Representing and comparing probabilities with kernels: Part 1

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

MLSS Madrid, 2018

A motivation: comparing two samples

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

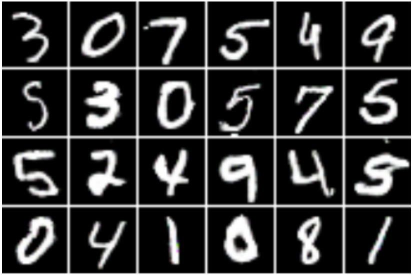
$$\sim P$$

$$\sim Q$$

A real-life example: two-sample tests

- **Have:** Two collections of samples X, Y from unknown distributions P and Q.
- **Goal:** do P and Q differ?

MNIST samples



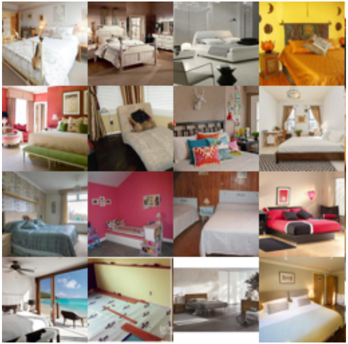
Samples from a GAN

Significant difference in GAN and MNIST?

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, Xi Chen, NIPS 2016
Sutherland, Tung, Strathmann, De, Ramdas, Smola, G., ICLR 2017.

Training generative models

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



LSUN bedroom samples P

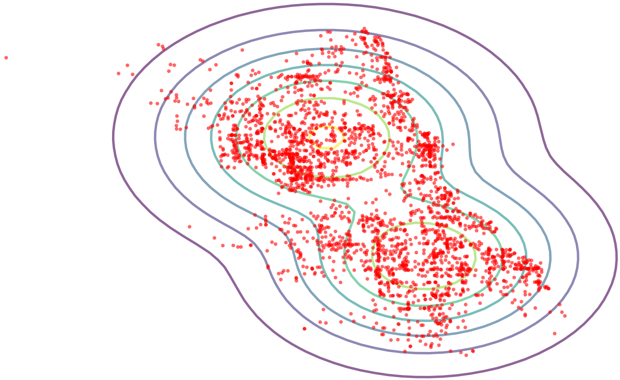
Generated Q, MMD GAN

Using MMD to train a GAN

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

Testing goodness of fit

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



Chicago crime data

Model is Gaussian mixture with **two** components.

Testing independence

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

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

Text from dogtime.com and petfinder.com

Outline: part 1

What is a reproducing kernel Hilbert space?

1. Hilbert space
2. Kernel (lots of examples: e.g. you can build kernels from simpler kernels)
3. Reproducing property
4. Using kernels to enforce smoothness

Classical results

1. Representer theorem
2. Kerrnel ridge regression

Outline: part 2

The maximum mean discrepancy (MMD)

- ...as a difference in feature means
- ...as an integral probability metric (not just a technicality!)

Statistical testing with the MMD

- How to choose the best kernel

Training GANs with MMD

- Learning kernel features with gradient regularisation

Characteristic kernels: "is my feature space rich enough?"

Outline: part 3

Goodness of fit testing

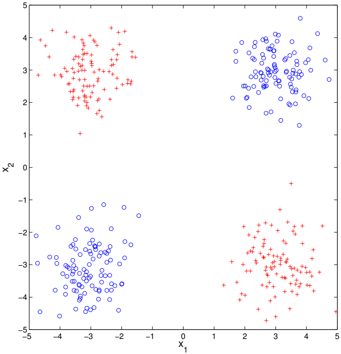
- The kernel Stein discrepancy

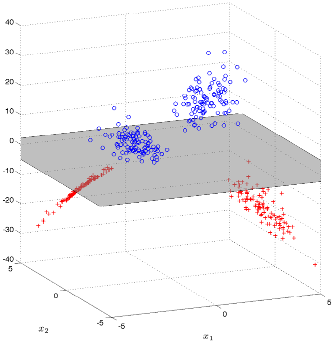
Dependence testing

- Dependence using the MMD
- Depenence using feature covariances
- Statistical testing

