Reproducing Kernel Hilbert Spaces in Machine Learning

Arthur Gretton, Gatsby Unit, CSML, UCL

October 25, 2017

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)
4. Further applications of kenels (feature selection, clustering, ICA)
5. Support vector machines for classification, regression
6. Theory of reproducing kernel Hilbert spaces (optional, not assessed)

Lecture notes will be put online at:
http://www.gatsby.ucl.ac.uk/~gretton/rkhscourse.html

Assessment and locations

The course has the following assessment components:

- Written Examination (2.5 hours, 50%)
- Coursework (50%)

To pass this course, you must pass *both* the exam and the coursework

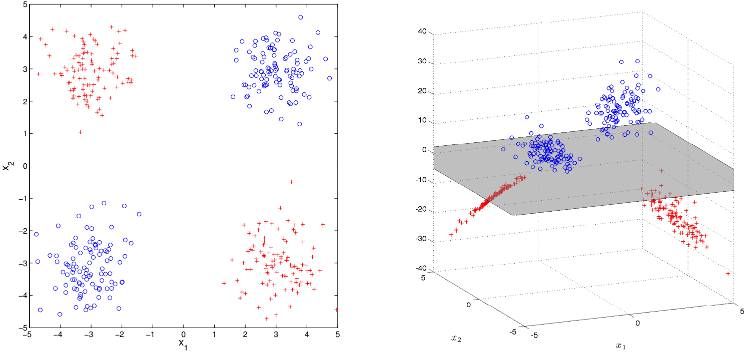
Course times, locations

Lectures will be at the Ground Floor Lecture Theatre, Sainsbury Wellcome Centre (with a couple of exceptions late in the term)

- Kernel lectures are Wednesday, 11:30 -13:00,
- Theory lectures are Friday 14:00 -15:30

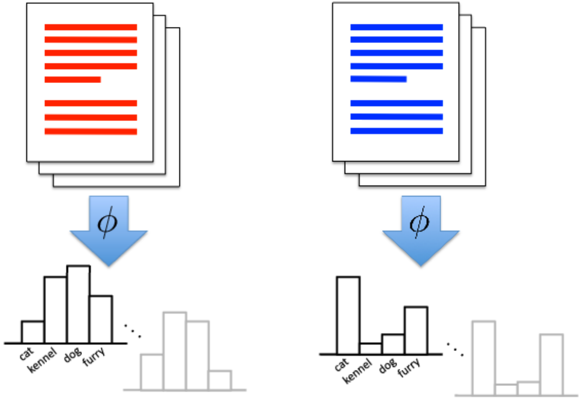
(with a couple of exceptions!)
There will be lectures during reading week, due to clash with NIPS conference.
The tutor for the kernels part is `Michael Arbel`.

Why kernel methods (1): XOR example



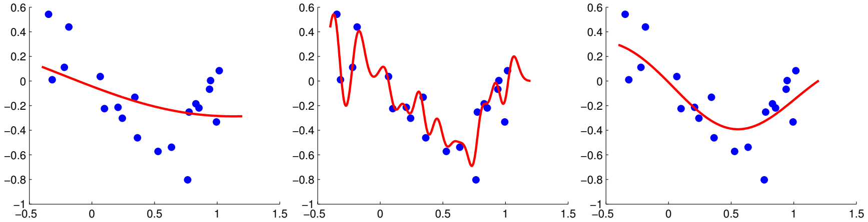
- No linear classifier separates red from blue
- Map points to **higher dimensional feature space**:
  $\phi(x) = \left[\begin{array}{ccc} x_1 & x_2 & x_1 x_2 \end{array} \right] \in \mathbb{R}^3$

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 feature 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 feature maps. 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 $\mathcal{X}$.

(Proof via positive definiteness: later!) 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 $\mathcal{X}$.

(Proof via positive definiteness: later!) 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!

k_1 is a kernel between **shapes**,

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

k_2 is a kernel between **colors**,

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

New kernels from old: products

"Natural" feature space for **colored shapes**:

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

Kernel is:

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

$$= \mathrm{tr}\left(\underbrace{\phi_1^\top(x')\phi_1(x)}_{k_1(x,x')}\right) k_2(x,x') = k_1(x,x')k_2(x,x')$$

New kernels from old: products

"Natural" feature space for **colored shapes**:

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

Kernel is:

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

$$= \mathrm{tr}\left(\underbrace{\phi_1^\top(x')\phi_1(x)}_{k_1(x,x')}\right) k_2(x,x') = k_1(x,x')k_2(x,x')$$

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.

$$k(x, y) = \left[\begin{array}{ccc} \sin(x) & x^3 & \log x \end{array} \right]^\top \left[\begin{array}{ccc} \sin(y) & y^3 & \log y \end{array} \right]$$

where $\phi(x) = \left[\begin{array}{ccc} \sin(x) & x^3 & \log x \end{array} \right]$

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

Infinite sequences

Definition

The space ℓ_2 (**square** summable sequences) comprises all sequences $(a_i)_{i\geq 1}$ for which

$$\sum_{i=1}^{\infty} a_i^2 < \infty.$$

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 $\phi(x)$. A well-defined kernel k on $\mathcal{X}$ is

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

Infinite sequences

Definition

The space ℓ_2 (**square** summable sequences) comprises all sequences $(a_i)_{i\geq 1}$ for which

$$\sum_{i=1}^{\infty} a_i^2 < \infty.$$

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 $\phi(x)$. A well-defined kernel k on $\mathcal{X}$ is

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

Infinite sequences (proof)

Proof: We just need to check that inner product remains finite. Norm $\|a\|_{\ell_2}$ associated with inner product (1)

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

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

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

By Cauchy-Schwarz,

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

so the sum converges.

Exponentiated quadratic kernel

Example (Exponentiated quadratic kernel)

This 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, like 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 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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 $\phi(x')$. □

Sum of kernels is a kernel

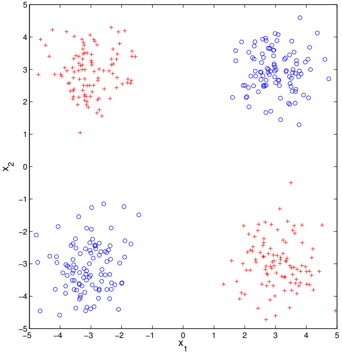
Consider two kernels $k_1(x, x')$ and $k_2(x, x')$. Then

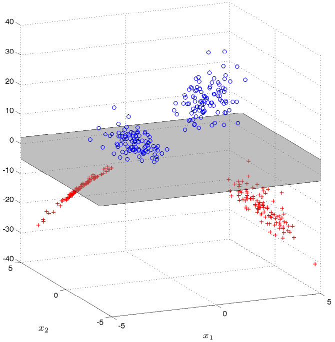
$$
\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \left[k_1(x_i, x_j) + k_2(x_i, x_j)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k_1(x_i, x_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k_2(x_i, x_j) \\
&\geq 0
\end{aligned}
$$

The reproducing kernel Hilbert space

First example: finite space, polynomial features

Reminder: XOR example:





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 $\mathbb{R}$.

What if we have infinitely many features?

Exponentiated quadratic kernel,

$$k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) = \sum_{i=1}^{\infty} \phi_i(x)\phi_i(y)$$

$$f(x) = \sum_{i=1}^{\infty} f_i \phi_i(x) \qquad \sum_{i=1}^{\infty} f_i^2 < \infty.$$

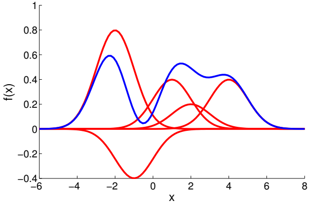
What if we have infinitely many features?

