

# Reproducing kernel Hilbert spaces in Machine Learning

Arthur Gretton

Gatsby Computational Neuroscience Unit,  
University College London

Advanced topics in Machine Learning, 2023

## Overview

The course has two parts:

- Kernel methods (Arthur Gretton)
- Convex optimization

The course has the following assessment components:

- Exam (50%)
- Coursework (50%)

For **non-Gatsby** students: only the first part of the coursework is mandatory - the group parts are optional.

For **Gatsby** students: the full coursework is required

## Course times, locations

Lecture times:

- Optimization lectures are Wednesday, 14:00 -15:30
- Kernel lectures will be on Friday 15:00 -16:30

There will be lectures during reading week, due to clash with NeurIPS conference.

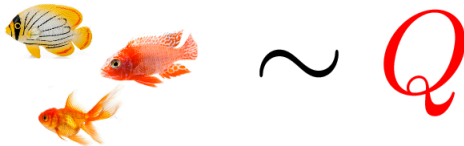
The tutor for the kernels part is Dimitri Meunier.

Lecture notes are online:

<http://www.gatsby.ucl.ac.uk/~gretton/rkhscourse.html>

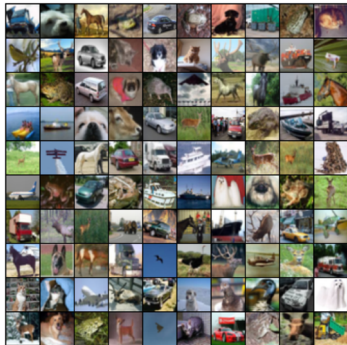
## A motivation: comparing two samples

- Given: Samples from unknown distributions  $P$  and  $Q$ .
- Goal: do  $P$  and  $Q$  differ?

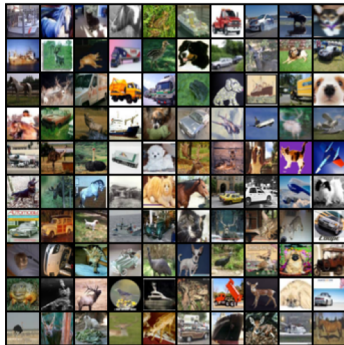


## A real-life example: two-sample tests

- Goal: do  $P$  and  $Q$  differ?



CIFAR 10 samples



Cifar 10.1 samples

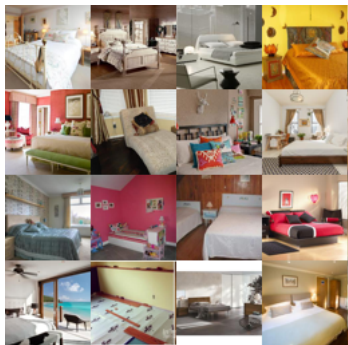
Significant difference?

Feng, Xu, Lu, Zhang, G., Sutherland, Learning Deep Kernels for Non-Parametric Two-Sample Tests, ICML 2020

Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., ICLR 2017.

## Training generative models

- Have: One collection of samples  $X$  from unknown distribution  $P$ .
- Goal: **generate** samples  $Q$  that look like  $P$



LSUN bedroom samples  $P$



Generated  $Q$ , MMD GAN

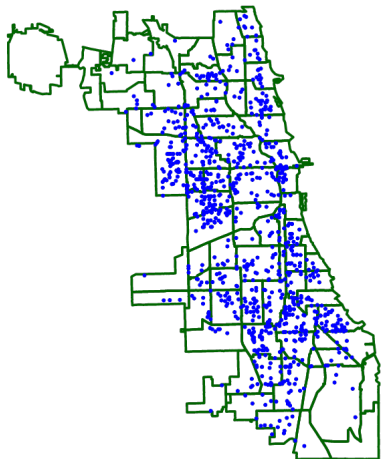
## Training a Generative Adversarial Network

(Binkowski, Sutherland, Arbel, G., ICLR 2018),  
(Arbel, Sutherland, Binkowski, G., NeurIPS 2018)

## Testing goodness of fit

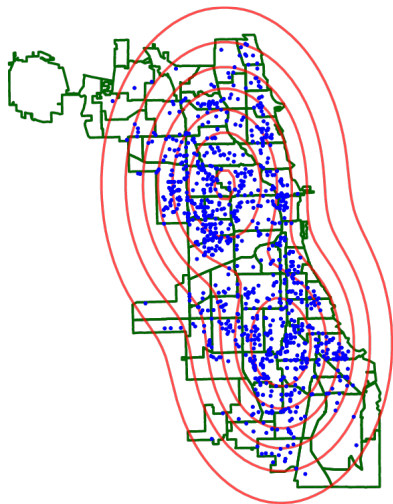
- Given: a model  $P$  and samples  $Q$ .
- Goal: is  $P$  a good fit for  $Q$ ?

Chicago crime data



## Testing goodness of fit

- Given: a model  $P$  and samples  $Q$ .
- Goal: is  $P$  a good fit for  $Q$ ?






### Chicago crime data

**Model** is Gaussian mixture with **two** components. Is this a good model?



# Testing independence

- Given: Samples from a distribution  $P_{XY}$
- Goal: Are  $X$  and  $Y$  independent?

X	Y
	A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose.
	Their noses guide them through life, and they're never happier than when following an interesting scent.
	A responsive, interactive pet, one that will blow in your ear and follow you everywhere.

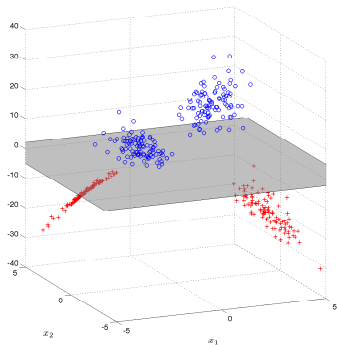
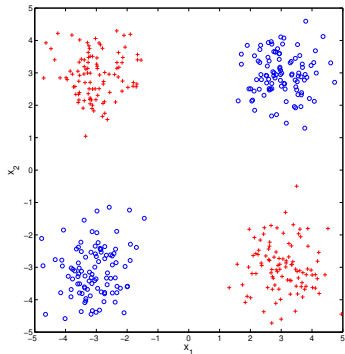
Text from dogtime.com and petfinder.com

## Course overview (kernels part)

- 1 Construction of RKHS,
- 2 Simple linear algorithms in RKHS (e.g. PCA, ridge regression)
- 3 Kernel methods for hypothesis testing (two-sample, independence, goodness-of-fit)
- 4 Further applications of kernels (feature selection, clustering,...)
- 5 Support vector machines for classification, regression
- 6 Cutting-edge kernel algorithms (to be a surprise)

# Reproducing Kernel Hilbert Spaces

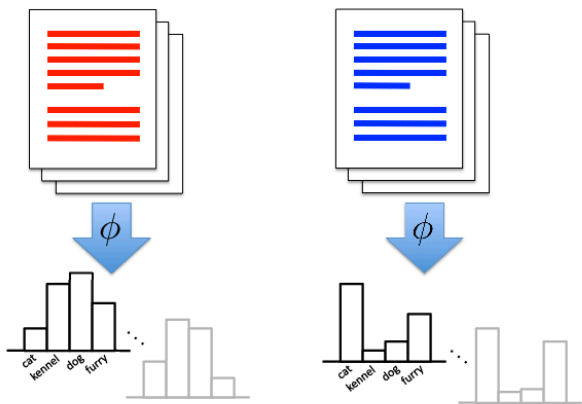
## Kernels and feature space (1): XOR example



- No linear classifier separates red from blue
- Map points to higher dimensional feature space:

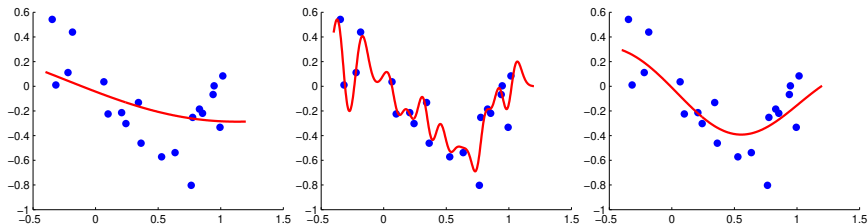
$$\phi(x) = \begin{bmatrix} x_1 & x_2 & x_1 x_2 \end{bmatrix} \in \mathbb{R}^3$$

## Kernels and feature space (2): document classification



Kernels let us compare objects on the basis of features

## Kernels and feature space (3): smoothing



Kernel methods can control smoothness and avoid overfitting/underfitting.

## Outline: reproducing kernel Hilbert space

We will describe in order:

- 1 Hilbert space (very simple)
- 2 Kernel (lots of examples: e.g. you can build kernels from simpler kernels)
- 3 Reproducing property

## Hilbert space

### Definition (Inner product)

Let  $\mathcal{H}$  be a vector space over  $\mathbb{R}$ . A function  $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$  is an **inner product** on  $\mathcal{H}$  if

- 1 Linear:  $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
- 2 Symmetric:  $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
- 3  $\langle f, f \rangle_{\mathcal{H}} \geq 0$  and  $\langle f, f \rangle_{\mathcal{H}} = 0$  if and only if  $f = 0$ .

Norm induced by the inner product:  $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$

### Definition (Hilbert space)

Inner product space containing Cauchy sequence limits.



## Hilbert space

### Definition (Inner product)

Let  $\mathcal{H}$  be a vector space over  $\mathbb{R}$ . A function  $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$  is an **inner product** on  $\mathcal{H}$  if

- 1 Linear:  $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
- 2 Symmetric:  $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
- 3  $\langle f, f \rangle_{\mathcal{H}} \geq 0$  and  $\langle f, f \rangle_{\mathcal{H}} = 0$  if and only if  $f = 0$ .

**Norm** induced by the inner product:  $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$

### Definition (Hilbert space)

Inner product space containing Cauchy sequence limits.

## Hilbert space

### Definition (Inner product)

Let  $\mathcal{H}$  be a vector space over  $\mathbb{R}$ . A function  $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$  is an **inner product** on  $\mathcal{H}$  if

- 1 Linear:  $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
- 2 Symmetric:  $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
- 3  $\langle f, f \rangle_{\mathcal{H}} \geq 0$  and  $\langle f, f \rangle_{\mathcal{H}} = 0$  if and only if  $f = 0$ .

**Norm** induced by the inner product:  $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$

### Definition (Hilbert space)

Inner product space containing Cauchy sequence limits.

## Kernel

### Definition

Let  $\mathcal{X}$  be a non-empty set. A function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is a kernel if there exists a Hilbert space  $\mathcal{H}$  and a **feature** map  $\phi : \mathcal{X} \rightarrow \mathcal{H}$  such that  $\forall x, x' \in \mathcal{X}$ ,

$$k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}.$$

- Almost no conditions on  $\mathcal{X}$  (eg,  $\mathcal{X}$  itself doesn't need an inner product, eg. documents).
- A single kernel can correspond to several possible features. A trivial example for  $\mathcal{X} := \mathbb{R}$ :

$$\phi_1(x) = x \quad \text{and} \quad \phi_2(x) = \begin{bmatrix} x/\sqrt{2} \\ x/\sqrt{2} \end{bmatrix}$$

## New kernels from old: sums, transformations

Theorem (Sums of kernels are kernels)

*Given  $\alpha > 0$  and  $k, k_1$  and  $k_2$  all kernels on  $\mathcal{X}$ , then  $\alpha k$  and  $k_1 + k_2$  are kernels on  $\mathcal{X}$ .*

(Proof via positive definiteness: **later!**) A difference of kernels may not be a kernel (why?)

