Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional

Unidad Zacatenco

Departamento de Matemáticas

Espacios de Banach con Kernel Reproductor y su Aplicación a Máquinas de Vectores Soporte

T E S I S

Que presenta

ERICK ALAN SILVA SALAZAR

para obtener el Grado de

MAESTRO EN CIENCIAS

en la especialidad de

MATEMÁTICAS

Director de Tesis

Dr. Carlos Gabriel Pacheco González

Ciudad de México **Febrero de 2022**

CENTER FOR RESEARCH AND ADVANCED STUDIES OF THE NATIONAL POLYTECHNIC INSTITUTE

Campus Zacatenco

Department of Mathematics

Reproducing Kernel Banach Spaces and its Application to Support Vector Machines.

T H E S I S

PRESENTED BY

ERICK ALAN SILVA SALAZAR

TO OBTAIN THE DEGREE OF

MASTER IN SCIENCE

IN THE SPECIALITY OF

MATHEMATICS

THESIS ADVISOR

Dr. Carlos Gabriel Pacheco González

Ciudad de México **Febrero de 2022**

Abstract

The purpose of this work is to bring the latest results about Reproducing Kernel Banach Spaces to a non-specialist public with some mathematical maturity.
A Reproducing Kernel Hilbert Space H is a space of functions defined on a fixed set X with an associated reproducing kernel k, i.e.:

$$f(x) = \langle f, k(\cdot, x) \rangle_H \; \forall f \in H.$$

The extension of the reproducing property to Banach spaces is done by bilinear forms and feature maps. We chose the framework given in [27] due to its generality and conciseness.

The first chapter begins with the essentials of Reproducing Kernel Hilbert Spaces and the equivalent ways to construct one. The second part consists of extending some results to Reproducing Kernel Banach Spaces, like the way to construct them, continuity of the functions, and boundedness properties. The third part fills in the details of constructions that have appeared in literature ([45] [16] [38] [37] [43] [27]) and expands on some of the concrete examples.

The second chapter starts with an explanation of SVM applied to classification tasks and its extension through reproducing kernels. The second part is about the Representer theorem for RKHS and its extensions to some classes of RKBS.

Resumen

El propósito de este trabajo es acercar resultados recientes sobre los Espacios de Banach con Núcleo Reproductor a un público no especializado con cierta madurez matemática. Un Espacio de Hilbert con Núcleo Reproductor es un espacio de funciones definidas sobre un conjunto fijo X con un núcleo reproductor k, es decir:

$$f(x) = \langle f, k(\cdot, x) \rangle_H \; \forall f \in H.$$

La extensión a espacios de Banach se realiza mediante formas bilineales. Elegimos la estructura dada en [27] debido a su generalidad y brevedad.

El primer capítulo comienza con los fundamentos de los Espacios de Hilbert con Núcleo Reproductor formas equivalentes para construir uno. La segunda parte consiste en extender algunos resultados a los Espacios de Banach de Núcleo Reproductor, como maneras de construirlos, la continuidad de las funciones y propiedades de ser acotadas. La tercera parte rellena los detalles faltantes en las construcciones que han aparecido en la literatura ([45] [16] [38] [37] [43] [27]) , y desarrolla algunos de los ejemplos concretos.

El segundo capítulo comienza con una explicación de la SVM aplicada a tareas simples de clasificación y su extensión por medio de núcleos reproductores. La segunda parte del capítulo trata sobre el teorema de Representación para RKHS y sus extensiones a algunas clases de RKBS.

Agradecimientos

Un enorme agradecimiento a mi familia que me ha apoyado en todo lo posible. A todos mis amigos que han sido un soporte emocional en estos tiempos. A mi asesor Carlos Pacheco, quien con infinita paciencia me guió durante mi estadía en el CINVESTAV. También quiero agradecer al personal del departamento que hicieron muy placentera mi estadía.
Y finalmente mi agradecimiento al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico, el cual me fue indispensable y me permitió terminar el posgrado.

Introduction

A problem that has appeared repeatedly through every branch of science is to find a function which satisfies some constraints. A common example is forcing a function f to take certain prescribed target values $\{y_i\}_{i=1}^N$ at some fixed input points $\{x_i\}_{i=1}^N$. This can be resumed in the following equations:

$$f(x_i) = y_i,\ i = 1, \cdots N.$$

Machine Learning methods attempt to solve these problems in multiple ways. The family of algorithms collectively known as *Support Vector Machines* (SVM) is made of tools for solving this kind of problems.

The simplest form of a SVM solves a binary classification problem in $\mathbb{R}^n$ by separating the points using hyperplanes. The SVM algorithm originally consists of finding an optimal hyperplane that separates points $\{x_i\} \subseteq \mathbb{R}^n$ into two categories. To classify them we want a function which is positive if applied to one class, and negative if applied to the other class. The key to find the solution is knowing the inner products

$$\langle x_i, x_j \rangle.$$

This comes with a caveat: hyperplanes are often not enough to correctly classify into two classes. But the idea of using separating hyperplanes can be rescued by mapping the original points to another space by using *feature maps*. The feature map sends the samples in a non-linear way to a new normed space where we can classify correctly. A convenient tool that allows us to exploit the full potential of the new space is *reproducing kernels*.

The ideas described above take place in Reproducing Kernel Hilbert Spaces (RKHS) and Reproducing Kernel Banach Spaces (RKBS). These are spaces of functions $\mathcal{F}$ defined on a set X. This means that we want to be able to evaluate the objects $f \in \mathcal{F}$ on points of X. These points thus induce *evaluation functionals* which we will ask to be continuous with respect to the space's norm. Moreover, these spaces also are coupled with another space, its dual space, and a bilinear form defined on these two.

When working with Hilbert spaces, due to the Riesz Representation theorem we can associate a vector to each continuous functional. Using this fact and associating the evaluation functional $eval_x$ to a point x, we define a feature map from the set X to the Hilbert space. The descriptive "Reproducing kernel" comes from this, since for every point x, we associate a vector $k_x \in H$ such that for every function $f \in H$ the following holds:

$$f(x) = eval_x(f) = \langle f, k_x \rangle,$$

In particular, for two points $x, y \in X$ we have the function

$$k(x, y) = eval_y(k_x) = \langle k_x, k_y \rangle.$$

Therefore we can extend the algorithm by constructing a new space with its inner product induced by a reproducing kernel function k. By mapping the original points to a more appropriate space, we can then apply the separation algorithm with the new points. And now the inner product between the new points is then represented by the evaluations $k(x_i, x_j)$. This fact is sometimes known as the *kernel trick*, since it allows to compute the inner product by just evaluating the reproducing kernel.

Since we can do the mapping to a Hilbert space beforehand, there is no need for the original set X to have a particular structure. This means that we can interpret the mapping translating a inner product structure to the original set. Or if there is a prior structure, it can be seen as a deformation of the space in a non-linear way to make classification possible.

Reproducing kernels and their induced Hilbert spaces have been thoroughly studied since last century. More recently, Boser, Guyon and Vapnik [8] found that they could also be used in the context of Machine Learning. They noticed that the Support Vector Machine algorithm, designed by Vapnik [40], could be extended by replacing the inner product with an arbitrary reproducing kernel.

The study of general reproducing kernel spaces go back to the last century. Positive definite kernels have been studied due to their relation to Hilbert spaces. The Reproducing Kernel Banach Spaces (RKBS) for Machine Learning purposes were proposed first in the context maximal margin classification for metric spaces [21]. Later in [45], they proposed using semi-inner products to generalize RKHS. The semi-inner product approach was not implemented until [16] constructed a special class of RKBS specifically made for this purpose. Another road that took place in RKBS research was developing a RKBS with a ℓ^1-type norm for the purpose of ℓ^1 regularization [37, 38, 26]. There is also current research aiming to explain other Machine Learning achievements for approximating seemingly arbitrary functions, like the Transformer model [42], as well as other Neural Networks [5].

This work is intended to be an introduction to some results from the past decade about Reproducing Kernel Banach Spaces in Machine Learning. As such, we will focus on its theory, some particular constructions and how they can be used for SVM's. We will be omitting essential related topics such as error bounds for the error or its probabilistic treatment [9].

The work is divided into two ideas: abstracting the useful points of the RKHS framework and making a RKBS version of them. These are distributed over three chapters.

The first and second chapters are an introduction to the theory of RKHS and RKBS respectively. We set up some key ideas of a RKHS in the first chapter, these being the relation between the reproducing kernel, the feature maps and the inner product. These ideas will be further developed when we generalize them to RKBS. We then approach the construction of RKBS spaces using feature maps and bilinear forms. This mimics closely the way RKHS are constructed, and later we show that this idea arises naturally by pairing vectors and functionals with the bilinear maps.

The second chapter goes further into particular constructions. This is to emphasize the flexibility of RKBS:

- If convenient, the choice of the bilinear form can be changed to a large degree, an inner product does not allow that.
- For a RKHS we need the Hilbert space H, its dual $H' = H$ and the inner product as a bilinear form. A RKBS needs a Banach space B and a space C just "big enough" to separate the functions in B. And a bilinear form between them. So C does not need to be made of functions defined on the same set as B's functions.

The freedom of choosing these two give place to multiple ways to find reproducing kernels for one fixed space.

The third chapter is concerned with the application of reproducing kernel spaces to SVM. We briefly give the derivation that justifies the algorithm of the original

SVM. We then highlight some parts that let us connect this algorithm to a non-linear version where RKHS are used. Then we show the reason the kernel trick found its place within Machine Learning algorithms: the Representer theorem. This theorem shows that a large class of problems can be solved by functions living in a finite-dimensional subspace, independently of the dimension of the Hilbert space. More specifically, it says that a solution takes the form

$$f = \sum_j \alpha_j k(\cdot, x_j).$$

The second part of the chapter shows that this idea can be generalized to the RKBS studied previously. This is where the feature maps and bilinear forms find their use by translating some Hilbert space concepts to Banach spaces. Finally, we conclude the chapter with various versions of the Representer theorem for two classes of RKBS. The problem of finding "Representer theorems" for general spaces is not limited to these classes. To read more on it we recommend reading [4, 34].

Contents

Chapter 1

Reproducing kernel Banach space (RKBS)

1.1 Reproducing kernel Hilbert spaces (RKHS)

In this section we give an overview of Reproducing Kernel Hilbert Spaces to draw parallels from when we define the Reproducing Kernel Banach Spaces.

Definition 1.1.1 (RKHS)**.** A Hilbert space of functions over a set X is called a **Reproducing Kernel Hilbert Space** (RKHS) if the evaluation functionals are continuous.

Now we define what it means for a RKHS to have a reproducing kernel function:

Definition 1.1.2. Let H be a Hilbert space of functions H over X with a function $k : X \times X \to \mathbb{C}$ such that

$$k_x := k(\cdot, x) \in H, \ \forall x \in X \tag{1.1}$$

$$f(x) = \langle f, k(\cdot, x) \rangle, \ \forall f \in H \tag{1.2}$$

then k is called a **reproducing kernel** for the Hilbert space H.

The Riesz representation theorem assures that every continuous functional is represented by the inner product with a fixed element of the Hilbert space. Then with the previous definition we can easily derive the so-called reproducing property of the kernel function for a RKHS. In a way this will motivate the coming definitions for the Banach spaces, despite the general absence of an inner product.

Definition 1.1.3. A mapping from a non-empty set X to a Banach space V will be called a **feature map** and V will be called a **feature space**.

As we mentioned before, every evaluation functional is represented by a unique vector, that means there is a mapping $x \mapsto v_x$ from the set X to the Hilbert space such that $eval_x = \langle \cdot, v_x \rangle$, thus this is an example of a feature map.

Theorem 1.1.1. Let H be a Hilbert space of functions. The following statements are equivalent [39].

(i) H is a RKHS.

(ii) H has a reproducing kernel function.

Proof. $(i) \implies (ii)$
Let $\Phi : X \to H$ be a feature map and $I : H' \to H$ the Riesz map. Note that $v_x = \Phi(x) = I(eval_x)$ is true for all $x \in X$. Next we define a complex-valued function by:

$$k(x, y) = \langle v_x, v_y \rangle, \forall x, y \in X.$$

Then by the observation above we have that

$$k(x,y) = \langle v_x, v_y \rangle = \langle I(eval_y), I(eval_x) \rangle = eval_x(I(eval_y)) = I(eval_y)(x) = v_y(x).$$

So we have that $k(\cdot, y) \in H$, moreover we showed that $v_y(\cdot) = k(\cdot, y)$. Next we prove that this function has property 1.2: Let $f \in H$ be an arbitrary function, then:

$$f(x) = eval_x(f) = \langle f, v_x \rangle = \langle f, k(\ , x) \rangle.$$

So $k(x,y)$ is a reproducing kernel for the RKHS.
$(ii) \implies (i)$
This follows from the reproducing property: take an arbitrary $x \in X$ and $f \in H$, then

$$|eval_x(f)| = |f(x)| = |\langle f, k(x,\ ) \rangle| \leq \|f\| \|k(x,\ )\|$$

An so we have that evaluation functionals are continuous, making H a RKHS. □

A function with property 1.2 is said to have the reproducing property. A reproducing kernel for a Hilbert space has many properties that come from its relationship with the inner product of the space itself.

Definition 1.1.4. A function $k : X \times X \to \mathbb{C}$ such that:

$$\sum_{i,j=0}^{n} a_i \bar{a}_j k(x_i, x_j) \geq 0 \text{ for all } x_i \in X, a_i \in \mathbb{C} \tag{1.3}$$

will be called a **positive definite kernel**.

Theorem 1.1.2. Let H be a RKHS and k be its reproducing kernel, then:

1. k is conjugate symmetric, i.e. $k(x,y) = \overline{k(y,x)}$.
2. $k(x,y)$ is positive definite.

Proof. Let $x, y \in X$. From the properties in 1.1.2 we have

$$k(x,y) = k_y(x) = \langle k_y, k(\cdot, x) \rangle = \langle k_y, k_x \rangle = \overline{\langle k_x, k_y \rangle} = \overline{\langle k_x, k(\cdot, y) \rangle} = \overline{k_x(y)} = \overline{k(y,x)}.$$

Therefore $k(x,y) = \overline{k(y,x)}$. Consider arbitrary $a_i \in \mathbb{C}$ and $x_i \in X$ for $i = 1, \ldots, n$. The function $f_{a_1 \ldots a_n}(\cdot) = \sum_{i=1}^{n} a_i k(\cdot, x_i)$ is an element of H, and as such $0 \leq \langle f_{a_1 \ldots a_n}, f_{a_1 \ldots a_n} \rangle$, but this last expression is $\sum_{i,j=0}^{n} a_i \bar{a}_j k(x_i, x_j)$ because k has the reproducing property. The uniqueness is a consequence of the reproducing property and the first property:

$$k'(x,y) = k'_y(x) = \langle k'_y, k_x \rangle = \overline{\langle k_x, k'_y \rangle} = \overline{k_x(y)} = \overline{k(y,x)} = k(x,y).$$

□

Given a conjugate symmetric and positive definite function k, we can construct the associated RKHS by using finite linear combinations of the form $k(\cdot, x)$ see [33] for the details.

Theorem 1.1.3. Let

$$k : X \times X \to \mathbb{C}$$

be a conjugate symmetric, positive definite function, then the function space

$$H_0 = \{f(y) = \sum_{i=0}^{n} a_i k(x_i, y) : n \in \mathbb{N}, a_i \in \mathbb{C}\} \tag{1.4}$$

is a pre-Hilbert space with the inner product

$$\langle \sum_{i=0}^{n} a_i k_{x_i}, \sum_{j=0}^{m} b_i k_{z_i} \rangle := \sum_{i=0,j=0}^{n,m} a_i \bar{b}_j k(x_i, z_j). \tag{1.5}$$

And its completion H is a RKHS with kernel function $k(x, y)$ and feature map $x \mapsto k(x, \cdot)$.

Due to Theorems 1.1.2 and 1.1.3 we can say that there is a bijection between positive semi-definite functions and RKHS. When moving the theory from Hilbert spaces to Banach spaces such relationship is lost. The above construction could be used for a vector space which would be a first attempt of a reproducing kernel Banach space, but it would find the issue of which norm to give the vector space. This characteristic is what will give rise to non-isometric Banach spaces having the same reproduction kernel. Manipulating evaluation functionals in Hilbert spaces can be reduced to studying one particular bilinear form due to the Riesz representation Theorem, but for the Banach spaces there is no natural or unique mapping from the given space $\mathscr{B}$ to its dual $\mathscr{B}'$.

1.2 Reproducing Kernel Banach Spaces (RKBS)

In the previous section we only worked in the context of Hilbert spaces, which are endowed with an inner product. This gives us a way to represent linear functionals in an unique way by using vectors of the space. To generalize that idea of representing abstract linear functionals as a concrete relation between two vectors in Banach spaces, we will use bilinear maps between the two spaces. This will allow us to bring the reproducing property to more general Banach spaces, but accordingly some things will not be guaranteed like uniqueness of a reproducing kernel for a given Banach space, nor a norm induced by the kernel function.