Function with exponentiated quadratic kernel:

$$f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x)$$

$$= \sum_{i=1}^{m} \alpha_i \langle \phi(x_i), \phi(x) \rangle_{\mathcal{H}}$$

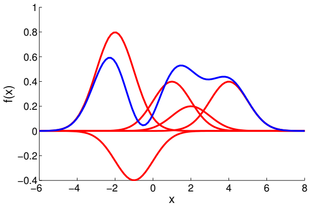
$$= \left\langle \sum_{i=1}^{m} \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}}$$



What if we have infinitely many features?

Function with exponentiated quadratic kernel:

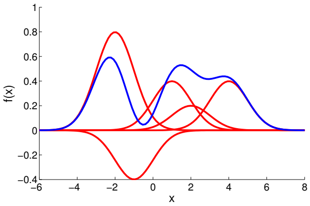
$$
\begin{aligned}
f(x) &= \sum_{i=1}^{m} \alpha_i k(x_i, x) \\
&= \sum_{i=1}^{m} \alpha_i \left\langle \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{m} \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}}
\end{aligned}
$$



What if we have infinitely many features?

Function with exponentiated quadratic kernel:

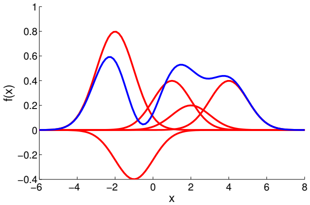
$$
\begin{aligned}
f(x) &= \sum_{i=1}^{m} \alpha_i k(x_i, x) \\
&= \sum_{i=1}^{m} \alpha_i \left\langle \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{m} \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}}
\end{aligned}
$$



What if we have infinitely many features?

Function with exponentiated quadratic kernel:

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



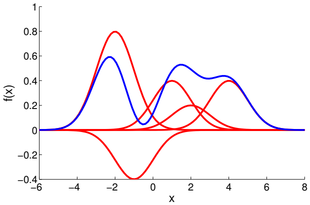
$f_\ell := \sum_{i=1}^{m} \alpha_i \phi_\ell(x_i)$

Possible to write functions of infinitely many features!

What if we have infinitely many features?

Function with exponentiated quadratic kernel:

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



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

Possible to write functions of infinitely many features!

The feature map is *also* a function

On previous page,

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

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

Then

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

The feature map is *also* a function

On previous page,

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

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

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

The feature map is *also* a function

On previous page,

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

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

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

....so the feature map is a (very simple) function!
We can write without ambiguity

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

The feature map is *also* a function

On previous page,

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

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

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

....so the feature map is a (very simple) function!
We can write without ambiguity

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

The reproducing property

This example illustrates the two defining features of an RKHS:

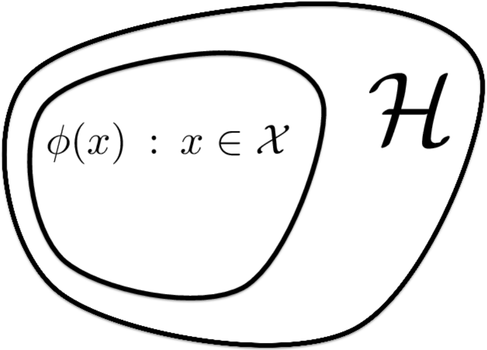
- **The reproducing property:**
  $\forall x \in \mathcal{X}, \forall f(\cdot) \in \mathcal{H}, \;\; \langle f(\cdot), k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$
  ...or use shorter notation $\langle f, \phi(x)\rangle_{\mathcal{H}}$.
- In particular, for any $x, y \in \mathcal{X}$,

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

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

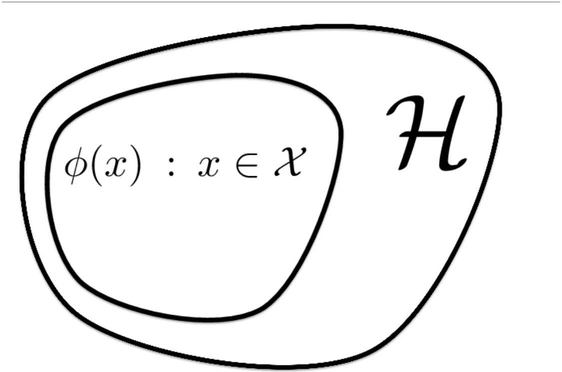
Another, more subtle point: $\mathcal{H}$ can be larger than all $\phi(x)$.



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 $\phi(x)$.



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

Second (infinite) example: fourier series

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

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

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

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

Example: "top hat" function,

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

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

Second (infinite) example: fourier series

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

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

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

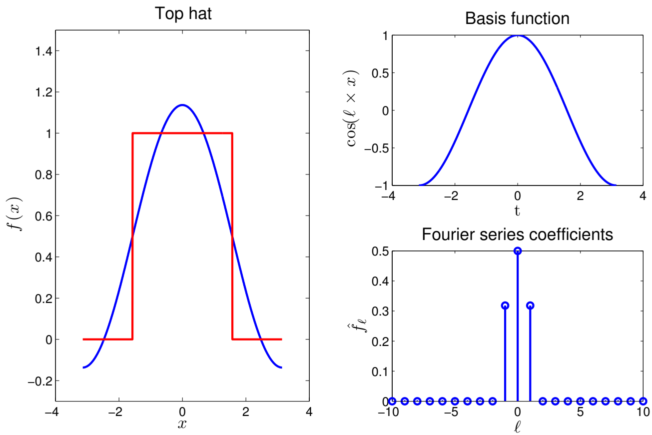
$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(\imath \ell x) \overline{\exp(\imath m x)} dx = \begin{cases} 1 & \ell = m, \\ 0 & \ell \neq m. \end{cases}$$

Example: "top hat" function,

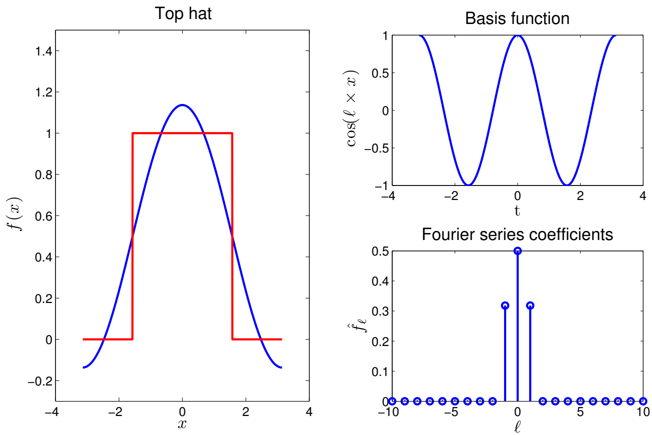
$$f(x) = \begin{cases} 1 & |x| < T, \\ 0 & T \leq |x| < \pi. \end{cases}$$