Reproducing Kernel Hilbert Spaces

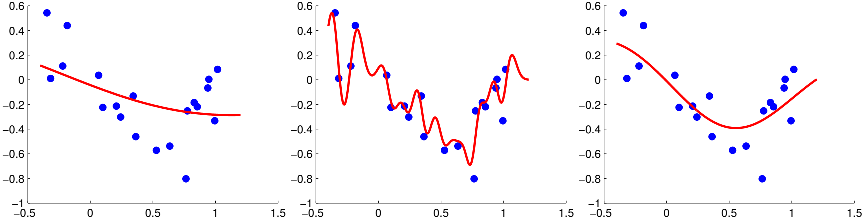
Kernels and feature space (1): XOR example





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

Kernels and feature space (2): smoothing



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

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 $\mathcal{X}$.

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

New kernels from old: products

Theorem (Products of kernels are kernels)

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

Proof: Main idea only!

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

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

$\mathcal{H}_2$ space of kernels 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:

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

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

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:

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

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

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

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

is a valid kernel.

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

Infinite sequences

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

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

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_2 (**square** summable sequences) comprises all sequences $a := (a_i)_{i\geq 1}$ for which

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

Definition

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

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

Infinite sequences

Definition

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

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

Definition

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

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

Infinite sequences (proof)

Why square summable? By Cauchy-Schwarz,

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

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

A famous infinite feature space kernel

Exponentiated quadratic kernel,

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

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

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

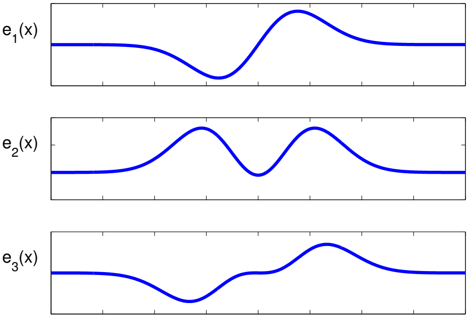
A famous infinite feature space kernel

Exponentiated quadratic kernel,

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

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

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



$$\lambda_\ell \propto b^\ell \qquad b < 1$$

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

a, b, c are functions of σ, and H_ℓ is ℓth order Hermite polynomial.

Positive definite functions

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

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

Positive definite functions

Definition (Positive definite functions)

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

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

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

Kernels are positive definite

Theorem

Let $\mathcal{H}$ be a Hilbert space, $\mathcal{X}$ a non-empty set and $\phi : \mathcal{X} \to \mathcal{H}$. 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 a unique $\mathcal{H}$ (**Moore-Aronsajn**: coming later!). □

Sum of kernels is a kernel

Proof by positive definiteness:

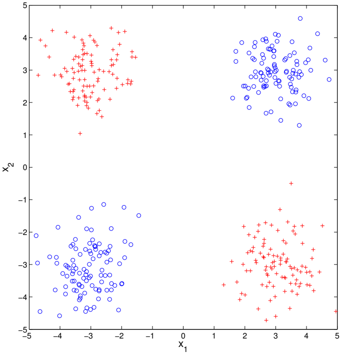
Consider two kernels $k_1(x, x')$ and $k_2(x, x')$. Then

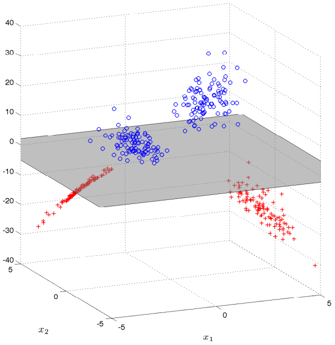
$$
\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \left[k_1(x_i, x_j) + k_2(x_i, x_j)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k_1(x_i, x_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k_2(x_i, x_j) \\
&\geq 0
\end{aligned}
$$

The reproducing kernel Hilbert space

First example: finite space, polynomial features

Reminder: XOR example:





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 $\mathcal{H}$.