We first define what it means for a space to be a Reproducing Kernel Banach Space [27].

Definition 1.2.1. [27] Let $\mathscr{B}$ be a Banach space of functions defined on a set X such that a function has zero norm if and only if it vanishes at every point. Then $\mathscr{B}$ is called a **Reproducing Kernel Banach Space (RKBS)** if every evaluation functional is continuous.

The only difference between the definitions of RKHS and RKBS for us is the choice of a Banach space $\mathscr{B}$ over a Hilbert space. That means there may not be an inner product to use and consequently there may not be one way to represent functionals with vectors from the Banach space. In some particular cases where the bilinear form has a stronger relation to the space's geometry we will recover some of the properties a kernel function has in the case of a RKHS. In each section we will be showing different ways to find these bilinear maps, and then we will show that the follow the next construction. For this general construction, we will require another Banach space and a non-degenerate[1] bilinear mapping, but we will not force the norms to be determined by the bilinear form, instead we only ask for the two previous conditions. For our purposes we will only consider bilinear forms which are both continuous and non-denegerate and will call them only bilinear forms or bilinear maps.

Definition 1.2.2. [27] Let $\mathscr{B}_1$ be a RKBS defined on a set X. Suppose there exists a second Banach space $\mathscr{B}_2$ defined on a set Y, a continuous[2] and non-degenerate bilinear form $\langle \cdot, \cdot \rangle : \mathscr{B}_1 \times \mathscr{B}_2 \to \mathbb{C}$, and a function $k : X \times Y \to \mathbb{C}$ such that $k(x, \cdot) \in \mathscr{B}_2$ for all $x \in X$ with the **right-sided reproducing property**:

$$f(x) = \langle f, k(x, \cdot) \rangle \text{ for all } x \in X,\ f \in \mathscr{B}_1. \tag{1.6}$$

Then k is called a **reproducing kernel** for the RKBS $\mathscr{B}_1$ with respect to the bilinear form $\langle \cdot, \cdot \rangle$.

Definition 1.2.3. Consider a Banach space $\mathscr{B}_2$ defined on a set Y and a RKBS $\mathscr{B}_1$ defined on a set X, with reproducing kernel $k : X \times Y \to \mathbb{C}$ with respect to a bilinear form $\langle \cdot, \cdot \rangle$ defined on $\mathscr{B}_1 \times \mathscr{B}_2$. Assume also that $k(\cdot, y) \in \mathscr{B}_1$ for all $y \in Y$ and that it has the **left-sided reproducing property**:

$$g(y) = \langle k(\cdot, y), g \rangle \text{ for all } y \in Y,\ g \in \mathscr{B}_2. \tag{1.7}$$

[1] We say that a bilinear form $B : V \times W \to \mathbb{C}$ is non-degenerate if for every pair of non-zero vectors $v \in V, w \in W$ the linear functionals $B(v, \cdot)$ and $B(\cdot, w)$ are not trivial.

[2] A bilinear form B defined over two normed spaces $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ will be called continuous if for every pair of vectors $(v, w) \in V \times W$ we have $|B(v, w)| \leq C_x \|v\|_V \|w\|_W$ for a constant $C_x > 0$.

Then the second condition implies that $\mathscr{B}_2$ is a RKBS and we call $\mathscr{B}_2$ an adjoint RKBS for $\mathscr{B}_1$. We will call $\mathscr{B}_1$ and $\mathscr{B}_2$ a **pair of adjoint RKBS's**. When there is no risk of ambiguity, we will refer to them as an **adjoint pair**.

We also have that a reproducing kernel for $\mathscr{B}_2$ can be defined by swapping the roles of both the sets X and Y and the spaces $\mathscr{B}_1$ and $\mathscr{B}_2$.

Theorem 1.2.1. Let $\mathscr{B}_1$ and $\mathscr{B}_2$ be an adjoint pair. Then the function $\tilde{k}(x,y) := k(y,x)$ is a reproducing kernel for $\mathscr{B}_2$.

To give intuition of how we will proceed, we see a Hilbert space and its inner product from another perspective. In Theorem 1.1.3, we see that the RKHS we defined consists of functions given by series expansions on the terms

$$x \mapsto \Phi(x) := k(\cdot, x),$$

and the inner product, which is defined by the kernel function [33]. On the other hand, the kernel function can also be given in terms of the inner product and the feature map by the equation $k(x,y) = \langle \Phi(x), \Phi(y) \rangle = \langle k(\cdot,x), k(\cdot,y) \rangle$. Since a Hilbert space is isometric to its dual, we can interpret the second argument $\Phi(y)$ as being a linear functional in H', which is a Banach space, and regard the inner product as the evaluation of the functional $\Phi(y)$. To make this function a bilinear form, we give another scalar product to H' by defining $\alpha \bullet v := \bar{\alpha} v$ for every $v \in H'$, bestowing it the structure of Banach space with the norm induced by H. Furthermore, by leaving the second argument fixed, $k(x,y) = \langle \Phi(x), \Phi(y) \rangle$ is a function defined on X which lies in the spanned space seen in Theorem 1.1.3. The next construction is inspired from such representation.

Definition 1.2.4. [27] Let V, W be Banach spaces and $\langle \cdot, \cdot \rangle_{V \times W} : V \times W \longrightarrow \mathbb{C}$ a bilinear form. We say that a subspace $E \subset V$ is dense in W with respect to the bilinear form $\langle \cdot, \cdot \rangle_{V \times W}$ if for a given $v \in W$ we have that

$$\langle u, v \rangle_{V \times W} = 0, \forall u \in E$$

then $v = 0$.

A necessary and sufficient condition for a bilinear form to satisfy this property is that the subspaces $\text{Span}\{\langle x, \cdot \rangle_{W_1 \times W_2}\}_{x \in W_1}$, $\text{Span}\{\langle \cdot, y \rangle_{W_1 \times W_2}\}_{y \in W_2}$ are weak-* dense in W_2^* and W_1^* respectively [1]

Construction 1.2.1.1.

Construction of RKBS by feature maps. Let X and Y be sets, W_1 and W_2 be a couple of Banach spaces, $\langle \cdot, \cdot \rangle_{W_1 \times W_2} : W_1 \times W_2 \to \mathbb{C}$ a continuous bilinear form and $\Phi_1 : X \to W_1$ and $\Phi_2 : Y \to W_2$ a pair of feature mappings such that the span of their images are dense subspaces with respect to the bilinear form. Consider the following functions spaces and their norms:

$$\begin{aligned} \mathscr{B}_1 &:= \{f_w(\cdot) = \langle \Phi_1(\cdot), w \rangle \mid w \in W_2\}, \\ \|f_w\|_{\mathscr{B}_1} &:= \|w\|_{W_2}, \end{aligned} \tag{1.8}$$

$$\begin{aligned} \mathscr{B}_2 &:= \{g_v(\cdot) = \langle v, \Phi_2(\cdot) \rangle \mid v \in W_1\}, \\ \|g_v\| &:= \|v\|_{W_1}. \end{aligned} \tag{1.9}$$

And the bilinear map is

$$\langle f_w, g_v \rangle_{\mathscr{B}_1 \times \mathscr{B}_2} := \langle v, w \rangle.$$

The next result is a direct consequence of how these spaces were constructed.

Theorem 1.2.2. Let $W_1, W_2, \mathscr{B}_1$ and $\mathscr{B}_2$ be as in Construction 1.2.1.1. Then W_1 is isometrically isomorphic to $\mathscr{B}_2$ and W_2 is isometrically isomorphic to $\mathscr{B}_1$.

From this point onward, whenever we say that two Banach spaces form an adjoint pair, we will assume they come from a construction like above with their respective feature maps.

When we showed that a kernel function defines a Hilbert space in Theorem 1.1.3, we derived the uniqueness of the norm by the properties of the positive definite function. In contrast, we only ask for continuity from the bilinear form, and this does not force a relation with the norms besides continuity. This makes up for one of the differences between RKHS and RKBS along the following:

- A second Banach space W_2 which does not necessarily have a relation with the space W_1, besides the bilinear map between them. Functions in it may be defined on an entirely different set Y. With Hilbert spaces this role was taken by the dual space, and both were spaces of functions defined on the same set, with the anti-isomorphism given thanks to the Hilbertian structure.

- A continuous bilinear map between the spaces $W_1$and W_2 which will serve the role the inner product did with RKHS. We ask for two things: continuity, and that it defines two monomorphisms, one from W_1 to W_2' and the other from W_2 to W_1'. The second condition makes it so there is a copy of W_1 in w_2' and a copy of W_2 in w_1'. The first condition makes it so the respective copies have a norm that is weaker [3] than their original norm. In other words, the conditions imposed on the bilinear form translate to relations between spaces and their duals.

[3] We say that a norm $\|\cdot\|_1$ defined on a Banach space B is weaker than the norm $\|\cdot\|_B$ if there exists a constant c> 0 such that $\|x\|_1 \leq \|x\|_B$ for all $x \in B$.

The next result says that for RKBS, convergence of a Cauchy sequence to a function is equivalent to pointwise convergence. This is due to the requirement of the continuity of evaluation functionals and the properties of the bilinear form.

Theorem 1.2.3. Let $\mathscr{B}_1$ and $\mathscr{B}_2$ be as above, then a Cauchy sequence in $\mathscr{B}_1$ with respect to its norm converges to 0 if and only if the limit is 0 pointwise.

Proof. The fact that convergence in norm implies pointwise limit is clear.
As for the other implication, let $\{\langle \Phi_1(\cdot), v_m\rangle\}_{m\in\mathbb{N}}$ be a Cauchy sequence which converges pointwise to 0. Since it is a Cauchy sequence there is a $v \in \mathscr{B}_2$ such that

$$\langle \Phi_1(\cdot), v_m\rangle \longrightarrow \langle \Phi_1(\cdot), v\rangle$$

in norm. Since evaluation functionals are continuous, then $\langle \Phi_1(\mathrm{x}), v\rangle = 0$ at every point, as such we have that $\langle \Phi_1(\cdot), v\rangle = 0$. But this is impossible since we asked for the bilinear form to be non-degenerate and Span $(\Phi_1(x)_{x\in X})$ to be a dense subspace of $\mathscr{B}_1$ □

We show next that the spaces constructed in Construction 1.2.1.1 are truly RKBS as we defined before:

Theorem 1.2.4. Let $\mathscr{B}_1, \mathscr{B}_2$ be as 1.8 and 1.9, then, together with Theorem 1.2.3 we can deduce that:

- $\mathscr{B}_1$ is a RKBS.
- $k(x,y) := \langle \Phi_1(x), \Phi_2(y)\rangle_{W_1\times W_2}$ is a reproducing kernel for $\mathscr{B}_1$ with respect to the continuous bilinear form

$$\langle f_v, g_w\rangle_{\mathscr{B}_1\times\mathscr{B}_2} := \langle w, v\rangle_{W_1\times W_2}.$$

- $\mathscr{B}_2$ is an adjoint RKBS for $\mathscr{B}_1$.

Proof. First we choose two arbitrary functions $f_{w_1}, f_{w_2} \in \mathscr{B}_1$ and a scalar $\lambda \in \mathbb{C}$. Since $\langle\cdot,\cdot\rangle_{W_1\times W_2}$ is a bilinear form we have for every $x \in X$:

$$\begin{aligned} f_{w_1}(x) + \lambda f_{w_2}(x) = \langle \Phi_1(x), w_1\rangle + \lambda\langle \Phi_1(x), w_1\rangle = \\ \langle \Phi_1(x), w_1\rangle + \langle \Phi_1(x), \lambda w_1\rangle = \langle \Phi_1(x), w_1 + \lambda w_2\rangle = f_{w_1+\lambda w_2}(x). \end{aligned} \tag{1.10}$$

So $\mathscr{B}_1$ is a vector space. The fact that $\mathscr{B}_1$ is a Banach space of functions comes from the non-degeneracy of the bilinear map together with the fact that W_2 is complete with the $\|\ \|_{W_2}$ norm, and it has the property we ask in definition 1.2.1. Now we need to show that all evaluation functionals are continuous: Let $x \in X$ be an arbitrary point and ψ_x the induced evaluation function. By the continuity of the bilinear form we have for an arbitrary $f_u \in \mathscr{B}_1$:

$$|\psi_x(f_u)| = |f_u(x)| = |\langle \Phi_1(x), u\rangle| \leq C\|\Phi_1(x)\|_{W_1}\|u\|_{W_2} = C\|\Phi_1(x)\|_{W_1}\|f_u\|_{\mathscr{B}_1}.$$

Therefore $\mathscr{B}_1$ is a RKBS. To show that k is a reproducing kernel for $\mathscr{B}_1$ we first prove that the bilinear form is continuous, take $f_u \in \mathscr{B}_1$, $g_v \in \mathscr{B}_2$, then

$$|\langle f_u, g_v\rangle_{\mathscr{B}_1\times\mathscr{B}_2}| = |\langle v, u\rangle_{W_1\times W_2}| \leq C\|v\|_{W_1}\|u\|_{W_2} = C\|g_v\|_{\mathscr{B}_1}\|f_u\|_{\mathscr{B}_1}.$$

Now we only need to verify that k is a reproducing kernel for $\mathscr{B}_1$, note first that

$$k(x,\cdot) = \langle \Phi_1(x), \Phi_2(\cdot)\rangle = g_{\Phi_1(x)}(\cdot) \in \mathscr{B}_2.$$

So we have the next equalities by definition of the bilinear form= $\langle\cdot,\cdot\rangle_{\mathscr{B}_1\times\mathscr{B}_2}$:

$$f_u(x) = \langle \Phi_1(x), u\rangle_{W_1\times W_2} = \langle f_u, g_{\Phi_1(x)}\rangle_{\mathscr{B}_1\times\mathscr{B}_2} = \langle f_u, k(x,\cdot)\rangle.$$

The proof that $\mathscr{B}_2$ is a RKBS with reproducing kernel $\tilde{k}(y,x) = k(x,y)$ is similar. □

Remark. It is worth mention that we can still obtain a RKBS if we weaken the hypothesis on the bilinear form $\langle\cdot,\cdot\rangle_{W_1\times W_2}$ by asking only non-degeneracy for the second space, meaning that for every $v \in W_2$ there exists $u \in W_1$ such that $\langle u, v\rangle \neq 0$. This is because this property ensures that the second space is embedded continuously in the dual space of W_1. If we weaken the hypothesis this way, the constructed space $\mathscr{B}_1$ continues to be a RKBS with the same space $\mathscr{B}_2$ and the associated bilinear form. It also retains the reproducing property for $\mathscr{B}_1$, which would coincide with the definition for right-sided reproducing property. Nevertheless $\mathscr{B}_2$ may have non-zero functions with zero norm. It could also lose its reproducing property with respect to the reproducing kernel.

When working with sets X, Y with extra structure, we can restrict their behavior accordingly, like asking for them to be bounded functions, as seen in [18].

Theorem 1.2.5. If $\mathscr{B}_1, \mathscr{B}_2$ are an adjoint pair RKBS and the feature maps $\Phi_1(x) = k(\cdot,x)$, $\Phi_2(y) = k(y,\cdot)$ are such that that $\sup_{x\in X}\{\|k(\cdot,x)\|_{\mathscr{B}_1}\}, \sup_{y\in Y}\{\|k(y,\cdot)\|_{\mathscr{B}_2}\}$ are both finite. Then every function in $\mathscr{B}_1$ and $\mathscr{B}_2$ is bounded and the identity maps $I_1 : \mathscr{B}_1 \to C_b(X)$, $I_2 : \mathscr{B}_2 \to C_b(Y)$, are bounded.

Proof. Let $f_v \in \mathscr{B}_1$ and $c \geq \sup_{x\in X}\{\|k(\cdot,x)\|_{\mathscr{B}_1}\}$, then

$$|f_v(x)| = |\langle f_v, k(\cdot,x)\rangle| \leq C\|f_v\|\|k(\cdot,x)\| \leq Cc\|f_v\|.$$

The proof for $\mathscr{B}_2$ is similar. □

Similarly, if the sets X and Y are topological spaces, we can impose conditions on the feature maps to exploit their structure. In applications, it is not unusual that the sets X, Y have a metric space structure, and since the range of the feature maps lie in Banach spaces, the following result becomes useful.

Theorem 1.2.6. [18] Let X and Y be topological spaces and $\mathscr{B}_1$ and $\mathscr{B}_2$ be as in construction 1.2.1.1. The function spaces $\mathscr{B}_1$ and $\mathscr{B}_2$ consist of continuous functions if any of these conditions hold:

- The feature maps $\Phi_1 : X \to \mathscr{B}_1$ and $\Phi_2 : Y \to \mathscr{B}_2$ are continuous.
- k is bounded and continuous with respect to each variable separately.

The first condition also implies that the reproducing kernel is continuous.