Theorem (Mappings between spaces)

*Let  $\mathcal{X}$  and  $\tilde{\mathcal{X}}$  be sets, and define a map  $A : \mathcal{X} \rightarrow \tilde{\mathcal{X}}$ . Define the kernel  $k$  on  $\tilde{\mathcal{X}}$ . Then the kernel  $k(A(x), A(x'))$  is a kernel on  $\mathcal{X}$ .*

Example:  $k(x, x') = x^2 (x')^2$ .

## New kernels from old: sums, transformations

Theorem (Sums of kernels are kernels)

*Given  $\alpha > 0$  and  $k, k_1$  and  $k_2$  all kernels on  $\mathcal{X}$ , then  $\alpha k$  and  $k_1 + k_2$  are kernels on  $\mathcal{X}$ .*

(Proof via positive definiteness: **later!**) A difference of kernels may not be a kernel (why?)

Theorem (Mappings between spaces)

*Let  $\mathcal{X}$  and  $\tilde{\mathcal{X}}$  be sets, and define a map  $A : \mathcal{X} \rightarrow \tilde{\mathcal{X}}$ . Define the kernel  $k$  on  $\tilde{\mathcal{X}}$ . Then the kernel  $k(A(x), A(x'))$  is a kernel on  $\mathcal{X}$ .*

Example:  $k(x, x') = x^2 (x')^2$ .

## New kernels from old: products

Theorem (Products of kernels are kernels)

Given  $k_1$  on  $\mathcal{X}_1$  and  $k_2$  on  $\mathcal{X}_2$ , then  $k_1 \times k_2$  is a kernel on  $\mathcal{X}_1 \times \mathcal{X}_2$ .  
If  $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}$ , then  $k := k_1 \times k_2$  is a kernel on  $\mathcal{X}$ .

**Proof:** Main idea only!

$\mathcal{H}_1$  space of kernels between shapes,

$$\phi_1(x) = \begin{bmatrix} \mathbb{I}_{\square} \\ \mathbb{I}_{\triangle} \end{bmatrix} \quad \phi_1(\square) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad k_1(\square, \triangle) = 0.$$

$\mathcal{H}_2$  space of kernels between colors,

$$\phi_2(x) = \begin{bmatrix} \mathbb{I}_{\bullet} \\ \mathbb{I}_{\color{blue}\bullet} \end{bmatrix} \quad \phi_2(\bullet) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad k_2(\bullet, \bullet) = 1.$$

## New kernels from old: products

“Natural” feature space for colored shapes:

$$\Phi(x) = \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \\ \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \begin{bmatrix} \mathbb{I}_{\bullet} \\ \mathbb{I}_{\bullet} \end{bmatrix} \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \phi_2(x)\phi_1^\top(x)$$

## New kernels from old: products

“Natural” feature space for colored shapes:

$$\Phi(x) = \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \\ \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \begin{bmatrix} \mathbb{I}_{\bullet} \\ \mathbb{I}_{\bullet} \end{bmatrix} \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \phi_2(x)\phi_1^\top(x)$$

Kernel is:

$$k(x, x') = \sum_{i \in \{\bullet, \bullet\}} \sum_{j \in \{\square, \triangle\}} \Phi_{ij}(x)\Phi_{ij}(x')$$



## New kernels from old: products

“Natural” feature space for colored shapes:

$$\Phi(x) = \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \\ \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \begin{bmatrix} \mathbb{I}_{\bullet} \\ \mathbb{I}_{\bullet} \end{bmatrix} \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \phi_2(x)\phi_1^\top(x)$$

Kernel is:

$$k(x, x') = \sum_{i \in \{\bullet, \bullet\}} \sum_{j \in \{\square, \triangle\}} \Phi_{ij}(x)\Phi_{ij}(x') = \text{tr} \left( \underbrace{\phi_1(x)\phi_2^\top(x)}_{\Phi^\top(x)} \underbrace{\phi_2(x')\phi_1^\top(x')}_{\Phi(x')} \right)$$

## New kernels from old: products

“Natural” feature space for colored shapes:

$$\Phi(x) = \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \\ \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \begin{bmatrix} \mathbb{I}_{\bullet} \\ \mathbb{I}_{\bullet} \end{bmatrix} \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \phi_2(x)\phi_1^\top(x)$$

Kernel is:

$$k(x, x') = \sum_{i \in \{\bullet, \bullet\}} \sum_{j \in \{\square, \triangle\}} \Phi_{ij}(x)\Phi_{ij}(x') = \text{tr} \left( \phi_1(x)\underbrace{\phi_2^\top(x)\phi_2(x')\phi_1^\top(x')}_{k_2(x, x')} \right)$$

## New kernels from old: products

“Natural” feature space for colored shapes:

$$\Phi(x) = \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \\ \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \begin{bmatrix} \mathbb{I}_{\bullet} \\ \mathbb{I}_{\bullet} \end{bmatrix} \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \phi_2(x)\phi_1^\top(x)$$

Kernel is:

$$\begin{aligned} k(x, x') &= \sum_{i \in \{\bullet, \bullet\}} \sum_{j \in \{\square, \triangle\}} \Phi_{ij}(x)\Phi_{ij}(x') = \text{tr} \left( \phi_1(x)\underbrace{\phi_2^\top(x)\phi_2(x')\phi_1^\top(x')}_{k_2(x, x')} \right) \\ &= \text{tr} \left( \underbrace{\phi_1^\top(x')\phi_1(x)}_{k_1(x, x')} k_2(x, x') \right) \end{aligned}$$

## New kernels from old: products

“Natural” feature space for colored shapes:

$$\Phi(x) = \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \\ \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \begin{bmatrix} \mathbb{I}_{\bullet} \\ \mathbb{I}_{\bullet} \end{bmatrix} \begin{bmatrix} \mathbb{I}_{\square} & \mathbb{I}_{\triangle} \end{bmatrix} = \phi_2(x)\phi_1^\top(x)$$

Kernel is:

$$\begin{aligned} k(x, x') &= \sum_{i \in \{\bullet, \bullet\}} \sum_{j \in \{\square, \triangle\}} \Phi_{ij}(x)\Phi_{ij}(x') = \text{tr} \left( \phi_1(x)\underbrace{\phi_2^\top(x)\phi_2(x')\phi_1^\top(x')}_{k_2(x, x')} \right) \\ &= \text{tr} \left( \underbrace{\phi_1^\top(x')\phi_1(x)}_{k_1(x, x')} k_2(x, x') \right) = k_1(x, x')k_2(x, x') \end{aligned}$$

## Sums and products $\implies$ polynomials

### Theorem (Polynomial kernels)

*Let  $x, x' \in \mathbb{R}^d$  for  $d \geq 1$ , and let  $m \geq 1$  be an integer and  $c \geq 0$  be a positive real. Then*

$$k(x, x') := (\langle x, x' \rangle + c)^m$$

*is a valid kernel.*

To prove: expand into a sum (with non-negative scalars) of kernels  $\langle x, x' \rangle$  raised to integer powers. These individual terms are valid kernels by the product rule.

## Infinite sequences

The kernels we've seen so far are dot products between **finitely** many features. E.g.

$$k(x, y) = [ \sin(x) \quad x^3 \quad \log x ]^\top [ \sin(y) \quad y^3 \quad \log y ]$$

where  $\phi(x) = [ \sin(x) \quad x^3 \quad \log x ]$

Can a kernel be a dot product between **infinitely many features**?

## Taylor series kernels

Definition (Taylor series kernel)

For  $r \in (0, \infty]$ , with  $a_n \geq 0$  for all  $n \geq 0$

$$f(z) = \sum_{n=0}^{\infty} a_n z^n \quad |z| < r, z \in \mathbb{R},$$

Define  $\mathcal{X}$  to be the  $\sqrt{r}$ -ball in  $\mathbb{R}^d$ , so  $\|x\| < \sqrt{r}$ ,

$$k(x, x') = f(\langle x, x' \rangle) = \sum_{n=0}^{\infty} a_n \langle x, x' \rangle^n.$$

Exponential kernel:

$$k(x, x') := \exp(\langle x, x' \rangle).$$

## Taylor series kernel (proof)

**Proof:** Non-negative weighted sums of kernels are kernels, and products of kernels are kernels, so the following is a kernel if it converges:

$$k(x, x') = \sum_{n=0}^{\infty} a_n (\langle x, x' \rangle)^n$$

By Cauchy-Schwarz,

$$|\langle x, x' \rangle| \leq \|x\| \|x'\| < r,$$

so the sum converges.



## Exponentiated quadratic kernel

**Exponentiated quadratic kernel:** This kernel on  $\mathbb{R}^d$  is defined as

$$k(x, x') := \exp\left(-\gamma^{-2} \|x - x'\|^2\right).$$

Proof: an exercise! Use product rule, mapping rule, exponential kernel.

## Infinite sequences

### Definition

The space  $\ell_2$  (square summable sequences) comprises all sequences  $a := (a_i)_{i \geq 1}$  for which

$$\|a\|_{\ell_2}^2 = \sum_{\ell=1}^{\infty} a_{\ell}^2 < \infty.$$

### Definition

Given sequence of functions  $(\phi_{\ell}(x))_{\ell \geq 1}$  in  $\ell_2$  where  $\phi_{\ell} : \mathcal{X} \rightarrow \mathbb{R}$  is the  $i$ th coordinate of  $\phi(x)$ . Then

$$k(x, x') := \sum_{\ell=1}^{\infty} \phi_{\ell}(x) \phi_{\ell}(x') \quad (1)$$

## Infinite sequences

### Definition

The space  $\ell_2$  (square summable sequences) comprises all sequences  $a := (a_i)_{i \geq 1}$  for which

$$\|a\|_{\ell_2}^2 = \sum_{\ell=1}^{\infty} a_{\ell}^2 < \infty.$$

### Definition

Given sequence of functions  $(\phi_{\ell}(x))_{\ell \geq 1}$  in  $\ell_2$  where  $\phi_{\ell} : \mathcal{X} \rightarrow \mathbb{R}$  is the  $i$ th coordinate of  $\phi(x)$ . Then

$$k(x, x') := \sum_{\ell=1}^{\infty} \phi_{\ell}(x)\phi_{\ell}(x') \quad (1)$$

## Infinite sequences (proof)

Why square summable? By Cauchy-Schwarz,

$$\left| \sum_{\ell=1}^{\infty} \phi_{\ell}(x)\phi_{\ell}(x') \right| \leq \|\phi(x)\|_{\ell_2} \|\phi(x')\|_{\ell_2},$$

so the sequence defining the inner product converges for all  $x, x' \in \mathcal{X}$

## Positive definite functions

If we are given a function of two arguments,  $k(x, x')$ , how can we determine if it is a valid kernel?

- 1 Find a feature map?
  - 1 Sometimes this is not obvious (eg if the feature vector is infinite dimensional, e.g. the exponentiated quadratic kernel in the last slide)
  - 2 The feature map is not unique.
- 2 A direct property of the function: **positive definiteness**.