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 $\mathbb{R}$.

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**).

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

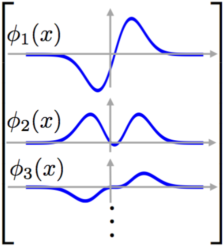
Evaluation of f at x is an **inner product in feature space** (here standard inner product in $\mathbb{R}^3$)

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

Functions of infinitely many features

Functions are linear combinations of features:

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



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

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

Expressing the functions with kernels

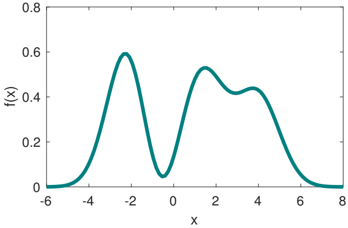
Function with **exponentiated quadratic kernel:**

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

$$= \sum_{\ell=1}^{\infty} \underbrace{\left(\sum_{i=1}^{m} \alpha_i \phi_\ell(x_i) \right)}_{f_\ell} \phi_\ell(x)$$

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

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



Expressing the functions with kernels

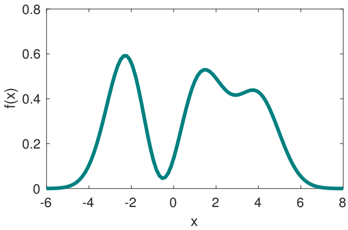
Function with **exponentiated quadratic kernel:**

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

$$= \sum_{\ell=1}^{\infty} \underbrace{\left(\sum_{i=1}^{m} \alpha_i \phi_\ell(x_i)\right)}_{f_\ell} \phi_\ell(x)$$

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

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



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

Expressing the functions with kernels

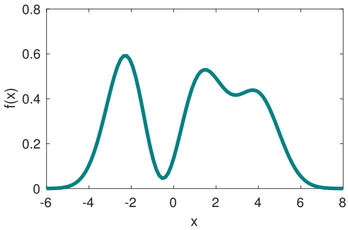
Function with **exponentiated quadratic kernel:**

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

$$= \sum_{\ell=1}^{\infty} \underbrace{\left(\sum_{i=1}^{m} \alpha_i \phi_\ell(x_i) \right)}_{f_\ell} \phi_\ell(x)$$

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

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

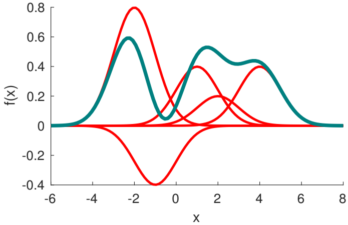


$f_\ell := \sum_{i=1}^{m} \alpha_i \phi_\ell(x_i)$

Expressing the functions with kernels

Function with **exponentiated quadratic kernel:**

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



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

Function of infinitely many features expressed using m coefficients.

The feature map is *also* a function

On previous page,

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

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

Then

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

The feature map is *also* a function

On previous page,

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

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

Then

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

The feature map is *also* a function

On previous page,

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

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

Then

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

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

....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 feature map is *also* a function

On previous page,

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

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

Then

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

....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 reproducing property

This example illustrates the two defining features of an RKHS:

- **The reproducing property:** (kernel trick)
  $\forall x \in \mathcal{X}, \forall f(\cdot) \in \mathcal{H}, \ \ \langle f(\cdot), k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$
  ...or use shorter notation $\langle f, \phi(x)\rangle_{\mathcal{H}}$.
- The feature map of every point is a function: $k(\cdot, x) = \phi(x) \in \mathcal{H}$ for any $x \in \mathcal{X}$, and