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

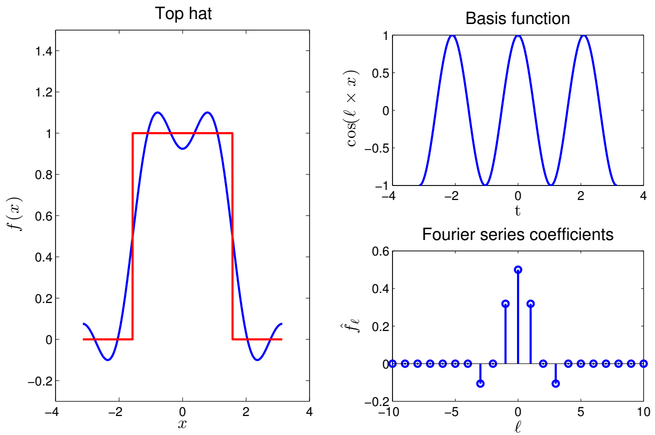
Fourier series for top hat function



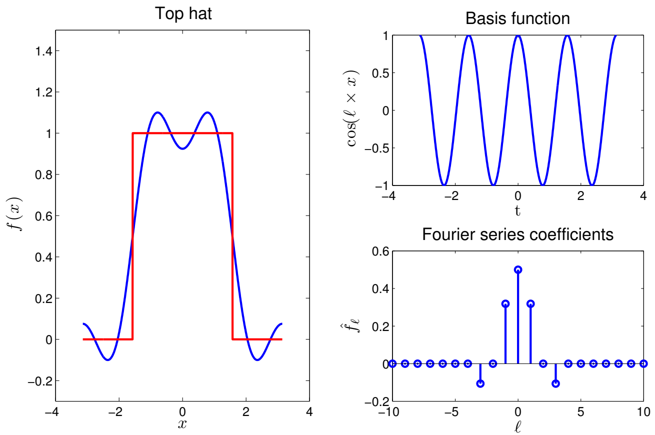
Fourier series for top hat function



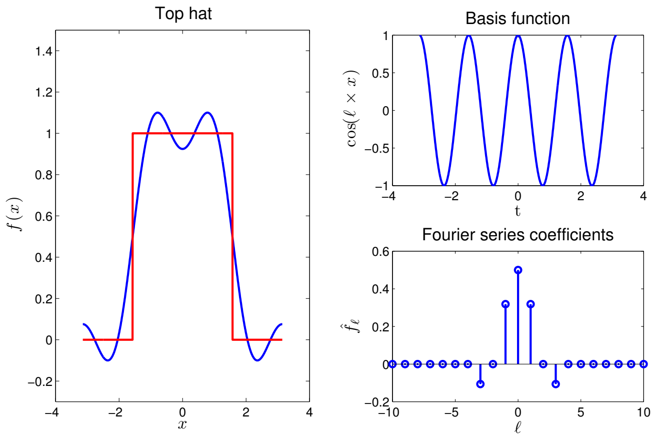
Fourier series for top hat function



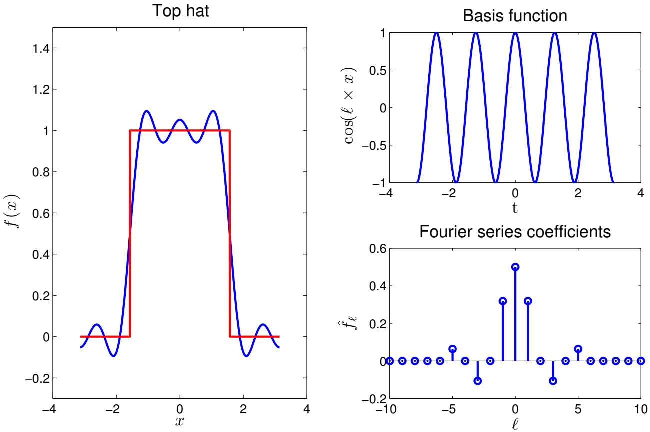
Fourier series for top hat function



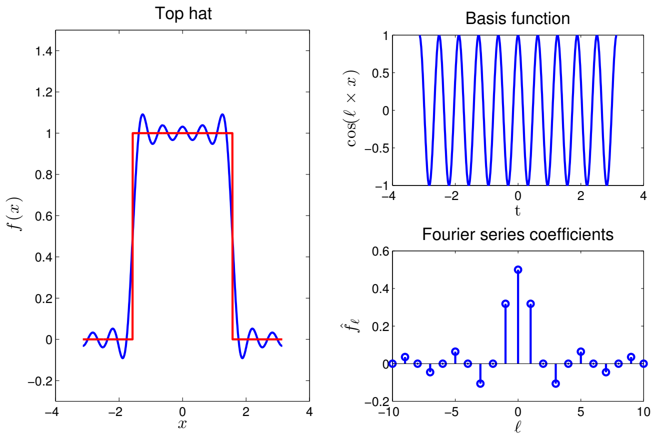
Fourier series for top hat function



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

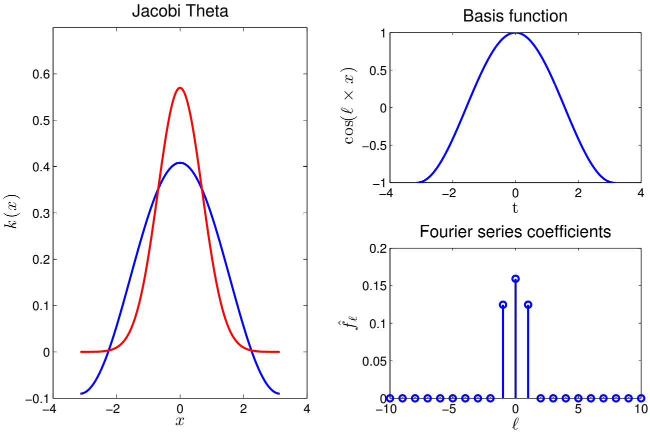
$$k(x) = \sum_{\ell=-\infty}^{\infty} \hat{k}_\ell \exp\left(\imath \ell x\right),$$

k and its Fourier transform are real and symmetric. For example,

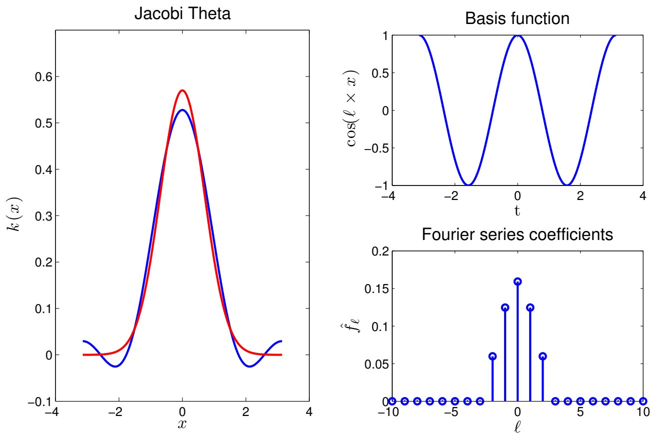
$$k(x) = \frac{1}{2\pi}\vartheta\left(\frac{x}{2\pi}, \frac{\imath\sigma^2}{2\pi}\right), \qquad \hat{k}_\ell = \frac{1}{2\pi}\exp\left(\frac{-\sigma^2\ell^2}{2}\right).$$

ϑ is the Jacobi theta function, close to exponentiated quadratic when σ^2 sufficiently narrower than $[-\pi, \pi]$.

