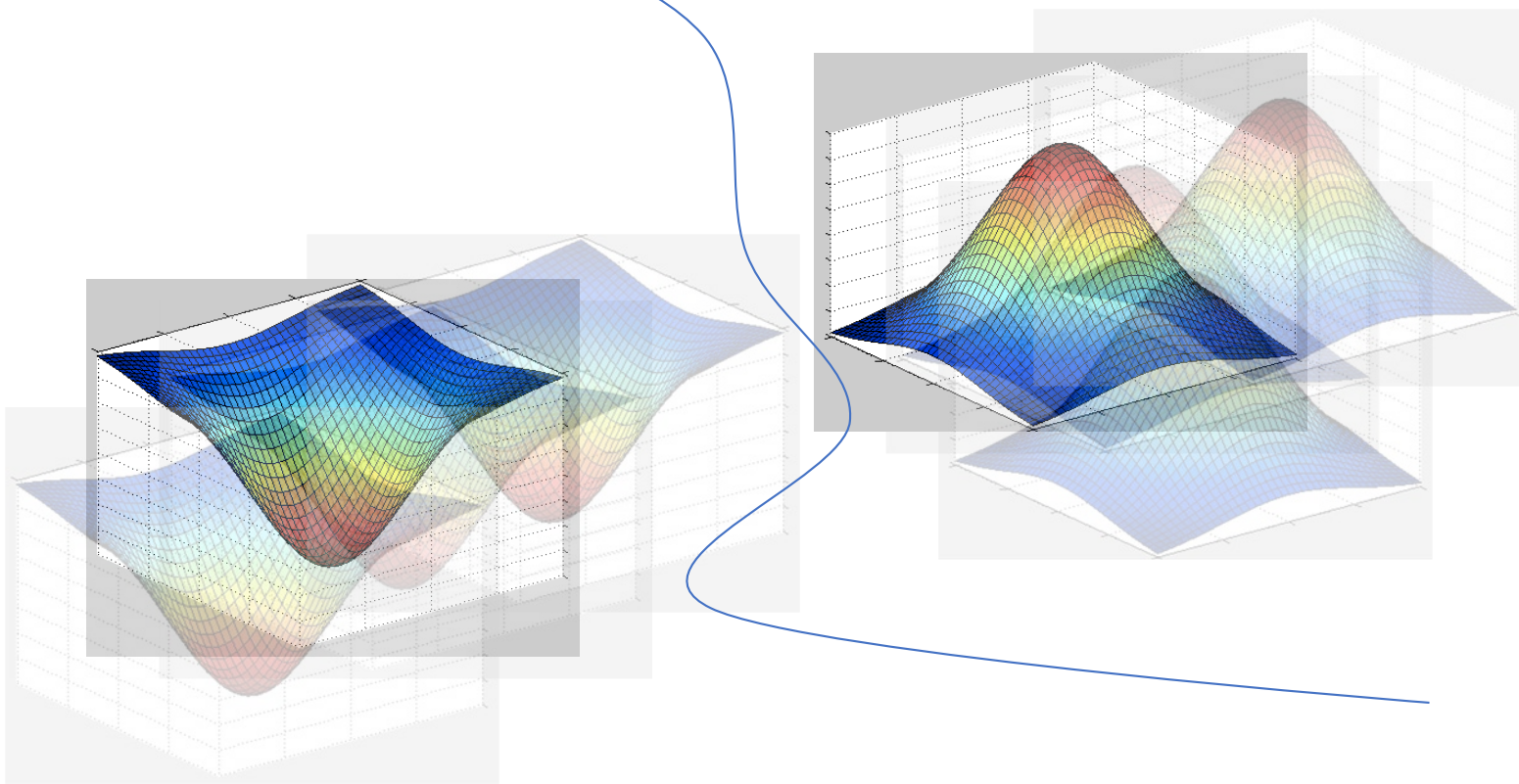


$\phi(\mathbf{x}_i)$





$$f(x)=0$$



## Notes:

- The width of the gaussian kernel controls regularization:
  - Too small kernel = memorizing the data = overfitting!
  - Too large kernel = too flat = underfitting!
- Either tune it using CV or set it to the default.
- It is not clear how to choose which kernel to use (linear, poly, gaussian). Usually try a few. Or just use gaussian!
- Again, beware of bad solvers. Don't expect it to work in higher dimensions!

Gaussian kernel demo in the next video!