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

Understanding smoothness in the RKHS

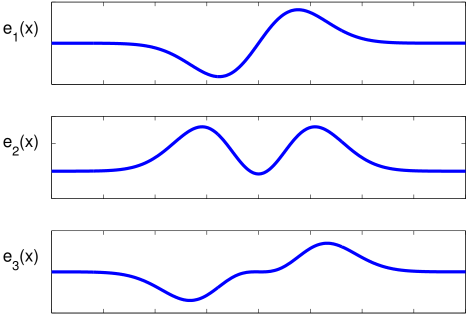
Smoothness in RKHS with exp. quad. kernel

Reminder, **exponentiated quadratic kernel**,

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

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

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

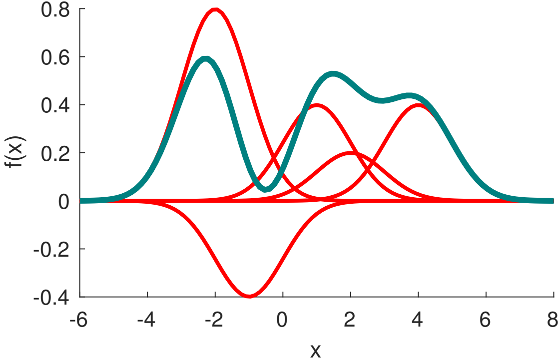


Smoothness in RKHS with exp. quad. kernel

RKHS function, exponentiated quadratic kernel:

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

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



NOTE that this enforces smoothing:

λ_ℓ decay as e_ℓ become rougher, f_ℓ decay since $\sum_\ell f_\ell^2 < \infty$.

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,

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

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

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 $[-\pi, \pi]$,

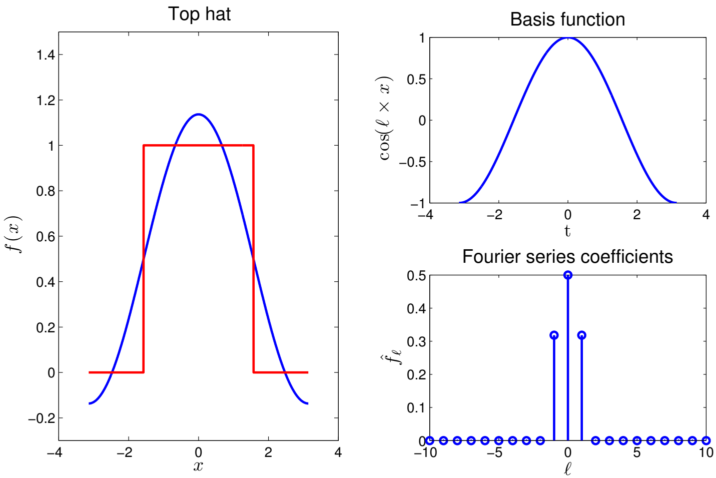
$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(\imath \ell x) \overline{\exp(\imath m x)}\, dx = \begin{cases} 1 & \ell = m, \\ 0 & \ell \neq m. \end{cases}$$

Example: "top hat" function,

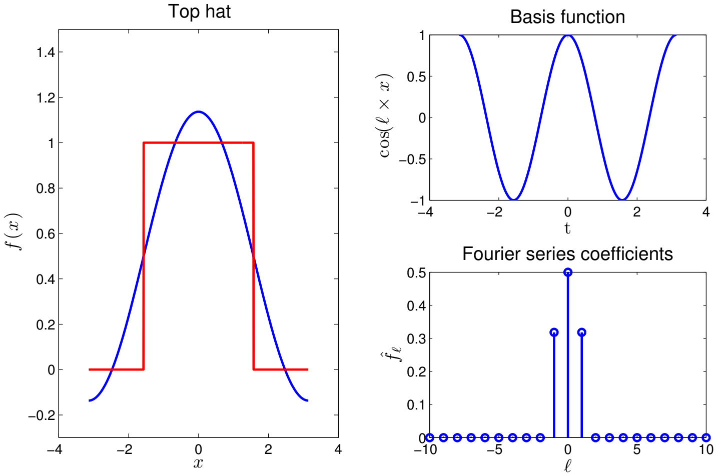
$$f(x) = \begin{cases} 1 & |x| < T, \\ 0 & T \leq |x| < \pi. \end{cases}$$

