Lecture 1: Introduction to RKHS

Lille, 2014

Gatsby Unit, CSML, UCL

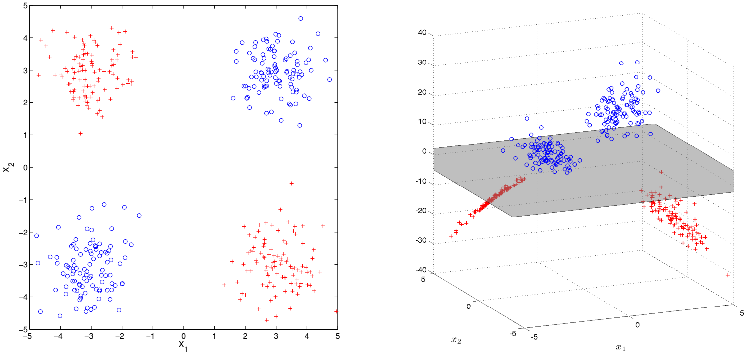
April 1, 2014

Overview

1. Construction of RKHS:
    1. Definition of a kernel as an inner product between feature space mappings of individual points,
    2. Construction of kernels on the basis of simpler kernels,
    3. Introduction of the reproducing kernel Hilbert space (RKHS) induced by positive definite kernels.
2. Mapping of probabilities to RKHS
    1. characteristic kernels
    2. two-sample tests
    3. independence tests
3. Further applications (if time): large-scale testing, three-way interaction testing, Bayesian inference, link with energy distance/distance covariance

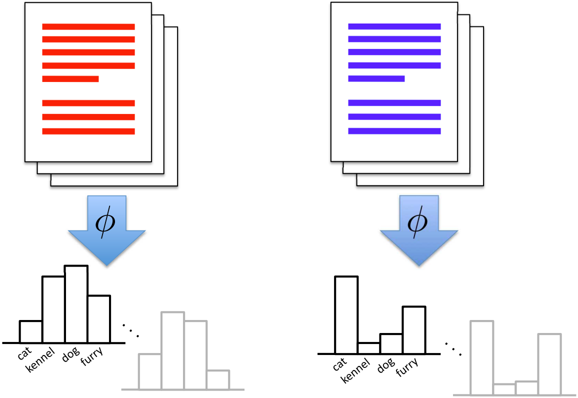
Kernel methods

Why kernel methods (1): XOR example



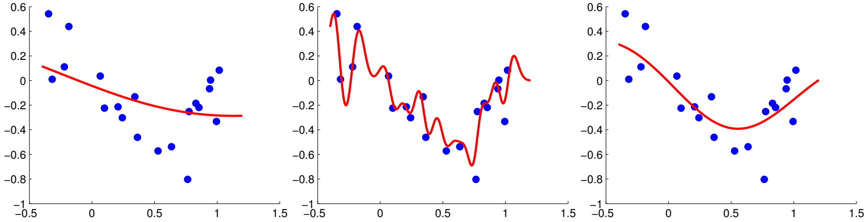
- No linear classifier separates red from blue
- Map points to **higher dimensional feature space**:
  $\phi(x) = \begin{bmatrix} x_1 & x_2 & x_1 x_2 \end{bmatrix} \in \mathbb{R}^3$

Why kernel methods (2): document classification



Kernels let us compare **objects** on the basis of **features**

Why kernel methods(3): smoothing



Kernel methods can control **smoothness** and **avoid overfitting/underfitting**.

Basics of reproducing kernel Hilbert spaces

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 $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ 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 $\langle\cdot,\cdot\rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ 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 $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ 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 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** if there exists an $\mathbb{R}$-Hilbert space and a map $\phi : \mathcal{X} \to \mathcal{H}$ 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 $\mathcal{X} := \mathbb{R}$:

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

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 αk and $k_1 + k_2$ are kernels on $\mathcal{X}$.

To prove this, just check inner product definition. A difference of kernels may not be a kernel (**why?**)

Theorem (Mappings between spaces)

Let $\mathcal{X}$ and $\widetilde{\mathcal{X}}$ be sets, and define a map $A : \mathcal{X} \to \widetilde{\mathcal{X}}$. Define the kernel k on $\widetilde{\mathcal{X}}$. 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 αk and $k_1 + k_2$ are kernels on $\mathcal{X}$.

To prove this, just check inner product definition. A difference of kernels may not be a kernel (**why?**)

Theorem (Mappings between spaces)

Let $\mathcal{X}$ and $\widetilde{\mathcal{X}}$ be sets, and define a map $A : \mathcal{X} \to \widetilde{\mathcal{X}}$. Define the kernel k on $\widetilde{\mathcal{X}}$. 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! $\mathcal{H}_1$ corresponding to k_1 is $\mathbb{R}^m$, and $\mathcal{H}_2$ corresponding to k_2 is $\mathbb{R}^n$. Define:

- $k_1 := u^\top v$ for $u, v \in \mathbb{R}^m$ (e.g.: kernel between two images)
- $k_2 := p^\top q$ for $p, q \in \mathbb{R}^n$ (e.g.: kernel between two captions)

Is the following a kernel?

$$K\left[(u,p);(v,q)\right] = k_1 \times k_2$$

(e.g. kernel between one image-caption pair and another)

New kernels from old: products

Proof.

(continued)

$$
\begin{aligned}
k_1 k_2 &= k_1 \left(q^\top p\right) \\
&= k_1 \mathrm{trace}(q^\top p) \\
&= k_1 \mathrm{trace}(pq^\top) \\
&= \mathrm{trace}(p\underbrace{u^\top v}_{k_1} q^\top) \\
&= \langle A, B \rangle ,
\end{aligned}
$$

where $A := up^\top$ and $B := vq^\top$.

Thus $k_1 k_2$ is valid inner product, since I.P. between $A, B \in \mathbb{R}^{m \times n}$ is

$$\langle A, B \rangle = \mathrm{trace}(A^\top B). \tag{1}$$

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

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

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.

$$k(x, y) = \begin{bmatrix} \sin(x) & x^3 & \log x \end{bmatrix}^\top \begin{bmatrix} \sin(y) & y^3 & \log y \end{bmatrix}$$

where $\phi(x) = \begin{bmatrix} \sin(x) & x^3 & \log x \end{bmatrix}$

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

Infinite sequences

Definition

The space ℓ_p of p-summable sequences is defined as all sequences $(a_i)_{i\geq 1}$ for which

$$\sum_{i=1}^{\infty} a_i^p < \infty.$$

Kernels can be defined in terms of sequences in ℓ_2.