Proof. We will show that any function $f \in \mathscr{B}_1$ is continuous. First suppose that the feature maps are continuous. Then any function $f_v \in \mathscr{B}_1$ has the form:

$$f_v(x) = \langle \Phi_1(x), v\rangle.$$

So f_v is the composition of continuous functions, hence it is continuous. Likewise for the kernel we have that:

$$k(x,y) = \langle \Phi_1(x), \Phi_2(y) \rangle,$$

making it also continuous on each variable.

Now assume the second set of conditions. Since k is continuous separately on each variable, each function $k(\cdot, y) = \langle \Phi_1(\cdot), \Phi_2(y) \rangle \in \mathscr{B}_1$ is continuous for each $y \in Y$. Remember that Span $(\Phi_2(Y))$ is a dense subspace, so when we construct $\mathscr{B}_1$ we induce another dense subspace of functions of the form $\langle \Phi_1(\cdot), \Phi_2(y) \rangle$ which are continuous. That means that if we show that functions in $\mathscr{B}_1$ are uniform limits of these, then we are done. But this is a consequence of the previous theorem and Theorem 1.2.2. □

Since we ask for $\text{Span}(\Phi_1(X)) \subset W_1$ and $\text{Span}(\Phi_2(Y)) \subset W_2$ to be dense subspaces, due to the isometry, these images also conserve separability [31].

Theorem 1.2.7. Let X and Y be non-empty sets, $\mathscr{B}_1$ and $\mathscr{B}_2$ an adjoint pair defined on them. Suppose that the feature maps $\Phi_1 : X \to \mathscr{B}_1$ and $\Phi_2 : Y \to \mathscr{B}_2$ are continuous and that both X and Y are separable. Then $\mathscr{B}_1$ and $\mathscr{B}_2$ are also separable Banach spaces.

Proof. Due to the continuity of the feature maps, both $\Phi_1(X)$ and $\Phi_2(Y)$ are separable subsets of W_1 and W_2 respectively. Choose countable, dense subsets $D_1 \subset \Phi_1(X)$, $D_2 \subset \Phi_2(Y)$ and consider linear combinations of their elements where the coefficients are complex numbers with rational coordinates, then we obtain countable sets which are dense in W_1 and W_2. And from Theorem 1.2.2 we know that these spaces are isometrically isomorphic to $\mathscr{B}_2$ and $\mathscr{B}_1$ respectively. □

As the first example of a reproducing kernel Banach space, we show that the concept generalizes a RKHS as expected.

Theorem 1.2.8. Let $\mathscr{H}$ be a RKHS defined on a set X, with feature map $\Phi : X \to \mathscr{H}$, and with reproducing kernel k. Then H is a RKBS of the form 1.2.1.1 with the same reproducing kernel.

Proof. Since there is no change in the chosen norm the valuation functionals are continuous and $\mathscr{H}$ is a RKBS. To see that the it fits scheme 1.2.1.1, let $R : \mathscr{H}' \to \mathscr{H}$ be the Riesz mapping, which is anti-linear and:

$$\begin{aligned} X &= Y. \\ W_1 = \mathscr{H}, W_2 &= \mathscr{H}'. \\ \Phi_1 = \Phi, \Phi_2 &= R^{-1} \circ \Phi. \\ \langle u, v \rangle_{W_1 \times W_2} &= \langle u, R(v) \rangle_{\mathscr{H}} \end{aligned}$$

So the spaces $\mathscr{B}_1$ and $\mathscr{B}_2$ end up being:

$$\begin{aligned} \mathscr{B}_1 &= \{ f_u(\cdot) = \langle \Phi_1(\cdot), u \rangle_{W_1 \times W_2} = \langle \Phi_1(\cdot), R(u) \rangle_{\mathscr{H}} \mid u \in \mathscr{H} \}. \\ \mathscr{B}_2 &= \{ g_v(\cdot) = \langle v, \Phi_2(\cdot) \rangle_{W_1 \times W_2} = \langle v, R(R^{-1}(\Phi(\cdot))) \rangle_{\mathscr{H}} = \langle v, \Phi(\cdot) \rangle_{\mathscr{H}} \mid v \in H \}. \end{aligned}$$

If we show that $\mathscr{H} \cong \mathscr{B}_2$, since $\mathscr{B}_1$ and $\mathscr{B}_2$ form an adjoint pair, we are done. But this is a consequence of how the space was constructed. □

Chapter 2

RKBS Constructions

This section will be devoted to examples and constructions of diverse RKBS. Starting from the candidates to RKBS, we will explain the particular way each approach choose to represent evaluation functionals, which will involve Banach spaces that can be embedded into the dual space of the RKBS candidates, explain the particular construction given and some concrete examples, and finally show that they fit the construction outlined 1.8 and 1.9 for some of them.

2.1 Reflexive spaces

The first class of spaces we develop are the reflexive spaces. A Banach space B is said to be reflexive if it is isometric to its double dual B'' [45].

Definition 2.1.1. [45] Let $\mathscr{B}$ be a reflexive Banach space of functions defined on non-empty set X. If its dual $\mathscr{B}'$ is isometric to some Banach space of functions with non-empty domain Y [1], and evaluation functionals are continuous for both spaces then we will call $\mathscr{B}$ a **Reflexive reproducing kernel Banach space**.

The main tool here is the natural bilinear form $\langle f, \phi \rangle_{\mathscr{B}} := \langle f, \phi \rangle_{\mathscr{B} \times \mathscr{B}'} = \phi(f)$, which is continuous in both arguments. By the Hahn-Banach theorem, we also know that this form is non-degenerate. Furthermore, since $\mathscr{B} = (\mathscr{B}')'$ we automatically obtain that $\mathscr{B}'$ is also a reflexive RKBS because if we identify f with its evaluation $eval_f$, then $\langle f, \phi \rangle_{\mathscr{B}} = \phi(f) = f(\phi) = \langle \phi, f \rangle_{\mathscr{B}'}$ also defines a continuous nondegenerate bilinear form. For a given reflexive RKBS, we will assume that $\mathscr{B}'$ is already under the identification from the definition, so $\mathscr{B}'$ is another space of functions defined on the same set as $\mathscr{B}$.

Since the second space and its norm are fixed in this case, the existence and uniqueness of the kernel function follows from that, just as in the RKHS case.

Theorem 2.1.1. Let $\mathscr{B}$ be a reflexive RKBS defined on X. Then:

- There exists an unique kernel function $k : X \times X \to \mathbb{C}$ such that

  $$k(x, \cdot) \in \mathscr{B}', k(\cdot, y) \in \mathscr{B},$$

  and k has the reproducing property on both $\mathscr{B}$ and its dual, in other words $f(x) = \langle f, k(x, \cdot) \rangle_{\mathscr{B}} \forall f \in \mathscr{B}$ and $g(y) = \langle k(\cdot, y), g \rangle_{\mathscr{B}'} \forall g \in \mathscr{B}'$.

- Moreover, the subspaces Span $\{k(\cdot, x) | x \in X\} \subset \mathscr{B}$ and Span $\{k(y, \cdot) | y \in X\} \subset \mathscr{B}'$ are dense.

[1] In [45] the authors asked for functions in both $\mathscr{B}$ and its dual to have the same domain but as pointed out in [16] such a requirement is unnecessary. Despite that, most worked examples we will see are of this kind so we will assume they both share the same domain unless otherwise noted.

Proof. Note that evaluation functionals for both spaces lie in their respective dual space, that is to say, if $\delta_x \in \mathscr{B}'$ is an evaluation functional for $\mathscr{B}$ and $\delta^y \in \mathscr{B}'' = \mathscr{B}$ is one for $\mathscr{B}$, so they really are function of two variables, one on the Banach space where they act on, and the other elements from X. So the natural way to define the kernel function is by using the natural bilinear form between a space and its dual:

$$k(x,y) := \langle \delta^x, \delta_y \rangle_{\mathscr{B}}.$$

Going by the observation above we can change notation to $\delta_y = g_y$ and that way our kernel is written like

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

We only show the reproducing property for $\mathscr{B}$, because the argument for its dual follows the same reasoning. Let $f \in \mathscr{B}$ be a function, then

$$f(x) = \delta_x(f), \ \forall x \in X,$$

but since the evaluation functionals are continuous and the dual is isometric to a Banach space of functions, then the evaluation functional δ_x can be though of as a function $= g(\cdot)$ with domain X, then

$$f(x) = \delta_x(f) = \langle f, \delta_x \rangle = \langle f, g_x \rangle = \langle f, k(x,\cdot) \rangle.$$

That show that k has the reproducing property. So if there were another function $l : X \times X \to \mathbb{C}$ with the reproducing property, then:

$$f(x) = \langle f, k(x,\cdot) \rangle = \langle f, l(x,\cdot) \rangle.$$

From where we can deduce that for every $f \in \mathscr{B}$

$$\langle f, k(x,\cdot) - l(x,\cdot) \rangle = 0.$$

So $k(x,\cdot) - l(x,\cdot)$ is the 0 functional, but since $\mathscr{B}'$ is a Banach space of functions, $k(x,y) - l(x,y) = 0$ for all $y \in X$. So k is unique. Now suppose that Span $\{k(\cdot,x) | x \in X\}$ is not dense in $\mathscr{B}$. Then by the Hahn-Banach Theorem we can choose a nontrivial continuous functional $h \in \mathscr{B}'$ such that $h(k(\cdot,x)) = 0$ for all $x \in X$. But since h is also a function of X, and by the reproducing property, that means that $h(x) = \langle k(\cdot,x), h \rangle = h(k(\cdot,x)) = 0$, so $h = 0$, which contradicts why we chose it. And consequently we have $\overline{\text{Span}\,\{k(\cdot,x) | x \in X\}} = \mathscr{B}$. □

If Construction 1.2.1.1 is done with a reflexive Banach space and its dual, the result falls within the class of reflexive RKBS's.

Theorem 2.1.2. Let W be a reflexive Banach with continuous dual space W'. Then the space $\mathscr{B}_1$ constructed with W and W' as in construction 1.2.1.1 is a reflexive RKBS with reproducing kernel $k(x,y) = \langle \Phi_1(x), \Phi_2(y) \rangle$.

Proof. We already know that $\mathscr{B}_1$ is a Banach space of functions, if its dual space is isometric to $\mathscr{B}_2$ then we are done. But this is exactly the result from Theorem 1.2.2. □

On the other hand, one kernel function can act as a reproducing kernel for multiple Banach spaces.

Example. [45] Let $p, q \in (1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$, $X = \mathbb{R}$, $\mathbb{I} = [-\frac{1}{2}, \frac{1}{2}]$ and fix the spaces $W_1 = L^p(\mathbb{I})$, $W_2 = L^q(\mathbb{I})$ and the bilinear map $\langle f, g \rangle_W = \int_{\mathbb{I}} fg$. Consider the feature maps $\Phi_1(x)(\cdot) := \exp^{-2\pi i x \cdot} \in L^p(\mathbb{I})$ and $\Phi_2(x)(\cdot) := \exp^{2\pi i x \cdot} \in L^q(\mathbb{I})$. For a function $f \in L^1$ denote its Fourier Transform by [41] [32]

$$\hat{f}(x) := (2\pi)^{\frac{-1}{2}} \int_{-\infty}^{\infty} f(t) e^{-itx} dx,$$

and its inverse transform by

$$\check{f}(t) := (2\pi)^{\frac{-1}{2}} \int_{-\infty}^{\infty} f(t) e^{2\pi i t x} dt.$$

The linear combinations of functions $\Phi_1(x)$ form a self-adjoint algebra of $C(\mathbb{I})$ which separates points, so by the Stone-Weierstrass theorem, it is dense in $C(\mathbb{I})$, and since $C(\mathbb{I})$ is dense in L^p then the density condition is fulfilled. By the previous theorem, the space

$$\mathscr{B}_1 = \{f(x) = \check{h}(x) = \langle h, e^{2\pi i x \cdot} \rangle \in \mathcal{C}(\mathbb{R}) : h \in L^p(\mathbb{I})\}$$

is a reflexive RKBS with dual

$$\mathscr{B}_2 = \{g(x) = \hat{j}(x) = \langle e^{-2\pi i x \cdot}, j \rangle \in \mathcal{C}(\mathbb{R}) : j \in L^q(\mathbb{I})\}.$$

And its reproducing kernel is

$$k(x, y) = \langle e^{-2\pi i x \cdot}, e^{2\pi i y \cdot} \rangle = \int_{\mathbb{I}} e^{-2\pi i x t} e^{2\pi i y t} dt = \frac{sin\pi(x - y)}{\pi(x - y)}.$$

Given two different exponents, the resulting Banach spaces are not isometric but share the same reproducing kernel nevertheless.

If the set X is finite, then any non-zero function on $X \times X$ can be a reproducing kernel for a finite-dimensional Banach space.

Theorem 2.1.3. Let X be a finite set and $k : X \times X \to \mathbb{C}$ such that it is not the zero function. Then there exists a RKBS $\mathscr{B}$ such that k is its reproducing kernel.

Proof. Let $X = \{x_1, \ldots, x_n\}$ and $k : X \times X \to \mathbb{C}$ be as above. Then the set $S := \{k(\cdot, x_1), \ldots, k(\cdot, x_n)\}$ is a set of functions defined on X, which generates a linear subspace of $\mathbb{C}^X$. We take $\mathscr{B} = \text{Span}\,(S)$, which is isomorphic to a $\mathbb{C}^m$ for some $m \leq n$, and choose a $p \in [1, \infty]$ to endow $\mathscr{B}$ with the p-norm of $\mathbb{C}^m$. This is a reflexive Banach space, moreover its dual $\mathscr{B}'$ is itself and the evaluation functionals are also continuous. We define a bilinear for $\mathscr{B}$ by selecting a basis $B = \{k(\cdot, x_1), \ldots, k(\cdot, x_m)\} \subset S$: Given $k(\cdot, x_{i_1}), k(\cdot, x_{i_2})$ we define

$$\langle k(\cdot, x_{i_1}), k(\cdot, x_{i_2}) \rangle_{\mathscr{B} \times \mathscr{B}'} = k(x_2, x_1),$$

and extend by linearity to $\mathscr{B}$. By the linear independence B this is a non-degenerate bilinear form which is also continuous. Moreover, for an arbitrary $f(\cdot) = \sum_{j=0}^{l} \alpha_j k(\cdot, x_j)$ we have the reproducing property:

$$\langle f, k(\cdot, y) \rangle_{\mathscr{B} \times \mathscr{B}'} = \sum_{j=0}^{l} \alpha_j \langle k(\cdot, x_j), k(\cdot, y) \rangle_{\mathscr{B} \times \mathscr{B}'} = \sum_{j=0}^{l} \alpha_j k(y, x_j) = f(y).$$

□

Summary

Let B be a reflexive Banach space and $\Phi : X \to B$, $\Phi^* : X \to B^*$ two feature maps such that their images have a dense span. Then an adjoint pair of RKBS can be constructed as in 1.2.1.1 by setting

$$W_1 := B^*, \; W_2 := B,$$

and

$$\langle u, v \rangle_{W_1 \times W_2} := \langle v, u \rangle_B = u(v).$$

This choice yields the adjoint pair

$$\mathscr{B}_1 = \{\langle \Phi^*(\cdot), v \rangle | v \in B\} \simeq B,$$

$$\mathscr{B}_2 = \{\langle u, \Phi(\cdot) \rangle | u \in B^*\} \simeq B^*.$$

And the induced two-sided reproducing kernel is

$$k(x, y) := \langle \Phi^*(x), \Phi(y) \rangle_{W_1 \times W_2} = \Phi^*(x)(\Phi(y)).$$

2.1.1 RKBS of slowly increasing functions with measures induced by positive definite functions.

The following approach to producing an adjoint pair of RKBS was for the purpose of developing adequate spaces to use for computations for machine learning [16]. In this subsection, the non-empty sets X and Y will be subsets of Euclidean $\mathbb{R}^d$ spaces.

First we define what is a positive definite function.

Definition 2.1.2. A continuous even function $\Psi : X \subset \mathbb{R}^d \longrightarrow \mathbb{R}$ will be called **positive definite** if for every finite subset of pairwise distinct points $\{x_1, \cdots, x_n\} \subset X$ and for every choice of scalars $\{\alpha_1, \cdots, \alpha_n\}$ we have that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \overline{\alpha_j} \Psi(x_i - x_j) > 0.$$

For this section we require the use of the distributional Fourier transform, properties of which are briefly discussed in the appendix. One of the reasons to use it is because it is useful for the next two results which characterize positive definite functions, one shows how to get one from a finite Borel measure and the next one how to tell if a function is positive definite [41].

Theorem 2.1.4. A continuous function $\Psi : \mathbb{R}^d \longrightarrow \mathbb{R}$ is positive definite if and only if it is bounded and its Fourier transform is nonnegative and nonvanishing.

To ensure that the Fourier transform works as an isometry of Banach spaces, we will restrict the kind of functions we work with.