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

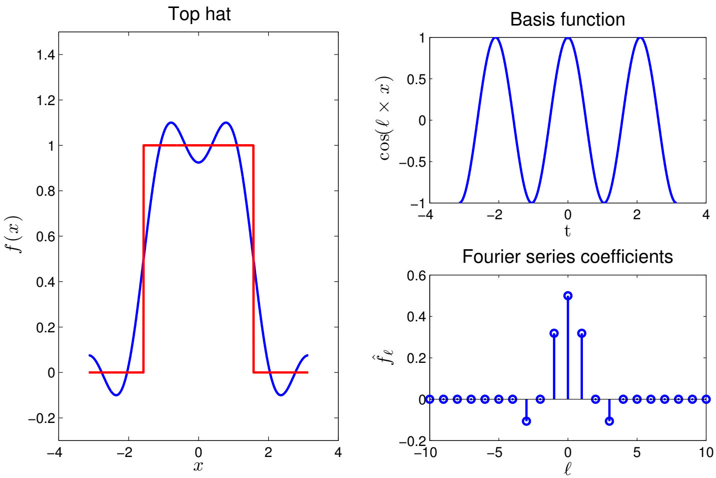
Fourier series for top hat function



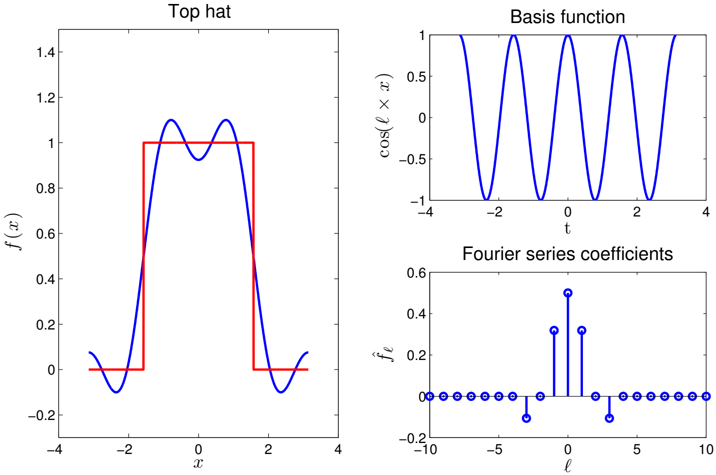
Fourier series for top hat function



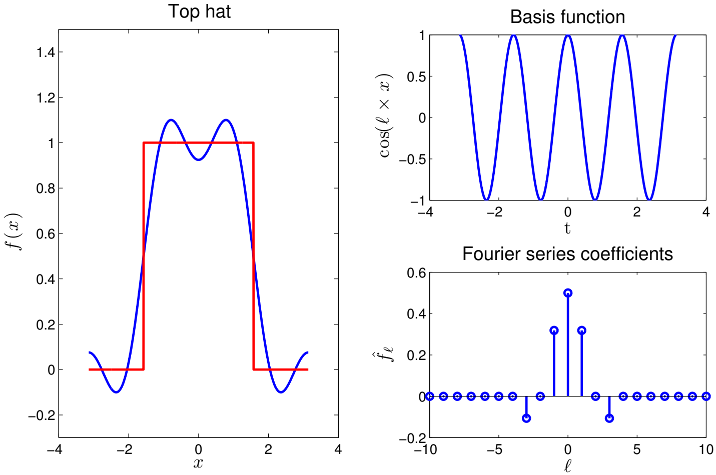
Fourier series for top hat function



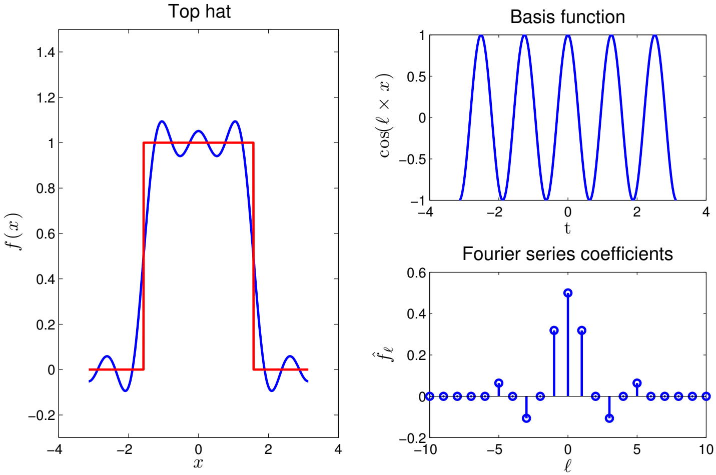
Fourier series for top hat function



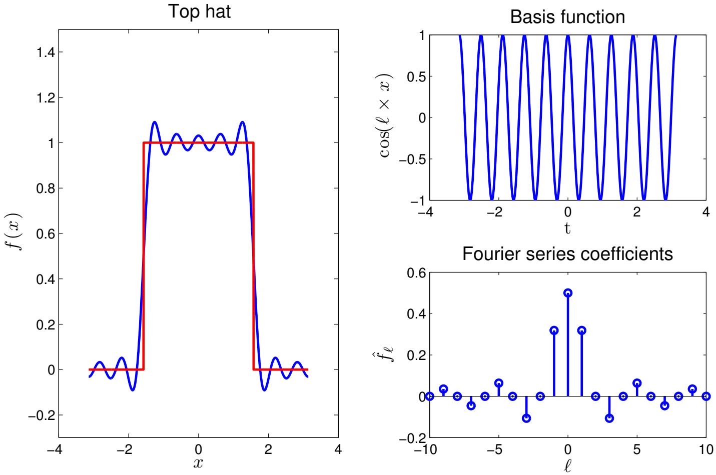
Fourier series for top hat function



Fourier series for top hat function



Fourier series for top hat function



Fourier series for kernel function

Assume kernel translation invariant,

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

Fourier series representation of k

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

$$= \sum_{\ell=-\infty}^{\infty} \Bigg[\underbrace{\sqrt{\hat{k}_\ell} \exp\left(\imath \ell (x)\right)}_{e_\ell(x)} \Bigg] \Bigg[\underbrace{\sqrt{\hat{k}_\ell} \exp\left(-\imath \ell y\right)}_{\overline{e_\ell(y)}} \Bigg].$$