Theorem

Given sequence of functions $(\phi_i(x))_{i\geq 1}$ in ℓ_2 where $\phi_i : \mathcal{X} \to \mathbb{R}$ is the ith coordinate of $\phi(x)$. Then

$$k(x, x') := \sum_{i=1}^{\infty} \phi_i(x)\phi_i(x') \tag{2}$$

Infinite sequences

Definition

The space ℓ_p of p-summable sequences is defined as all sequences $(a_i)_{i\geq 1}$ for which

$$\sum_{i=1}^{\infty} a_i^p < \infty.$$

Kernels can be defined in terms of sequences in ℓ_2.

Theorem

Given sequence of functions $(\phi_i(x))_{i\geq 1}$ in ℓ_2 where $\phi_i : \mathcal{X} \to \mathbb{R}$ is the ith coordinate of $\phi(x)$. Then

$$k(x, x') := \sum_{i=1}^{\infty} \phi_i(x)\phi_i(x') \tag{2}$$

Infinite sequences (proof)

Proof: We just need to check that inner product remains finite. Norm $\|a\|_{\ell_2}$ associated with inner product (2)

$$\|a\|_{\ell_2} := \sqrt{\sum_{i=1}^{\infty} a_i^2},$$

where a represents sequence with terms a_i. Via Cauchy-Schwarz,

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

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

Taylor series kernels

Definition (Taylor series kernel)

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

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

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

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

Example (Exponential kernel)

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

Taylor series kernel (proof)

Proof: By Cauchy-Schwarz,

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

so the Taylor series converges. Define $c_{j_1 \ldots j_d} = \frac{n!}{\prod_{i=1}^{d} j_i!}$

$$\begin{aligned}
k(x, x') &= \sum_{n=0}^{\infty} a_n \left(\sum_{j=1}^{d} x_j x_j'\right)^n \\
&= \sum_{n=0}^{\infty} a_n \sum_{\substack{j_1 \ldots j_d \geq 0 \\ j_1 + \ldots + j_d = n}} c_{j_1 \ldots j_d} \prod_{i=1}^{d} (x_i, x_i')^{j_i} \\
&= \sum_{j_1 \ldots j_d > 0} a_{j_1 + \ldots + j_d} c_{j_1 \ldots j_d} \prod_{i=1}^{d} x_i^{j_i} \prod_{i=1}^{d} (x_i')^{j_i}.
\end{aligned}$$

Gaussian kernel

Example (Gaussian kernel)

The Gaussian kernel on $\mathbb{R}^d$ is defined as

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

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

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 Gaussian 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 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if $\forall n \geq 1,\ \forall (a_1, \ldots a_n) \in \mathbb{R}^n,\ \forall (x_1, \ldots, x_n) \in \mathcal{X}^n$,

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

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 $\phi : \mathcal{X} \to \mathcal{H}$. Then $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} =: k(x, y)$ is positive definite.

Proof.

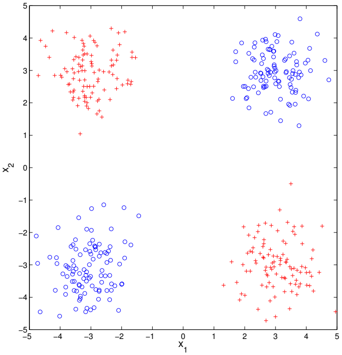
$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \langle a_i\phi(x_i), a_j\phi(x_j)\rangle_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^{n} a_i\phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0.\end{aligned}$$

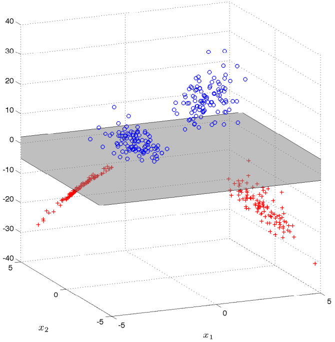
Reverse also holds: positive definite $k(x, x')$ is inner product in $\mathcal{H}$ between $\phi(x)$ and $\phi(x')$. □

The reproducing kernel Hilbert space

First example: finite space, polynomial features

Reminder: XOR example:





First example: finite space, polynomial features

Reminder: Feature space from XOR motivating example:

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

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 $\mathcal{H}$.

First 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,

$$f(\cdot) = \begin{bmatrix} f_1 & f_2 & f_3 \end{bmatrix}^\top.$$

$f(\cdot)$ 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 $\mathbb{R}$.

First 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,

$$f(\cdot) = \begin{bmatrix} f_1 & f_2 & f_3 \end{bmatrix}^\top.$$

$f(\cdot)$ 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 $\mathbb{R}$.

First example: finite space, polynomial features

$\phi(y)$ is a mapping from $\mathbb{R}^2$ to $\mathbb{R}^3$...
...which also parametrizes a function mapping $\mathbb{R}^2$ to $\mathbb{R}$.

$$k(\cdot, y) := \begin{bmatrix} y_1 & y_2 & y_1 y_2 \end{bmatrix}^\top = \phi(y),$$

Given y, there is a vector $k(\cdot, y)$ in $\mathcal{H}$ such that

$$\langle k(\cdot, y), \phi(x) \rangle_{\mathcal{H}} = a x_1 + b x_2 + c x_1 x_2,$$

where $a = y_1$, $b = y_2$, and $c = y_1 y_2$
Due to symmetry,

$$\begin{aligned} \langle k(\cdot, x), \phi(y) \rangle &= u y_1 + v y_2 + w y_1 y_2 \\ &= k(x, y). \end{aligned}$$

We can write $\phi(x) = k(\cdot, x)$ and $\phi(y) = k(\cdot, y)$ without ambiguity: canonical feature map

First example: finite space, polynomial features

$\phi(y)$ is a mapping from $\mathbb{R}^2$ to $\mathbb{R}^3$...
...which also parametrizes a function mapping $\mathbb{R}^2$ to $\mathbb{R}$.

$$k(\cdot, y) := \left[\begin{array}{ccc} y_1 & y_2 & y_1 y_2 \end{array}\right]^\top = \phi(y),$$