## Positive definite functions

### Definition (Positive definite functions)

A symmetric function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is **positive definite** if  $\forall n \geq 1, \forall (a_1, \dots, a_n) \in \mathbb{R}^n, \forall (x_1, \dots, x_n) \in \mathcal{X}^n,$

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) \geq 0.$$

The function  $k(\cdot, \cdot)$  is **strictly positive definite** if for mutually distinct  $x_i$ , the equality holds only when all the  $a_i$  are zero.

## Kernels are positive definite

### Theorem

Let  $\mathcal{H}$  be a Hilbert space,  $\mathcal{X}$  a non-empty set and  $\phi : \mathcal{X} \rightarrow \mathcal{H}$ .  
Then  $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} =: k(x, y)$  is positive definite.

### Proof.

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) &= \sum_{i=1}^n \sum_{j=1}^n \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^n a_i \phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0. \end{aligned}$$

**Reverse also holds:** positive definite  $k(x, x')$  is inner product in a unique  $\mathcal{H}$  (**Moore-Aronsjajn**: coming later!). □

## Sum of kernels is a kernel

Proof by positive definiteness:

Consider two kernels  $k_1(x, x')$  and  $k_2(x, x')$ . Then

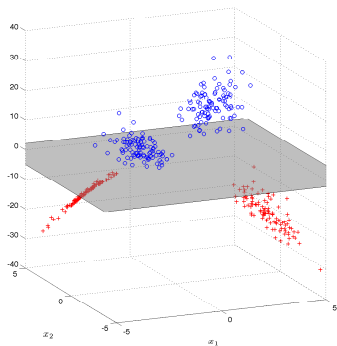
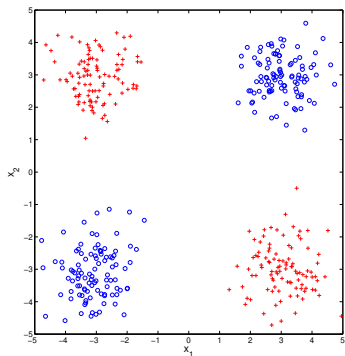
$$\begin{aligned} & \sum_{i=1}^n \sum_{j=1}^n a_i a_j [k_1(x_i, x_j) + k_2(x_i, x_j)] \\ &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j k_1(x_i, x_j) + \sum_{i=1}^n \sum_{j=1}^n a_i a_j k_2(x_i, x_j) \\ &\geq 0 \end{aligned}$$



# The reproducing kernel Hilbert space

## First example: finite space, polynomial features

Reminder: XOR example:



## Example: finite space, polynomial features

**Reminder:** Feature space from XOR motivating example:

$$\begin{aligned} \phi : \mathbb{R}^2 &\rightarrow \mathbb{R}^3 \\ x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &\mapsto \phi(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}, \end{aligned}$$

with kernel

$$k(x, y) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}^\top \begin{bmatrix} y_1 \\ y_2 \\ y_1 y_2 \end{bmatrix}$$

(the standard inner product in  $\mathbb{R}^3$  between features). Denote this feature space by  $\mathcal{H}$ .

## Example: finite space, polynomial features

Define a **linear function** of the inputs  $x_1, x_2$ , and their product  $x_1 x_2$ ,

$$f(x) = f_1 x_1 + f_2 x_2 + f_3 (x_1 x_2).$$

$f$  in a space of functions mapping from  $\mathcal{X} = \mathbb{R}^2$  to  $\mathbb{R}$ . Equivalent representation for  $f$ ,

$$f(\cdot) = [ f_1 \quad f_2 \quad f_3 ]^\top.$$

$f(\cdot)$  or  $f$  refers to the function as an object (here as a **vector** in  $\mathbb{R}^3$ )

$f(x) \in \mathbb{R}$  is function evaluated at a point (a **real number**).

$$f(x) = f(\cdot)^\top \phi(x) = \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}}$$

Evaluation of  $f$  at  $x$  is an inner product in feature space (here standard inner product in  $\mathbb{R}^3$ )

$\mathcal{H}$  is a space of functions mapping  $\mathbb{R}^2$  to  $\mathbb{R}$ .

## Example: finite space, polynomial features

Define a **linear function** of the inputs  $x_1, x_2$ , and their product  $x_1 x_2$ ,

$$f(x) = f_1 x_1 + f_2 x_2 + f_3 (x_1 x_2).$$

$f$  in a space of functions mapping from  $\mathcal{X} = \mathbb{R}^2$  to  $\mathbb{R}$ . Equivalent representation for  $f$ ,

$$f(\cdot) = [ f_1 \quad f_2 \quad f_3 ]^\top.$$

$f(\cdot)$  or  $f$  refers to the function as an object (here as a **vector** in  $\mathbb{R}^3$ )

$f(x) \in \mathbb{R}$  is function evaluated at a point (a **real number**).

$$f(x) = f(\cdot)^\top \phi(x) = \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}}$$

Evaluation of  $f$  at  $x$  is an inner product in feature space (here standard inner product in  $\mathbb{R}^3$ )

$\mathcal{H}$  is a space of functions mapping  $\mathbb{R}^2$  to  $\mathbb{R}$ .

## Functions of infinitely many features

Functions are linear combinations of features:

$$f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(x) = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{bmatrix}^{\top} \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \\ \phi_3(x) \\ \vdots \end{bmatrix}$$

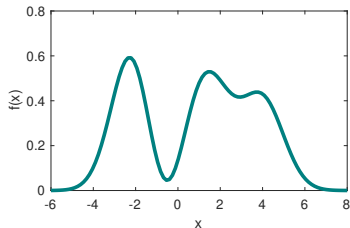
$$k(x, y) = \sum_{\ell=1}^{\infty} \phi_{\ell}(x) \phi_{\ell}(x')$$

$$f(x) = \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(x) \quad \sum_{\ell=1}^{\infty} f_{\ell}^2 < \infty.$$

## Expressing the functions with kernels

Function with **exponentiated quadratic kernel**:

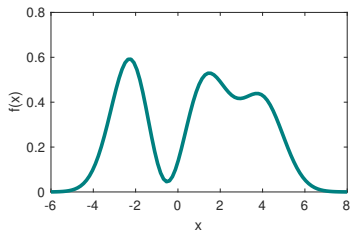
$$\begin{aligned} f(x) &= \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(x) \\ &= \sum_{\ell=1}^{\infty} \underbrace{\left( \sum_{i=1}^m \alpha_i \phi_{\ell}(x_i) \right)}_{f_{\ell}} \phi_{\ell}(x) \\ &= \left\langle \underbrace{\sum_{i=1}^m \alpha_i \phi(x_i)}_f, \phi(x) \right\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^m \alpha_i k(x_i, x) \end{aligned}$$



## Expressing the functions with kernels

Function with **exponentiated quadratic kernel**:

$$\begin{aligned} f(x) &= \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(x) \\ &= \sum_{\ell=1}^{\infty} \underbrace{\left( \sum_{i=1}^m \alpha_i \phi_{\ell}(x_i) \right)}_{f_{\ell}} \phi_{\ell}(x) \\ &= \left\langle \underbrace{\sum_{i=1}^m \alpha_i \phi(x_i)}_f, \phi(x) \right\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^m \alpha_i k(x_i, x) \end{aligned}$$



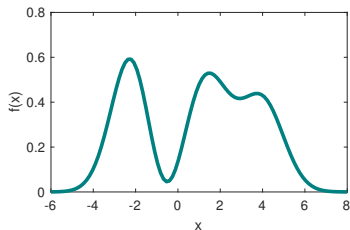
$$f_{\ell} := \sum_{i=1}^m \alpha_i \phi_{\ell}(x_i)$$



## Expressing the functions with kernels

Function with **exponentiated quadratic kernel**:

$$\begin{aligned} f(x) &= \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(x) \\ &= \sum_{\ell=1}^{\infty} \underbrace{\left( \sum_{i=1}^m \alpha_i \phi_{\ell}(x_i) \right)}_{f_{\ell}} \phi_{\ell}(x) \\ &= \left\langle \underbrace{\sum_{i=1}^m \alpha_i \phi(x_i)}_f, \phi(x) \right\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^m \alpha_i k(x_i, x) \end{aligned}$$

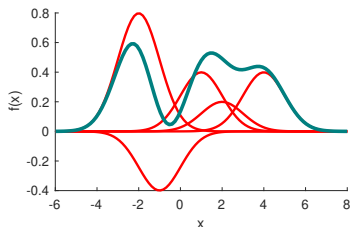


$$f := \sum_{i=1}^m \alpha_i \phi(x_i)$$

## Expressing the functions with kernels

Function with **exponentiated quadratic kernel**:

$$\begin{aligned} f(x) &= \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(x) \\ &= \sum_{\ell=1}^{\infty} \underbrace{\left( \sum_{i=1}^m \alpha_i \phi_{\ell}(x_i) \right)}_{f_{\ell}} \phi_{\ell}(x) \\ &= \left\langle \underbrace{\sum_{i=1}^m \alpha_i \phi(x_i)}_f, \phi(x) \right\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^m \alpha_i k(x_i, x) \end{aligned}$$



$$f := \sum_{i=1}^m \alpha_i \phi(x_i)$$

Function of **infinitely many features** expressed using  $\{(\alpha_i, x_i)\}_{i=1}^m$ .

## The feature map is *also* a function

On previous page,

$$f(x) := \sum_{i=1}^m \alpha_i k(x_i, x) = \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}} \quad \text{where} \quad f_{\ell} = \sum_{i=1}^m \alpha_i \phi_{\ell}(x_i).$$

What if  $m = 1$  and  $\alpha_1 = 1$ ?

Then

$$f(x) = k(x_1, x) = \left\langle \underbrace{k(x_1, \cdot)}_{f(\cdot)}, \phi(x) \right\rangle_{\mathcal{H}}$$

## The feature map is *also* a function

On previous page,

$$f(\mathbf{x}) := \sum_{i=1}^m \alpha_i k(\mathbf{x}_i, \mathbf{x}) = \langle f(\cdot), \phi(\mathbf{x}) \rangle_{\mathcal{H}} \quad \text{where} \quad f_{\ell} = \sum_{i=1}^m \alpha_i \phi_{\ell}(\mathbf{x}_i).$$

What if  $m = 1$  and  $\alpha_1 = 1$ ?

Then

$$f(\mathbf{x}) = k(\mathbf{x}_1, \mathbf{x}) = \left\langle \underbrace{k(\mathbf{x}_1, \cdot)}_{f(\cdot)}, \phi(\mathbf{x}) \right\rangle_{\mathcal{H}}$$

## The feature map is *also* a function

On previous page,

$$f(\mathbf{x}) := \sum_{i=1}^m \alpha_i k(\mathbf{x}_i, \mathbf{x}) = \langle f(\cdot), \phi(\mathbf{x}) \rangle_{\mathcal{H}} \quad \text{where} \quad f_{\ell} = \sum_{i=1}^m \alpha_i \phi_{\ell}(\mathbf{x}_i).$$

What if  $m = 1$  and  $\alpha_1 = 1$ ?

Then

$$\begin{aligned} f(\mathbf{x}) = k(\mathbf{x}_1, \mathbf{x}) &= \left\langle \underbrace{k(\mathbf{x}_1, \cdot)}_{f(\cdot)}, \phi(\mathbf{x}) \right\rangle_{\mathcal{H}} \\ &= \langle k(\mathbf{x}, \cdot), \phi(\mathbf{x}_1) \rangle_{\mathcal{H}} \end{aligned}$$

....so the feature map is a (very simple) function!