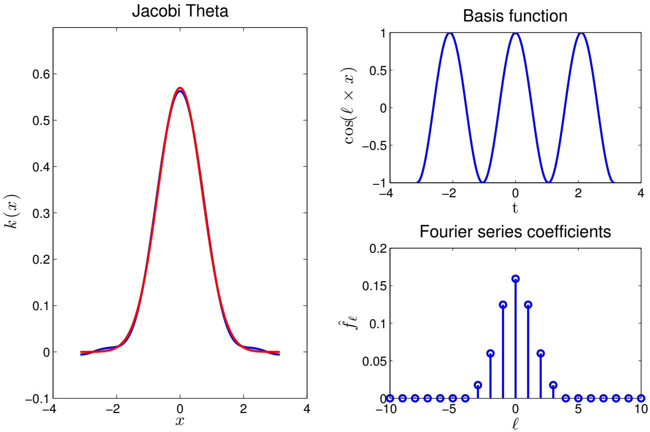
Fourier series for "Gaussian spectrum" kernel



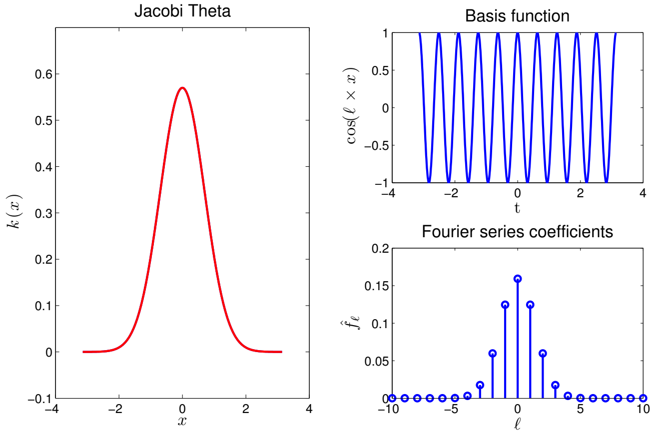
Fourier series for "Gaussian spectrum" kernel



Fourier series for "Gaussian spectrum" kernel



Fourier series for "Gaussian spectrum" kernel



RKHS via fourier series

Recall standard dot product in L_2:

$$\begin{aligned}
\langle f, g \rangle_{L_2} &= \left\langle \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell x), \sum_{m=-\infty}^{\infty} \overline{\hat{g}_m \exp(\imath m x)} \right\rangle_{L_2} \\
&= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}}_\ell \left\langle \exp(\imath \ell x), \exp(-\imath m x) \right\rangle_{L_2} \\
&= \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}}_\ell .
\end{aligned}$$

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

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

RKHS via fourier series

Recall standard dot product in L_2:

$$\begin{aligned}
\langle f, g \rangle_{L_2} &= \left\langle \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell x), \sum_{m=-\infty}^{\infty} \overline{\hat{g}_m \exp(\imath m x)} \right\rangle_{L_2} \\
&= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}}_\ell \left\langle \exp(\imath \ell x), \exp(-\imath m x) \right\rangle_{L_2} \\
&= \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}}_\ell .
\end{aligned}$$

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:

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

If $\hat{k}_\ell$ decays fast, then so must $\hat{f}_\ell$ if we want $\|f\|^2_{\mathcal{H}} < \infty$.
Recall $f(x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \left(\cos(\ell x) + \imath \sin(\ell x)\right)$.

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

Warning: need stronger conditions on kernel than L_2 convergence: Mercer's theorem (later).

Roughness penalty explained

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

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

If $\hat{k}_\ell$ decays fast, then so must $\hat{f}_\ell$ if we want $\|f\|_{\mathcal{H}}^2 < \infty$.
Recall $f(x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \left(\cos(\ell x) + \imath \sin(\ell x)\right)$.

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

Warning: need stronger conditions on kernel than L_2 convergence: Mercer's theorem (later).

Roughness penalty explained

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

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

If $\hat{k}_\ell$ decays fast, then so must $\hat{f}_\ell$ if we want $\|f\|_{\mathcal{H}}^2 < \infty$.
Recall $f(x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \left(\cos(\ell x) + \imath \sin(\ell x)\right)$.
Question: is the **top hat** function in the "Gaussian spectrum" RKHS?

Warning: need stronger conditions on kernel than L_2 convergence: Mercer's theorem (later).

Feature map and reproducing property

Reproducing property: define a function

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

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

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

Feature map and reproducing property

Reproducing property for the kernel:
Recall kernel definition:

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

Define two functions

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

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

Feature map and reproducing property

Check the reproducing property:

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

Link back to original RKHS definition

Original form of a function in the RKHS was (detail: sum now from $-\infty$ to ∞, complex conjugate)

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

We've defined the RKHS dot product as

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

Link back to original RKHS definition

Original form of a function in the RKHS was (detail: sum now from $-\infty$ to ∞, complex conjugate)

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

We've defined the RKHS dot product as

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

Link back to original RKHS definition

Original form of a function in the RKHS was (detail: sum now from $-\infty$ to ∞, complex conjugate)

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

We've defined the RKHS dot product as

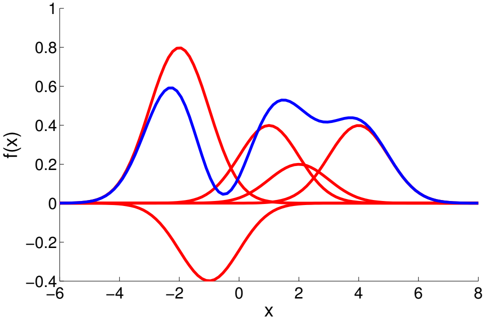
$$\langle f, g \rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_\ell \overline{\hat{g}_\ell}}{\hat{k}_\ell} \qquad \langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell \left(\overline{\hat{k}_\ell \exp(-\imath \ell z)} \right)}{\left(\sqrt{\hat{k}_\ell}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

By inspection

$$f_\ell = \hat{f}_\ell / \sqrt{\hat{k}_\ell} \qquad \phi_\ell(x) = \sqrt{\hat{k}_\ell} \exp(-\imath \ell x).$$

Third example: infinite feature space on $\mathbb{R}$

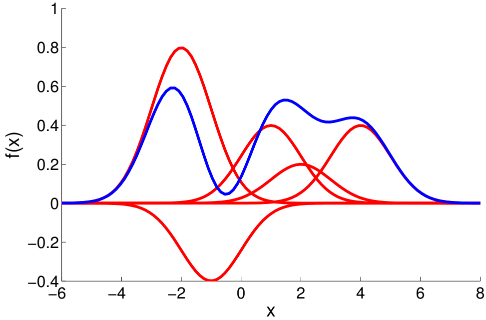
Reproducing property for function with exponentiated quadratic kernel on $\mathbb{R}$: $f(x) := \sum_{i=1}^{m} \alpha_i k(x_i, x) = \left\langle \sum_{i=1}^{m} \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}}$.



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

Third example: infinite feature space on $\mathbb{R}$

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



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

Third example: infinite feature space on $\mathbb{R}$

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

$$d\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2\right) dx$$

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

$$\lambda_i e_i(x) = \int k(x, x') e_i(x') d\mu(x'), \qquad \int_{L_2(\mu)} e_i(x) e_j(x) d\mu(x) = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

We can write

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

which converges in $L_2(\mu)$.
Warning: again, need stronger conditions on kernel than L_2 convergence.

Third example: infinite feature space on $\mathbb{R}$

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

$$d\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2\right) dx$$

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

$$\lambda_i e_i(x) = \int k(x, x') e_i(x') d\mu(x'), \qquad \int_{L_2(\mu)} e_i(x) e_j(x) d\mu(x) = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

We can write

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

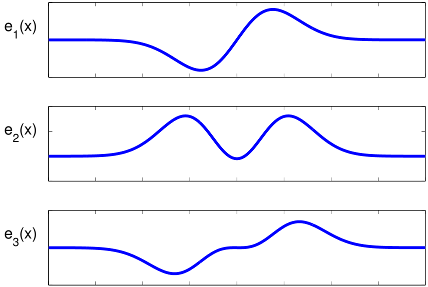
which converges in $L_2(\mu)$.
Warning: again, need stronger conditions on kernel than L_2 convergence.