Given y, there is a vector $k(\cdot, y)$ in $\mathcal{H}$ such that

$$\langle k(\cdot, y), \phi(x)\rangle_{\mathcal{H}} = ax_1 + bx_2 + cx_1x_2,$$

where $a = y_1$, $b = y_2$, and $c = y_1 y_2$
Due to symmetry,

$$\begin{aligned} \langle k(\cdot, x), \phi(y)\rangle &= uy_1 + vy_2 + wy_1y_2 \\ &= k(x, y). \end{aligned}$$

We can write $\phi(x) = k(\cdot, x)$ and $\phi(y) = k(\cdot, y)$ without ambiguity: canonical feature map

The reproducing property

This example illustrates the two defining features of an RKHS:

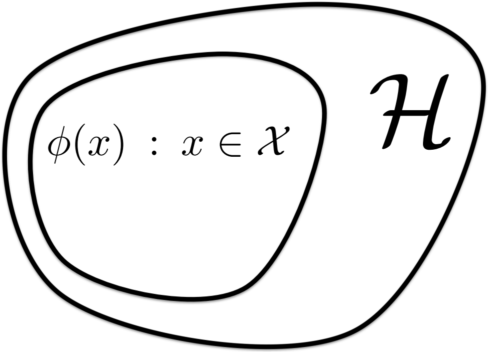
- **The reproducing property:**
  $\forall x \in \mathcal{X}, \forall f(\cdot) \in \mathcal{H}, \ \ \langle f(\cdot), k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$
  ...or use shorter notation $\langle f, \phi(x)\rangle_{\mathcal{H}}$.
- In particular, for any $x, y \in \mathcal{X}$,

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

Note: the feature map of every point is in the feature space:
$\forall x \in \mathcal{X}, \ \ k(\cdot, x) = \phi(x) \in \mathcal{H},$

First example: finite space, polynomial features

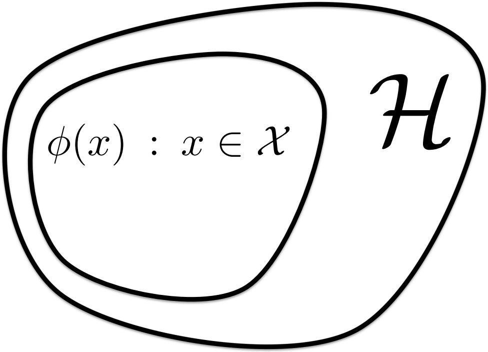
Another, more subtle point: $\mathcal{H}$ can be larger than all $\phi(x)$.
Why?



E.g. $f = [1\,1\,-1] \in \mathcal{H}$ cannot be obtained by $\phi(x) = [x_1\;x_2\;(x_1x_2)]$.

First example: finite space, polynomial features

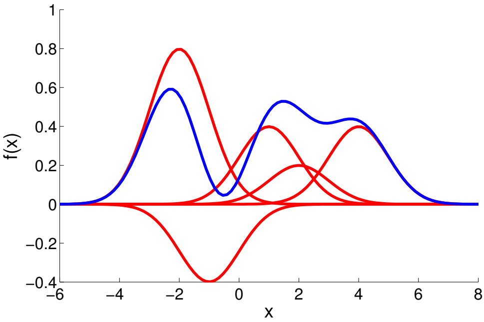
Another, more subtle point: $\mathcal{H}$ can be larger than all $\phi(x)$.
Why?



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

Second example: infinite feature space

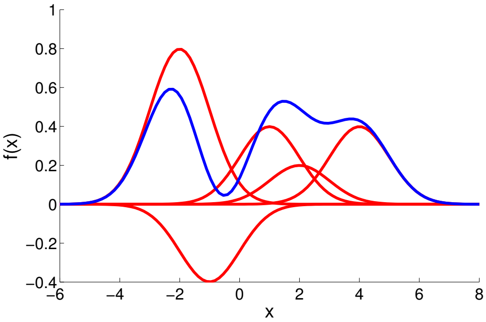
Reproducing property for function with Gaussian kernel:
$f(x) := \sum_{i=1}^{m} \alpha_i k(x_i, x) = \left\langle \sum_{i=1}^{m} \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}}$.



- What do the features $\phi(x)$ look like (warning: there are infinitely many of them!)
- What do these features have to do with smoothness?

Second example: infinite feature space

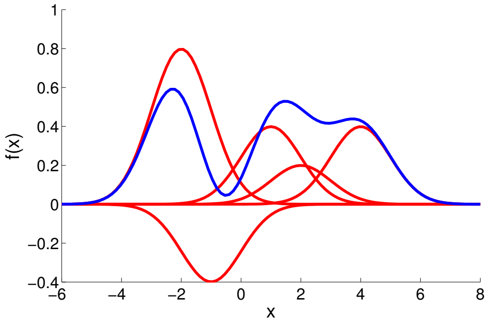
Reproducing property for function with Gaussian kernel:
$f(x) := \sum_{i=1}^{m} \alpha_i k(x_i, x) = \langle \sum_{i=1}^{m} \alpha_i \phi(x_i), \phi(x) \rangle_{\mathcal{H}}$.



- What do the features $\phi(x)$ look like (warning: there are infinitely many of them!)
- What do these features have to do with smoothness?

Second example: infinite feature space

Reproducing property for function with Gaussian kernel:
$f(x) := \sum_{i=1}^{m} \alpha_i k(x_i, x) = \left\langle \sum_{i=1}^{m} \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}}$.



- What do the features $\phi(x)$ look like (warning: there are infinitely many of them!)
- What do these features have to do with smoothness?

Second example: infinite feature space

Under certain conditions (e.g Mercer's theorem), we can write

$$k(x,x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x'), \qquad \int_{\mathcal{X}} e_i(x) e_j(x) d\mu(x) = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

where this sum is guaranteed to converge whatever the x and x'.

Infinite dimensional feature map: $$\phi(x) = \begin{bmatrix} \vdots \\ \sqrt{\lambda_i} e_i(x) \\ \vdots \end{bmatrix} \in \ell_2.$$

Define $\mathcal{H}$ to be the space of functions: for $\{f_i\}_{i=1}^{\infty} \in \ell_2$,

$$f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} f_i \sqrt{\lambda_i} e_i(x).$$

Does this work? Is $f(x) < \infty$ despite the infinite feature space?

Second example: infinite feature space

Under certain conditions (e.g Mercer's theorem), we can write

$$k(x,x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x'), \qquad \int_{\mathcal{X}} e_i(x) e_j(x) d\mu(x) = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

where this sum is guaranteed to converge whatever the x and x'.