We can write without ambiguity

$$k(\mathbf{x}, \mathbf{y}) = \langle k(\cdot, \mathbf{x}), k(\cdot, \mathbf{y}) \rangle_{\mathcal{H}}.$$

## The feature map is *also* a function

On previous page,

$$f(\mathbf{x}) := \sum_{i=1}^m \alpha_i k(\mathbf{x}_i, \mathbf{x}) = \langle f(\cdot), \phi(\mathbf{x}) \rangle_{\mathcal{H}} \quad \text{where} \quad f_{\ell} = \sum_{i=1}^m \alpha_i \phi_{\ell}(\mathbf{x}_i).$$

What if  $m = 1$  and  $\alpha_1 = 1$ ?

Then

$$\begin{aligned} f(\mathbf{x}) = k(\mathbf{x}_1, \mathbf{x}) &= \left\langle \underbrace{k(\mathbf{x}_1, \cdot)}_{f(\cdot)}, \phi(\mathbf{x}) \right\rangle_{\mathcal{H}} \\ &= \langle k(\mathbf{x}, \cdot), \phi(\mathbf{x}_1) \rangle_{\mathcal{H}} \end{aligned}$$

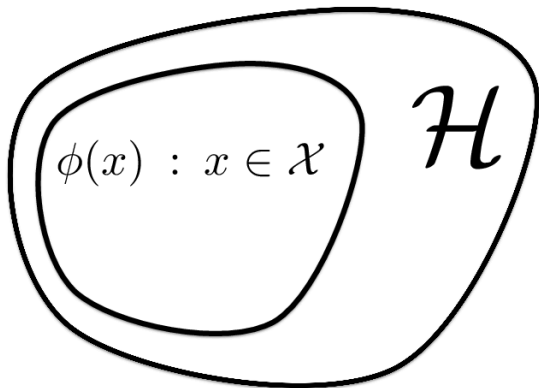
....so the feature map is a (very simple) function!

We can write without ambiguity

$$k(\mathbf{x}, \mathbf{y}) = \langle k(\cdot, \mathbf{x}), k(\cdot, \mathbf{y}) \rangle_{\mathcal{H}}.$$

## Features vs functions

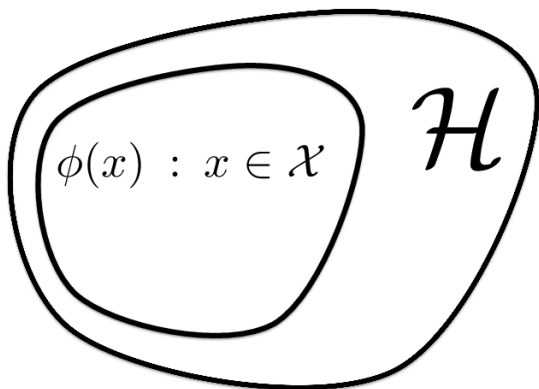
A subtle point:  $\mathcal{H}$  can be larger than all  $\phi(x)$ .



E.g.  $f(\cdot) = [1 \ 1 \ -1] \in \mathcal{H}$  cannot be obtained by  $\phi(x) = [x_1 \ x_2 \ (x_1 x_2)]$ .

## Features vs functions

A subtle point:  $\mathcal{H}$  can be larger than all  $\phi(x)$ .



E.g.  $f(\cdot) = [1 \ 1 \ -1] \in \mathcal{H}$  cannot be obtained by  $\phi(x) = [x_1 \ x_2 \ (x_1 x_2)]$ .



## The reproducing property

This example illustrates the two defining features of an RKHS:

- **The reproducing property:** (kernel trick)

$$\forall x \in \mathcal{X}, \forall f(\cdot) \in \mathcal{H}, \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$$

...or use shorter notation  $\langle f, \phi(x) \rangle_{\mathcal{H}}$ .

- The feature map of every point is a function:  $k(\cdot, x) = \phi(x) \in \mathcal{H}$  for any  $x \in \mathcal{X}$ , and

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} = \langle k(\cdot, x), k(\cdot, x') \rangle_{\mathcal{H}}.$$

# Understanding smoothness in the RKHS

## Infinite feature space via fourier series

Function on the interval  $[-\pi, \pi]$  with periodic boundary.

Fourier series:

$$f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l \exp(\imath l x) = \sum_{l=-\infty}^{\infty} \hat{f}_l (\cos(lx) + \imath \sin(lx)).$$

using the orthonormal basis on  $[-\pi, \pi]$ ,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(\imath l x) \overline{\exp(\imath m x)} dx = \begin{cases} 1 & l = m, \\ 0 & l \neq m. \end{cases}$$

Example: “top hat” function,

$$f(x) = \begin{cases} 1 & |x| < T, \\ 0 & T \leq |x| < \pi. \end{cases}$$

$$\hat{f}_l := \frac{\sin(\ell T)}{\ell \pi} \quad f(x) = \sum_{l=0}^{\infty} 2\hat{f}_l \cos(\ell x).$$

## Infinite feature space via fourier series

Function on the interval  $[-\pi, \pi]$  with periodic boundary.

Fourier series:

$$f(x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \exp(i\ell x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} (\cos(\ell x) + i \sin(\ell x)).$$

using the orthonormal basis on  $[-\pi, \pi]$ ,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(i\ell x) \overline{\exp(imx)} dx = \begin{cases} 1 & \ell = m, \\ 0 & \ell \neq m. \end{cases}$$

Example: “top hat” function,

$$f(x) = \begin{cases} 1 & |x| < T, \\ 0 & T \leq |x| < \pi. \end{cases}$$

$$\hat{f}_{\ell} := \frac{\sin(\ell T)}{\ell\pi} \quad f(x) = \sum_{\ell=0}^{\infty} 2\hat{f}_{\ell} \cos(\ell x).$$

## Infinite feature space via fourier series

Function on the interval  $[-\pi, \pi]$  with periodic boundary.

Fourier series:

$$f(x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \exp(i\ell x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} (\cos(\ell x) + i \sin(\ell x)).$$

using the orthonormal basis on  $[-\pi, \pi]$ ,

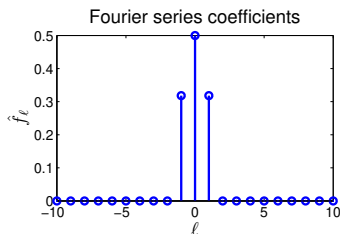
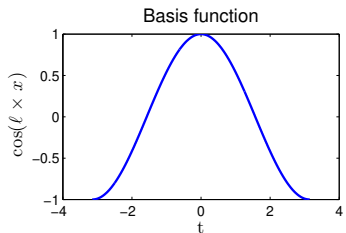
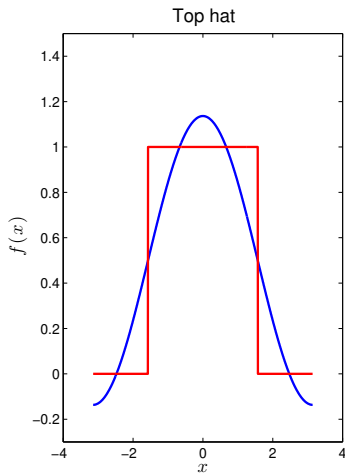
$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(i\ell x) \overline{\exp(imx)} dx = \begin{cases} 1 & \ell = m, \\ 0 & \ell \neq m. \end{cases}$$

Example: “top hat” function,

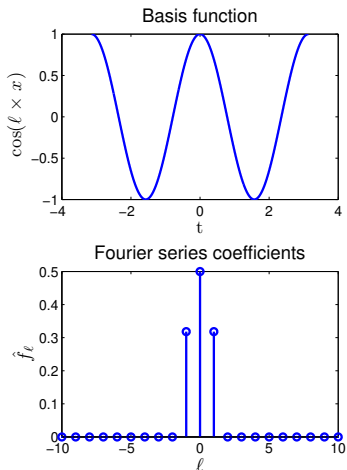
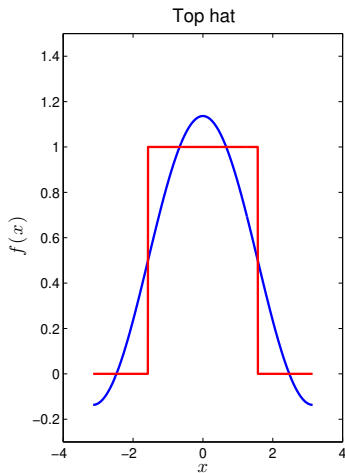
$$f(x) = \begin{cases} 1 & |x| < T, \\ 0 & T \leq |x| < \pi. \end{cases}$$

$$\hat{f}_{\ell} := \frac{\sin(\ell T)}{\ell\pi} \quad f(x) = \sum_{\ell=0}^{\infty} 2\hat{f}_{\ell} \cos(\ell x).$$

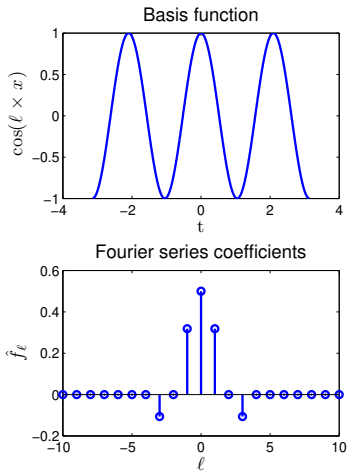
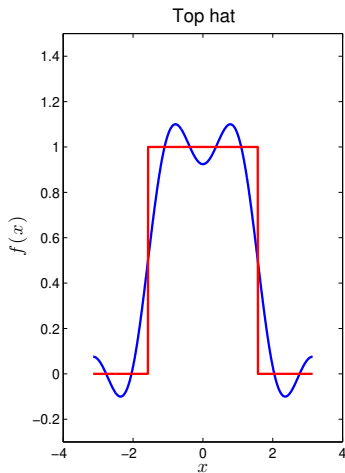
# Fourier series for top hat function



# Fourier series for top hat function

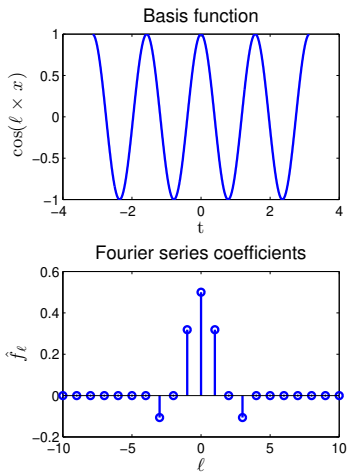
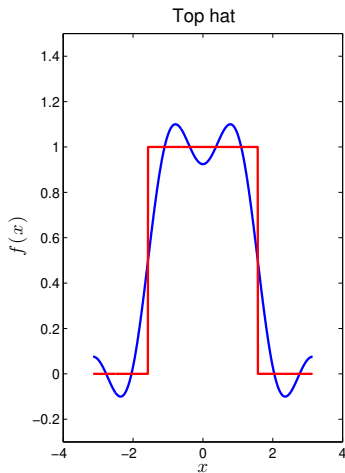