Third example: infinite feature space on $\mathbb{R}$

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

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

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



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

Result from Rasmussen and Williams (2006, Section 4.3)

Third example: infinite feature space

Reminder: for two functions f, g in $L_2(\mu)$,

$$f(x) = \sum_{\ell=1}^{\infty} \hat{f}_\ell e_\ell(x) \qquad g(x) = \sum_{\ell=1}^{\infty} \hat{g}_\ell e_\ell(x),$$

dot product is

$$\begin{aligned}\langle f, g \rangle_{L_2(\mu)} &= \left\langle \sum_{\ell=1}^{\infty} \hat{f}_\ell e_\ell(x), \sum_{\ell=1}^{\infty} \hat{g}_\ell e_\ell(x) \right\rangle_{L_2(\mu)} \\ &= \sum_{\ell=1}^{\infty} \hat{f}_\ell \hat{g}_\ell.\end{aligned}$$

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

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

Third example: infinite feature space

Reminder: for two functions f, g in $L_2(\mu)$,

$$f(x) = \sum_{\ell=1}^{\infty} \hat{f}_\ell e_\ell(x) \qquad g(x) = \sum_{\ell=1}^{\infty} \hat{g}_\ell e_\ell(x),$$

dot product is

$$\begin{aligned}\langle f, g\rangle_{L_2(\mu)} &= \left\langle \sum_{\ell=1}^{\infty} \hat{f}_\ell e_\ell(x), \sum_{\ell=1}^{\infty} \hat{g}_\ell e_\ell(x) \right\rangle_{L_2(\mu)} \\ &= \sum_{\ell=1}^{\infty} \hat{f}_\ell \hat{g}_\ell.\end{aligned}$$

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

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

Link back to the original RKHS definition

Original form of a function in the RKHS was

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

Link back to the original RKHS definition

Original form of a function in the RKHS was

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

We've defined the RKHS dot product as

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

Link back to the original RKHS definition

Original form of a function in the RKHS was

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

We've defined the RKHS dot product as

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

Link back to the original RKHS definition

Original form of a function in the RKHS was

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

We've defined the RKHS dot product as

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

Link back to the original RKHS definition

Original form of a function in the RKHS was

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

We've defined the RKHS dot product as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{l=1}^{\infty} \frac{\hat{f}_\ell \hat{g}_\ell}{\lambda_\ell} \qquad \langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell \left(\lambda_\ell e_\ell(z)\right)}{\left(\sqrt{\lambda_\ell}\right)^2}$$

Link back to the original RKHS definition

Original form of a function in the RKHS was

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

We've defined the RKHS dot product as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{l=1}^{\infty} \frac{\hat{f}_\ell \hat{g}_\ell}{\lambda_\ell} \qquad \langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell \left(\lambda_\ell e_\ell(z)\right)}{\left(\sqrt{\lambda_\ell}\right)^2}$$

By inspection

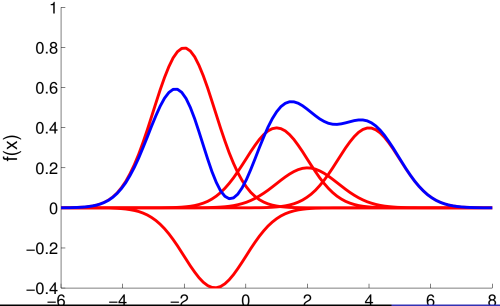
$$f_\ell = \hat{f}_\ell / \sqrt{\lambda_\ell} \qquad \phi_\ell(x) = \sqrt{\lambda_\ell} e_\ell(x).$$

Writing RKHS functions without explicit features

Example RKHS function from earlier:

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

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 $\sum_j f_j^2 < \infty$.

Explicit feature space as element of ℓ_2

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

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

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

and by triangle inequality,

$$\begin{aligned}
\|f\|_{\ell_2} &= \left\| \sum_{i=1}^{m} \alpha_i \phi(x_i) \right\| \\
&\le \sum_{i=1}^{m} |\alpha_i| \, \|\phi(x_i)\| < \infty.
\end{aligned}$$

Explicit feature space as element of ℓ_2

Does this work? Is $f(x) < \infty$ despite the infinite feature space?
Finiteness of $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}$ obtained by Cauchy-Schwarz,

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

and by triangle inequality,

$$
\begin{aligned}
\|f\|_{\ell_2} &= \left\| \sum_{i=1}^{m} \alpha_i \phi(x_i) \right\| \\
&\leq \sum_{i=1}^{m} |\alpha_i| \, \|\phi(x_i)\| < \infty.
\end{aligned}
$$

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 $x, y \in \mathcal{X}$,

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

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 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

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

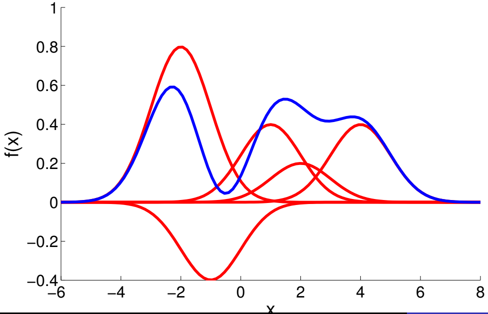
Cauchy-Schwarz in 3rd line . Consequently, $\delta_x : \mathcal{F} \to \mathbb{R}$ bounded with $\lambda_x = k(x,x)^{1/2}$ (other direction: Riesz theorem).

Moore-Aronsajn

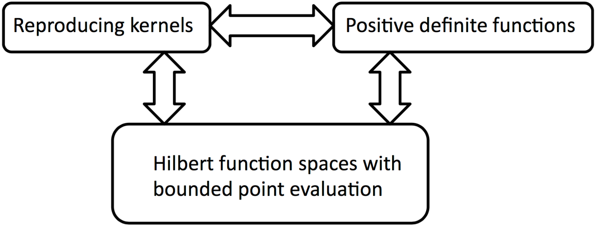
Theorem (Moore-Aronszajn)

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, exponentiated quadratic kernel:
$f(\cdot) := \sum_{i=1}^{m} \alpha_i k(x_i, \cdot)$.



Correspondence



Simple Kernel Algorithms

Distance between means (1)

Sample $(x_i)_{i=1}^m$ from p and $(y_i)_{i=1}^m$ from q. What is the distance between their means *in feature space*?

$$
\begin{aligned}
&\left\| \frac{1}{m}\sum_{i=1}^m \phi(x_i) - \frac{1}{n}\sum_{j=1}^n \phi(y_j) \right\|_{\mathcal{H}}^2 \\
=&\left\langle \frac{1}{m}\sum_{i=1}^m \phi(x_i) - \frac{1}{n}\sum_{j=1}^n \phi(y_j), \frac{1}{m}\sum_{i=1}^m \phi(x_i) - \frac{1}{n}\sum_{j=1}^n \phi(y_j) \right\rangle_{\mathcal{H}} \\
=&\frac{1}{m^2}\left\langle \sum_{i=1}^m \phi(x_i), \sum_{i=1}^m \phi(x_i) \right\rangle + \ldots \\
=&\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m k(x_i,x_j) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^m k(x_i,y_j).
\end{aligned}
$$

Distance between means (1)

Sample $(x_i)_{i=1}^m$ from p and $(y_i)_{i=1}^m$ from q. What is the distance between their means *in feature space*?

$$
\begin{aligned}
&\left\| \frac{1}{m}\sum_{i=1}^m \phi(x_i) - \frac{1}{n}\sum_{j=1}^n \phi(y_j) \right\|_{\mathcal{H}}^2 \\
=&\left\langle \frac{1}{m}\sum_{i=1}^m \phi(x_i) - \frac{1}{n}\sum_{j=1}^n \phi(y_j), \frac{1}{m}\sum_{i=1}^m \phi(x_i) - \frac{1}{n}\sum_{j=1}^n \phi(y_j) \right\rangle_{\mathcal{H}} \\
=&\frac{1}{m^2}\left\langle \sum_{i=1}^m \phi(x_i), \sum_{i=1}^m \phi(x_i) \right\rangle + \ldots \\
=&\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m k(x_i, x_j) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(y_i, y_j) - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^m k(x_i, y_j).
\end{aligned}
$$

Distance between means (2)

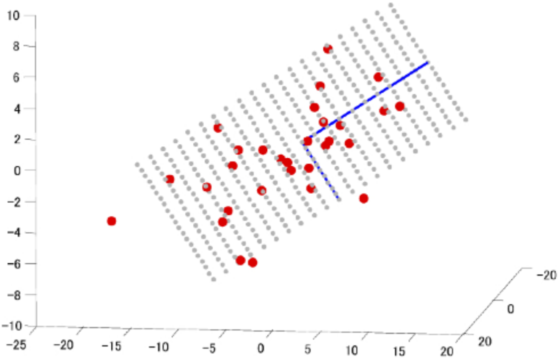
Sample $(x_i)_{i=1}^m$ from p and $(y_i)_{i=1}^m$ from q. What is the distance between their means *in feature space*?