Infinite dimensional feature map: $\phi(x) = \begin{bmatrix} \vdots \\ \sqrt{\lambda_i} e_i(x) \\ \vdots \end{bmatrix} \in \ell_2.$

Define $\mathcal{H}$ to be the space of functions: for $\{f_i\}_{i=1}^{\infty} \in \ell_2$,

$$f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} f_i \sqrt{\lambda_i} e_i(x).$$

Does this work? Is $f(x) < \infty$ despite the infinite feature space?

Second example: infinite feature space

Reminder: for the kernel, we obtained by Cauchy-Schwarz that if $\phi(x) \in \ell_2$ for all x, then

$$\left|k(x,x')\right| = \left|\sum_{i=1}^{\infty} \phi_i(x)\phi_i(x')\right| \leq \|\phi_i(x)\| \left\|\phi_i(x')\right\| < \infty$$

Finiteness of $f(x) = \langle f, \phi(x)\rangle_{\mathcal{H}}$ also obtained by Cauchy-Schwarz,

$$\begin{aligned}|\langle f, \phi(x)\rangle_{\mathcal{H}}| &= \left|\sum_{i=1}^{\infty} f_i \sqrt{\lambda_i} e_i(x)\right| \leq \left(\sum_{i=1}^{\infty} f_i^2\right)^{1/2} \left(\sum_{i=1}^{\infty} \lambda_i e_i^2(x)\right)^{1/2} \\ &= \|f\|_{\ell_2} \sqrt{k(x,x)}\end{aligned}$$

Second example: infinite feature space

Reminder: for the kernel, we obtained by Cauchy-Schwarz that if $\phi(x) \in \ell_2$ for all x, then

$$|k(x,x')| = \left| \sum_{i=1}^{\infty} \phi_i(x)\phi_i(x') \right| \leq \|\phi_i(x)\| \, \|\phi_i(x')\| < \infty$$

Finiteness of $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}$ also obtained by Cauchy-Schwarz,

$$|\langle f, \phi(x) \rangle_{\mathcal{H}}| = \left| \sum_{i=1}^{\infty} f_i \sqrt{\lambda_i} e_i(x) \right| \leq \left(\sum_{i=1}^{\infty} f_i^2 \right)^{1/2} \left(\sum_{i=1}^{\infty} \lambda_i e_i^2(x) \right)^{1/2}$$
$$= \|f\|_{\ell_2} \sqrt{k(x,x)}$$

Second example: infinite feature space

We can also define inner product in $\mathcal{H}$ between two functions f (represented by f_i) and g (represented by g_i) as

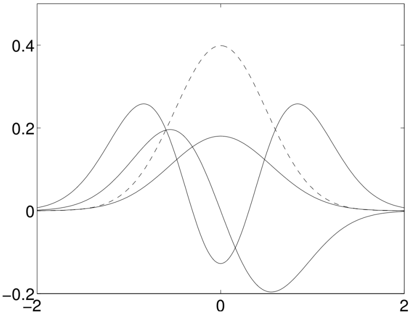
$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} f_i g_i .$$

Second example: infinite feature space

Gaussian kernel, $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$,

$$\begin{aligned}\lambda_k &\propto b^k \qquad b<1\\ e_k(x) &\propto \exp(-(c-a)x^2)H_k(x\sqrt{2c}),\end{aligned}$$

a, b, c are functions of σ, and H_k is kth order Hermite polynomial.



$$k(x,x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x')$$

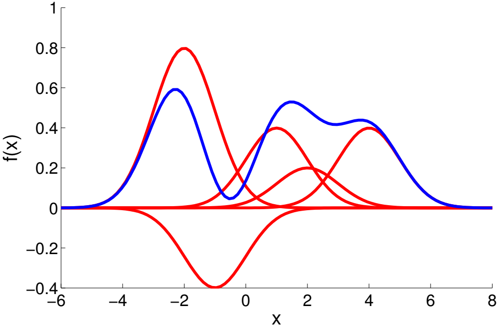
(Figure from Rasmussen and Williams)

Second example: infinite feature space

Example RKHS function, Gaussian kernel:

$$f(x) := \sum_{i=1}^{m} \alpha_i k(x_i, x) = \sum_{i=1}^{m} \alpha_i \left[\sum_{j=1}^{\infty} \lambda_j e_j(x_i) e_j(x) \right] = \sum_{j=1}^{\infty} f_j \left[\sqrt{\lambda_j} e_j(x) \right]$$

where $f_j = \sum_{i=1}^{m} \alpha_i \sqrt{\lambda_j} e_j(x_i)$.



NOTE that this enforces smoothing: λ_j decay as e_j become rougher, f_j decay since $\sum_j f_j^2 < \infty$.

Third (infinite) example: fourier series

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

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

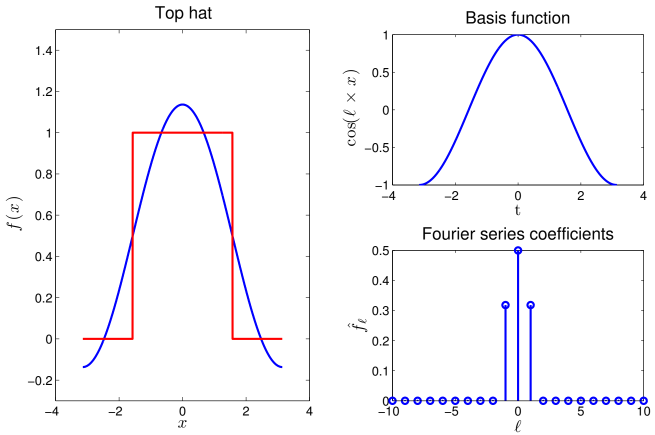
Example: "top hat" function,

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

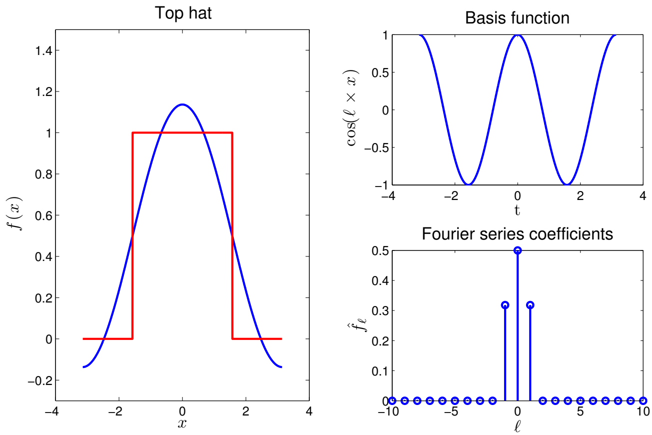
Fourier series:

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

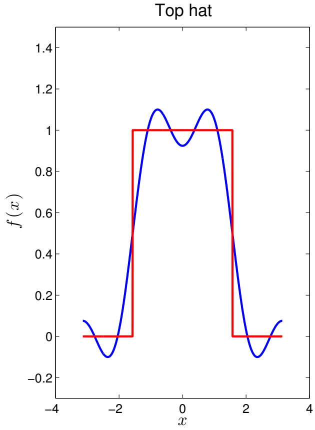
Fourier series for top hat function

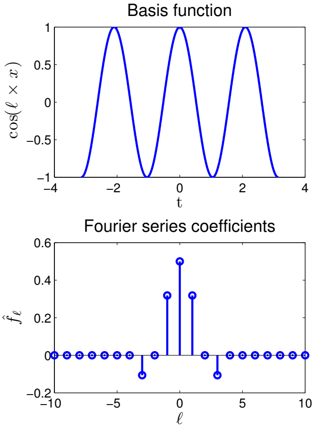


Fourier series for top hat function

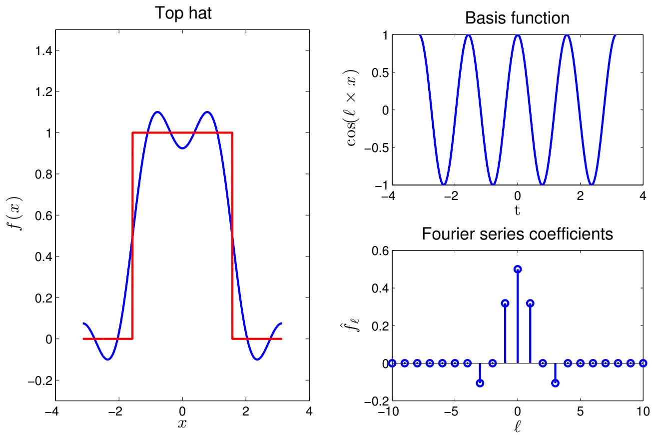


Fourier series for top hat function

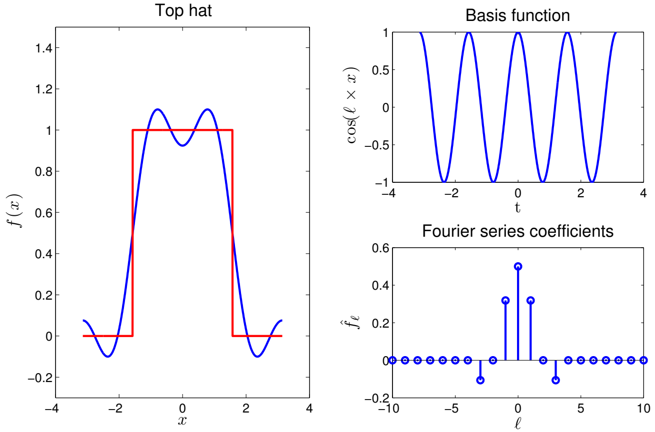




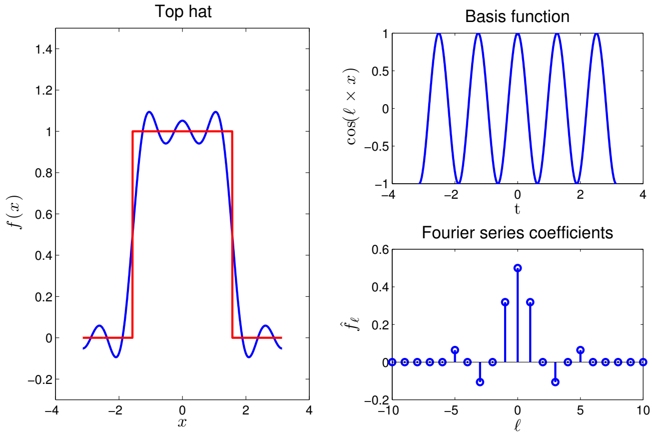
Fourier series for top hat function



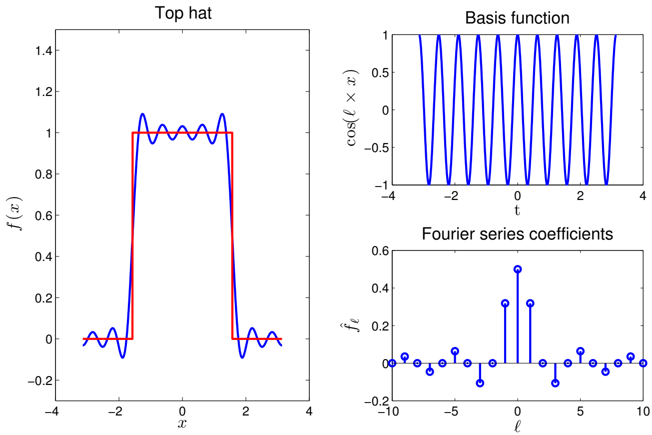
Fourier series for top hat function



Fourier series for top hat function



Fourier series for top hat function



Fourier series for kernel function

Kernel takes a single argument,

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

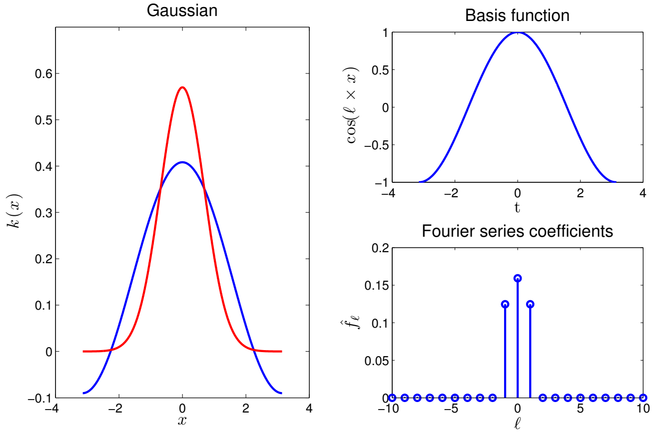
Define the Fourier series representation of k

