

# Reproducing Kernel Hilbert Space

Northeastern University

Machine Learning and Statistical Learning Theory II

Feb 27, 2023

# Overview

- 1 Background
- 2 Kernels and RKHS
- 3 Regularization
- 4 Feature Map and Feature Space

## Background

- Functional Analysis
- Linear Algebra
- Examples

## Definitions

**Definition 1** A **Function Space**  $\mathcal{F}$  is a space whose elements are functions, i.e.  $f : \mathbb{R}^d \rightarrow \mathbb{R} \in \mathcal{F}$

**Definition 2** An **Inner Product** is a function  $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}$  that satisfies the following properties for every  $f, g \in \mathcal{F}$ , and  $\alpha \in \mathbb{R}$

- 1.) Symmetry:  $\langle f, g \rangle = \langle g, f \rangle$
- 2.) Linear:  $\langle r_1 f_1 + r_2 f_2, g \rangle = r_1 \langle f_1, g \rangle + r_2 \langle f_2, g \rangle$
- 3.) Positive-definite:  $\langle f, f \rangle \geq 0$  for all  $f \in \mathcal{F}$  and  $\langle f, f \rangle = 0$  iff  $f = 0$

Note that the dot product in  $\mathbb{R}^n$  is an inner product, but an inner product is a more general operator

**Definition 3** An inner product space  $V$  is **complete** if any sequence  $\{x_i\}_{i=1}^{\infty} \subset V$  which is **Cauchy**, i.e.  $\|x_i - x_j\|_{i,j \rightarrow \infty} \rightarrow 0$  converges to some  $x \in V$ .

In a Cauchy sequence the elements become arbitrarily close to one another as the sequence progresses

**Definition 4** A **Hilbert Space** is a complete, (possibly) infinite-dimensional linear space endowed with an inner product  
Note that we operate in a Hilbert Space because completeness guarantees convergence and the inner product allows for the use of projection and orthogonality

# Hilbert Space Examples

**Example 1**  $H = \mathbb{R}^3 = \{v = (v_1, v_2, v_3) \mid v_i \in \mathbb{R}\}$   
with the standard inner product

$$\langle v, w \rangle = \sum_{i=1}^3 v_i w_i$$

**Example 2**  $H = L_2[a, b]$  which denotes the space of square integrable functions on the interval  $[a, b]$

$$\begin{aligned}\langle f, g \rangle &= \int_a^b f(x)g(x)dx \\ \|f\| &= \int_a^b f^2(x)dx\end{aligned}$$

which is the correct norm. Just need to check completeness

# Kernels and RKHS

- What is a kernel
- What is an RKHS
- Examples of reproducing kernels

## Definitions

**Definition 5** A **kernel** is a function  $k : X \times X \rightarrow \mathbb{R}$  which is symmetric

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

for  $x, y \in X$

We say that  $k$  is **positive** if for any fixed collection

$$\{x_1, \dots, x_n\} \subset X$$

the corresponding kernel matrix

$$(K_{ij}) = k(x_i, x_j) \geq 0$$

**Definition 6** A **Mercer Kernel** is a positive definite kernel  $k(x, y)$  that is also continuous as a function of  $x, y$  and bounded.



**Definition 7** An **evaluation functional** over the Hilbert space of functions  $\mathcal{H}$  is a linear functional  $\mathcal{F}_t : \mathcal{H} \rightarrow \mathbb{R}$  that evaluates each functional in the space at the point  $t$ , i.e.

$$\mathcal{F}_t|f| = f(t) \quad \forall f \in \mathcal{H}$$

**Definition 8** A Hilbert space  $\mathcal{H}$  is a **reproducing kernel Hilbert space (RKHS)** if the evaluation functionals are bounded, i.e. if for all  $t$  there exists some  $M > 0$  such that

$$|\mathcal{F}_t|f|| = |f(t)| \leq M\|f\|_{\mathcal{H}}$$

## RKHS Importance

**Example 3** Consider two functions in the space and find the inner product

$$\begin{aligned}\mathcal{H} &:= \{f(\cdot) = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \cdot)\} \\ \langle f, g \rangle_k &= \left\langle \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \cdot), \sum_{j=1}^n \beta_j k(\mathbf{y}_j, \cdot) \right\rangle_k \\ &= \left\langle \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \cdot), \sum_{j=1}^n \beta_j k(\cdot, \mathbf{y}_j) \right\rangle_k \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{y}_j)\end{aligned}$$

What does this do?