Definition 2.1.3. The space of **Slowly increasing functions** $(\mathcal{SI})$ is the set of functions f such that there exists a constant $m \in \mathbb{N}_0$ such that $f(x) = \mathcal{O}(\|x\|_2^m)$ for $\|x\|_2 \longrightarrow \infty$. In other words, there exists $c, M > 0$ such that

$$\frac{|f(x)|}{\|x\|_2{}^m} < c$$

if $\|x\|_2 > M$.

Let p and q be conjugate exponents, i.e., $\frac{1}{p} + \frac{1}{q} = 1$, and suppose that $\Psi \in C(\mathbb{R}^d) \bigcap L^1(\mathbb{R}^d)$ is positive definite. We define the spaces which we will work with by:

$$\mathscr{B}_\Psi^p(\mathbb{R}^d) := \{f \in C(\mathbb{R}^d) \bigcap \mathcal{SI} : \frac{\hat{f}}{\hat{\Psi}^{\frac{1}{q}}} \in L^q(\mathbb{R}^d),\}$$

with norm defined by

$$\|f\|^q_{\mathscr{B}_\Psi^p(\mathbb{R}^d)} := \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \frac{|\hat{f}|^q}{\hat{\Psi}} dx.$$

Since $\hat{\Psi}$ is nonnegative and nonvanishing, $\frac{1}{\hat{\Psi}}$ defines a positive measure μ by integrating with respect to Lebesgue measure:

$$\mu(A) := \int_{\mathbb{R}^d} \frac{dx}{\hat{\Psi}(x)}.$$

What comes next is showing that the spaces defined above are isometrically isomorphic to the spaces $L^q(\mathbb{R}^d, \mu)$ by using the Fourier transform.

Theorem 2.1.5. Let Ψ be a positive definite function such that $\Psi^{min(p,q)-1} \in L^1(\mathbb{R}^d)$. Then the space $\mathscr{B}^p_\Psi(\mathbb{R}^d)$ is isometrically isomorphic to $L^q(\mathbb{R}^d, \mu)$ and the isomorphism is the distributional Fourier transform (See appendix). Moreover, its dual $\mathscr{B}^p_\Psi(\mathbb{R}^d)^*$ is isometrically isomorphic to $\mathscr{B}^q_\Psi(\mathbb{R}^d)$

Proof. From the definition of $\mathscr{B}^p_\Psi(\mathbb{R}^d)$ we see that the Fourier transform defines a monomorphism to $L^q(\mathbb{R}^d)$. If $f \in \mathscr{B}^p_\Psi(\mathbb{R}^d)$ then we have that its norm is:

$$\|f\|^q_{\mathscr{B}^p_\Psi(\mathbb{R}^d)} = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \frac{|\hat{f}(x)|^q}{\hat{\Psi}(x)} dx. = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} |\hat{f}(x)|^q d\mu(x) = \|\hat{f}\|^q_{L^q(\mathbb{R}^d,\mu)}.$$

So $\hat{}$ is an isometric isomorphism, the only thing left to show is the surjectivity. Given an element $h \in L^q(\mathbb{R}^d, \mu)$, its inverse Fourier transform is well defined since $h \in L^1(\mathbb{R}^d)$, because

$$\|h\|_{L^1(\mathbb{R}^d)} = \int_{\mathbb{R}^d} |h(x)| dx = \int_{\mathbb{R}^d} \frac{|h(x)|}{\hat{\Psi}(x)^{\frac{1}{q}}} \hat{\Psi}(x)^{\frac{1}{q}}(x) dx \leq (\int_{\mathbb{R}^d} \frac{|h(x)|^q}{\hat{\Psi}(x)})^{\frac{1}{q}} (\int_{\mathbb{R}^d} \hat{\Psi}(x)^{\frac{p}{q}})^{\frac{1}{p}} > \infty.$$

Since $\check{h} \in C(\mathbb{R}^d) \bigcup \mathcal{SI} \subset Ł^1_{loc}(\mathbb{R}^m) \bigcup \mathcal{S} \subset \mathcal{S}'$, we can take its distributional Fourier transform to deduce that

$$\hat{\check{h}} = h,$$

and since $\check{h} \in \mathscr{B}^p_\Psi(\mathbb{R}^d)$ we conclude that $\hat{}$ is surjective. We can prove the equivalent statement for $\mathscr{B}^q_\Psi(\mathbb{R}^d)$ because $\hat{\Psi}^{\frac{q}{p}} = \hat{\Psi}^{q-1} \in L^1(\mathbb{R}^d)$. □

Thus we conclude that $\mathscr{B}^q_\Psi(\mathbb{R}^d)$ is isometric to the dual space of $\mathscr{B}^p_\Psi(\mathbb{R}^d)$ by means of these isometries.

Corollary 2.1.5.1. $\mathscr{B}^q_\Psi(\mathbb{R}^d)$ is isometrically isomorphic to $\mathscr{B}^p_\Psi(\mathbb{R}^d)^*$

Now we show that the positive definite function works as a reproducing kernel for these spaces with feature maps $\Phi_i(x)(\cdot) := \Psi(\cdot + x)$.

Theorem 2.1.6. Let Ψ be a positive definite function such that $\Psi^{\min p,q-1} \in L^1(\mathbb{R}^d)$, then the spaces $\mathscr{B}^p_\Psi(\mathbb{R}^d)$ and $\mathscr{B}^q_\Psi(\mathbb{R}^d)$ form an adjoint pair of RKBS with with respect to the bilinear form

$$\langle f, g \rangle_{\mathscr{B}^p_\Psi \times \mathscr{B}^q_\Psi} := \int_{\mathbb{R}^d} \hat{f}(x)\hat{g}(x) d\mu(x)$$

and the feature maps $\Phi_1(y)(\cdot) := \Psi(\cdot + y)$, $\Phi_1(y)(\cdot) := \Psi(y + \cdot)$ induce the reproducing kernel

$$k(x, y) := \Psi(x + y).$$

Proof. Due to the isometry with $L^p(\mathbb{R}^d, \mu)$, we know that the bilinear form above has what we require in the definition, it suffices to show that the feature maps' images fall in the respective spaces and the reproducing property of $\Psi(x+y)$. Since $\Phi_1(x) = \Psi(\cdot + x)$, from the properties of the Fourier transform we get that $\widehat{\Phi_1(x)}(y) = \hat{\Psi}(y)e^{i\langle x,y\rangle_{\mathbb{R}^d}}$. Therefore

$$\|\Phi_1(x)\|_{L^q(\mathbb{R}^d,\mu)} = \int_{\mathbb{R}^d} \frac{\widehat{\Phi_1(x)}^q(y)}{\hat{\Psi}(y)} = \int_{\mathbb{R}^d} \frac{\hat{\Psi}^q(y)}{\hat{\Psi}(y)} = \int_{\mathbb{R}^d} \hat{\Psi}^{q-1}(y) < \infty.$$

So $\Phi_1(x) \in \mathscr{B}^p_\Psi$ and similarly $\Phi_2(y) \in \mathscr{B}^q_\Psi$. To see that $\Psi(x + y)$ has the reproducing property, let $f \in \mathscr{B}^p_\Psi$ be a function, then:

$$\langle f, \Phi_2(x) \rangle_{\mathscr{B}^p_\Psi \times \mathscr{B}^q_\Psi} = \int_{\mathbb{R}^d} \hat{f}(t)\widehat{\Phi_2(x)}(t) d\mu(t)$$

$$= \int_{\mathbb{R}^d} \frac{\hat{f}(t)\hat{\Psi}(t)e^{i\langle x,t\rangle_{\mathbb{R}^d}}}{\hat{\Psi}(t)}dt = \int_{\mathbb{R}^d} \hat{f}(t)e^{i\langle x,t\rangle_{\mathbb{R}^d}}dt = f(x).$$

Following the same steps, we have that

$$g(x) = \langle k(x,\cdot), g\rangle_{\mathscr{B}^p_\Psi \times \mathscr{B}^q_\Psi},$$

therefore we have that they form an adjoint pair of RKBS.

Since $\Phi_1(x) = k(x,\cdot) \in \mathscr{B}^p_\Psi$ and $k(\cdot,y) = \Phi_2(y)$ then $k(x,y) = \langle k(x,\cdot), \Phi_2(y)\rangle = \langle \Phi_2(x), \Phi_2(y)\rangle$ is the reproducing kernel.

□

In the same way we can construct spaces of functions defined on non-empty subsets $\Omega \subset \mathbb{R}^d$. We consider the subspace $\mathtt{N}_0$ of functions in $\mathscr{B}^p_\Psi(\mathbb{R}^d)$ which vanish on Ω. Since convergent Cauchy sequences also converge pointwise, we know that this is a closed subspace of $\mathscr{B}^p_\Psi(\mathbb{R}^d)$. Moreover, by considering the quotient space $\mathscr{B}^p_\Psi(\mathbb{R}^d)/\mathtt{N}_0$ we can show that the space

$$\mathscr{B}^p_\Psi(\Omega) = \{f : \text{ there exists } F \in \mathscr{B}^p_\Psi(\mathbb{R}^d) \text{ such that } F|_\Omega = f\}$$

is a Banach space. Using the isomorphism between the dual space of $\mathscr{B}^p_\Psi(\mathbb{R}^d)/\mathtt{N}_0$ and $\mathtt{N}_0^\perp$, we will show that $k(x,y)|_{\mathbb{R}^d\times\Omega}$ is a reproducing kernel for $\mathscr{B}^p_\Psi(\Omega)$. Since the reproducing kernel k worked for functions defined on $\mathbb{R}^d$, this property is kept while restricting it to the domain Ω. We just need to check that it is well defined on the equivalency classes. We first define what it means for a space to be uniformly convex.

Definition 2.1.4. [32] A Banach space B is **uniformly convex** if for every two sequences $\{x_n\}, \{y_n\} \subset B$ such that $\|x_n\|, \|y_n\| \leq 1$ and $\|x_n + y_n\| \longrightarrow 2$, we have that $\lim_{n\longrightarrow\infty}\|x_n - y_n\| = 0$.

Example. If μ is a positive measure defined on the measurable space Ω, then the space $L_p(\Omega,\mu)$ is uniformly convex [12].

Theorem 2.1.7. $\mathscr{B}^p_\Psi(\Omega)$ is a RKBS with reproducing kernel $k(x,y)|_{\mathbb{R}^d\times\Omega}$.

Proof. We will show that the evaluation functionals are well defined by using the kernel function. For this, we need to see first that they can be regarded as an element of $\mathscr{B}^p_\Psi(\Omega)'$. Let $y \in \Omega$ and $f \in \mathtt{N}_0$, from the reproducing property of k we know that

$$0 = f(y) = \langle f, k(\cdot,y)\rangle_{\mathscr{B}^p_\Psi(\mathbb{R}^d)},$$

therefore $k(\cdot,y) \in \mathtt{N}_0^\perp$ for every $y \in \Omega$. Since for every function $f \in \mathscr{B}^p_\Psi(\Omega)$ the set $\{F \in \mathscr{B}^p_\Psi(\mathbb{R}^d) \mid F|_\Omega = f\} \neq \emptyset$. Moreover, it is convex, closed subset of $\mathscr{B}^p_\Psi(\mathbb{R}^d)$ which is uniformly convex and reflexive, therefore there exists a unique $F \in f + \mathtt{N}_0$ such that $\|f\|_{\mathscr{B}^p_\Psi(\Omega)} = \|F\|_{\mathscr{B}^p_\Psi(\mathbb{R}^d)}$ by 3.2.2. We use the isomorphism discussed above and the existence of F for every $f \in \mathscr{B}^p_\Psi(\Omega)$ to define a function T by $Tf := F$. We propose the next bilinear form [15] [20] between $\mathscr{B}^p_\Psi(\Omega)$ and the subspace of $\mathtt{N}_0^\perp$ generated by the functions $k(\cdot,y)$:

$$\langle f, k(\cdot,z)\rangle_{\mathscr{B}^p_\Psi(\Omega)} := \langle Tf, k(\cdot,z)\rangle_{\mathscr{B}^p_\Psi(\mathbb{R}^d)} = Tf(z) = f(z).$$

With this we proved that evaluation functionals are continuous and that the pair $(\mathscr{B}^p_\Psi(\Omega), \overline{\text{Span}\,\{k(\cdot,x)\}_{x\in X}}\}$ form an adjoint pair of RKBS with reproducing kernel $k|_{\mathbb{R}^d\times\Omega}$. □

2.2 Spaces with semi-inner products defined by Gateaûx differentials.

The authors in [45] started exploring RKBS by subtracting the conjugate symmetry property from the inner product, to obtain a semi-inner product.

Definition 2.2.1. Let $\mathscr{B}$ be a Banach space. A function $[\cdot,\cdot] : \mathscr{B} \times \mathscr{B} \to \mathbb{C}$ is called a **semi-inner product** (SIP) if it has the next properties:

- $[x,x] \geq 0 \; \forall x \in \mathscr{B}$ and $[x,x] = 0 \iff x = 0$
- $[\lambda x, y] = \lambda [x,y]$ and $[x, \lambda y] = \overline{\lambda}[x,y] \; \forall x,y \in \mathscr{B}, \lambda \in \mathbb{C}$.
- $[x+z,y] = [x,y] + [z,y] \; \forall x,y,z \in \mathscr{B}$.
- $|[x,y]| \leq [x,x]^2 [y,y]^2$

In general a semi-inner product is not necessarily conjugate symmetric. In fact, being conjugate symmetric is equivalent to being an inner product [16]. The first and second property implies that a norm can be defined by a semi-inner product like with the inner product case [45]. Moreover, given a Banach space B, one can always choose a semi-inner product which induces an equivalent norm by mapping each $x \in B$ to a x^*, where $x^* \in B'$ is a functional such that $x^*(x) = \|x\|^2$, which always exists by the Hahn-Banach theorem. A semi-inner product is therefore defined by $[x,y] := y^*(x)$ [28] [19]. However, such a selection may not be unique since the set $J_x := \{\phi \in B' : \phi(x) = \|x\|^2\}$ may contain more than one point for some x. For a fixed choice of a mapping $x \mapsto x^*$, we will call such a function a dual mapping, duality mapping or duality map.

From Construction 1.2.1.1, we see that the resulting spaces' properties and behavior are dependent on the bilinear form between spaces W_1 and W_2. In other words, if we find a way to get a reproducing kernel starting from a semi-inner product, the way we chose the functionals for the semi-inner product will affect the reproducing kernel induced by it. If what we seek is to make it so there is a unique kernel function related to the semi-inner product, we must give sufficient conditions to make this mapping unique.

The next definitions are about the geometry induced by the norm.

Definition 2.2.2. [23] A Banach space is called **smooth** if for every point on the unit sphere there is a unique support hyperplane.

J_x being a one-point set is equivalent to the space being smooth, and the latter is also equivalent to the next property [23].

Definition 2.2.3. A normed vector space B is called **Gâteaux differentiable** if for every fixed pair $x, y \in B \setminus \{0\}$ and for $t \in \mathbb{R}$ the next limit exists:

$$\lim_{t \longrightarrow 0} \frac{\|x+ty\|_B - \|x\|_B}{t}. \tag{2.1}$$

A space will be called uniformly **Fréchet differentiable** if the limit above is uniform on $\{(x,y) \in B \times B : \|x\|_B = \|y\|_B = 1\}$.

When restricted to finite dimensional spaces, this just says that for the unit ball to be smooth at a point there must be a unique tangential plane that touches the sphere at that point [29].

The next theorem says that Gateaux differentiable spaces are spaces where there is a unique semi-inner product which induces the norm, and it is given in terms of the limit (2.1) [45] [28].

Theorem 2.2.1. Let B be a Gateaux differentiable Banach space, then the semi-inner product which induces its norm is unique and it is given by:

$$[y,x] := \|x\| \left(\lim_{t \longrightarrow 0} \frac{\|x+ty\|_B - \|x\|_B}{t} + i \lim_{t \longrightarrow 0} \frac{\|ix+ty\|_B - \|x\|_B}{t} \right).$$

With these we can show that if we define a reproducing kernel using the unique semi-inner product, this will also be unique. Up until here the semi-inner product has not been explicitly connected with the geometry. To connect them we will work towards proving a version of the Riesz Representation Theorem.

On Hilbert spaces, when exposing an element to represent a given linear functional, an essential concept is that of orthogonality. For the given functional, we choose a vector orthogonal to its kernel. Now since we lack an inner product, the following result shows that there is a way to tell when a vector is orthogonal without resorting to an inner product. For this we next show that a semi-inner product has a similar property.

Theorem 2.2.2. Let H be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_H$. Then two vectors $x, y \in H$ are orthogonal if and only if $\|x + \lambda y\|_H \geq \|x\|_H$, for all scalars $\lambda \in \mathbb{C}$.

To show the analogous result for a semi-inner product space, we need first a technical lemma.