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

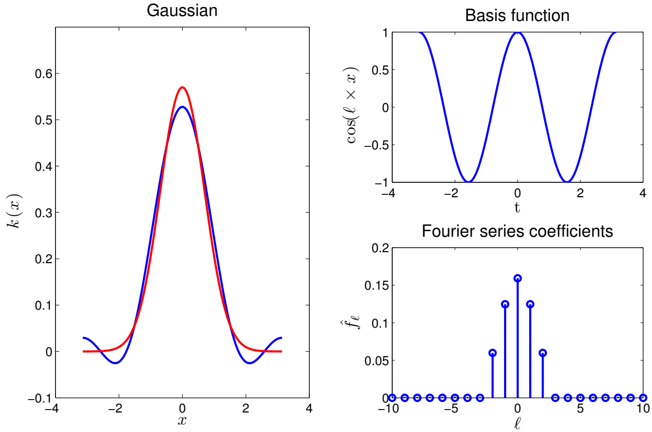
k and its Fourier transform are real and symmetric. E.g. Gaussian,

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

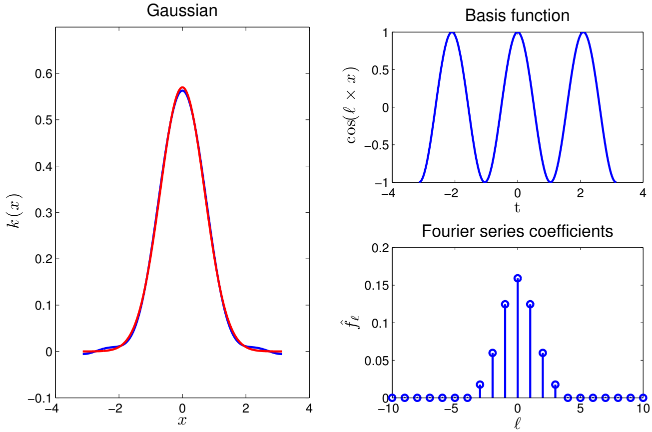
Fourier series for Gaussian kernel



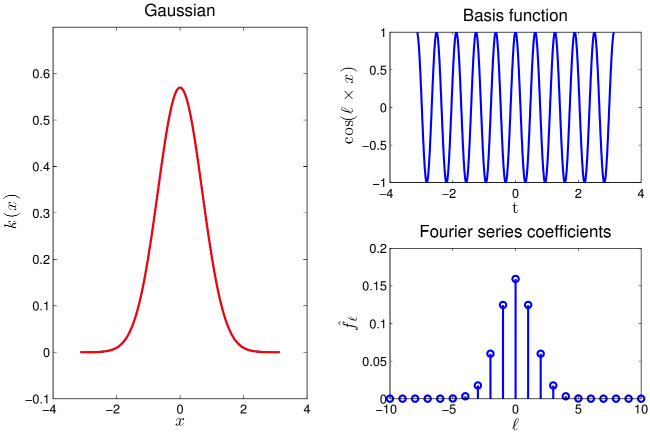
Fourier series for Gaussian kernel



Fourier series for Gaussian kernel



Fourier series for Gaussian kernel



Feature space via fourier series

Define $\mathcal{H}$ to be the space of functions with (infinite) feature space representation

$$f(\cdot) = \begin{bmatrix} \dots & \hat{f}_\ell / \sqrt{\hat{k}_\ell} & \dots \end{bmatrix}^\top .$$

The space $\mathcal{H}$ has an inner product:

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

Define the feature map

$$k(\cdot, x) = \phi(x) = \begin{bmatrix} \dots & \sqrt{\hat{k}_\ell} \exp(-\imath \ell x) & \dots \end{bmatrix}^\top$$

Feature space via fourier series

Define $\mathcal{H}$ to be the space of functions with (infinite) feature space representation

$$f(\cdot) = \begin{bmatrix} \dots & \hat{f}_\ell / \sqrt{\hat{k}_\ell} & \dots \end{bmatrix}^\top.$$

The space $\mathcal{H}$ has an inner product:

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

Define the feature map

$$k(\cdot, x) = \phi(x) = \begin{bmatrix} \dots & \sqrt{\hat{k}_\ell} \exp(-\imath \ell x) & \dots \end{bmatrix}^\top$$

Feature space via fourier series

The reproducing theorem holds,

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

...including for the kernel itself,

$$
\begin{aligned}
\langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}} &= \sum_{\ell=-\infty}^{\infty} \left(\sqrt{\hat{k}_\ell} \exp(-\imath \ell x)\right) \left(\overline{\sqrt{\hat{k}_\ell} \exp(-\imath \ell y)}\right) \\
&= \sum_{\ell=-\infty}^{\infty} \hat{k}_\ell \exp(\imath \ell (y - x)) = k(x - y).
\end{aligned}
$$

Feature space via fourier series

The reproducing theorem holds,

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

...including for the kernel itself,

$$\begin{aligned}\langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}} &= \sum_{\ell=-\infty}^{\infty} \left(\sqrt{\hat{k}_\ell}\exp(-\imath\ell x)\right)\left(\overline{\sqrt{\hat{k}_\ell}\exp(-\imath\ell y)}\right) \\ &= \sum_{\ell=-\infty}^{\infty} \hat{k}_\ell \exp(\imath\ell(y-x)) = k(x-y).\end{aligned}$$

Fourier series: what does it achieve?

The squared norm of a function f in $\mathcal{H}$ is:

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

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

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

Enforces smoothness.

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

Fourier series: what does it achieve?

The squared norm of a function f in $\mathcal{H}$ is:

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

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

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

Enforces smoothness.

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

Fourier series: what does it achieve?

The squared norm of a function f in $\mathcal{H}$ is:

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

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

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

Enforces smoothness.

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

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 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel of $\mathcal{H}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space, if

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

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

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}}. \tag{3}$$

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 δ_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 δ_x is bounded: $\forall x \in \mathcal{X}$ there exists $\lambda_x \geq 0$ such that for all $f \in \mathcal{H}$,

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

$\Longrightarrow$ two functions identical in RHKS 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 δ_x)

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

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

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

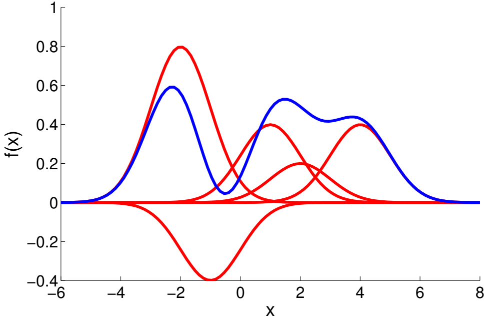
Cauchy-Schwarz in 3rd line . Consequently, $\delta_x : \mathcal{F} \to \mathbb{R}$ bounded with $\lambda_x = k(x,x)^{1/2}$ (other direction: Riesz theorem).