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(y_j) \right\|_{\mathcal{H}}^2$$

- When $\phi(x) = x$, distinguish means. When $\phi(x) = [x\, x^2]$, distinguish means and variances.
- There are kernels that can distinguish *any* two distributions

PCA (1)

Goal of classical PCA: to find a d-dimensional subspace of a higher dimensional space (D-dimensional, $\mathbb{R}^D$) containing the directions of maximum variance.



(Figure by K. Fukumizu)

Applicationof kPCA: image denoising

What is the purpose of kernel PCA?

We consider the problem of **denoising** hand-written digits.

We are given a noisy digit x^*.

$$P_d\phi(x^*) = P_{f_1}\phi(x^*) + \ldots + P_{f_d}\phi(x^*)$$

is the projection of $\phi(x^*)$ onto one of the first d eigenvectors $\{f_\ell\}_{\ell=1}^d$ from kernel PCA (these are orthogonal).

Define the nearest point $y^* \in \mathcal{X}$ to this feature space projection as

$$y^* = \arg\min_{y\in\mathcal{X}} \|\phi(y) - P_d\phi(x^*)\|^2_{\mathcal{H}}.$$

In many cases,not possible to reduce the squared error to zero, as no single y^* corresponds to exact solution.

Applicationof kPCA: image denoising

What is the purpose of kernel PCA?

We consider the problem of **denoising** hand-written digits.

We are given a noisy digit x^*.

$$P_d\phi(x^*) = P_{f_1}\phi(x^*) + \ldots + P_{f_d}\phi(x^*)$$

is the projection of $\phi(x^*)$ onto one of the first d eigenvectors $\{f_\ell\}_{\ell=1}^d$ from kernel PCA (these are orthogonal).

Define the nearest point $y^* \in \mathcal{X}$ to this feature space projection as

$$y^* = \arg\min_{y\in\mathcal{X}} \|\phi(y) - P_d\phi(x^*)\|^2_{\mathcal{H}}.$$

In many cases,not possible to reduce the squared error to zero, as no single y^* corresponds to exact solution.

Applicationof kPCA: image denoising

What is the purpose of kernel PCA?

We consider the problem of **denoising** hand-written digits.
We are given a noisy digit x^*.

$$P_d\phi(x^*) = P_{f_1}\phi(x^*) + \ldots + P_{f_d}\phi(x^*)$$

is the projection of $\phi(x^*)$ onto one of the first d eigenvectors $\{f_\ell\}_{\ell=1}^d$ from kernel PCA (these are orthogonal).
Define the nearest point $y^* \in \mathcal{X}$ to this feature space projection as

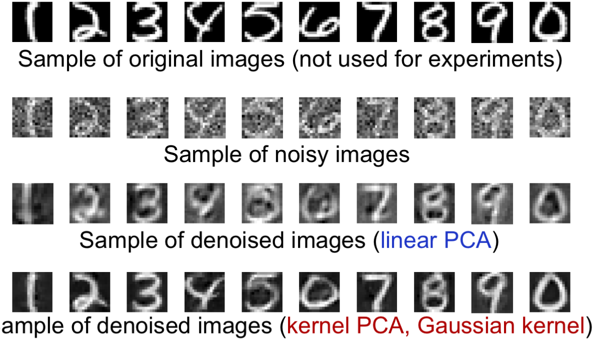
$$y^* = \arg\min_{y \in \mathcal{X}} \|\phi(y) - P_d\phi(x^*)\|^2_{\mathcal{H}}.$$

In many cases,not possible to reduce the squared error to zero, as no single y^* corresponds to exact solution.

Applicationof kPCA: image denoising

Projection onto PCA subspace for denoising. kPCA: data may not be Gaussian distributed, but can lie in a submanifold in input space.

USPS hand-written digits data:
7191 images of hand-written digits of 16 × 16 pixels.



Generated by Matlab Stprtool (by V. Franc). (Figure: K. Fukumizu)

What is PCA? (reminder)

First principal component (max. variance)

$$
\begin{aligned}
u_1 &= \arg\max_{\|u\|\leq 1} \frac{1}{n}\sum_{i=1}^{n}\left(u^\top\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)\right)^2 \\
&= \arg\max_{\|u\|\leq 1} u^\top C u
\end{aligned}
$$

where

$$
C = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^\top = \frac{1}{n} X H X^\top,
$$

$X = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}$, $H = I - n^{-1}\mathbf{1}_{n\times n}$, $\mathbf{1}_{n\times n}$ a matrix of ones.

Definition (Principal components)

The pairs (λ_i, u_i) are the eigensystem of $n\lambda_i u_i = C u_i$.

PCA in feature space

Kernel version, first principal component:

$$\begin{aligned} f_1 &= \arg\max_{\|f\|_{\mathcal{H}} \leq 1} \frac{1}{n} \sum_{i=1}^{n} \left(\left\langle f, \phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(x_j) \right\rangle_{\mathcal{H}} \right)^2 \\ &= \arg\max_{\|f\|_{\mathcal{H}} \leq 1} \mathrm{var}(f). \end{aligned}$$

We can write

$$f = \sum_{i=1}^{n} \alpha_i \left(\phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(x_j) \right) = \sum_{i=1}^{n} \alpha_i \tilde{\phi}(x_i),$$

since any component orthogonal to the span of $\tilde{\phi}(x_i) := \phi(x_i) - \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$ vanishes.

PCA in feature space

Kernel version, first principal component:

$$
\begin{aligned}
f_1 &= \arg\max_{\|f\|_{\mathcal{H}}\leq 1} \frac{1}{n}\sum_{i=1}^{n}\left(\left\langle f, \phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\phi(x_j)\right\rangle_{\mathcal{H}}\right)^2 \\
&= \arg\max_{\|f\|_{\mathcal{H}}\leq 1} \operatorname{var}(f).
\end{aligned}
$$

We can write

$$
f = \sum_{i=1}^{n}\alpha_i\left(\phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\phi(x_j)\right) = \sum_{i=1}^{n}\alpha_i\tilde{\phi}(x_i),
$$

since any component orthogonal to the span of $\tilde{\phi}(x_i) := \phi(x_i) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$ vanishes.

How to solve kernel PCA

We can also define an infinite dimensional analog of the covariance:

$$
\begin{aligned}
C &= \frac{1}{n}\sum_{i=1}^{n}\left(\phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\phi(x_j)\right) \otimes \left(\phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\phi(x_j)\right), \\
&= \frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(x_i) \otimes \tilde{\phi}(x_i)
\end{aligned}
$$

where we use the definition

$$
(a \otimes b)c := \langle b, c \rangle_{\mathcal{H}}\, a \tag{3}
$$

this is analogous to the case of finite dimensional vectors, $(ab^\top)c = (b^\top c)a$.