Example: **Jacobi theta kernel:**

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

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

Fourier series for kernel function

Assume kernel translation invariant,

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

Fourier series representation of k

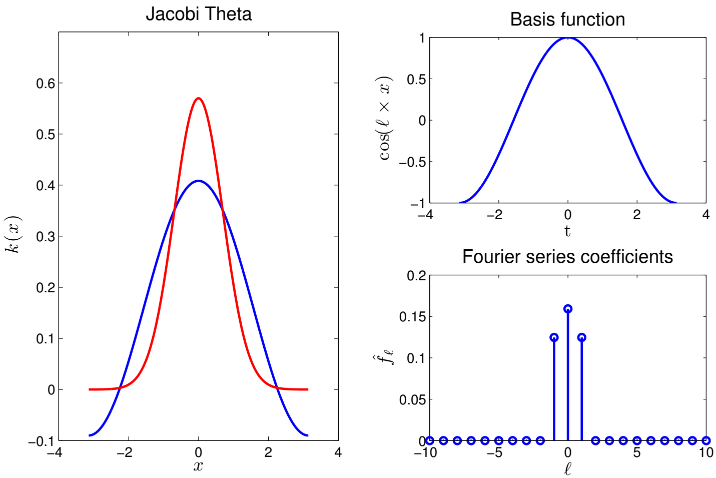
$$\begin{aligned} k(x-y) &= \sum_{\ell=-\infty}^{\infty} \hat{k}_\ell \exp\left(\imath \ell (x-y)\right) \\ &= \sum_{\ell=-\infty}^{\infty} \Bigg[\underbrace{\sqrt{\hat{k}_\ell} \exp\left(\imath \ell (x)\right)}_{e_\ell(x)}\Bigg]\Bigg[\underbrace{\sqrt{\hat{k}_\ell} \exp\left(-\imath \ell y\right)}_{\overline{e_\ell(y)}}\Bigg].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

Example: **Jacobi theta kernel:**

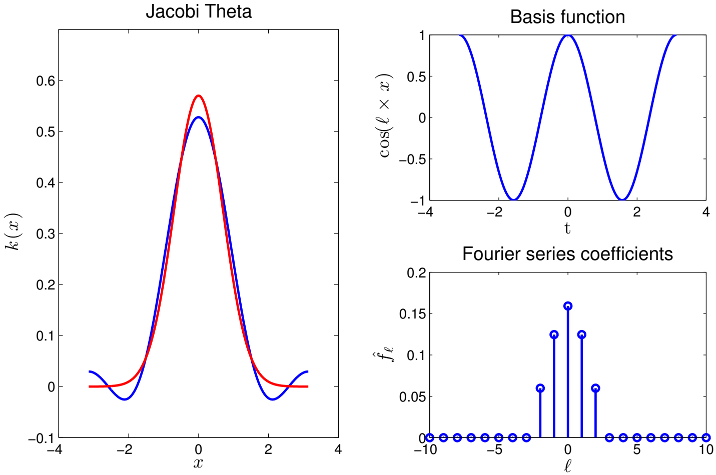
$$k(x-y) = \frac{1}{2\pi}\vartheta\left(\frac{(x-y)}{2\pi}, \frac{\imath\sigma^2}{2\pi}\right), \qquad \hat{k}_\ell = \frac{1}{2\pi}\exp\left(\frac{-\sigma^2\ell^2}{2}\right).$$

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

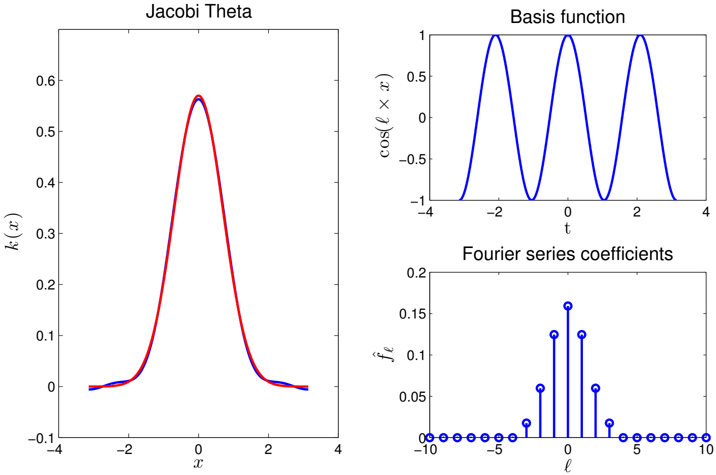
Fourier series for Gaussian-spectrum kernel