Moore-Aronsajn

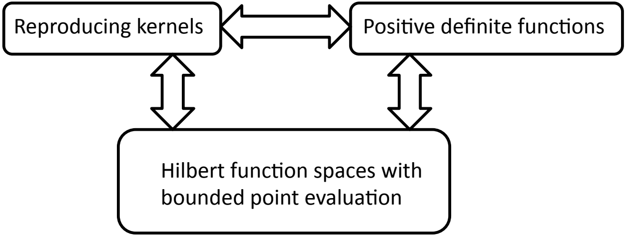
Theorem (Moore-Aronszajn)

Every positive definite kernel k uniquely associated with RKHS $\mathcal{H}$.

Recall feature map is *not* unique (as we saw earlier): only kernel is.
Example RKHS function, Gaussian kernel: $f(\cdot) := \sum_{i=1}^{m} \alpha_i k(x_i, \cdot)$.

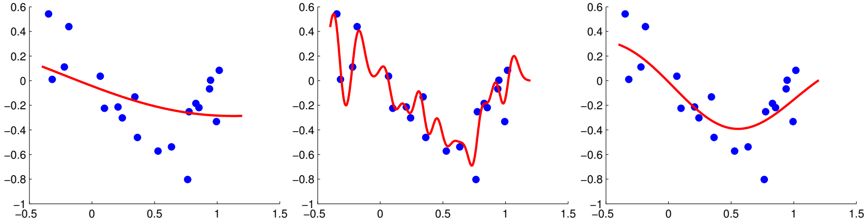


Correspondence



Kernel Ridge Regression

Kernel ridge regression



Very simple to implement, works well when no outliers.

Ridge regression: case of $\mathbb{R}^D$

We are given n training points in $\mathbb{R}^D$:

$$X = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix} \in \mathbb{R}^{D \times n} \quad y := \begin{bmatrix} y_1 & \dots & y_n \end{bmatrix}^\top$$

Define some $\lambda > 0$. Our goal is:

$$\begin{aligned} f^* &= \arg\min_{f \in \mathbb{R}^d} \left(\sum_{i=1}^n (y_i - x_i^\top f)^2 + \lambda \|f\|^2 \right) \\ &= \arg\min_{f \in \mathbb{R}^d} \left(\left\| y - X^\top f \right\|^2 + \lambda \|f\|^2 \right), \end{aligned}$$

The second term $\lambda \|f\|^2$ is chosen to avoid problems in high dimensional spaces (see below).

Ridge regression: case of $\mathbb{R}^D$

We are given n training points in $\mathbb{R}^D$:

$$X = \left[\begin{array}{ccc} x_1 & \dots & x_n \end{array} \right] \in \mathbb{R}^{D \times n} \quad y := \left[\begin{array}{ccc} y_1 & \dots & y_n \end{array} \right]^\top$$

Define some $\lambda > 0$. Our goal is:

$$\begin{aligned} a^* &= \arg\min_{f \in \mathbb{R}^d} \left(\sum_{i=1}^n (y_i - x_i^\top f)^2 + \lambda \|f\|^2 \right) \\ &= \arg\min_{f \in \mathbb{R}^d} \left(\left\| y - X^\top f \right\|^2 + \lambda \|f\|^2 \right), \end{aligned}$$

Solution is:

$$f^* = \left(XX^\top + \lambda I \right)^{-1} Xy,$$

which is the classic regularized least squares solution.

Kernel ridge regression

Use features of $\phi(x_i)$ in the place of x_i:

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

E.g. for finite dimensional feature spaces,

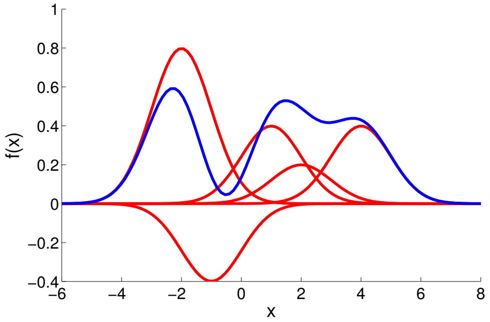
$$\phi_p(x) = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^\ell \end{bmatrix} \qquad \phi_s(x) = \begin{bmatrix} \sin x \\ \cos x \\ \sin 2x \\ \vdots \\ \cos \ell x \end{bmatrix}$$

a is a vector of length ℓ giving weight to each of these features so as to find the mapping between x and y. Feature vectors can also have *infinite* length (more soon).

Kernel ridge regression

Solution easy if we already know f is a linear combination of feature space mappings of points: representer theorem.

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



Representer theorem

Given a set of paired observations $(x_1, y_1), \ldots (x_n, y_n)$ (regression or classification).
Find the function f^* in the RKHS $\mathcal{H}$ which satisfies

$$J(f^*) = \min_{f \in \mathcal{H}} J(f), \qquad (4)$$

where

$$J(f) = L_y(f(x_1), \ldots, f(x_n)) + \Omega\left(\|f\|^2_{\mathcal{H}}\right),$$

Ω is non-decreasing, and y is the vector of y_i.

- Classification: $L_y(f(x_1), \ldots, f(x_n)) = \sum_{i=1}^n \mathbb{I}_{y_i f(x_i) \leq 0}$
- Regression: $L_y(f(x_1), \ldots, f(x_n)) = \sum_{i=1}^n (y_i - f(x_i))^2$

Representer theorem

The representer theorem:(simple version) solution to

$$\min_{f \in \mathcal{H}} \left[L_y(f(x_1), \ldots, f(x_n)) + \Omega \left(\|f\|_{\mathcal{H}}^2 \right) \right]$$

takes the form

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

If Ω is strictly increasing, all solutions have this form.