- ensures existence of inner product
- ability to evaluate each function in the space at every point in the domain
- Bases of the RKHS are kernels
- Every function in RKHS can be written as a linear combination
- Given a kernel, the corresponding RKHS is unique (up to isometric isomorphisms). Given an RKHS, the corresponding kernel is unique. In other words, each kernel generates a new RKHS.

## Kernel Reproduction

**Theorem 1** Given a reproducing kernel Hilbert space  $\mathcal{H}$  of functions on  $X \subset \mathbb{R}^d$  there exists a unique symmetric positive kernel function  $K(\mathbf{x}, \mathbf{y})$  such that for all  $f \in \mathcal{H}$ ,

$$f(\mathbf{x}) = \langle \mathbf{f}(\cdot), \mathbf{K}(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}$$

**Proof** Each evaluation function  $\mathcal{F}_{\mathbf{x}}$  on  $\mathcal{H}$  is a bounded linear functional for a fixed  $\mathbf{x} \in \mathbf{X}$ . By the [Riesz Representation Theorem](#), there exists a fixed function  $K_{\mathbf{x}}(\cdot)$  such that for all  $f \in \mathcal{H}$

$$f(\mathbf{x}) = \mathcal{F}_{\mathbf{x}}[\mathbf{f}(\mathbf{x})] = \langle \mathbf{f}(\cdot), \mathbf{K}_{\mathbf{x}}(\cdot) \rangle$$

Where  $K_{\mathbf{x}}(\cdot) = K(\mathbf{x}, \cdot)$  and note

$$\langle K_{\mathbf{x}}(\cdot), K_{\mathbf{y}}(\cdot) \rangle = K_{\mathbf{y}}(\mathbf{x}) = \mathbf{K}_{\mathbf{x}}(\mathbf{y})$$

So the kernel is symmetric. Now to prove it is positive

Let  $\{x_1, \dots, x_n\}$  be a fixed collection. Then we have the corresponding kernel matrix  $K_{ij}$  as defined earlier. Let  $\mathbf{c}^T = [c_1 \ \cdots \ c_n]$ . Now taking the inner product

$$\begin{aligned} \langle c, Kc \rangle &\equiv c^T Kc = \sum_{i,j=1}^n c_i c_j K(x_i, x_j) = \sum_{i,j=1}^n c_i c_j \langle K_{x_i}(\cdot), K_{x_j}(\cdot) \rangle \\ &= \left\langle \sum_{i=1}^n c_i K_{x_i}(\cdot), \sum_{j=1}^n c_j K_{x_j}(\cdot) \right\rangle = \left\| \sum_{i=1}^n c_i K_{x_i}(\cdot) \right\|_{\mathcal{H}}^2 \geq 0 \quad \square \end{aligned}$$

Note this means that evaluation of  $f$  at  $x$  is equivalent to taking inner product of  $f$  with the fixed function  $K_x(\cdot)$

**Definition 8** We call the above kernel  $K(\mathbf{x}, \mathbf{y})$  the **reproducing kernel** of  $\mathcal{H}$

This allows us to define an RKHS in terms of its reproducing kernel, rather than attempting to derive the kernel from the definition of the function space directly

## Reproducing Kernels

**Example 4** The Linear Kernel

$$k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$$

**Example 5** The RBF/Gaussian Kernel

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

**Example 6** Polynomial Kernel

$$k(x, y) = (\gamma x^T y + c)^d$$

where  $\gamma > 0$  is the slope,  $c$  is the intercept, and  $d$  is the dimensionality of the data. Typical values for these parameters are  $\gamma = \frac{1}{d}$  and  $c = 1$

# Regularization

- More kernel examples
- Tikhonov Regularization
- Representation Theorem



# Purpose

- The goal with regularization is to make the result depend more smoothly on the data (restore well-posedness of the empirical risk minimization)
- $ERR(f) + \lambda pen(f)$
- Armed with the knowledge of kernels and the properties of RKHS, we can examine how different choices of RKHS and their norms provide different forms of regularization.