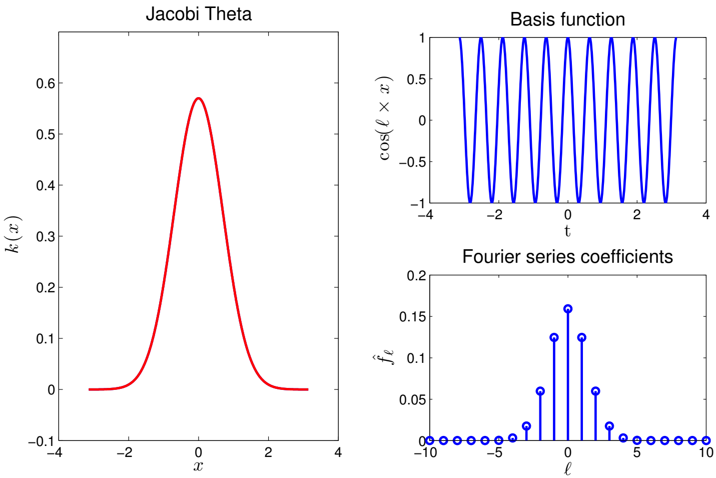
Fourier series for Gaussian-spectrum kernel



Fourier series for Gaussian-spectrum kernel



Fourier series for Gaussian-spectrum kernel



RKHS via fourier series

Recall standard dot product in L_2:

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

$$= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}_\ell} \langle \exp(\imath \ell x), \exp(-\imath m x) \rangle_{L_2}$$

$$= \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}_\ell}.$$

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:

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

$$= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}_\ell} \langle \exp(\imath \ell x), \exp(-\imath m x)\rangle_{L_2}$$

$$= \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}_\ell}.$$

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**.

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**.

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**.

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 $f(\cdot) \in \mathcal{H}$,

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

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

$$= \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell z) = f(z).$$

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 $f(\cdot) \in \mathcal{H}$,

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

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

$$\sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell z) = f(z).$$

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 $f(\cdot) \in \mathcal{H}$,

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

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

$$\sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell z) = f(z).$$

Feature map and reproducing property

Reproducing property for the **kernel**:

You can also show

$$\langle k(\cdot, y), k(\cdot, z)\rangle_{\mathcal{H}} = k(y - z)$$

This is an exercise!

Hint: define a second function

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

Feature map and reproducing property

Reproducing property for the **kernel**:

You can also show

$$\langle k(\cdot, y), k(\cdot, z) \rangle_{\mathcal{H}} = k(y - z)$$

This is an exercise!

Hint: define a second function

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

Link back to original RKHS function 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 function 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 function 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

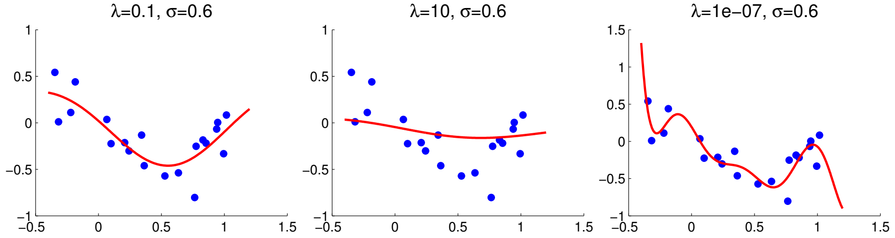
$$f_\ell = \hat{f}_\ell / \sqrt{\hat{k}_\ell} \qquad \phi_\ell(x) = \sqrt{\hat{k}_\ell} \exp(-\imath \ell x).$$

Main message

Small RKHS norm results in **smooth functions**.

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

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



Some reproducing kernel Hilbert space theory

Reproducing kernel Hilbert space (1)

Definition

$\mathcal{H}$ a Hilbert space of $\mathbb{R}$-valued functions on non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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 $\lambda_x = k(x, x)^{1/2}$.

RKHS definitions equivalent

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

We use...

Theorem

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

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

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

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

Moore-Aronszajn Theorem

Theorem (Moore-Aronszajn)

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

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

Main message