Lemma 2.2.3. [19] A Banach space B with a SIP is Gâteaux differentiable if and only if for any pair $x, y \in B$ we have that

$$\lim_{\lambda \to 0} [x, y + \lambda x] = [x, y].$$

Since we work with a semi-inner product, the orthogonality relation may not be symmetric.

Theorem 2.2.4. [19] Let B be a S.I.P. space such that $(\forall x, y \in B)$

$$\lim_{\lambda \to 0} [x, y + \lambda x] = [x, y].$$

Then $[y, x] = 0$ if and only if $\|x + \lambda y\| \geq \|x\|$ for all $\lambda \in \mathbb{C}$.

Proof. Let $x, y \in B$ be such that $[y, x] = 0$. Then

$$\|x\|^2 = |[x,x]| = [x,x] + [\lambda y, x] = [x + \lambda y, x] \leq \|x + \lambda y\| \|x\|.$$

Conversely, let $x, y \in B$ be such that $\|x + \lambda y\| \geq \|x\|$ for all $\lambda \in \mathbb{C}$. We know that

$$[y,x] = \|x\| (\lim_{t \longrightarrow 0} \frac{\|x+ty\|_B - \|x\|_B}{t} + i \lim_{t \longrightarrow 0} \frac{\|ix+ty\|_B - \|x\|_B}{t})$$

because of the preceding lemma. The limit

$$\mathcal{R}e([y,x]) = \|x\| (\lim_{t \longrightarrow 0} \frac{\|x+ty\|_B - \|x\|_B}{t})$$

must be the same whether we approximate from below or above 0, and since

$$\|x + ty\|_B - \|x\|_B \geq 0$$

for $t \in \mathbb{R}$, it can only be 0. Similarly, we have that $\mathcal{I}m([y,x]) = 0$, then it must be that $[y,x] = 0$. □

Finally, to show the Riesz' representation Theorem for S.I.P. spaces we also need ensure the existence of orthogonal vectors to proper subspaces. To this end we give the next enunciate a lemma followed by the theorem.

Lemma 2.2.5. [19] Let B be a Banach space which is uniformly convex. Then for every proper subspace V there exists a vector $u \in B$ such that for every $v \in V$:

$$[v,u] = 0.$$

Theorem 2.2.6. Let $\mathscr{B}$ be a reflexive, uniformly convex and uniformly Fréchet differentiable space, then for every continuous functional ϕ there exists a unique $x_\phi \in \mathscr{B}$ such that:

$$\phi(y) = [y, x_\phi],$$

in other words $\phi = x_\phi^*$ Moreover $\|\phi\| = \|x_\phi\|$.

The Riesz representation Theorem for S.I.P. spaces says that the duality map $*$ is a bijection from the Banach space to its dual space. And as in Hilbert spaces, we can use the representation of functionals to make the dual space a S.I.P. space.

Corollary 2.2.6.1. With the same hypothesis as Theorem 2.2.6, $\mathscr{B}^*$ is a S.I.P. space, with the semi-inner product:

$$[x^*, y^*] := [y, x].$$

Example. We have seen that the Banach spaces $L^p(X,\mu)$ for $p \in (1,\infty)$ are uniformly convex and uniformly Fréchet differentiable spaces [12]. Therefore there is a unique semi-inner product, and it is given by the formula [12] [2] :

$$[f,g] = \frac{1}{\|g\|_p^{p-2}} \int f\hat{g}|g|^{p-2}d\mu, \tag{2.2}$$

for $f, g \in L^p(X,\mu)$. It is well known that every bounded linear functional on $L^p(X,\mu)$ can be represented by a formula like above. The Riesz representation Theorem for S.I.P. says that every function in $L^q(X,\mu)$ are exactly of this form.

In summary, this section will exclusively deal with uniformly convex and uniformly Fréchet differentiable spaces. The uniform differentiability makes it so that there is a unique choice of semi-inner product which induces the norm. With this, we have an unambiguous condition that can be interpreted as orthogonality. The uniform convexity implies the reflexivity and it is the last piece we needed for a version of Riesz' theorem for S.I.P. spaces [12].

Definition 2.2.4. We say that a RKBS $\mathscr{B}$ of functions defined on a non-empty set X is a S.I.P. RKBS if it is uniformly convex and uniformly Fréchet differentiable.

Back with the Hilbert case, the kernel could be recovered using the inner product. RKBS have the same property, but we deal with the extra condition that the bilinear form is defined on two possibly different spaces. So we use the dual mapping to make sense of the bilinear form which will give rise to our kernel function.

The condition of uniform convexity implies that the spaces here are all reflexive [23].

Theorem 2.2.7. Let $\mathscr{B}$ be a S.I.P. RKBS of functions defined on a set X, and consider its reproducing kernel as seen in Theorem 2.1.1. Then there exists a unique function $G(\cdot,\cdot) : X \times X \longrightarrow \mathbb{C}$ such that $\{G(x,\cdot) : x \in X\} \subset \mathscr{B}$ and

$$f(x) = [f, G(x,\cdot)],\ \forall f \in \mathscr{B}.$$

And this function G is related to the reproducing kernel k by the next equation:

$$k(\cdot,x) = (G(x,\cdot))^*,$$

which in turn implies that

$$f^*(x) = [G(x,\cdot), f]$$

where $*$ denotes the dual mapping.

Proof. We know from the Riesz Representation Theorem for S.I.P. spaces that for each evaluation functional δ_x there exists an element $G_x \in \mathscr{B}$ such that

$$\delta_x(\cdot) = [\cdot, G_x^*].$$

Since $\mathscr{B}$ is a functions space, we set $G(x,y) := G_x(y)$. Theorem 2.1.1 says that this space has a reproducing kernel $k(\cdot,\cdot)$, then for every $f \in \mathscr{B}$ we have

$$f(x) = \langle f, k(\cdot,x)\rangle_{\mathscr{B}}$$

but from the definition of G_x we get that

$$\langle f, k(\cdot,x)\rangle = [f, G(x,\cdot)^*].$$

Because of the uniqueness of the duality map, we have that $G(x,\cdot)^* = k(\cdot,x) \in \mathscr{B}$. Since k is a left and right-sided reproducing kernel, he have that for an element $f^* \in \mathscr{B}'$:

$$f^*(x) = \langle k(x,\cdot), f^*\rangle = [k(x,\cdot), f].$$

The last equality comes from the property of the unique duality map. □

Just as before, there is also the a way to construct a S.I.P. RKBS from a feature map.

Theorem 2.2.8. Let X be a non-empty set, W be a uniformly convex and uniformly Fréchet differentiable with semi-inner product $[\cdot,\cdot]_W$, and $\Phi : X \longrightarrow W$ such that Span $(\Phi(X))$ is dense in W and Span $((\Phi(X))^*)$ is dense in W^*. Then the spaces defined in Construction 1.2.1.1 take the form

$$\begin{aligned} \mathscr{B}_1 &= \{[u, \Phi(\cdot)] : u \in \mathscr{B}\}, \\ \mathscr{B}_2 &= \{[\Phi(\cdot), v] : v \in \mathscr{B}\}. \end{aligned} \tag{2.3}$$

And each of them has the respective semi-inner product:

$$[[u,\Phi(\cdot)],[v,\Phi(\cdot)]]_{\mathscr{B}_1} := [u,v]_W, \quad [[\Phi(\cdot),v],[\Phi(\cdot),u]]_{\mathscr{B}_2} := [u,v]_W. \tag{2.4}$$

Moreover the function

$$\langle [u,\Phi(\cdot)], [\Phi(\cdot),v]\rangle_{\mathscr{B}_1\times\mathscr{B}_2} := [u,v]_W$$

is a bilinear form on $\mathscr{B}_1$ and $\mathscr{B}_2$ and the associated reproducing kernel is

$$G(x,y) := [\Phi(x), \Phi(y)]_W.$$

Remark. The notation Φ^* indicates the composition $\Phi \circ *$, where $*$ denotes the duality map. The bilinear mapping above is linear on its second argument despite having that $[\Phi(\cdot), v] + [\Phi(\cdot), w] \neq [\Phi(\cdot), v + w]$ in general. What has to be done to show the linearity is taking $z \in \mathscr{B}$ such that $[\Phi(\cdot), v] + [\Phi(\cdot), w] = [\Phi(\cdot), z]$.

Proof. For an element $\langle v, \Phi(\cdot)^* \rangle_{\mathscr{B}_1}$, by the Riesz representation Theorem for S.I.P. spaces, we have for every $x \in X$:

$$\langle v, \Phi(x)^* \rangle = [v, \Phi(x)].$$

This is also the case for $\mathscr{B}_2$, since the mapping $*$ is a bijection. It is clear that the first equation in 2.4 defines a semi-inner product. The linearity of the second equation may not be so clear, but from the remark, we know that given $v_1, v_2 \in W$ there exists a $v_3 \in \mathscr{B}$ such that $[\cdot, v_1] + [\cdot, v_2] = [\cdot, v_3]$. Then we have that

$$\begin{aligned}[[\Phi(\cdot), v_1] + [\Phi(\cdot), v_2], [\Phi(\cdot), u]] =& [[\Phi(\cdot), v_3], [\Phi(\cdot), u]] = [u, v_3] = [u, v_1] + [u, v_2] = \\ & [[\Phi(\cdot), v_1], [\Phi(\cdot), u]] + [[\Phi(\cdot), v_2], [\Phi(\cdot), u]]\end{aligned}$$

As for the last claim, we know that $k(\cdot, x) = (G(x, \cdot))^*$, then

$$k(x,y) = \langle \Phi(x), \Phi(y)^* \rangle = [\Phi(x), \Phi(y)] = G(x,y).$$

□

Summary

Let B be a uniformly convex and uniformly Fréchet differentiable Banach space and $\Phi : X \to B$ a feature map such that $\operatorname{Span} \Phi(X)$ is dense in B. Assume also the same for the mapping Φ^*, where * denotes the bijective duality map $* : B \to B^*$ as defined by 2.2.6. Then an adjoint pair of RKBS can be constructed as in 1.2.1.1 by setting

$$W_1 := B, \; W_2 := B^*,$$

and

$$\langle u, v* \rangle_{W_1 \times W_2} := \langle u, v* \rangle_B = [u, v]$$

This choice yields the adjoint pair

$$\begin{aligned}\mathscr{B}_1 &= \{[u, \Phi(\cdot)] : u \in \mathscr{B}\}, \\ \mathscr{B}_2 &= \{[\Phi(\cdot), v] : v \in \mathscr{B}\}.\end{aligned}$$

And the two-sided reproducing kernel is

$$k(x,y) := \langle \Phi(x), \Phi(y)^* \rangle_{W_1 \times W_2} = \Phi^*(x)(\Phi(y)) = [\Phi(x), \Phi(y)].$$

2.3 RKBS from Borel measures.

The previous section dealt with reflexive spaces. In [38], the authors constructed Banach spaces with ℓ^1 norm, which makes them non-reflexive. The purpose of such constructions was to show that they could be used for machine learning, like the space constructed by positive definite functions.

Based on these results, the authors of [37] generalized the construction by using finite Borel measures, which contain a copy of the space with ℓ^1 norm constructed. They made this generalization to find a bound for the error that came from using a RKBS with ℓ^1 norm in the machine learning algorithm from [38]. The general construction in [37] started with a locally compact space X and the associated space of functions which vanish at infinity $C_0(X)$. Then they consider its dual space $M(X)$ which consist of the signed Borel measures with finite variation defined on X [13]. The proof of the next result is similar to the verification that Construction 1.2.1.1 yields a RKBS.

Theorem 2.3.1. If $k : X \times X \longrightarrow \mathbb{C}$ is a function such that Span $(k(\cdot, x))$ is dense in $C_0(X)$, the function space

$$\mathscr{B} := \{f_\mu := \langle k(\cdot, x), \mu\rangle_{C_0(X)} = \int_X k(t, x)d\mu(t) \mid t \in X, \mu \in M(X)\}$$

with norm $\|f_\mu\|_{\mathscr{B}} := \|\mu\|_{M(X)}$ is a RKBS.

It can be seen that $C_0(X)$ is isometric to a subspace (possibly a proper one) of $\mathscr{B}$. Likewise, the space ℓ^1 can be regarded as a subspace of $M(X)$ by considering an element as a measure supported on a countable subset.

One of the problems [37] tried to address finding conditions a kernel function needs to satisfy to construct a RKBS with ℓ^1 norm. The issue is that a space with the mentioned norm would not be uniformly convex due to the lack of reflexivity. Their solution was adding an extra condition which the reproducing kernel must abide to, condition which is listed at the end of the next definition. This property will not be exploited in this section but will appear in the next chapter.

Definition 2.3.1. We say that a function $k : X \times X \longrightarrow \mathbb{C}$ is an admissible kernel for a RKBS space with ℓ^1 norm if it has the following properties:

- *A1*- For every finite subset $\{x_1, \cdots, x_n\} \subset X$ of pairwise distinct elements, the matrix $k[\mathbf{x}] := (k(x_i, x_j))_{i,j=1,\cdots,n}$ is non-singular.
- *A2*- There exists a positive constant M such that $|k(x, y)| \leq M$ for all $x, y \in X$
- *A3*- For any sequence of pairwise distinct points $x_j \in X$ and any $(c_j)_{j\in\mathbb{N}} \in \ell^1(\mathbb{N})$, the fact that
$$\sum_{j=1}^{\infty} c_j k(x, x_j) = 0$$
for all $x \in X$ implies that $c_j = 0$ for all j.
- *A4*- For any finite subset of pairwise distinct points $x_1, \cdots, x_{n+1} \subset X$ we define the column vector $K_{\mathbf{x}}(x_{n+1}) := (k(x_{n+1}, x_j))_{j=1,\cdots,n}$. Then
$$\|(K[\mathbf{x}])^{-1} K_{\mathbf{x}}(x_{n+1})\|_{\ell^1} \leq 1.$$

This approach starts from a space defined by a kernel function k as seen in Theorem 1.1.3, and shows that the conditions above are sufficient to show that the ℓ^1 norm makes it a RKBS with k as its reproducing kernel.

For this space to be a RKBS we need the evaluation functionals to be continuous. The next results will be shown for a dense subspace, which means they can be extended to their closure.

Theorem 2.3.2. Let $\mathscr{B}_0$ be the space Span $\{k(\cdot, x)\}_{x\in X}$ with norm $\|\sum c_i k(\cdot, x_i)\|_{\mathscr{B}_0} := \sum_i |c_i|$ where k is bounded by a positive constant M. Then the evaluation functionals are continuous.

Proof. We have that

$$|f(x)| = |\sum_{i=1}^{m} c_i k(x, x_i)| \leq |\sum_{i=1}^{m} |c_i| M = \|f\|_{\mathscr{B}_0} M.$$

□

Next we prove that for a space with this norm, conditioning its kernel to verify condition *A3* is equivalent to our requirement of functions having zero norm if they vanish everywhere.

Theorem 2.3.3. Let $\mathscr{B}_0$ be the space Span $\{k(\cdot, x)\}_{x\in X}$ with norm $\|\sum_i c_i k(\cdot, x_i)\|_{\mathscr{B}_0} := \sum_i |c_i|$ where k is bounded. Then the next conditions are equivalent:

- The norm satisfies condition *A3* in Definition 2.3.1.
- A Cauchy sequence converging pointwise to zero implies the sequence of norms also converge to zero.