# Fourier series for top hat function

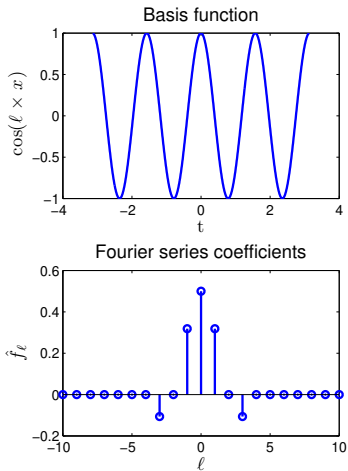
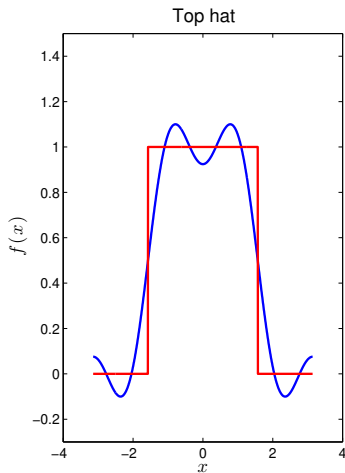




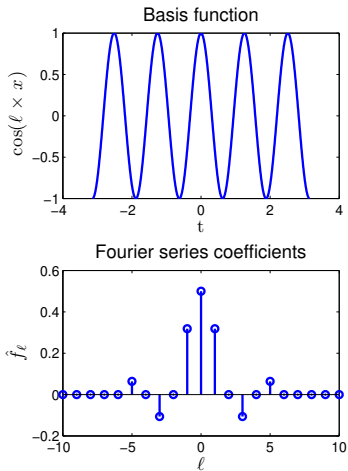
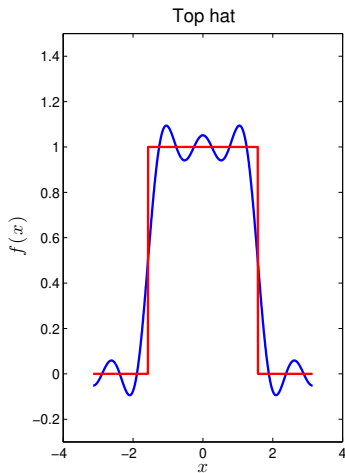
# Fourier series for top hat function



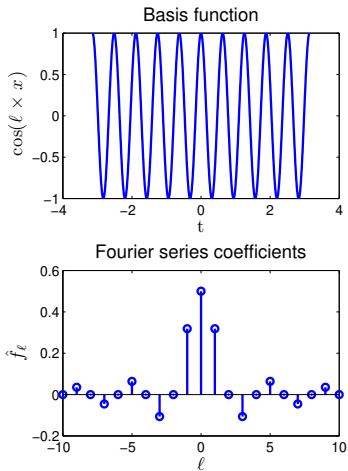
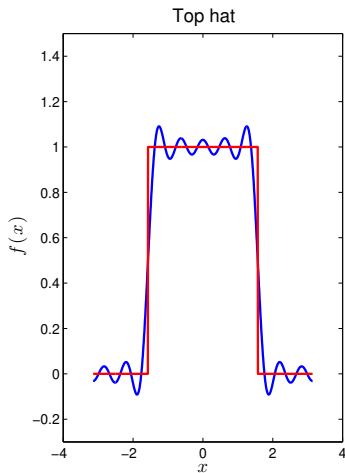
# Fourier series for top hat function



# Fourier series for top hat function



# Fourier series for top hat function



## Fourier series for kernel function

Assume kernel translation invariant,

$$k(x, y) = k(x - y),$$

Fourier series representation of  $k$

$$\begin{aligned} k(x - y) &= \sum_{\ell=-\infty}^{\infty} \hat{k}_{\ell} \exp(i\ell(x - y)) \\ &= \sum_{\ell=-\infty}^{\infty} \underbrace{\left[ \sqrt{\hat{k}_{\ell}} \exp(i\ell x) \right]}_{\phi_{\ell}(x)} \underbrace{\left[ \sqrt{\hat{k}_{\ell}} \exp(-i\ell y) \right]}_{\overline{\phi_{\ell}(y)}}. \end{aligned}$$

Example: Jacobi theta kernel:

$$k(x - y) = \frac{1}{2\pi} \vartheta \left( \frac{(x - y)}{2\pi}, \frac{i\sigma^2}{2\pi} \right), \quad \hat{k}_{\ell} = \frac{1}{2\pi} \exp \left( \frac{-\sigma^2 \ell^2}{2} \right).$$

$\vartheta$  is Jacobi theta function, close to Gaussian when  $\sigma^2$  much narrower than  $[-\pi, \pi]$ .

## Fourier series for kernel function

Assume kernel translation invariant,

$$k(x, y) = k(x - y),$$

Fourier series representation of  $k$

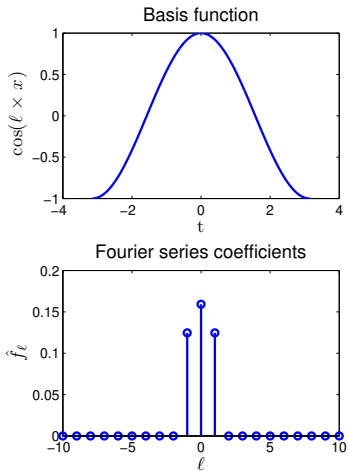
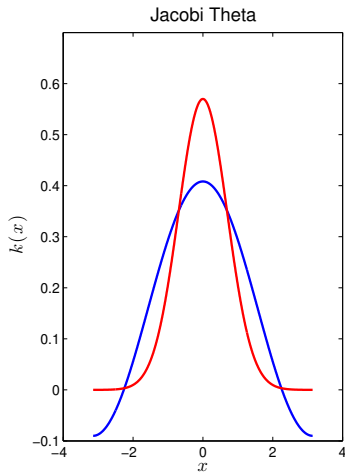
$$\begin{aligned}k(x - y) &= \sum_{\ell=-\infty}^{\infty} \hat{k}_{\ell} \exp(i\ell(x - y)) \\ &= \sum_{\ell=-\infty}^{\infty} \underbrace{\left[ \sqrt{\hat{k}_{\ell}} \exp(i\ell x) \right]}_{\phi_{\ell}(x)} \underbrace{\left[ \sqrt{\hat{k}_{\ell}} \exp(-i\ell y) \right]}_{\overline{\phi_{\ell}(y)}}.\end{aligned}$$

Example: **Jacobi theta kernel:**

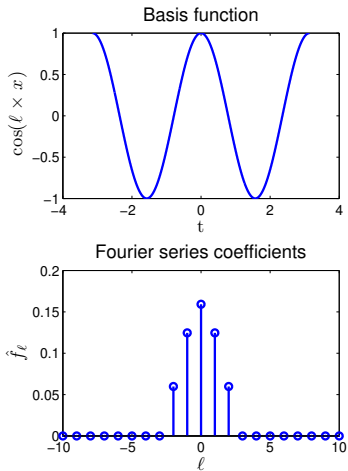
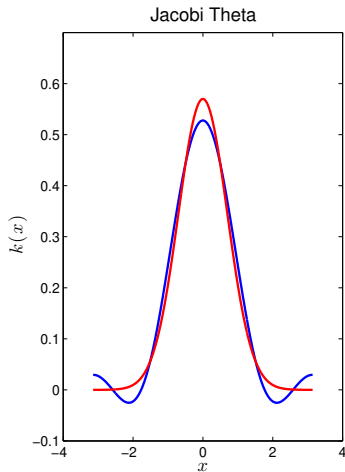
$$k(x - y) = \frac{1}{2\pi} \vartheta \left( \frac{(x - y)}{2\pi}, \frac{i\sigma^2}{2\pi} \right), \quad \hat{k}_{\ell} = \frac{1}{2\pi} \exp \left( \frac{-\sigma^2 \ell^2}{2} \right).$$

$\vartheta$  is Jacobi theta function, close to Gaussian when  $\sigma^2$  much narrower than  $[-\pi, \pi]$ .

# Fourier series for Gaussian-spectrum kernel

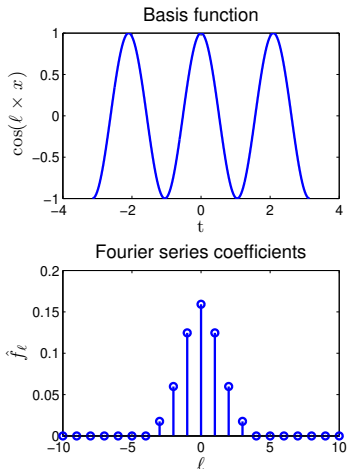
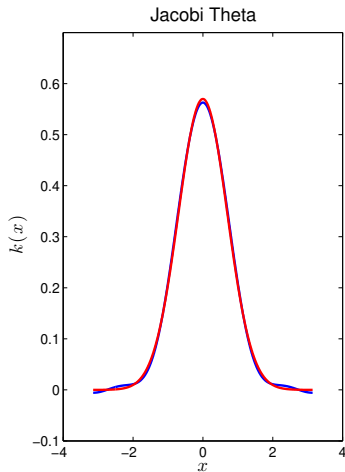


# Fourier series for Gaussian-spectrum kernel

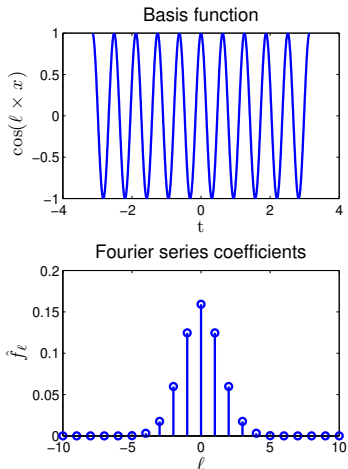
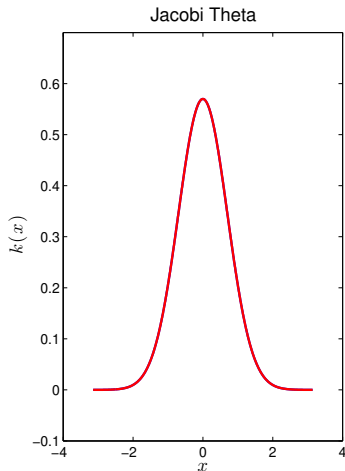




# Fourier series for Gaussian-spectrum kernel



# Fourier series for Gaussian-spectrum kernel



## RKHS via fourier series

Recall **standard dot product** in  $L_2$ :

$$\begin{aligned}\langle f, g \rangle_{L_2} &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \exp(i\ell x) \right] \left[ \sum_{m=-\infty}^{\infty} \overline{\hat{g}_m \exp(imx)} \right] dx \\ &= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{f}_{\ell} \overline{\hat{g}_m} \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(i\ell x) \exp(-imx) \\ &= \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \overline{\hat{g}_{\ell}}.\end{aligned}$$

## RKHS via fourier series

Recall **standard dot product** in  $L_2$ :

$$\begin{aligned}\langle f, g \rangle_{L_2} &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \exp(i\ell x) \right] \left[ \sum_{m=-\infty}^{\infty} \overline{\hat{g}_m \exp(imx)} \right] dx \\ &= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{f}_{\ell} \overline{\hat{g}_m} \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(i\ell x) \exp(-imx) \\ &= \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \overline{\hat{g}_{\ell}}.\end{aligned}$$

## RKHS via fourier series

Recall **standard dot product** in  $L_2$ :

$$\begin{aligned}\langle f, g \rangle_{L_2} &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \exp(i\ell x) \right] \left[ \sum_{m=-\infty}^{\infty} \overline{\hat{g}_m \exp(imx)} \right] dx \\ &= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{f}_{\ell} \overline{\hat{g}_m} \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(i\ell x) \exp(-imx) \\ &= \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \overline{\hat{g}_{\ell}}.\end{aligned}$$