How to solve kernel PCA (1)

Eigenfunctions of kernel covariance:

$$
\begin{aligned}
f_\ell \lambda_\ell &= C f_\ell \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \otimes \tilde{\phi}(x_i) \right) f_\ell \\
&= \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \left\langle \tilde{\phi}(x_i), \sum_{j=1}^{n} \alpha_{\ell j} \tilde{\phi}(x_j) \right\rangle_{\mathcal{H}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \left(\sum_{j=1}^{n} \alpha_{\ell j} \tilde{k}(x_i, x_j) \right)
\end{aligned}
$$

$\tilde{k}(x_i, x_j)$ is the (i,j)th entry of the matrix $\tilde{K} := HKH$ (exercise!).

How to solve kernel PCA (1)

Eigenfunctions of kernel covariance:

$$
\begin{aligned}
f_\ell \lambda_\ell &= C f_\ell \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \otimes \tilde{\phi}(x_i) \right) f_\ell \\
&= \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \left\langle \tilde{\phi}(x_i), \sum_{j=1}^{n} \alpha_{\ell j} \tilde{\phi}(x_j) \right\rangle_{\mathcal{H}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \left(\sum_{j=1}^{n} \alpha_{\ell j} \tilde{k}(x_i, x_j) \right)
\end{aligned}
$$

$\tilde{k}(x_i, x_j)$ is the (i,j)th entry of the matrix $\tilde{K} := HKH$ (exercise!).

How to solve kernel PCA (2)

We can now project both sides of

$$f_\ell \lambda_\ell = C f_\ell$$

onto all of the $\tilde{\phi}(x_q)$:

$$\left\langle \tilde{\phi}(x_q), \mathrm{LHS} \right\rangle_{\mathcal{H}} = \lambda_\ell \left\langle \tilde{\phi}(x_q), f_\ell \right\rangle_{\mathcal{H}} = \lambda_\ell \sum_{i=1}^{n} \alpha_{\ell i} \tilde{k}(x_q, x_i) \qquad \forall q \in \{1 \ldots n$$

$$\left\langle \tilde{\phi}(x_q), \mathrm{RHS} \right\rangle_{\mathcal{H}} = \left\langle \tilde{\phi}(x_q), C f_\ell \right\rangle_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^{n} \tilde{k}(x_q, x_i) \left(\sum_{j=1}^{n} \alpha_{\ell j} \tilde{k}(x_i, x_j) \right)$$

Writing this as a matrix equation,

$$n \lambda_\ell \widetilde{K} \alpha_\ell = \widetilde{K}^2 \alpha_\ell \qquad n \lambda_\ell \alpha_\ell = \widetilde{K} \alpha_\ell.$$

How to solve kernel PCA (2)

We can now project both sides of

$$f_\ell \lambda_\ell = C f_\ell$$

onto all of the $\tilde{\phi}(x_q)$:

$$\left\langle \tilde{\phi}(x_q), \mathrm{LHS} \right\rangle_{\mathcal{H}} = \lambda_\ell \left\langle \tilde{\phi}(x_q), f_\ell \right\rangle_{\mathcal{H}} = \lambda_\ell \sum_{i=1}^{n} \alpha_{\ell i} \tilde{k}(x_q, x_i) \qquad \forall q \in \{1 \ldots n$$

$$\left\langle \tilde{\phi}(x_q), \mathrm{RHS} \right\rangle_{\mathcal{H}} = \left\langle \tilde{\phi}(x_q), C f_\ell \right\rangle_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^{n} \tilde{k}(x_q, x_i) \left(\sum_{j=1}^{n} \alpha_{\ell j} \tilde{k}(x_i, x_j) \right)$$

Writing this as a matrix equation,

$$n \lambda_\ell \widetilde{K} \alpha_\ell = \widetilde{K}^2 \alpha_\ell \qquad n \lambda_\ell \alpha_\ell = \widetilde{K} \alpha_\ell .$$

Eigenfunctions f have unit norm in feature space?

$$
\begin{aligned}
&\|f\|^2_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^n \alpha_i \tilde{\phi}(x_i), \sum_{i=1}^n \alpha_i \tilde{\phi}(x_i) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_i \left\langle \tilde{\phi}(x_i), \tilde{\phi}(x_j) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_i \tilde{k}(x_i, x_j) \\
&= \alpha^\top \widetilde{K} \alpha = n\lambda \alpha^\top \alpha = n\lambda \|\alpha\|^2.
\end{aligned}
$$

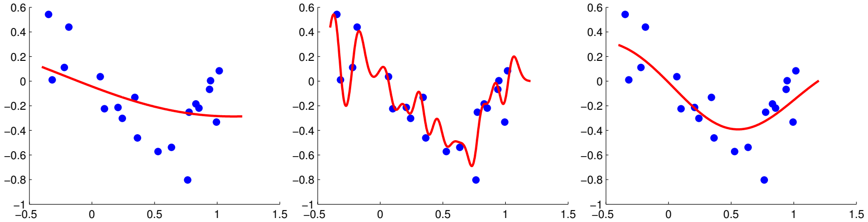
Thus $\alpha \leftarrow \alpha / \sqrt{n\lambda}$ (assumed: original eigenvector solution has $\|\alpha\| = 1$)

Projection onto kernel PC

How do you project a new point x^* onto the principal component f? Assuming $\|f\|_{\mathcal{H}} = 1$, the projection is

$$
\begin{aligned}
P_f \phi(x^*) &= \langle \phi(x^*), f \rangle_{\mathcal{H}} f \\
&= \sum_{i=1}^{n} \alpha_i \left(\sum_{j=1}^{n} \alpha_j \left\langle \phi(x^*), \tilde{\phi}(x_i) \right\rangle_{\mathcal{H}} \right) \tilde{\phi}(x_i) \\
&= \sum_{i=1}^{n} \alpha_i \left(\sum_{j=1}^{n} \alpha_j \left(k(x^*, x_j) - \frac{1}{n} \sum_{\ell=1}^{n} k(x^*, x_\ell) \right) \right) \tilde{\phi}(x_i).
\end{aligned}
$$

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 $\mathbb{R}^D$:

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

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

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

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

Ridge regression: solution (1)

Expanding out the above term, we get

$$\begin{aligned}\left\|y - X^\top a\right\|^2 + \lambda\|a\|^2 &= y^\top y - 2y^\top X a + a^\top X X^\top a + \lambda a^\top a \\ &= y^\top y - 2y^\top X^\top a + a^\top \left(XX^\top + \lambda I\right) a = (*)\end{aligned}$$

- Define $b = \left(XX^\top + \lambda I\right)^{1/2} a$
- Square root defined since matrix positive definite
- $XX^\top$ may not be invertible eg when $D > n$, adding λI means we can write $a = \left(XX^\top + \lambda I\right)^{-1/2} b$).

Ridge regression: solution (2)

Complete the square:

$$
\begin{aligned}
(*) =& y^\top y - 2y^\top X^\top \left(XX^\top + \lambda I\right)^{-1/2} b + b^\top b \\
=& y^\top y + \left\| \left(XX^\top + \lambda I\right)^{-1/2} Xy - b \right\|^2 - \left\| y^\top X^\top \left(XX^\top + \lambda I\right)^{-1/2} \right\|^2
\end{aligned}
$$

This is minimized when

$$
\begin{aligned}
b^* &= \left(XX^\top + \lambda I\right)^{-1/2} Xy \quad \text{or} \\
a^* &= \left(XX^\top + \lambda I\right)^{-1} Xy,
\end{aligned}
$$

which is the classic regularized least squares solution.