Proof. Suppose that the norm satisfies condition *A3*, and consider a Cauchy sequence $\{f_n\}_{n\in\mathbb{N}}$. By construction, we have that for every $n \in \mathbb{N}$ there exists scalars $c_j^n \in \mathbb{C}$ and pairwise distinct $x_j \in X$ such that $f_n(x) = \sum_{j\in\mathbb{N}} c_j^n k(x, x_j)$ where only finitely many c_j^n are not zero. Since the sequence $f_n$is a Cauchy sequence, from the definition of the norm for functions in $\mathscr{B}_0$, for a fixed j we can define another Cauchy sequence $\{c_j^n\}_n$. These sequences are also Cauchy sequences in $\mathbb{C}$ because

$$|c_j^n - c_j^m| \leq \sum_{i\in\mathbb{N}} |c_i^n - c_i^m| = \|f_n - f_m\|_{\mathscr{B}_0} \xrightarrow[n,m\to\infty]{} 0$$

For every $j \in \mathbb{N}$ let $c_j = \lim_{n\to\infty} c_j^n$. Since k is bounded, we can define for every $x \in X$ the pointwise limit of the Cauchy sequence as

$$f(x) := \sum_n c_j k(x, x_j).$$

This is the pointiwise limit of the Cauchy sequence, this follows from

$$|f_n(x) - f(x)| = |\sum_{i\in\mathbb{N}} (c_j^n - c_j) k(x_j, x)| \leq M\|\mathbf{c}_n - \mathbf{c}\|_{\ell^1} \to 0.$$

Thus we can conclude that $f(x) = 0$ everywhere, since the evaluation functionals are continuous. From condition *A3* we know that this means $c_n = 0$ for every $n \in \mathbb{N}$, therefore

$$\lim_{n\to\infty} \|f_n\|_{\mathscr{B}_0} = \lim_{n\to\infty} \|c^n\|_{\mathscr{B}_0} = \|c\|_{\ell^1(\mathbb{N})} = 0.$$

Conversely, suppose that for a Cauchy sequence, converging to zero pointwise is equivalent to convergence to zero in norm and let

$$f(x) = \sum_{n=1}^{\infty} a_n k(x, x_n) = 0$$

for every $x \in X$, where $= \{c_n\}_{n\in\mathbb{N}} \in \ell^1(\mathbb{N})$. We define the sequence f_j by

$$f_j(x) := \sum_{n=1}^{j} a_n k(x, x_n).$$

This is a Cauchy sequence because $\in \ell^1(\mathbb{N})$ and by fixing an arbitrary x, we get that

$$\lim_{j\to\infty} f_j(x) = 0.$$

By hypothesis, this means that

$$\lim_{j\to\infty} \|f_j(x)\|_{\mathscr{B}_0} = \lim_{j\to\infty} \sum_{n=1}^{j} |c_n| = \sum_{n=1}^{\infty} |c_n| = 0.$$

Thus we conclude that the kernel has property *A3*. □

Next we will consider the space $\mathscr{B}^{\#}$ spanned by the linear combinations $\sum a_i k(x_i, \cdot)$. To define the norm we will consider a bilinear form that takes $f(\cdot) = \sum_i c_i k(x_i, \cdot) \in \mathscr{B}_0$ and $g(\cdot) = \sum_j d_j k(\cdot, y_j) \in \mathscr{B}^{\#}$ to

$$\langle f, g\rangle := \sum_i \sum_j c_i d_j k(x_i, y_j). \tag{2.5}$$

We consider them as linear functionals on $\mathscr{B}_0$ and bestow it with the norm defined on the dual space:

$$\|\sum_j d_j k(\cdot, x_j)\|_{\mathscr{B}^{\#}} := \sup_{f\in\mathscr{B}_0} \frac{|\langle f, \sum_j d_j k(\cdot, x_j)\rangle|}{\|f\|_{\mathscr{B}_0}}.$$

This space is a subspace of the space of bounded functions, since k is bounded on X. The supremum norm inherited from the bounded functions coincides with the norm here defined.

Theorem 2.3.4. For a function $h \in \mathscr{B}^{\#}$ we have that $\|h\|_{\mathscr{B}_0} = \|h\|_{\infty}$.

Proof. From the definition of the norms $\|\ \|_{\mathscr{B}_0}$ and $\|\cdot\|_{\mathscr{B}^{\#}}$ we see that $\|h\|_{\mathscr{B}^{\#}} \leq \|h\|_{\infty}$ because if $f(\cdot) = \sum_i a_i k(x_i, \cdot)$ then:

$$|\langle f, h\rangle| = |\sum_i a_i h(x_i)| \leq \sum_i |a_i||h(x_i)| \leq \|f\|_{\mathscr{B}_0}\|h\|_{\infty}.$$

For the other inequality, we have that for an arbitrary point x_0, the function $k(x_0, \cdot)$ has norm 1, and from Equation 2.5, we know its action on an arbitrary function $h \in \mathscr{B}^{\#}$ is an evaluation on x_0, therefore:

$$\|h\|_{\mathscr{B}^{\#}} = \sup_{f\neq 0} \frac{\langle f, h\rangle}{\|f\|_{\mathscr{B}_0}} \geq \frac{|\langle k(x_0, \cdot), h\rangle|}{\|k(x_0, \cdot)\|_{\mathscr{B}_0}} = |h(x_0)|.$$

□

We can prove that evaluation functionals are continuous by adapting the proof for $\mathscr{B}_0$. Due to condition *A3*, we have that $\overline{\mathscr{B}_0}$ and $\overline{\mathscr{B}^\#}$ form an adjoint pair of RKBS. The reproducing kernel clearly is $k(x,y)$ and this is proved by showing it has the reproducing property for the dense subsets Span $\{k(x,\cdot)\}_{x\in X}$ and Span $\{k(\cdot,y)\}_{y\in X}$.

Theorem 2.3.5. The spaces $\overline{\mathscr{B}_0}$ and $\overline{\mathscr{B}^\#}$, defined from a kernel function $k : X \times X \longrightarrow \mathbb{C}$ with properties listed in Definition 2.3.1, form an adjoint pair of RKBS with respect to the bilinear form defined in Equation 2.5 and the corresponding reproducing kernel is the function k.

We constructed here an adjoint pair where one of them was a space of continuous functions. The authors of [27] took this approach of taking subspaces of dual capable of separating points. This way, they proved that the space of continuous functions could be realized as a RKBS. Previous definitions of RKBS avoided this space on purpose by considering exclusively reflexive spaces [45].

Consider again the space $C[0,1]$, and the following families of functions:

$$|x-\bullet|, \quad e^{x\bullet}$$

with $x \in X = [0,1]$. We know that every function in $e^{x\bullet}$ is holomorphic as a complex function. Therefore, the bilinear form applied to any one of its members cannot be zero for every element in $[0,1]$, thus it spans a weak$-^*$ set. The span of the family $|x-\bullet|$ contains the piecewise linear functions, therefore it is also a dense subspace of $C[0,1]$. For the next results, we will label both either of them as η whenever it makes no difference considering one or the other.

In [27] they considered the subspace $\ell^1([0,1])$ of $M([0,1])$ as W_2. This space is seen as a subspace of the finite Borel measure, therefore they doted these two spaces with the integral as the bilinear form, which is non-degenerate.

Lemma 2.3.6. The bilinear form

$$\langle\cdot,\cdot\rangle : C([0,1]) \times D \longrightarrow \mathbb{C}$$

where $D = \ell^1([0,1])$ or $\ell^1(s_n)$, is non-degenerate, where $\{s_n\} \subset [0,1]$ is a dense subset of $[0,1]$.

Proof. It is enough to show that the lemma is true for $\ell^1(s_n)$, since this space can be regarded as a subspace of $\ell^1([0,1])$. Suppose that for a fixed $f \in C([0,1])$ we have that

$$\langle f, a\rangle = 0$$

for any given $a = \{a_{s_n}\}$. Consider the elements of a^j of $\ell^1(s_n)$, where

$$a^j_{s_n} := \delta_{j,n}.$$

From the hypothesis we have that $\langle f, a^j\rangle = 0$, but from the definition of the bilinear form we get that

$$\langle f, a\rangle = f(s_n) = 0.$$

Since s_n is dense in $[0,1]$, this implies that $f = 0$.

Conversely, for a fixed sequence $\{a_{s_n}\}$, what we want to prove comes directly from the choice of space, since we consider $\ell^1(s_n)$ as a subspace of the finite Borel

measures, which is the dual space of $C([0,1])$. Therefore we have proved that the bilinear form is non-degenerate. □

Consequently we can proceed with Construction 1.2.1.1 with these two spaces.

Theorem 2.3.7. Let $W_1 = \ell^1([0,1])$ and $W_2 = C([0,1])$ along with the bilinear form above and the following feature maps:

$$\Phi_1 : [0,1] \longrightarrow W_1,\ t \mapsto \{\delta_t\}$$

$$\Phi_2 : [0,1] \longrightarrow W_2,\ x \mapsto \eta(x)(\cdot).$$

Then the spaces

$$\mathscr{B}_1 := \{h_f(t) := \langle \Phi_1(t), f\rangle = f(t) \mid t \in [0,1],\ f \in W_2\} = C([0,1]),$$
$$\mathscr{B}_2 := \{g_a(x) := \langle a, \Phi_2(x)\rangle = \sum_{t \in supp(a)} a_t \Phi_2(x)(t) \mid a = \{a_s\}_{s\in[0,1]} \in W_1,\ x \in [0,1]\}$$

form an adjoint pair of RKBS with $\eta(\cdot)(\cdot) : [0,1]^2 \longrightarrow \mathbb{C}$ as its reproducing kernel.

If we choose $W_1 = \ell^1(s_n)$ instead, we recover the continuous functions defined on the dense subset, which we know are isometrically isomorphic to $C([0,1])$ because of Theorem 1.2.2.

Theorem 2.3.8. Let $W_1 = \ell^1(s_n)$, $W_2 = C([0,1])$, $\Phi_1 : [0,1] \longrightarrow W_1$, be as in the previous result, and define $\Phi_2 : \{s_n\} \longrightarrow C([0,1])$ by

$$\Phi_2(s_j)(t) := \eta(s_j)(t).$$

Then $\mathscr{B}_1$ and $\mathscr{B}_2$ form an adjoint pair of RKBS with the reproducing kernel $k(x, s_n) := \eta(s_n)(x)$.

With these last examples we show that the definition of RKBS can cover the case of continuous functions spaces, by using the Construction 1.2.1.1. And also we showed that it is enough to consider subspaces of the dual space, or spaces embedded in it, to make use of Construction 1.2.1.1. As we will see in the next section, if we consider the whole dual space we may run into a degenerate bilinear form, which would fall into the case stated in Remark 1.2.

Summary

Let X be a locally compact Hausdorff space and $\Phi : X \to C_0(X)$ a feature map such that the image has dense span. Let $\Phi^* : X \to M(X)$ such that its image's span is a total subspace, call the closure of its span B. Then an adjoint pair of RKBS can be constructed as in 1.2.1.1 by setting

$$W_1 := C_0(X),\ W_2 := B,$$

and

$$\langle u, v\rangle_{W_1 \times W_2} := \langle v, u\rangle_B = u(v).$$

This choice yields the adjoint pair

$$\mathscr{B}_1 = \{\langle \Phi^*(\cdot), v\rangle | v \in B\} \simeq B,$$

$$\mathscr{B}_2 = \{\langle u, \Phi(\cdot)\rangle | u \in B^*\} \simeq B^*.$$

And the two-sided reproducing kernel is

$$k(x,y) := \langle \Phi^*(x), \Phi(y)\rangle_{W_1 \times W_2} = \Phi^*(x)(\Phi(y)).$$

Let X be a locally compact Hausdorff space and $k : X \times X \longmapsto \mathbb{C}$ a continuous kernel which satisfies assumptions 2.3.1 and $C_0(X) = \overline{\text{Span}\,\{k(\cdot, x) : x \in X\}}$. Then we set $\Phi_1(x) := k(\cdot, x) \in C_0(X)$, $\Phi_2(y) := \delta_y \in \ell^1(X)$ and

$$W_1 := C_0(X),\ W_2 := \ell^1(X),$$

and

$$\langle f, \mu\rangle_{W_1 \times W_2} := \int_X f d\mu = \sum_{u(x)\neq 0} u(x) f(x).$$

This choice yields the following adjoint pair of RKBS

$$\mathscr{B}_1 := \{\langle \Phi_1(x), a\rangle_{W_1} = \sum_{s \in supp a} \Phi_1(x)(s)a(s) |\ a \in \ell^1(X), x \in X\},$$

$$\mathscr{B}_2 = \{\langle u, \Phi_2(y)\rangle_{W_1} = u(y) |\ u \in C_0(X), y \in X\} = C_0(X)$$

with bilinear form

$$\langle \langle \Phi_1(\cdot), a\rangle_{W_1}, \langle u, \Phi_2(y)\rangle_{W_1}\rangle := \sum_{t \in \text{supp}\, a} a(t)u(t).$$

The two-sided reproducing kernel coincides with the continuous kernel $k(x,y)$.

2.4 RKBS with p-norm

Previous sections can be divided between constructions of reflexive spaces and non-reflexive spaces. The way reflexive RKBS were constructed cannot be used to obtain a non-reflexive space, and similarly for the non-reflexive space constructions. However, in [43] both reflexive and ℓ^1 cases are unified in one construction. We will explain in this section the relevant results from their work.

When working with a finite dimensional subspace of the Hilbert space of functions $L^2([-1,1])$, say H, with an orthonormal basis $\vartheta_1, \cdots, \vartheta_m$, one can easily check that the kernel function

$$k(x,y) := \sum_{j=1}^{m} \vartheta_j(x)\vartheta_j(y)$$

has the reproducing property for functions in H with its inner product. Since every norm is equivalent in finite dimensions, we can choose another p-norm for the space H, and this will not affect the reproducing properties of the kernel k. Moreover, the bilinear forms associated with the new spaces will be consistent with the integral form

$$\int_{-1}^{1} f(x)g(x)dx = \sum_j \sum_k a_j b_k \int_{-1}^{1} \vartheta_j(x)\vartheta_k(x)dx = \langle f, g\rangle_{\ell^p},$$

where a_k and b_k are the coefficients that represent f and g respectively. Therefore, if we want to study spaces of functions with ℓ^p norm in general, we must impose conditions on the reproducing kernel so

- The infinite sum representation $\sum_k \vartheta_k(x)\vartheta_k(y)$ is well defined at every pair of points.
- The infinite sums of the kind $\sum_j a_j\vartheta_j(x)$ and $\sum_j b_j\vartheta'_j(x)$ for some sequences $\{a_n\}, \{b_n\}$ are both well defined and only vanish everywhere if every coefficient is zero.

For the rest of the section we will assume the sets Ω and Ω' are locally compact Hausdorff spaces, both with a regular Borel measure μ and μ' respectively.

Definition 2.4.1. [43] Let $\mathsf{S}_k = \{\vartheta_n\}$ and $\mathsf{S}'_k = \{\vartheta'_n\}$ be families of measurable functions. A measurable function $k : \Omega \times \Omega' \longrightarrow \mathbb{C}$ is called a generalized Mercer kernel induced by the expansion-sets S_k and S'_k, if k can be represented by:

$$k(x,y) := \sum_{n\in\mathbb{N}} \vartheta_n(x)\vartheta'_n(y).$$

Where the convergence of the sum is assumed to be pointwise. We assume the expansion terms are countably infinite so we can work with infinite dimensional spaces of functions.

To define a RKBS starting from the kernel function induced by two families of functions, we first need to make sure that the kernel is well-defined for every pair of points. To construct these kind of spaces, it is enough for the sum of the expansion-sets to verify some kind of convergence everywhere. A sufficient condition would be imposing that the expansion-sets send the spaces Ω, Ω' to some ℓ^p and ℓ^q respectively.

Assumption. (A-p) Let $1 < p, q < \infty$ be conjugate exponents. Assume the expansion-sets are linearly independent sets and that for every pair $(x,y) \in \Omega \times \Omega'$, the next

conditions can be verified:

$$\sum_{n\in\mathbb{N}} |\vartheta_n(x)|^q\,, \qquad \sum_{n\in\mathbb{N}} |\vartheta'_n(y)|^p < \infty.$$

Assumption. (*A-1*) Suppose that the expansion-sets are linearly independent sets and that for every pair $(x, y) \in \Omega \times \Omega'$, the next conditions can be verified:

$$\sum_{n\in\mathbb{N}} |\vartheta_n(x)|\,, \qquad \sum_{n\in\mathbb{N}} |\vartheta'_n(y)| < \infty.$$

To simplify notations, we name the sums from the assumptions above by:

$$\Theta_q(x) := \sum_{n\in\mathbb{N}} |\vartheta_n(x)|^q$$

$$\Theta'_p(y) := \sum_{n\in\mathbb{N}} |\vartheta'_n(y)|^p$$

for $1 \leq p, q < \infty$. These conditions ensure that the induced generalized Mercer kernel can be evaluated on every point.

The next result is consequence of applying Hölder's inequality to the expansion-sets.

Theorem 2.4.1. Let $\mathsf{S}_k, \mathsf{S}'_k$ be sets that satisfy either assumption *A-p* or *A-1*. Then function

$$k(x, y) := \sum_{n\in\mathbb{N}} \vartheta_n(x)\vartheta_n(y)$$

is a generalized Mercer kernel.

If the expansion-sets have the property *A-p*, we define the following spaces with the help of the expansion-sets by

$$B_k^p(\Omega) := \{f(\cdot) = \sum_j a_j\vartheta_j(\cdot) \mid \{a_j\} \in \ell^p\},$$

$$B_{k'}^q(\Omega') := \{g(\cdot) = \sum_j b_j\vartheta'_j(\cdot) \mid \{b_j\} \in \ell^q\}.$$

Each of them accompanied by their respective norm

$$\|f\|_{B_k^p(\Omega)} := \|\{a_j\}\|_{\ell^p} = (\sum_j |a_j|^p)^{\frac{1}{p}},$$

$$\|g\|_{B_{k'}^q(\Omega')} := \|\{b_j\}\|_{\ell^q} = (\sum_j |b_j|^q)^{\frac{1}{q}}.$$

If they satisfy assumption *A-1* then the spaces

$$B_k^1(\Omega) := \{f(\cdot) = \sum_j a_j\vartheta_j(\cdot) \mid \{a_j\} \in \ell^1\},$$

$$B_{k'}^\infty(\Omega') := \{g(\cdot) = \sum_j b_j\vartheta_j(\cdot) \mid \{b_j\} \in c_0 \subset \ell^\infty\}$$

are also Banach spaces if equipped with the norms

$$\|f\|_{B_k^1(\Omega)} := \|\{a_j\}\|_{\ell^1} = \sum_j |a_j|,$$

$$\|g\|_{B_{k'}^\infty(\Omega')} := \|\{b_j\}\|_{\ell^\infty} = \sup_{j\in\mathbb{N}} |b_j|.$$

Since the space ℓ^∞ does not have a Schauder basis, we restricted the coefficients used for the space $B_{k'}^\infty(\Omega')$. This does not hinder the reproducing properties of the kernel functions.