## Sobolev Kernel

Consider functions  $f : [0, 1] \rightarrow \mathbb{R}$  where  $f(0) = f(1) = 0$  and kernel

$$k(x, y) = \Theta(y - x)(1 - y)x + \Theta(x - y)(1 - x)y$$

Induces the norm

$$\|f\|_{\mathcal{H}}^2 = \int \omega^2 |F(\omega)|^2 d\omega$$

Where

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt$$

Note this is the Fourier Transform. The norm will be infinite for any function whose frequency magnitudes do not decay faster than  $\frac{1}{\omega}$ . This imposes a condition on the smoothness of the functions, since a high derivative gives rise to high frequency components.

# Gaussian Kernel

We can also see how the Gaussian kernel leads to a penalization on high frequency terms. The norm

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{2\pi^d} \int |F(\omega)|^2 e^{\frac{\sigma^2 \omega^2}{2}} d\omega$$

leads to a harsher penalization

## Linear Kernel

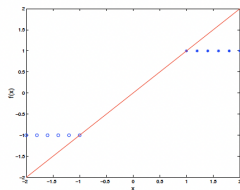
The linear kernel provides a simple way to visualize how regularization controls complexity. Consider the one-dimensional case

$$f(x) = wx$$
$$k(x, x_i) = x^T x_i$$

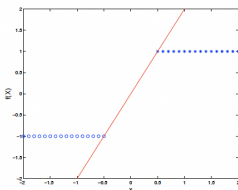
With norm

$$\begin{aligned}\|f\|_{\mathcal{H}}^2 &= \langle f, f \rangle_{\mathcal{H}} \\ &= \langle K_w, K_w \rangle_{\mathcal{H}} \\ &= K(w, w) = w^2\end{aligned}$$

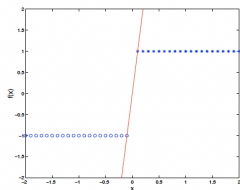
So complexity is controlled by the slope. A classification problem can be thought of as “harder” when the distinctions between the two classes are less pronounced



(a) Wide margin



(b) Moderate margin



(c) Small margin

Figure 1: Three different training sets to demonstrate that higher slopes are necessary to describe the data as the class distinctions become finer

# Tikhonov Regularization

The goal is to use RKHS to derive the general solution of Tikhonov regularization in RKHS, known as the representer theorem.

$$\hat{f} = \arg \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n V(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

where

$\lambda > 0$  is a regularization parameter

$V(f(x); y)$  is the loss function, that is the price we pay when we predict  $f(x)$  in place of  $y$

$\|\cdot\|_{\mathcal{H}}^2$  is the norm in the function space  $\mathcal{H}$

Potential problems to well-posedness?

- Regularization imposes stability on the problem
- The loss functions has to be convex to ensure there is a solution to the minimization problem
- The function space  $\mathcal{H}$  can be infinite-dimensional

## Representation Theorem

**Theorem 2** A solution to the Tikhonov optimization problem can be written as

$$\hat{f} = \sum_{i=1}^n \alpha_i K(x, x_i)$$

**Proof** Assume we project the function  $f$  onto a subspace spanned by the representer of the training set

$$\mathcal{H}_0 = \{f \in \mathcal{H} | f = \sum_{i=1}^n \alpha_i K_{x_i}\} \quad (1)$$

Decomposing  $f$  into parallel and perpendicular components

$$f = f_{\parallel} + f_{\perp} \quad (2)$$

$$\implies \|f\|^2 = \|f_{\parallel}\|^2 + \|f_{\perp}\|^2 \geq \|f_{\parallel}\|^2 \quad (3)$$