## RKHS via fourier series

Recall **standard dot product** in  $L_2$ :

$$\begin{aligned}\langle f, g \rangle_{L_2} &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \exp(i\ell x) \right] \left[ \sum_{m=-\infty}^{\infty} \overline{\hat{g}_m \exp(imx)} \right] dx \\ &= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{f}_{\ell} \overline{\hat{g}_m} \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(i\ell x) \exp(-imx) \\ &= \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \overline{\hat{g}_{\ell}}.\end{aligned}$$

## RKHS via fourier series

Recall **standard dot product** in  $L_2$ :

$$\begin{aligned}\langle f, g \rangle_{L_2} &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \exp(i\ell x) \right] \left[ \sum_{m=-\infty}^{\infty} \overline{\hat{g}_m \exp(imx)} \right] dx \\ &= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{f}_{\ell} \overline{\hat{g}_m} \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(i\ell x) \exp(-imx) \\ &= \sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \overline{\hat{g}_{\ell}}.\end{aligned}$$

Define the **dot product** in  $\mathcal{H}$  to have a **roughness penalty**,

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} \overline{\hat{g}_{\ell}}}{\hat{k}_{\ell}}.$$

## Roughness penalty explained

The **squared norm** of a function  $f$  in  $\mathcal{H}$  **enforces smoothness**:

$$\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \bar{\hat{f}}_l}{\hat{k}_l} = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_l|^2}{\hat{k}_l}.$$

If  $\hat{k}_l$  decays fast, then so must  $\hat{f}_l$  if we want  $\|f\|_{\mathcal{H}}^2 < \infty$ .

Recall  $f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l (\cos(lx) + i \sin(lx))$ .

**Question:** is the top hat function in the “Gaussian spectrum” RKHS?

**Warning:** need stronger conditions on kernel than  $L_2$  convergence: **Mercer's theorem**.



## Roughness penalty explained

The **squared norm** of a function  $f$  in  $\mathcal{H}$  **enforces smoothness**:

$$\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \bar{\hat{f}}_l}{\hat{k}_l} = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_l|^2}{\hat{k}_l}.$$

If  $\hat{k}_l$  decays fast, then so must  $\hat{f}_l$  if we want  $\|f\|_{\mathcal{H}}^2 < \infty$ .

Recall  $f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l (\cos(lx) + i \sin(lx))$ .

**Question:** is the top hat function in the “Gaussian spectrum” RKHS?

**Warning:** need stronger conditions on kernel than  $L_2$  convergence: **Mercer's theorem**.

## Roughness penalty explained

The **squared norm** of a function  $f$  in  $\mathcal{H}$  **enforces smoothness**:

$$\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \bar{\hat{f}}_l}{\hat{k}_l} = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_l|^2}{\hat{k}_l}.$$

If  $\hat{k}_l$  decays fast, then so must  $\hat{f}_l$  if we want  $\|f\|_{\mathcal{H}}^2 < \infty$ .

Recall  $f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l (\cos(lx) + i \sin(lx))$ .

**Question:** is the top hat function in the “Gaussian spectrum” RKHS?

**Warning:** need stronger conditions on kernel than  $L_2$  convergence: **Mercer's theorem**.

## Feature map and reproducing property

Reproducing property: define a function

$$g(x) := k(x - z) = \sum_{\ell=-\infty}^{\infty} \exp(i\ell x) \underbrace{\hat{k}_{\ell} \exp(-i\ell z)}_{\hat{g}_{\ell}}$$

Then for a function  $f(\cdot) \in \mathcal{H}$ ,

$$\langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \langle f(\cdot), g(\cdot) \rangle_{\mathcal{H}}$$

$$\sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} \overbrace{\hat{k}_{\ell} \exp(i\ell z)}^{\hat{g}_{\ell}}}{\hat{k}_{\ell}}$$
$$\sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \exp(i\ell z) = f(z).$$

## Feature map and reproducing property

Reproducing property: define a function

$$g(x) := k(x - z) = \sum_{\ell=-\infty}^{\infty} \exp(i\ell x) \underbrace{\hat{k}_{\ell} \exp(-i\ell z)}_{\hat{g}_{\ell}}$$

Then for a function  $f(\cdot) \in \mathcal{H}$ ,

$$\langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \langle f(\cdot), g(\cdot) \rangle_{\mathcal{H}}$$

$$\sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} \overbrace{\hat{k}_{\ell} \exp(i\ell z)}^{\hat{g}_{\ell}}}{\hat{k}_{\ell}}$$
$$\sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \exp(i\ell z) = f(z).$$

## Feature map and reproducing property

Reproducing property: define a function

$$g(x) := k(x - z) = \sum_{\ell=-\infty}^{\infty} \exp(i\ell x) \underbrace{\hat{k}_{\ell} \exp(-i\ell z)}_{\hat{g}_{\ell}}$$

Then for a function  $f(\cdot) \in \mathcal{H}$ ,

$$\langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \langle f(\cdot), g(\cdot) \rangle_{\mathcal{H}}$$

$$\sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} \overbrace{\hat{k}_{\ell} \exp(i\ell z)}^{\hat{g}_{\ell}}}{\hat{k}_{\ell}}$$
$$\sum_{\ell=-\infty}^{\infty} \hat{f}_{\ell} \exp(i\ell z) = f(z).$$

## Feature map and reproducing property

Reproducing property for the kernel:

Recall kernel definition:

$$k(x - y) = \sum_{\ell=-\infty}^{\infty} \hat{k}_{\ell} \exp(i\ell(x - y)) = \sum_{\ell=-\infty}^{\infty} \hat{k}_{\ell} \exp(i\ell x) \exp(-i\ell y)$$

Define two functions

$$\begin{aligned} f(x) &:= k(x - y) = \sum_{\ell=-\infty}^{\infty} \hat{k}_{\ell} \exp(i\ell(x - y)) \\ &= \sum_{\ell=-\infty}^{\infty} \exp(i\ell x) \underbrace{\hat{k}_{\ell} \exp(-i\ell y)}_{\hat{f}_{\ell}} \\ g(x) &:= k(x - z) = \sum_{\ell=-\infty}^{\infty} \exp(i\ell x) \underbrace{\hat{k}_{\ell} \exp(-i\ell z)}_{\hat{g}_{\ell}} \end{aligned}$$

## Feature map and reproducing property

Reproducing property for the kernel:

Recall kernel definition:

$$k(x - y) = \sum_{\ell=-\infty}^{\infty} \hat{k}_{\ell} \exp(i\ell(x - y)) = \sum_{\ell=-\infty}^{\infty} \hat{k}_{\ell} \exp(i\ell x) \exp(-i\ell y)$$

Define two functions

$$\begin{aligned} f(x) &:= k(x - y) = \sum_{\ell=-\infty}^{\infty} \hat{k}_{\ell} \exp(i\ell(x - y)) \\ &= \sum_{\ell=-\infty}^{\infty} \exp(i\ell x) \underbrace{\hat{k}_{\ell} \exp(-i\ell y)}_{\hat{f}_{\ell}} \\ g(x) &:= k(x - z) = \sum_{\ell=-\infty}^{\infty} \exp(i\ell x) \underbrace{\hat{k}_{\ell} \exp(-i\ell z)}_{\hat{g}_{\ell}} \end{aligned}$$

## Feature map and reproducing property

Check the **reproducing property**:

$$\begin{aligned}\langle k(\cdot, y), k(\cdot, z) \rangle_{\mathcal{H}} &= \langle f(\cdot), g(\cdot) \rangle_{\mathcal{H}} \\ &= \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} \bar{\hat{g}}_{\ell}}{\hat{k}_{\ell}} \\ &= \sum_{\ell=-\infty}^{\infty} \frac{(\hat{k}_{\ell} \exp(-i\ell y)) (\overline{\hat{k}_{\ell} \exp(-i\ell z)})}{\hat{k}_{\ell}} \\ &= \sum_{\ell=-\infty}^{\infty} \hat{k}_{\ell} \exp(i\ell(z - y)) = k(z - y).\end{aligned}$$



## Feature map and reproducing property

Check the **reproducing property**:

$$\begin{aligned}\langle k(\cdot, y), k(\cdot, z) \rangle_{\mathcal{H}} &= \langle f(\cdot), g(\cdot) \rangle_{\mathcal{H}} \\ &= \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} \bar{\hat{g}}_{\ell}}{\hat{k}_{\ell}} \\ &= \sum_{\ell=-\infty}^{\infty} \frac{\left( \hat{k}_{\ell} \exp(-i\ell y) \right) \left( \overline{\hat{k}_{\ell} \exp(-i\ell z)} \right)}{\hat{k}_{\ell}} \\ &= \sum_{\ell=-\infty}^{\infty} \hat{k}_{\ell} \exp(i\ell(z - y)) = k(z - y).\end{aligned}$$

## Feature map and reproducing property

Check the **reproducing property**:

$$\begin{aligned}\langle k(\cdot, y), k(\cdot, z) \rangle_{\mathcal{H}} &= \langle f(\cdot), g(\cdot) \rangle_{\mathcal{H}} \\ &= \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} \bar{\hat{g}}_{\ell}}{\hat{k}_{\ell}} \\ &= \sum_{\ell=-\infty}^{\infty} \frac{(\hat{k}_{\ell} \exp(-\imath \ell y)) (\overline{\hat{k}_{\ell} \exp(-\imath \ell z)})}{\hat{k}_{\ell}} \\ &= \sum_{\ell=-\infty}^{\infty} \hat{k}_{\ell} \exp(\imath \ell (z - y)) = k(z - y).\end{aligned}$$

## [Link back to original RKHS function definition](#)

**Original form** of a function in the RKHS was

(detail: sum now from  $-\infty$  to  $\infty$ , complex conjugate)

$$f(z) = \sum_{\ell=-\infty}^{\infty} f_{\ell} \overline{\phi_{\ell}(z)} = \langle f(\cdot), \phi(z) \rangle_{\mathcal{H}}.$$

We've defined the **RKHS dot product** as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} \overline{\hat{g}_{\ell}}}{\hat{k}_{\ell}} \qquad \langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} \left( \overline{\hat{k}_{\ell} \exp(-\nu_{\ell} z)} \right)}{\hat{k}_{\ell}}$$

## [Link back to original RKHS function definition](#)

Original form of a function in the RKHS was

(detail: sum now from  $-\infty$  to  $\infty$ , complex conjugate)

$$f(z) = \sum_{\ell=-\infty}^{\infty} f_{\ell} \overline{\phi_{\ell}(z)} = \langle f(\cdot), \phi(z) \rangle_{\mathcal{H}}.$$

We've defined the RKHS dot product as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} \hat{g}_{\ell}}{\hat{k}_{\ell}} \quad \langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} \left( \overline{\hat{k}_{\ell} \exp(-\nu \ell z)} \right)}{\left( \sqrt{\hat{k}_{\ell}} \right)^2}$$

## [Link back to original RKHS function definition](#)

Original form of a function in the RKHS was

(detail: sum now from  $-\infty$  to  $\infty$ , complex conjugate)

$$f(z) = \sum_{\ell=-\infty}^{\infty} f_{\ell} \overline{\phi_{\ell}(z)} = \langle f(\cdot), \phi(z) \rangle_{\mathcal{H}}.$$

We've defined the RKHS dot product as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} \overline{\hat{g}_{\ell}}}{\hat{k}_{\ell}} \quad \langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} \left( \overline{\hat{k}_{\ell} \exp(-\nu_{\ell} z)} \right)}{\left( \sqrt{\hat{k}_{\ell}} \right)^2}$$

By inspection

$$f_{\ell} = \hat{f}_{\ell} / \sqrt{\hat{k}_{\ell}} \quad \phi_{\ell}(z) = \sqrt{\hat{k}_{\ell}} \exp(-\nu_{\ell} z).$$

## Second infinite feature space (on $\mathbb{R}$ )

Define a probability measure on  $\mathcal{X} := \mathbb{R}$ . We'll use the Gaussian density,

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2)$$

Define the eigenexpansion of  $k(x, x')$  wrt this measure:

$$\lambda_\ell e_\ell(x) = \int k(x, x') e_\ell(x') p(x') dx' \quad \int e_i(x) e_j(x) p(x) dx = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

We can write

$$k(x, x') = \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x) e_\ell(x'),$$

which converges in  $L_2(p)$ .