To prove that these constructions yield adjoint pairs of RKBS we first show that the spaces above defined are isometric to ℓ^p spaces. This property is attained even if the expansion-sets are only linearly independent and not necessarily verify conditions *A-p* or *A-1*. We begin by showing that the evaluation functionals are continuous on Span $\{\varphi_n\}$.

Theorem 2.4.2. Let S_k and S'_k be expansion-sets that satisfy either assumption *A-p* or *A-1*. Then the evaluation functionals are continuous on Span $\{\varphi_n\}$ endowed with its corresponding norm.

Proof. If the expansion-sets satisfy assumption *A-p*, then by Hölder's inequality we the following for any $f(\cdot) = \sum_{n\in\mathbb{N}} a_n\varphi_n(\cdot) \in B_k^p(\Omega)$ and $g(\cdot) = \sum_{n\in\mathbb{N}} b_n\varphi'_n(\cdot) \in B_{k'}^p(\Omega')$:

$$|f(x)| \leq \sum_{n\in\mathbb{N}} |a_n\varphi_n(x)| \leq \Big(\sum_{n\in\mathbb{N}} |a_n|^p\Big)^{\frac{1}{p}} (\Theta_q(x))^{\frac{1}{q}} = \|f\|_{B_k^p(\Omega)} \Theta_q^{\frac{1}{q}}(x)$$

$$|g(y)| \leq \Big(\sum_{n\in\mathbb{N}} |b_n|^q\Big)^{\frac{1}{q}} (\Theta'_p(x))^{\frac{1}{p}}.$$

Similarly, if they satisfy *A-1* and we choose functions $f(\cdot) = \sum_{n\in\mathbb{N}} a_n\varphi_n(\cdot) \in B_k^1(\Omega)$ and $g(\cdot) = \sum_{n\in\mathbb{N}} b_n\varphi'_n(\cdot) \in B_{k'}^\infty(\Omega')$, then we have for every $x \in \Omega$ that

$$\|f\|_{B_k^1(\Omega)}\Theta_1(x), \|g\|_{B_{k'}^\infty(\Omega')}\Theta'_1(y) < \infty.$$

Then we get that its evaluation functional is continuous because

$$|f(x)| \leq \sum_{n\in\mathbb{N}} |a_n\varphi_n(x)| \leq \|f\|_{B_k^1(\Omega)}\Theta_1(x).$$

The inequalities for g come from the uniform norm applied to the coefficients in c_0. □

Theorem 2.4.3. Let $p \in [1,\infty)$ and the expansion-sets be linearly independent. Then the space $B_k^p(\Omega)$ is isometrically isomorphic to the space ℓ^p.

Proof. Consider the standard Schauder basis of ℓ^p consisting of the elements $\{\mathbf{e}_n\}$. We set up an isomorphism by sending each $\varphi_n \in \mathsf{S}_k$ to the element $\mathbf{e}_n$ and extending linearly to their spans. In other words, we define the next operator for every element $f \in$ Span $\{\varphi_n\}$:

$$f(\cdot) = \sum_n a_n\varphi_n(\cdot) \longmapsto \sum_n a_n\mathbf{e}_n,$$

where only a finite amount of the a_n are not zero. From the definition of the norm for $B_k^p(\Omega)$, we know that this defines an isometry. Furthermore, every element $\sum_n b_n\mathbf{e}_n$ is

the image of an element $g \in B_k^p(\Omega)$ under this isometry, namely, $g(\cdot) = \sum_n b_n \varphi_n(\cdot)$. Therefore the claim is proved by extending the isometry to their respective closures. □

Due to these isomorphisms, the pairs $B_k^p(\Omega)$ and $B_{k'}^q(\Omega\prime)$ form an adjoint pair of RKBS.

The spaces $\mathscr{B}_k^1(\Omega)$ and $\mathscr{B}_{k'}^\infty(\Omega')$ can also be shown to be isomorphic to sequence spaces, but these are not reflexive spaces, unlike the ones above.

Theorem 2.4.4. The space $\mathscr{B}_{k'}^\infty(\Omega')$ is isometrically isomorphic to the subspace $c_0 \subset \ell^\infty(\Omega')$ of sequences which converge to 0.

Proof. The idea of the proof is to follow the same steps as in Theorem 2.4.3, with the space c_0 replacing the space ℓ^p. □

Theorem 2.4.5. Let $p, q \in (1, \infty)$ be conjugate exponents and k be a generalized Mercer kernel induced by the expansion-sets $\mathsf{S}_k = \{\vartheta_n\}$ and $\mathsf{S}_k' = \{\vartheta_n'\}$ which satisfy condition *A-p*. Then the spaces $\mathscr{B}_k^p(\Omega)$ and $\mathscr{B}_{k'}^q(\Omega')$ form an adjoint pair of RKBS with reproducing kernel $k(x, y) = \sum_{n\in\mathbb{N}} \varphi_n(x)\varphi_n'(y)$.

Proof. The previous results show that the spaces $\mathscr{B}_k^p(\Omega)$ and $\mathscr{B}_{k'}^q(\Omega')$ are RKBS, moreover they are dual to each other. We use the isometries to define the bilinear form between $f(\cdot) = \sum a_n \varphi_n(\cdot)$ and $g(\cdot) = \sum b_n \varphi_n'(y)$ as

$$\langle f, g\rangle = \langle f, g\rangle_{\mathscr{B}_k^p(\Omega)\times\mathscr{B}_{k'}^q(\Omega')} := \sum_{n\in\mathbb{N}} a_n b_n.$$

We just need to verify that the kernel function has the reproducing property with respect with this bilinear form. Fix an element $f = \sum_n a_n \varphi_n \in \mathscr{B}_k^p(\Omega)$ and for a $x \in \Omega$, consider the element $k(x, \cdot) = \sum_n \varphi_n(x)\varphi_n'(\cdot)$. Given that $\sum_n |\varphi_n(x)|^q < \infty$, this element lies in $\mathscr{B}_{k'}^q(\Omega')$. Therefore the next equality follows:

$$\langle f, k(x, \cdot)\rangle = \sum_n a_n \varphi_n(x) = f(x).$$

The same line of reasoning leads us to conclude that k also has the right-handed reproducing property. □

We know that the space ℓ^1 is the dual space of c_0, therefore we know that the same reasoning as above will give us that the kernel function k has the reproducing property for functions in $\mathscr{B}_{k'}^q(\Omega')$ which is isometrically isomorphic to c_0. But as we have seen in the previous section, we do not need the whole dual space to construct an adjoint pair of RKBS. This is the case for these two spaces.

Theorem 2.4.6. Let k be a generalized Mercer kernel induced by the expansion-sets $\mathsf{S}_k = \{\vartheta_n\}$ and $\mathsf{S}_k' = \{\vartheta_n'\}$ which satisfy condition *A-1*. Then the spaces $\mathscr{B}_k^1(\Omega)$ and $\mathscr{B}_{k'}^\infty(\Omega')$ form an adjoint pair of RKBS.

Proof. Since $\mathscr{B}_k^1(\Omega)$ and $\mathscr{B}_{k'}^\infty(\Omega')$ are isomorphic to ℓ^1 and $c_0 \subset \ell^\infty$, we define the bilinear form in the same way as the previous theorem. Both spaces are RKBS because of Theorem 2.4.2, and the deduction of reproducing property from both sides follows the same argument as in the previous theorem. □

The definition of RKBS in [43] was limited by forcing a condition on the whole dual space. The space $\mathscr{B}_k^1(\Omega)$ was only a right-sided RKBS because of this restrictive condition. Definition 1.2.1 and Construction 1.2.1.1 do not have this limitation.

Summary

Let Ω_1, Ω_2 be locally compact Hausdorff spaces and $k : \Omega_1 \times \Omega_2 \longmapsto \mathbb{C}$ be a generalized Mercer kernel induced by the families $\mathsf{S}_k = \{\vartheta_n\}$ and $\mathsf{S}'_k = \{\vartheta'_n\}$ which satisfy assumption *A-p* for $1 < p < \infty$. Then the induced adjoint pair of RKBS by setting

$$W_1 := \ell^q \; W_2 := \ell^p$$

with the feature maps $\Phi_1 : \Omega_1 \longmapsto W_1$, $\Phi_2 : \Omega_2 \longmapsto W_2$ and the bilinear form

$$\langle \{a_n\}, \{b_n\} \rangle_{W_1 \times W_2} := \langle \{a_n\}, \{b_n\} \rangle_{\ell^q} = \sum_{n \in \mathbb{N}} a_n b_n.$$

The induced adjoint pair of RKBS is

$$B_k^p(\Omega) := \{f(\cdot) = \sum_j a_j \vartheta_j(\cdot) \mid \{a_j\} \in \ell^p\},$$

$$B_{k'}^q(\Omega') := \{g(\cdot) = \sum_j b_j \vartheta'_j(\cdot) \mid \{b_j\} \in \ell^q\}.$$

and the associated reproducing kernel

$$k(x, y) := \sum_{n \in \mathbb{N}} \vartheta_n(x) \vartheta'_n(y)$$

Let $g : \Omega_1 \times \Omega_2 \longmapsto \mathbb{C}$ be a generalized Mercer kernel induced by the families $\mathsf{T}_g = \{\vartheta_n\}$ and $\mathsf{T}_{g'} = \{\vartheta'_n\}$ which satisfy assumption (*A-1*). Then the induced adjoint pair of RKBS by setting

$$W_1 := c_0 \; W_2 := \ell^1$$

with the feature maps $\Phi_1 : \Omega_1 \longmapsto W_1$, $\Phi_2 : \Omega_2 \longmapsto W_2$ and the bilinear form

$$\langle \{a_n\}, \{b_n\} \rangle_{W_1 \times W_2} := \langle \{a_n\}, \{b_n\} \rangle_{c_0} = \sum_{n \in \mathbb{N}} a_n b_n.$$

The induced adjoint pair of RKBS is

$$\mathscr{B}_1 = B_g^1(\Omega) := \{f(\cdot) = \sum_j a_j \vartheta_j(\cdot) \mid \{a_j\} \in \ell^1\},$$

$$\mathscr{B}_2 = B_{g'}^\infty(\Omega') := \{h(\cdot) = \sum_j b_j \vartheta_j(\cdot) \mid \{b_j\} \in c_0 \subset \ell^\infty\}$$

and the associated reproducing kernel

$$g(x, y) := \sum_{n \in \mathbb{N}} \vartheta_n(x) \vartheta'_n(y).$$

Chapter 3

Applications: Support vector machines.

In this section, we introduce an application of RKBS to Machine Learning. We will use some results from the previous chapter but applied to spaces of real-valued functions. Those results can be stated and proven for real Banach spaces with the appropriate modifications.

First, we introduce some standard terminology from Machine Learning literature. Through this section, the input set denoted by $\mathcal{X}$, will be the set from where patterns, inputs or observations $\{x_i\}$ will be taken. The output set $\mathcal{Y}$ will be the set of *predictions*. The functions will map inputs to $\{y_j\} \subset \mathcal{Y}$ called labels, outputs or targets. As we will focus on real-valued functions, we will assume that the labels are a subset of $\mathbb{R}$.

A loss function will be understood as any function that "measures" how good a solution candidate f is for a given task. This measure of "goodness" is not only about counting the mistakes a function makes. It usually can be seen as a function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \longrightarrow \mathbb{R}^+$ but the arguments can be grouped together in different ways. We refer the reader to [35] and [39] for more details on loss functions.

The set of functions from where we will pick our solutions will be occasionally referred to as the hypothesis space. For example, we can restrict our search to a Banach space of functions B, a Hilbert space H or a more specific subset of them.

The results presented here are the essentials which we will need to develop an application of RKBS to Machine Learning. To read further on this topic, we recommend reading [35] [39] [30]. We will only develop the parts which explicitly connect with the application of RKBS to the algorithm.

A typical problem in machine learning is classifying. We want a classifier that gives the correct labels to the samples and at the same time can label correctly future observations. We illustrate this point with the following example. Suppose that you want to predict whether a person will contract a certain disease or not. What is given to you is a database of the clinical data of a number m of patients. Part of this data is given as real numbers, so we can associate to each patient a vector $x_p \in \mathbb{R}^m$.

You can categorize the patients into those who have contracted the disease and those who have not contracted it. With the clinical data available, one must find a reliable classifier. In this context, reliable means whether it can predict if a new patient is in risk of contracting the disease or not with enough accuracy.

An important part of modeling these problems is the loss function. The loss function will depend on how the model is proposed. A simple way to construct

a loss function for classification would be to count the amount of mistakes a classifier f makes, i.e., $L(\mathbf{x}, \mathbf{y}, f(\mathbf{x})) := \sum_{i=1}^{n} 1_{f(x_i) \neq y_i}$. But depending on our hypothesis space and the approach to solve it, it could be a difficult function to optimize. We could for example choose Lagrange polynomials if we were looking for functions of one real variable, or we could extend the search to a bigger hypothesis set. For this loss function, any function that interpolates those points is the best it can find. Another example is the ϵ insensitive loss: $\sum_i \max\{|f(x_i) - y_i| - \epsilon, 0\}$. This loss allows a certain error as long as it does not pass a threshold defined beforehand. This is to accommodate the cases where the data is noisy. So the problem of classification could very well begin from the choice of our loss function. Another thing that must be considered, not every choice of loss function will be adequate from a computational perspective.

The problem of how to choose a loss function adequate for the problem will not be covered in this work. From here onwards we will either give a specific loss function, it will appear while working through an example, or we will assume one is already given.

Support vector machines were introduced by Vapnik [8] for classifying, assuming the points can be classified by a hyperplane. That means that given points $x_1, \cdots, x_n \in \mathbb{R}^m$, each one labeled by $y_1, \cdots, y_n \in \{-1, 1\}$, there exists a linear functional f such that all the points with the same label all lie on the same half-plane determined by $\{x | f(x) = a\}$, for some $a \in \mathbb{R}$. Thus the linear functional classifies them with their signs. Given the existence of this functional, a classifier can be taken to be $\text{sign}\,(f(x) + a)$.

Suppose that we have a set of sample points $x_1, \cdots, x_n \in \mathbb{R}^m$ and their labels $y_1, \cdots, y_n \in \{-1, 1\}$. A support vector machine returns a classifier of the form

$$h(x) := \text{sign}\,(w \cdot x + b)\,,$$

where $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$. To find the parameters w and b of this function, we will assume first that there is at least one linear classifier. For this classification problem with m samples, we will optimize the loss function

$$L(\mathbf{x}, \mathbf{y}, h(\mathbf{x})) := \frac{1}{m} \sum_i \max\{0, 1 - h(x_i) y_i\},$$

where $\mathbf{x} = (x_1, \cdots, x_m)$, $\mathbf{y} = (y_1, \cdots, y_m)$ and $h(\mathbf{x}) = (h(x_1), \cdots, h(x_m))$.

Given the existence of one linear classifier, we can change it by transforming the defining linear functional, and still manage to classify correctly the patterns. But the purpose of a classifier does not exclusively lies on classifying correctly the training points, but to make as few mistakes as possible when classifying new ones. Therefore if the hyperplane is too close to some labeled points, it may incur in some errors if given a new pattern x_{n+1}. A way to make this less likely to happen is to find a classifier which does it correctly with the current samples, and at the same time maximizes its distance to all the pattern points. Simply speaking, we are looking for a hyperplane which lies right in the middle of the two classes. This is sometimes called the maximal margin classifier [14].