## Using the reproduction property of the RKHS

$$f(x_i) = \langle f(\cdot), k(x_i, \cdot) \rangle \quad (4)$$

By (2)

$$= \langle f_{\parallel}, k(x_i, \cdot) \rangle + \langle f_{\perp}, k(x_i, \cdot) \rangle \quad (5)$$

$$= \langle f_{\parallel}, k(x_i, \cdot) \rangle \quad (6)$$

$$= f_{\parallel}(x_i) \quad (7)$$

The perpendicular has inner product zero with the bases of the subspace. The optimization problem becomes

$$\hat{f} = \arg \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n V(f_{\parallel}(x_i), y_i) + \lambda \|f_{\parallel}\|_{\mathcal{H}}^2 \quad (8)$$



## Results

- For the minimization problem, we only require the components lying in the space spanned by the kernels of the RKHS
- the minimizer can be written as a linear combination of kernel terms from these points guarantees that we can represent the minimizer as a vector in  $\mathbb{R}^n$
- We only require  $n$  numbers,  $\alpha_i$  to optimize the infinite-dimension problem

# Feature Map and Feature Space

- What is a feature space
- Distance between means
- Model Selection
- Ridge Regression

# Feature Map

**Definition 9** We define the mapping

$$\phi : \mathcal{X} \rightarrow \mathcal{H}$$

to transform data from the input space to the feature space. In other words, the mapping pulls data into the feature space

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

Here  $\phi(\mathbf{x})$  is the **feature map**, which is a vector with components

$$\phi(\mathbf{x}) := (\sqrt{\lambda_1}\psi_1(\mathbf{x}), \sqrt{\lambda_2}\psi_2(\mathbf{x}), \dots)$$

Where  $\{\lambda_i\}$  are eigenvalues of the kernel operation and  $\{\psi_i(\mathbf{x})\}$  are eigenvectors.

Consider the kernel

$$\begin{aligned}k(x, y) &= \langle \phi(x), \phi(y) \rangle \\ &= \phi(x)^T \phi(y)\end{aligned}$$

Consider a data matrix  $X \in \mathbb{R}^{n \times d}$ . We then have a feature map of all points

$$\Phi(x) = (\phi(x_1), \phi(x_2), \dots, \phi(x_n))$$

The Kernel matrix is therefore

$$\begin{aligned}K &= \langle \Phi(x), \Phi(x) \rangle \\ &= \Phi(x)^T \Phi(x)\end{aligned}$$

There is no need to compute kernel using eigenfunctions but a simple inner product suffices for kernel computation

# Spaces

**Definition 10** The space in which data  $X$  exists is called the **input space** and is usually an  $\mathbb{R}^d$  Euclidean space. The RKHS to which the data is pulled is called the **feature space**.

Pulling data to the feature space is performed using kernels which is the inner product of points in RKHS.

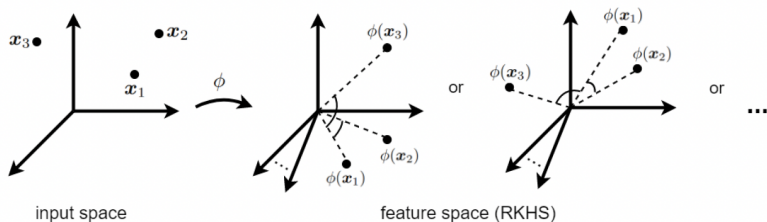


Figure 2: Pulling data from the input space to the feature space (RKHS). The explicit locations of pulled points are not necessarily known but the relative similarity (inner product) of pulled data points is known in the feature space.

## Distance Between Means

Suppose we have two distributions  $p$ ,  $q$  and we sample  $(x_i)_{i=1}^m$  from  $p$  and  $(y_i)_{i=1}^n$  from  $q$ . What is the distance between their means in feature space?

$$\begin{aligned} \left\| \frac{1}{m} \sum_{i=1}^m \phi(x_i) - \frac{1}{n} \sum_{y=1}^n \phi(y_j) \right\|_{\mathcal{H}}^2 &= \left\langle \frac{1}{m} \sum_{i=1}^m \phi(x_i) - \frac{1}{n} \sum_{y=1}^n \phi(y_j) \right. \\ &\quad \left. , \frac{1}{m} \sum_{i=1}^m \phi(x_i) - \frac{1}{n} \sum_{y=1}^n \phi(y_j) \right\rangle_{\mathcal{H}} \\ &= \frac{1}{m^2} \left\langle \sum_{i=1}^m \phi(x_i), \sum_{i=1}^m \phi(x_i) \right\rangle + \dots \end{aligned}$$



$$= \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m k(x_i, x_j) + \frac{1}{n^2} \sum_{i=1}^m \sum_{j=1}^n k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n k(x_i, y_j)$$