Ridge regression solution as sum of training points (1)

We may rewrite this expression in a way that is more informative, $a^* = \sum_{i=1}^{n} \alpha_i^* x_i$.
The solution is a linear combination of training points x_i.

Proof: Assume $D > n$ (in feature space case D can be very large or even infinite).
Perform an SVD on X, i.e.

$$X = USV^\top,$$

where

$$U = \begin{bmatrix} u_1 & \dots & u_D \end{bmatrix} \quad S = \begin{bmatrix} \tilde{S} & 0 \\ 0 & 0 \end{bmatrix} \quad V = \begin{bmatrix} \tilde{V} & 0 \end{bmatrix}.$$

Here U is $D \times D$ and $U^\top U = UU^\top = I_D$ (subscript denotes unit matrix size), S is $D \times D$, where $\tilde{S}$ has n non-zero entries, and V is $n \times D$, where $\tilde{V}^\top \tilde{V} = \tilde{V}\tilde{V}^\top = I_n$.

Ridge regression solution as sum of training points (1)

We may rewrite this expression in a way that is more informative, $a^* = \sum_{i=1}^n \alpha_i^* x_i$.

The solution is a linear combination of training points x_i.

Proof: Assume $D > n$ (in feature space case D can be very large or even infinite).

Perform an SVD on X, i.e.

$$X = USV^\top,$$

where

$$U = \left[\begin{array}{ccc} u_1 & \ldots & u_D \end{array} \right] \quad S = \left[\begin{array}{cc} \tilde{S} & 0 \\ 0 & 0 \end{array} \right] \quad V = \left[\begin{array}{cc} \tilde{V} & 0 \end{array} \right].$$

Here U is $D \times D$ and $U^\top U = UU^\top = I_D$ (subscript denotes unit matrix size), S is $D \times D$, where $\tilde{S}$ has n non-zero entries, and V is $n \times D$, where $\tilde{V}^\top \tilde{V} = \tilde{V}\tilde{V}^\top = I_n$.

Ridge regression solution as sum of training points (2)

Proof (continued):

$$
\begin{aligned}
a^* &= \left(XX^\top + \lambda I_D\right)^{-1} Xy \\
&= \left(US^2U^\top + \lambda I_D\right)^{-1} USV^\top y \\
&= U\left(S^2 + \lambda I_D\right)^{-1} U^\top USV^\top y \\
&= U\left(S^2 + \lambda I_D\right)^{-1} SV^\top y \\
&= US\left(S^2 + \lambda I_D\right)^{-1} V^\top y \\
&= U\underbrace{SV^\top V}_{\text{(a)}}\left(S^2 + \lambda I_D\right)^{-1} V^\top y \\
&\underset{\text{(b)}}{=} X(X^\top X + \lambda I_n)^{-1} y
\end{aligned}
\tag{4}
$$

Ridge regression solution as sum of training points (3)

Proof (continued):
(a): both S and $V^\top V$ are non-zero in same sized top-left block, and $V^\top V$ is I_n in that block.
(b): since

$$
\begin{aligned}
& V\left(S^2+\lambda I_D\right)^{-1} V^\top \\
=& \begin{bmatrix} \tilde{V} & 0 \end{bmatrix} \begin{bmatrix} \left(\tilde{S}^2+\lambda I_n\right)^{-1} & 0 \\ 0 & (\lambda I_{D-n})^{-1} \end{bmatrix} \begin{bmatrix} \tilde{V}^\top \\ 0 \end{bmatrix} \\
=& \tilde{V}\left(\tilde{S}^2+\lambda I_n\right)^{-1} \tilde{V}^\top \\
=& \left(X^\top X+\lambda I_n\right)^{-1}.
\end{aligned}
$$

Kernel ridge regression

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

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

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: proof

Use previous proof!

$$X = \left[\begin{array}{ccc} \phi(x_1) & \dots & \phi(x_n) \end{array}\right].$$

All of the steps that led us to $a^* = X(X^\top X + \lambda I_n)^{-1} y$ follow.

$$XX^\top = \sum_{i=1}^{n} \phi(x_i) \otimes \phi(x_i)$$

(using tensor notation from kernel PCA), and

$$(X^\top X)_{ij} = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} = k(x_i, x_j).$$

Making these replacements, we get

$$\begin{aligned} a^* &= X(K + \lambda I_n)^{-1} y \\ &= \sum_{i=1}^{n} \alpha_i^* \phi(x_i) \qquad \alpha^* = (K + \lambda I_n)^{-1} y. \end{aligned}$$

Kernel ridge regression: easier proof

We *begin* knowing a is a linear combination of feature space mappings of points (representer theorem: later in course)

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

Then

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

Differentiating wrt α and setting this to zero, we get

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

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

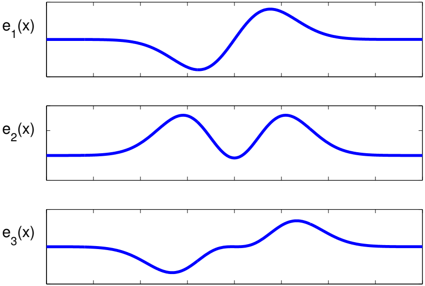
Reminder: smoothness

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

Example 1: The exponentiated quadratic kernel. Recall

$$f(x) = \sum_{i=1}^{\infty} \hat{f}_\ell e_\ell(x), \qquad \langle e_i, e_j \rangle_{L_2(\mu)} = \int_{\mathcal{X}} e_i(x) e_j(x) d\mu(x) = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

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



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

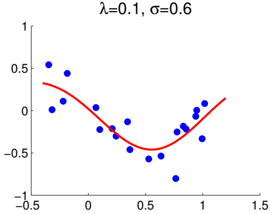
$$a^* = \arg\min_{a \in \mathcal{H}} \left(\sum_{i=1}^{n} \left(y_i - \langle a, \phi(x_i) \rangle_{\mathcal{H}}\right)^2 + \lambda \|a\|_{\mathcal{H}}^2 \right).$$

How do we choose

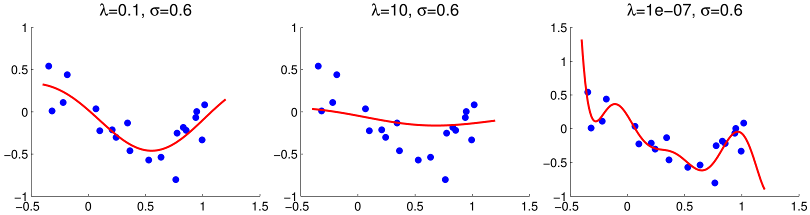
- The regularization parameter λ?
- The kernel parameter: for exponentiated quadratic kernel, σ in

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

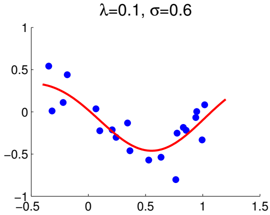
Choice of λ



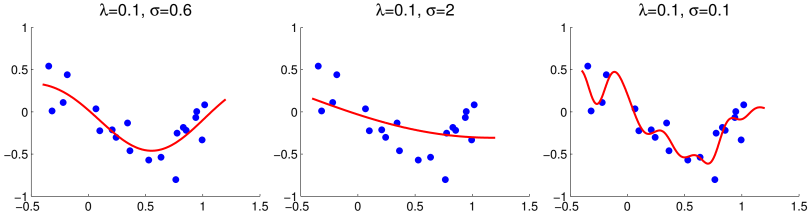
Choice of λ



Choice of σ



Choice of σ



Cross validation

- Split n data into training set size n_{tr} and **test set** size $n_{\mathrm{te}} = n - n_{\mathrm{tr}}$.
- Split training 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 $X_{\mathrm{te}}, Y_{\mathrm{te}}$.