Representer theorem: proof

Proof: Denote f_s projection of f onto the subspace

$$\mathrm{span}\left\{k(x_i,\cdot):\ 1 \le i \le n\right\}, \tag{5}$$

such that

$$f = f_s + f_\perp,$$

where $f_s = \sum_{i=1}^n \alpha_i k(x_i,\cdot)$.
Regularizer:

$$\|f\|_{\mathcal{H}}^2 = \|f_s\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2 \ge \|f_s\|_{\mathcal{H}}^2,$$

then

$$\Omega\left(\|f\|_{\mathcal{H}}^2\right) \ge \Omega\left(\|f_s\|_{\mathcal{H}}^2\right),$$

so this term is minimized for $f = f_s$.

Representer theorem: proof

Proof (cont.): Individual terms $f(x_i)$ in the loss:

$$f(x_i) = \langle f, k(x_i, \cdot)\rangle_{\mathcal{H}} = \langle f_s + f_\perp, k(x_i, \cdot)\rangle_{\mathcal{H}} = \langle f_s, k(x_i, \cdot)\rangle_{\mathcal{H}},$$

so

$$L_y(f(x_1), \ldots, f(x_n)) = L_y(f_s(x_1), \ldots, f_s(x_n)).$$

Hence

- Loss $L(\ldots)$ only depends on the component of f in the data subspace,
- Regularizer $\Omega(\ldots)$ minimized when $f = f_s$.
- If Ω is strictly non-decreasing, then $\|f_\perp\|_{\mathcal{H}} = 0$ is required at the minimum.

Kernel ridge regression: proof

We *begin* knowing f is a linear combination of feature space mappings of points (representer theorem)

$$f = \sum_{i=1}^{n} \alpha_i \phi(x_i).$$

Then

$$\begin{aligned} \sum_{i=1}^{n} \left(y_i - \langle f, \phi(x_i) \rangle_{\mathcal{H}}\right)^2 + \lambda \|f\|_{\mathcal{H}}^2 &= \|y - K\alpha\|^2 + \lambda \alpha^\top K \alpha \\ &= y^\top y - 2y^\top K\alpha + \alpha^\top \left(K^2 + \lambda K\right) \alpha \end{aligned}$$

Differentiating wrt α and setting this to zero, we get

$$\alpha^* = (K + \lambda I_n)^{-1} y.$$

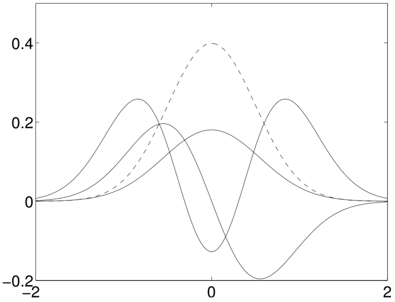
Recall: $\frac{\partial \alpha^\top U \alpha}{\partial \alpha} = (U + U^\top)\alpha, \qquad \frac{\partial v^\top \alpha}{\partial \alpha} = \frac{\partial \alpha^\top v}{\partial \alpha} = v$

Reminder: smoothness

What does $\|a\|_{\mathcal{H}}$ have to do with smoothing?

Example 1: The Gaussian kernel. Recall

$$f(x) = \sum_{i=1}^{\infty} a_i \sqrt{\lambda_i} e_i(x), \qquad \|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty} a_i^2.$$



Reminder: smoothness

What does $\|a\|_{\mathcal{H}}$ have to do with smoothing?
Example 2: The Fourier series representation:

$$f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l \exp(\imath l x),$$

and

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

Thus,

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

Parameter selection for KRR

Given the objective

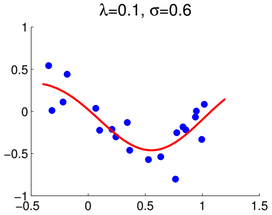
$$f^* = \arg\min_{f \in \mathcal{H}} \left(\sum_{i=1}^{n} (y_i - \langle f, \phi(x_i) \rangle_{\mathcal{H}})^2 + \lambda \|f\|_{\mathcal{H}}^2 \right).$$

How do we choose

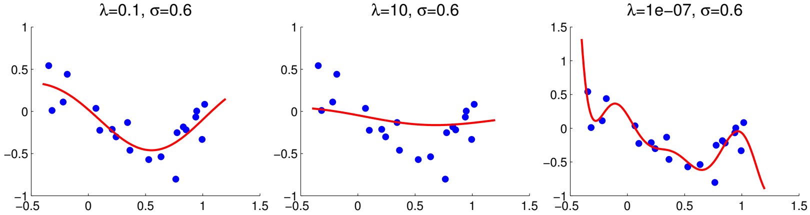
- The regularization parameter λ?
- The kernel parameter: for Gaussian kernel, σ in

$$k(x, y) = \exp\left(\frac{-\|x - y\|^2}{\sigma} \right).$$

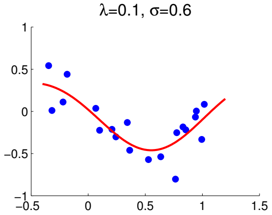
Choice of λ



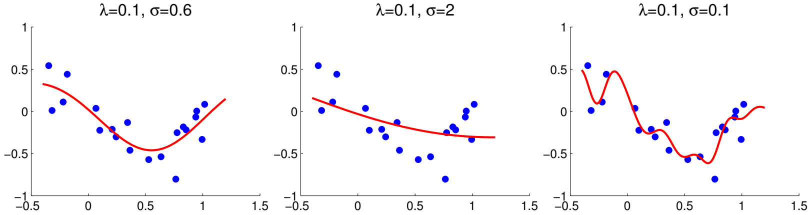
Choice of λ



Choice of σ



Choice of σ



Cross validation

- Split data into training set size n_{tr} and **test set** size $n_{\mathrm{te}} = 1 - n_{\mathrm{tr}}$.
- Split trainining set into m equal chunks of size $n_{\mathrm{val}} = n_{\mathrm{tr}}/m$. Call these $X_{\mathrm{val},i}, Y_{\mathrm{val},i}$ for $i \in \{1, \ldots, m\}$
- For each λ, σ pair
  - For each $X_{\mathrm{val},i}, Y_{\mathrm{val},i}$
    - Train ridge regression on remaining trainining set data $X_{\mathrm{tr}} \setminus X_{\mathrm{val},i}$ and $Y_{\mathrm{tr}} \setminus Y_{\mathrm{val},i}$,
    - Evaluate its error on the validation data $X_{\mathrm{val},i}, Y_{\mathrm{val},i}$
  - Average the errors on the validation sets to get the average validation error for λ, σ.
- Choose λ^*, σ^* with the lowest average validation error
- Measure the performance on the test set $X_{\mathrm{te}}, Y_{\mathrm{te}}$.