- How is this useful?
- In the case  $\phi(x) = x$ , we can use this statistic to distinguish distributions with different means.
- If we use the feature mapping  $\phi(x) = [x \ x^2]$  we can distinguish both means and variances.
- More complex feature spaces permit us to distinguish increasingly complex features of the distributions.

# Model Selection

In kernel regression methods we control two things

- The kernel controls the smoothness of the class of functions
- $\lambda$  controls the trade-off between function smoothness and fitting error
- A too large  $\lambda$  prioritises smoothness over getting a small prediction error on the points, resulting in a very smooth function which barely follows the shape of the underlying data: **underfitting**
- A too small  $\lambda$  gives too much priority to fitting small fluctuations in the data due to noise, at the expense of smoothness: **overfitting**

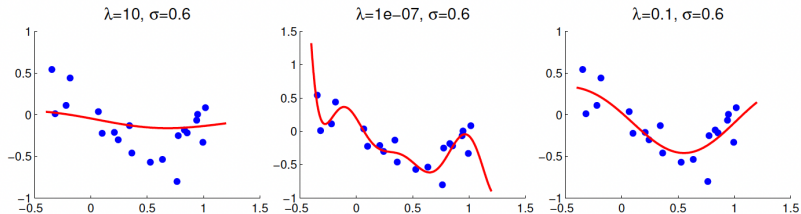


Figure 3: Effect of choice of  $\lambda$  on the fit of ridge regression. Here (a) represents underfitting, (b) represents overfitting, and (c) represents an appropriate choice for  $\lambda$

# Ridge Regression

Consider the solution to the Ridge Regression model

$$\hat{\theta}_{ridge} = \operatorname{argmin}_{\theta \in \mathcal{H}} \left( \sum_{i=1}^n (y_i - \langle \theta, \phi(x) \rangle)_{\mathcal{H}}^2 + \lambda \|\theta\|_{\mathcal{H}}^2 \right)$$

Consider different options for features

$$\phi_p(x) = \begin{pmatrix} x \\ x^2 \\ \vdots \\ x^\ell \end{pmatrix}$$

$$\phi_s(x) = \begin{pmatrix} \sin x \\ \cos x \\ \vdots \\ \cos \ell x \end{pmatrix}$$

Thus we have the new data matrix







$$X = (\phi(x_1) \quad \dots \quad \phi(x_n)) \quad X^T X = \sum_{i=1}^n \phi(x_i) \otimes \phi(x_i)$$

$$(X^T X)_{ij} = k(x_i, x_j)$$

Therefore the ridge regression solution is

$$\begin{aligned} \hat{\theta}_{ridge} &= X(K + \lambda I_n)^{-1} y \\ &= \sum_{i=1}^n \alpha_i \phi(x_i) \\ \alpha_i &= (K + \lambda I_n)^{-1} y \end{aligned}$$

## References

-  Benjamin Ghoggh, Ali Ghodsi, Fakhri Karray, and Mark Crowley, *Reproducing Kernel Hilbert Space, Mercer's Theorem, Eigenfunctions, Nystrom Method, and Use of Kernels in Machine Learning: Tutorial and Survey*, Scientific Reports (June 15, 2021)
-  Lorenzo Rosasco, Greg Durrett, *Reproducing Kernel Hilbert Spaces* MIT Feb 10, 2010
-  *Infinite Dimensional Vector Spaces* MA 751 Part 3
-  *Measurability and Hilbert Spaces* MA 751 Part 4
-  *Statistical machine learning and kernel methods* MA 751 Part 5
-  Arthur Gretton *Introduction to RKHS, and some simple kernel algorithms* Gatsby Computational Neuroscience Unit Oct 16, 2019

*Thank You!*