**Warning:** again, need stronger conditions on kernel than  $L_2$  convergence.

## Second infinite feature space (on $\mathbb{R}$ )

Define a probability measure on  $\mathcal{X} := \mathbb{R}$ . We'll use the Gaussian density,

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2)$$

Define the eigenexpansion of  $k(x, x')$  wrt this measure:

$$\lambda_\ell e_\ell(x) = \int k(x, x') e_\ell(x') p(x') dx' \quad \int e_i(x) e_j(x) p(x) dx = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

We can write

$$k(x, x') = \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x) e_\ell(x'),$$

which converges in  $L_2(p)$ .

**Warning:** again, need stronger conditions on kernel than  $L_2$  convergence.

## Second infinite feature space (on $\mathbb{R}$ )

Define a probability measure on  $\mathcal{X} := \mathbb{R}$ . We'll use the Gaussian density,

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2)$$

Define the eigenexpansion of  $k(x, x')$  wrt this measure:

$$\lambda_\ell e_\ell(x) = \int k(x, x') e_\ell(x') p(x') dx' \quad \int e_i(x) e_j(x) p(x) dx = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

We can write

$$k(x, x') = \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x) e_\ell(x'),$$

which converges in  $L_2(p)$ .

**Warning:** again, need stronger conditions on kernel than  $L_2$  convergence.



## Second infinite feature space (on $\mathbb{R}$ )

Exponentiated quadratic kernel,

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) = \sum_{\ell=1}^{\infty} \underbrace{\left(\sqrt{\lambda_{\ell}} e_{\ell}(x)\right)}_{\phi_{\ell}(x)} \underbrace{\left(\sqrt{\lambda_{\ell}} e_{\ell}(x')\right)}_{\phi_{\ell}(x')}$$

$$\lambda_{\ell} e_{\ell}(x) = \int k(x, x') e_{\ell}(x') p(x') dx',$$

$$p(x) = \mathcal{N}(0, \sigma^2).$$

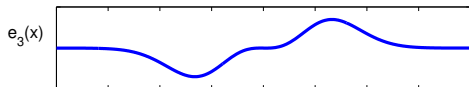
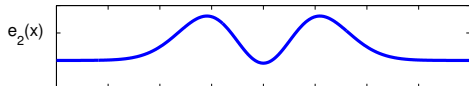
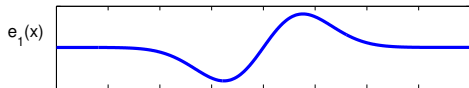
## Second infinite feature space (on $\mathbb{R}$ )

Exponentiated quadratic kernel,

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) = \sum_{\ell=1}^{\infty} \underbrace{\left(\sqrt{\lambda_{\ell}} e_{\ell}(x)\right)}_{\phi_{\ell}(x)} \underbrace{\left(\sqrt{\lambda_{\ell}} e_{\ell}(x')\right)}_{\phi_{\ell}(x')}$$

$$\lambda_{\ell} e_{\ell}(x) = \int k(x, x') e_{\ell}(x') p(x') dx',$$

$$p(x) = \mathcal{N}(0, \sigma^2).$$



$$\lambda_{\ell} \propto b^{\ell} \quad b < 1$$

$$e_{\ell}(x) \propto \exp(-(c - a)x^2) H_{\ell}(x\sqrt{2c}),$$

$a, b, c$  are functions of  $\sigma$ ,  
and  $H_{\ell}$  is  $\ell$ th order Hermite polynomial.

## Second infinite feature space (on $\mathbb{R}$ )

**Reminder:** for two functions  $f, g$  in  $L_2(p)$ ,

$$f(x) = \sum_{\ell=1}^{\infty} \hat{f}_{\ell} e_{\ell}(x) \quad g(x) = \sum_{m=1}^{\infty} \hat{g}_m e_m(x),$$

dot product is

$$\begin{aligned} \langle f, g \rangle_{L_2(p)} &= \int_{-\infty}^{\infty} f(x)g(x)p(x) dx \\ &= \int_{-\infty}^{\infty} \left( \sum_{\ell=1}^{\infty} \hat{f}_{\ell} e_{\ell}(x) \right) \left( \sum_{m=1}^{\infty} \hat{g}_m e_m(x) \right) p(x) dx \\ &= \sum_{\ell=1}^{\infty} \hat{f}_{\ell} \hat{g}_{\ell} \end{aligned}$$

Define the dot product in  $\mathcal{H}$  to have a roughness penalty,

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \hat{g}_{\ell}}{\lambda_{\ell}} \quad \|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell}^2}{\lambda_{\ell}}.$$

## Second infinite feature space (on $\mathbb{R}$ )

**Reminder:** for two functions  $f, g$  in  $L_2(p)$ ,

$$f(x) = \sum_{\ell=1}^{\infty} \hat{f}_{\ell} e_{\ell}(x) \quad g(x) = \sum_{m=1}^{\infty} \hat{g}_m e_m(x),$$

dot product is

$$\begin{aligned} \langle f, g \rangle_{L_2(p)} &= \int_{-\infty}^{\infty} f(x)g(x)p(x) dx \\ &= \int_{-\infty}^{\infty} \left( \sum_{\ell=1}^{\infty} \hat{f}_{\ell} e_{\ell}(x) \right) \left( \sum_{m=1}^{\infty} \hat{g}_m e_m(x) \right) p(x) dx \\ &= \sum_{\ell=1}^{\infty} \hat{f}_{\ell} \hat{g}_{\ell} \end{aligned}$$

Define the dot product in  $\mathcal{H}$  to have a roughness penalty,

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \hat{g}_{\ell}}{\lambda_{\ell}} \quad \|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell}^2}{\lambda_{\ell}}.$$

## Second infinite feature space (on $\mathbb{R}$ )

**Reminder:** for two functions  $f, g$  in  $L_2(p)$ ,

$$f(x) = \sum_{\ell=1}^{\infty} \hat{f}_{\ell} e_{\ell}(x) \quad g(x) = \sum_{m=1}^{\infty} \hat{g}_m e_m(x),$$

dot product is

$$\begin{aligned} \langle f, g \rangle_{L_2(p)} &= \int_{-\infty}^{\infty} f(x)g(x)p(x) dx \\ &= \int_{-\infty}^{\infty} \left( \sum_{\ell=1}^{\infty} \hat{f}_{\ell} e_{\ell}(x) \right) \left( \sum_{m=1}^{\infty} \hat{g}_m e_m(x) \right) p(x) dx \\ &= \sum_{\ell=1}^{\infty} \hat{f}_{\ell} \hat{g}_{\ell} \end{aligned}$$

Define the dot product in  $\mathcal{H}$  to have a roughness penalty,

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \hat{g}_{\ell}}{\lambda_{\ell}} \quad \|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell}^2}{\lambda_{\ell}}.$$

## Second infinite feature space (on $\mathbb{R}$ )

**Reminder:** for two functions  $f, g$  in  $L_2(p)$ ,

$$f(x) = \sum_{\ell=1}^{\infty} \hat{f}_{\ell} e_{\ell}(x) \quad g(x) = \sum_{m=1}^{\infty} \hat{g}_m e_m(x),$$

dot product is

$$\begin{aligned} \langle f, g \rangle_{L_2(p)} &= \int_{-\infty}^{\infty} f(x)g(x)p(x) dx \\ &= \int_{-\infty}^{\infty} \left( \sum_{\ell=1}^{\infty} \hat{f}_{\ell} e_{\ell}(x) \right) \left( \sum_{m=1}^{\infty} \hat{g}_m e_m(x) \right) p(x) dx \\ &= \sum_{\ell=1}^{\infty} \hat{f}_{\ell} \hat{g}_{\ell} \end{aligned}$$

Define the dot product in  $\mathcal{H}$  to have a roughness penalty,

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \hat{g}_{\ell}}{\lambda_{\ell}} \quad \|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell}^2}{\lambda_{\ell}}.$$

## Does the reproducing property hold?

Check the reproducing property:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{l=1}^{\infty} \frac{\hat{f}_l \hat{g}_l}{\lambda_l}$$

## Does the reproducing property hold?

Check the reproducing property:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \hat{g}_{\ell}}{\lambda_{\ell}} \qquad g(\cdot) = k(\cdot, z) = \sum_{\ell=1}^{\infty} \underbrace{\lambda_{\ell} e_{\ell}(z)}_{\hat{g}_{\ell}} e_{\ell}(\cdot)$$



## Does the reproducing property hold?

Check the reproducing property:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \hat{g}_{\ell}}{\lambda_{\ell}} \qquad g(\cdot) = k(\cdot, z) = \sum_{\ell=1}^{\infty} \underbrace{\lambda_{\ell} e_{\ell}(z)}_{\hat{g}_{\ell}} e_{\ell}(\cdot)$$

Then:

$$\langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \overbrace{(\lambda_{\ell} e_{\ell}(z))}^{\hat{g}_{\ell}}}{\lambda_{\ell}}$$

## Does the reproducing property hold?

Check the reproducing property:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \hat{g}_{\ell}}{\lambda_{\ell}} \qquad g(\cdot) = k(\cdot, z) = \sum_{\ell=1}^{\infty} \underbrace{\lambda_{\ell} e_{\ell}(z)}_{\hat{g}_{\ell}} e_{\ell}(\cdot)$$

Then:

$$\langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \cancel{\lambda_{\ell}} e_{\ell}(z)}{\cancel{\lambda_{\ell}}}$$

## Does the reproducing property hold?

Check the reproducing property:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{l=1}^{\infty} \frac{\hat{f}_l \hat{g}_l}{\lambda_l} \qquad g(\cdot) = k(\cdot, z) = \sum_{l=1}^{\infty} \underbrace{\lambda_l e_l(z)}_{\hat{g}_l} e_l(\cdot)$$

Then:

$$\begin{aligned} \langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} &= \sum_{l=1}^{\infty} \frac{\hat{f}_l \cancel{\lambda_l} e_l(z)}{\cancel{\lambda_l}} \\ &= \sum_{l=1}^{\infty} \hat{f}_l e_l(z) = f(z) \end{aligned}$$

## [Link back to the original RKHS definition](#)

Original form of a function in the RKHS was

$$f(z) = \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(z) = \langle f(\cdot), \phi(z) \rangle_{\mathcal{H}}$$

Expansion of  $f(\cdot)$  in terms of kernel eigenbasis:

$$f(\cdot) = \sum_{\ell=1}^{\infty} \hat{f}_{\ell} e_{\ell}(\cdot) \qquad k(x, z) = \sum_{\ell=1}^{\infty} \lambda_{\ell} e_{\ell}(x) e_{\ell}(z)$$

## [Link back to the original RKHS definition](#)

Original form of a function in the RKHS was

$$f(z) = \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(z) = \langle f(\cdot), \phi(z) \rangle_{\mathcal{H}}$$

Expansion of  $f(\cdot)$  in terms of kernel eigenbasis:

$$f(\cdot) = \sum_{\ell=1}^{\infty} \hat{f}_{\ell} e_{\ell}(\cdot) \qquad k(x, z) = \sum_{\ell=1}^{\infty} \lambda_{\ell} e_{\ell}(x) e_{\ell}(z)$$

Same expression with “roughness penalised” dot product:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \hat{g}_{\ell}}{\lambda_{\ell}} \qquad g(\cdot) = k(\cdot, z) = \sum_{\ell=1}^{\infty} \underbrace{\lambda_{\ell} e_{\ell}(z)}_{\hat{g}_{\ell}} e_{\ell}(\cdot)$$

## [Link back to the original RKHS definition](#)

Original form of a function in the RKHS was

$$f(z) = \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(z) = \langle f(\cdot), \phi(z) \rangle_{\mathcal{H}}$$

Expansion of  $f(\cdot)$  in terms of kernel eigenbasis:

$$f(\cdot) = \sum_{\ell=1}^{\infty} \hat{f}_{\ell} e_{\ell}(\cdot) \quad k(x, z) = \sum_{\ell=1}^{\infty} \lambda_{\ell} e_{\ell}(x) e_{\ell}(z)$$

Same expression with “roughness penalised” dot product:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \hat{g}_{\ell}}{\lambda_{\ell}} \quad g(\cdot) = k(\cdot, z) = \sum_{\ell=1}^{\infty} \underbrace{\lambda_{\ell} e_{\ell}(z)}_{\hat{g}_{\ell}} e_{\ell}(\cdot)$$

$$\text{Thus: } \langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \overbrace{(\lambda_{\ell} e_{\ell}(z))}^{\hat{g}_{\ell}}}{\lambda_{\ell}}$$

## [Link back to the original RKHS definition](#)

Original form of a function in the RKHS was

$$f(z) = \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(z) = \langle f(\cdot), \phi(z) \rangle_{\mathcal{H}}$$

Expansion of  $f(\cdot)$  in terms of kernel eigenbasis:

$$f(\cdot) = \sum_{\ell=1}^{\infty} \hat{f}_{\ell} e_{\ell}(\cdot) \quad k(x, z) = \sum_{\ell=1}^{\infty} \lambda_{\ell} e_{\ell}(x) e_{\ell}(z)$$

Same expression with “roughness penalised” dot product:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \hat{g}_{\ell}}{\lambda_{\ell}} \quad g(\cdot) = k(\cdot, z) = \sum_{\ell=1}^{\infty} \underbrace{\lambda_{\ell} e_{\ell}(z)}_{\hat{g}_{\ell}} e_{\ell}(\cdot)$$

Thus:  $\langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} (\lambda_{\ell} e_{\ell}(z))}{(\sqrt{\lambda_{\ell}})^2}$

## [Link back to the original RKHS definition](#)

Original form of a function in the RKHS was

$$f(z) = \sum_{\ell=1}^{\infty} f_{\ell} \phi_{\ell}(z) = \langle f(\cdot), \phi(z) \rangle_{\mathcal{H}}$$

Expansion of  $f(\cdot)$  in terms of kernel eigenbasis:

$$f(\cdot) = \sum_{\ell=1}^{\infty} \hat{f}_{\ell} e_{\ell}(\cdot) \quad k(x, z) = \sum_{\ell=1}^{\infty} \lambda_{\ell} e_{\ell}(x) e_{\ell}(z)$$

Same expression with “roughness penalised” dot product:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_{\ell} \hat{g}_{\ell}}{\lambda_{\ell}} \quad g(\cdot) = k(\cdot, z) = \sum_{\ell=1}^{\infty} \underbrace{\lambda_{\ell} e_{\ell}(z)}_{\hat{g}_{\ell}} e_{\ell}(\cdot)$$

Thus:  $\langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_{\ell} (\lambda_{\ell} e_{\ell}(z))}{(\sqrt{\lambda_{\ell}})^2}$

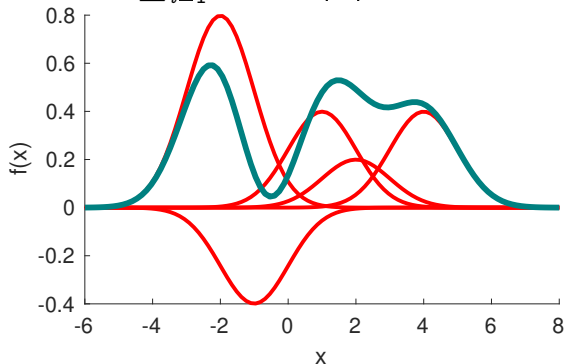
By inspection:  $f_{\ell} = \hat{f}_{\ell} / \sqrt{\lambda_{\ell}}$        $\phi_{\ell}(z) = \sqrt{\lambda_{\ell}} e_{\ell}(z)$ .



## RKHS function, exponentiated quadratic kernel:

$$f(x) = \sum_{i=1}^m \alpha_i k(x_i, x) = \sum_{i=1}^m \alpha_i \left[ \sum_{j=1}^{\infty} \lambda_j e_j(x_i) e_j(x) \right] = \sum_{\ell=1}^{\infty} f_{\ell} \underbrace{\left[ \sqrt{\lambda_{\ell}} e_{\ell}(x) \right]}_{\phi_{\ell}(x)}$$

where  $f_{\ell} = \sum_{i=1}^m \alpha_i \sqrt{\lambda_{\ell}} e_{\ell}(x_i)$ .



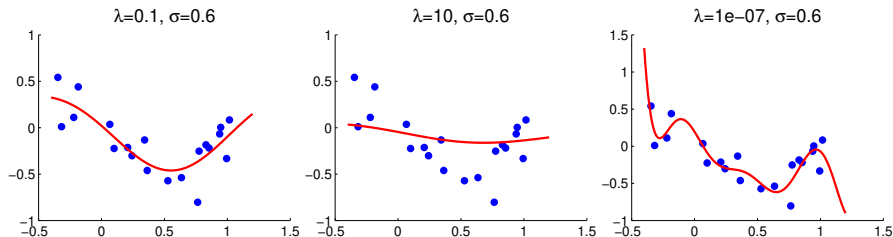
**NOTE** that this enforces smoothing:  $\lambda_{\ell}$  decay as  $e_{\ell}$  become rougher,  $f_{\ell}$  decay since  $\|f\|_{\mathcal{H}}^2 = \sum_{\ell} f_{\ell}^2 < \infty$ .

## Main message

Small RKHS norm results in smooth functions.

E.g. kernel ridge regression with exponentiated quadratic kernel:

$$f^* = \arg \min_{f \in \mathcal{H}} \left( \sum_{i=1}^n (y_i - \langle f, \phi(x_i) \rangle_{\mathcal{H}})^2 + \lambda \|f\|_{\mathcal{H}}^2 \right).$$



# Some reproducing kernel Hilbert space theory

## Reproducing kernel Hilbert space (1)

### Definition

$\mathcal{H}$  a Hilbert space of  $\mathbb{R}$ -valued functions on non-empty set  $\mathcal{X}$ . A function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is a **reproducing kernel** of  $\mathcal{H}$ , and  $\mathcal{H}$  is a **reproducing kernel Hilbert space**, if

- $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{H}$ ,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$  (the reproducing property).

In particular, for any  $x, y \in \mathcal{X}$ ,

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}. \quad (2)$$

Original definition: kernel an inner product between feature maps.  
Then  $\phi(x) = k(\cdot, x)$  a valid feature map.

## Reproducing kernel Hilbert space (2)

Another RKHS definition:

Define  $\delta_x$  to be the operator of evaluation at  $x$ , i.e.

$$\delta_x f = f(x) \quad \forall f \in \mathcal{H}, x \in \mathcal{X}.$$

Definition (Reproducing kernel Hilbert space)

$\mathcal{H}$  is an RKHS if the evaluation operator  $\delta_x$  is **bounded**:  $\forall x \in \mathcal{X}$  there exists  $\lambda_x \geq 0$  such that for all  $f \in \mathcal{H}$ ,

$$|f(x)| = |\delta_x f| \leq \lambda_x \|f\|_{\mathcal{H}}$$

$\implies$  two functions identical in RKHS norm agree at every point:

$$|f(x) - g(x)| = |\delta_x(f - g)| \leq \lambda_x \|f - g\|_{\mathcal{H}} \quad \forall f, g \in \mathcal{H}.$$

## RKHS definitions equivalent

Theorem (Reproducing kernel equivalent to bounded  $\delta_x$  )

$\mathcal{H}$  is a reproducing kernel Hilbert space (i.e., its evaluation operators  $\delta_x$  are bounded linear operators), if and only if  $\mathcal{H}$  has a reproducing kernel.

**Proof:** If  $\mathcal{H}$  has a reproducing kernel  $\implies \delta_x$  bounded

$$\begin{aligned} |\delta_x[f]| &= |f(x)| \\ &= |\langle f, k(\cdot, x) \rangle_{\mathcal{H}}| \\ &\leq \|k(\cdot, x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \\ &= \langle k(\cdot, x), k(\cdot, x) \rangle_{\mathcal{H}}^{1/2} \|f\|_{\mathcal{H}} \\ &= k(x, x)^{1/2} \|f\|_{\mathcal{H}} \end{aligned}$$

Cauchy-Schwarz in 3rd line . Consequently,  $\delta_x : \mathcal{F} \rightarrow \mathbb{R}$  bounded with  $\lambda_x = k(x, x)^{1/2}$ .

## RKHS definitions equivalent

**Proof:**  $\delta_x$  bounded  $\implies \mathcal{H}$  has a reproducing kernel

We use...

### Theorem

*(Riesz representation)* In a Hilbert space  $\mathcal{H}$ , all bounded linear functionals are of the form  $\langle \cdot, g \rangle_{\mathcal{H}}$ , for some  $g \in \mathcal{H}$ .

If  $\delta_x : \mathcal{F} \rightarrow \mathbb{R}$  is a bounded linear functional, by Riesz  $\exists f_{\delta_x} \in \mathcal{H}$  such that

$$\delta_x f = \langle f, f_{\delta_x} \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}.$$

Define  $k(\cdot, x) = f_{\delta_x}(\cdot)$ ,  $\forall x, x' \in \mathcal{X}$ . By its definition, both  $k(\cdot, x) = f_{\delta_x}(\cdot) \in \mathcal{H}$  and  $\langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = \delta_x f = f(x)$ . Thus,  $k$  is the reproducing kernel.

## Moore-Aronszajn Theorem

### Theorem (Moore-Aronszajn)

*Let  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  be positive definite. There is a unique RKHS  $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$  with reproducing kernel  $k$ .*

Recall feature map is *not unique* (as we saw earlier):  
*only kernel is unique.*



## Main message