We impose the extra condition that the hyperplane is as far as possible to all the sample points $x_1, \cdots, x_n$. The distance from a point x to a hyperplane $w \cdot x + b$ can

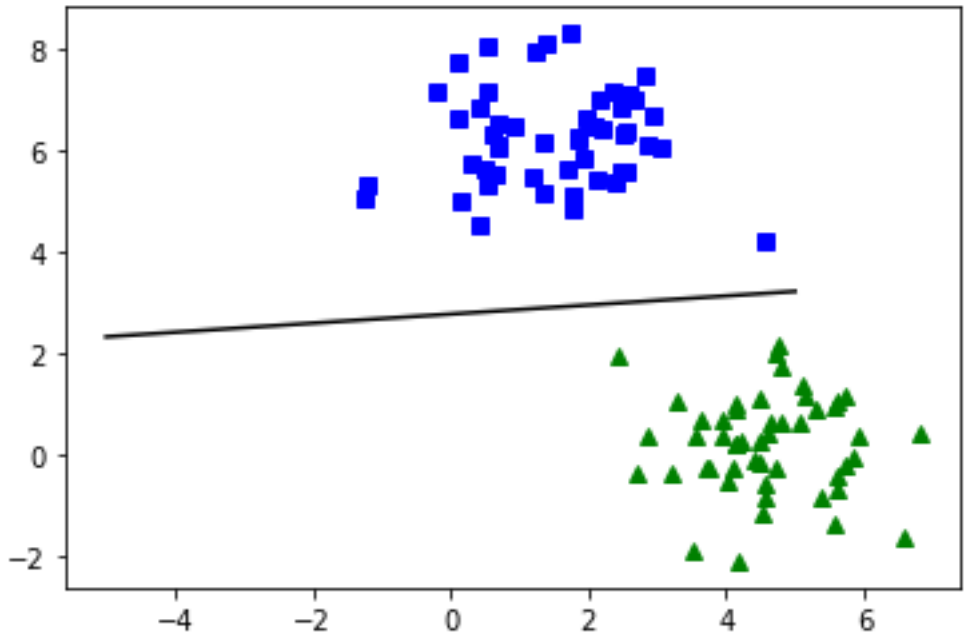

FIGURE 3.1: The points lying on the plane are labeled by squares or triangles. A linear classifier is chosen such that the hyperplane separates the points by their label.

be calculated by

$$\frac{|w \cdot x + b|}{\|w\|},$$

this is because w is perpendicular to the hyperplane it defines, so we can use orthogonal projection of the point x on w. Therefore, we find that the problem to optimize is

$$\text{Max margin.} = \max_{\substack{w,b\in\mathbb{R}^n: \\ \min_i y_i(w\cdot x_i+b)\geq 0}} \min_{x_i} \frac{|w \cdot x_i + b|}{\|w\|} = \max_{w,b\in\mathbb{R}^n} \min_{x_i} \frac{y_i\,(w \cdot x_i + b)}{\|w\|}.$$

The second equality comes from how the classifier separates the patterns with its sign, so the expression $y_i\,(w \cdot x_i + b)$ is positive if the problem is linearly separable. This quantity has the property of being invariant if we multiply the pair (w, b) by a positive constant β, so if we choose a $\beta_{w,b}$ for every expression of the form $\frac{y_i(w\cdot x_i+b)}{\|w\|}$ we can simplify the expression. In particular, since

$$\min_i y_i\,(w \cdot x_i + b)$$

is positive by our assumption, we can divide the expressions by it so we have that $\min y_i\,(w \cdot x_i + b) = 1$. This means that those points that reach the minimum lie on two "margins" that run parallel to the separating hyperplane, these margins being $w \cdot x + b = \pm 1$. Thus we end up with the following expressions:

$$\max_{\substack{w,b\in\mathbb{R}^n: \\ \min_i y_i(w\cdot x_i+b)=1}} \frac{1}{\|w\|} = \max_{\substack{w,b\in\mathbb{R}^n: \\ y_i(w\cdot x_i+b)\geq 1}} \frac{1}{\|w\|}.$$

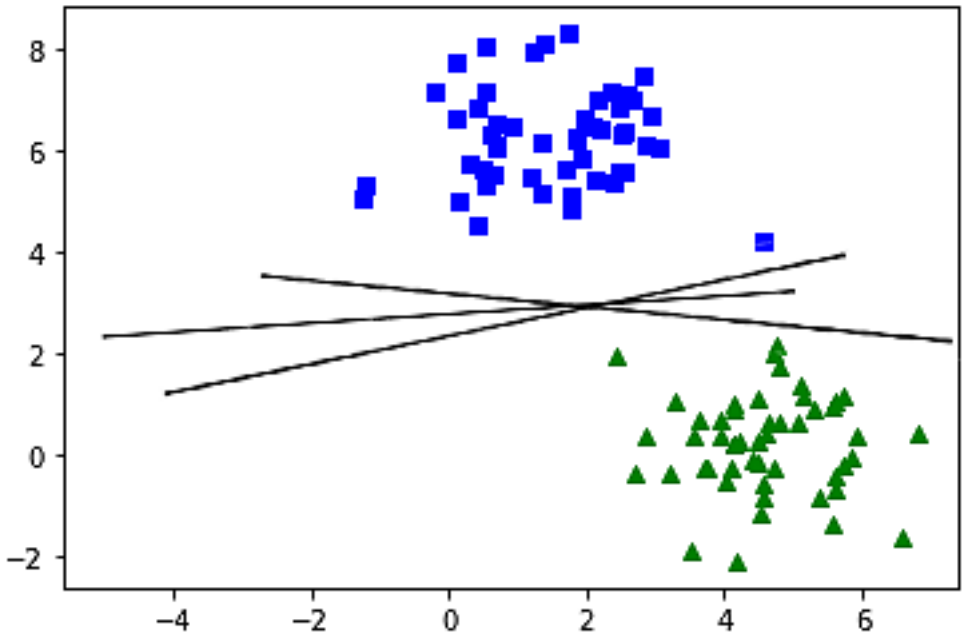

FIGURE 3.2: Different choices of linear classifiers for the same problem which do correct classification of the training points. Some hyperplanes lie closer to the sample points than others.

Furthermore, maximizing $\frac{1}{\|w\|}$ is equivalent to minimizing

$$\frac{1}{2}\|w\|^2.$$

This means that we are looking to solve the following optimization problem:

$$\min_{(w,b)} \frac{1}{2}\|w\|^2 \tag{3.1}$$

subject to:

$$y_i\,(w \cdot x_i + b) \geq 1,\ \forall i = 1, \cdots, n. \tag{3.2}$$

As this is a convex and quadratic optimization problem, we can look at its Lagrangian dual problem to solve it [30][35]. The Lagrangian of the primal problem 3.1 is

$$\mathcal{L}\,(w, b, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i \left[y_i(w \cdot x_i + b) - 1\right],$$

where the coefficients $\alpha_i \geq 0$ are the Lagrange multipliers. We are optimizing with respect to the primal variables, therefore we look for the saddle points:

$$\nabla_w \mathcal{L} = w - \sum_{i=1}^{m} \alpha_i x_i = 0, \tag{3.3}$$

$$\nabla_b \mathcal{L} = -\sum_{i=1}^{m} \alpha_i y_i = 0.$$

From these we derive the following relations:

$$w = \sum_{i=1}^{m} \alpha_i x_i, \tag{3.4}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{3.5}$$

And the conditions applied to the constraints end up being

$$\alpha_i \left[y_i \left(w \cdot x + b \right) - 1 \right] = 0.$$

From this last equation we get that those samples whose coefficients are non-zero are those who completely determine the solution, this is because they lie on the margin $y_i \left(w \cdot x + b \right) = 1$, so the remaining samples did not affect the solution. We call these vectors *support vectors*. Since for any support vector x_{i_0} we have that $w \cdot x_{i_0} + b = y_i$, we can solve for b by using the relations above:

$$b = y_i - \sum_{j=1}^{m} \alpha_j x_j \cdot x_{i_0}. \tag{3.6}$$

By plugging in the obtained vector w, applying the second relation 3.5 and simplifying, we obtain the following expression for the Laplacian:

$$\mathcal{L} = \frac{1}{2} \| \sum_{i=1}^{m} \alpha_i x_i \|^2 + \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle.$$

This leads to the following dual optimization problem to 3.1.

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j \langle x_i, x_j \rangle \tag{3.7}$$

subject to:

$$\alpha_i \geq 0 \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{3.8}$$

This expression shows that the solutions depend on the inner product between the representation $\Phi(x_j)$ of the sample points. In this case $\Phi(x_j) = x_j$, and the representation itself does not need to be explicitly known. This observation is what will motivate the use of kernel functions.

Remark. The loss function is only implicitly used here. At first we choose the hypothesis set as the set of affine functions. Over this set the loss function

$$\frac{1}{m} \sum_{i} \max\{0, 1 - h(x_i) y_i\}$$

can take multiple values. But we restrict the hypothesis set to affine functions which make no mistakes, i.e. we restrict it to where the loss function is 0. And finally the hard margin algorithm looks for the affine functions which maximize the margin. These considerations gave us a clue what form solutions can take with more general loss functions.

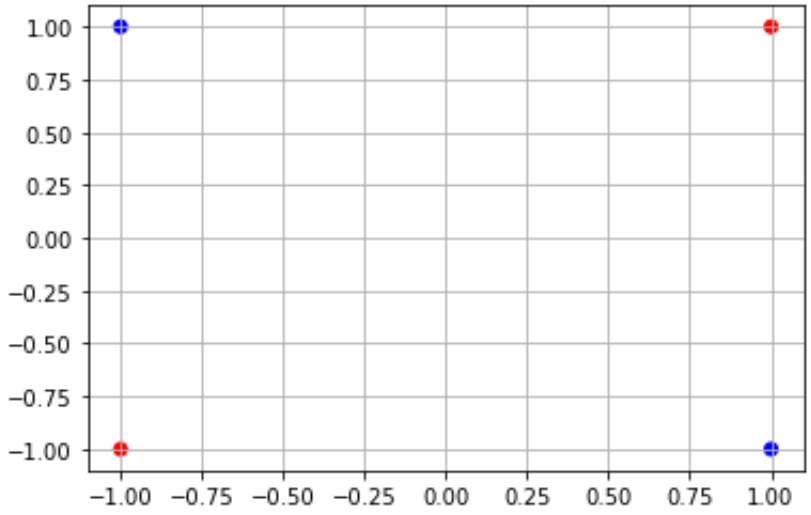

FIGURE 3.3: Four points, labeled by color. Any linear classifier will give a wrong label to at least a couple of them, therefore making the problem of classifying them a non-linearly separable one.

For solving real-life problems, the family of linear classifiers is not enough. As a trivial example, consider the colored four points in Figure 3.3. Any hyperplane will classify wrongly at least two of them. One way to surmount this problem would be mapping the points to a higher dimensional space where they can be separated by a hyperplane.

But mapping to a higher dimensional space would need finding the appropriate feature map, and then perform the separation after mapping the sample points. Equation 3.7 leads us to think that the formulation of the problem can put in terms of the inner product between the mapped points. This is where the theory of reproducing kernel spaces helps to solve this new problem.

We will see in the following sections that we can change the loss function to more general functions and the solutions obtained will retain a similar form to (3.3).

Note that with our assumptions of linear separability, the loss function reaches 0. Our derivation's purpose was to show that we could take a look at the dual problem for hints about the solution's properties. And what we found is that it takes the form (3.5). But in the following sections we will not restrict the results to a classification problem. Therefore we do not necessarily choose the loss function $\frac{1}{m}\sum_i \max\{0, 1 - h(x_i)y_i\}$.

3.1 Representer theorem for RKHS

The original purpose of the algorithm above is to find a classifier trained on a set of examples, and we want one that is less likely to make a mistake when classifying a new point. In this sense is that we refer about the generalization, the capacity of providing a sufficiently accurate prediction for unseen patterns. A way to improve the generalization ability of the algorithm is to expand the set of classifiers to a set with more complex functions. The problem is that, doing it blindly would possibly be overfitting, which means that the selected classifier has "memorized" the samples. New patterns fall outside of what it has memorized and its performance is poor on these new examples. On the other hand, a hypothesis set could be too small. In

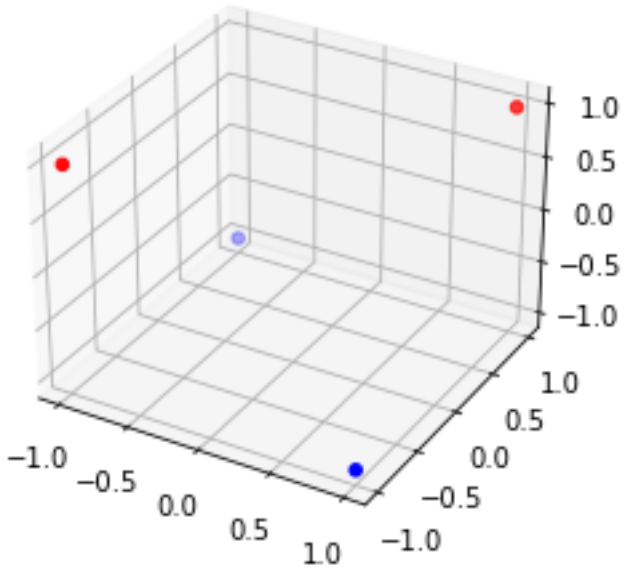

FIGURE 3.4: The points in Figure 3.3 are mapped to $\mathbb{R}^3$, transforming the original problem to one that is linearly separable in this new space.

this case there is no function with an acceptable accuracy even within the training examples. This is referred to as underfitting. So one must find a way to balance the complexity of the hypothesis set to avoid both. The model of maximal margin classification with kernels deals with the underfitting aspect. For overfitting we will work with regularization [39].

Regularization consists of restricting the search of the classifier to smaller subsets by adding a penalization term $\phi(f)$, sometimes accompanied by a coefficient λ as a weight. This is to add some extra information when doing the search of the solution [25].

Other forms of avoiding overfitting and the theory behind it is beyond the scope of this work, so we refer the reader to [35] [36].

The discussion at the end of the previous section argues that loss functions is identically zero on the hypothesis set. Therefore, we can think our solution $f_{(w,b)} = \langle \cdot, w \rangle + b$ is a minimizer of the function

$$L(\mathbf{x}, \mathbf{y}, f_{(w,b)}(\mathbf{x})) + \lambda\phi(f_{(w,b)}).$$

Thus the problem which we solved has the form

$$\min_{(w,b)\in\mathbb{R}^j+1} L(\mathbf{x}, \mathbf{y}, f_{(w,b)}(\mathbf{x})) + \lambda\phi(f_{(w,b)}).$$

We call a problem of this form a *regularized problem*.

Support vector machines, as formulated at the beginning of the chapter, find affine functions as solutions. This hypothesis set may not be enough to face a real problem, since linear separability is a very strict condition. But this approach can be adapted to deal with more general problems. To see how this can be achieved, we go back to the first expression in (3.7), where we see that the solution is given in terms of inner products between the sample points.

In the previous section, the resulting loss function came naturally after the classification with the geometrical margin. As we discussed at the beginning of the chapter, depending on the model and what you want out of it, you can use different loss functions. We also found that our solution could be expressed as

$$\sum_i \alpha_i k(\cdot, x_i) = \sum_i \alpha_i \langle \cdot, x_i \rangle.$$

We want this feature to be preserved when we change loss functions, and that is what we will work towards.

Remember that a RKHS with a reproducing kernel k and an associated feature map Φ has the property of representing the inner product between $\Phi(x_1)$ and $\Phi(x_2)$ by the evaluation $k(x_1, x_2)$. By the expression of the dual problem

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j \langle x_i, x_j \rangle \tag{3.9}$$

we can propose searching for solutions of this form. This is because the solution depends on the evaluations of the functionals $\langle \cdot, x_j \rangle$. More concretely, to find the solution to the dual problem we need the coefficients of the positive definite matrix $\left(k(x_i, k_j)\right)_{i,j=1,\cdots,n}$.

We know from Section 1.1 that a positive definite kernel function will map the sample points to a RKHS. Therefore we could use this idea to implement a SVM that returned functions more complex than affine functions. The issue is that by changing spaces to a higher dimensional one could make the solution much more computationally expensive to find, even unfeasible to compute.

The Representer theorem for RKHS says that this is not necessarily the case for a big class of loss functions. Moreover, this theorem extends the studied case to the regularized problems

$$L + \lambda \phi$$

where L is a loss function and the function ϕ is a function with certain properties we will explain later. This theorem is consequence solely of the convexity of the functions and the orthogonal decomposition of a Hilbert space. Solving the optimization problem with a regularizer biases the search of solutions towards smaller class, making it less likely to overfit [35]. The regularization term ϕ we use will be dependent on the norm of the function, this means $\phi(f) = \phi(\|f\|)$.

We assume the sets of m samples x_i with their m labels y_i are fixed.

Theorem 3.1.1. (Representer Theorem for RKHS) [33] Let $L : (X \times \mathbb{R} \times \mathbb{R})^m \longrightarrow \mathbb{R} \cup \{\infty\}$ be an arbitrary loss function, $G : X \times X \longrightarrow \mathbb{R}$ a strictly increasing function and $k : X \times X \longrightarrow \mathbb{R}$ a PDS kernel with associated Hilbert space H. Then a minimizer of the problem

$$\min_{h \in H} L((x_1, y_1, h(x_1)), \cdots, (x_m, y_m, h(x_m))) + G(\|h\|_H) \tag{3.10}$$

always has a representation of the form $h^* = \sum_{i=1}^{m} a_i k(\cdot, x_i)$.

Proof. Let $M = \text{Span}\{k(\cdot, x_i)\}_{i=1,\cdots n}$. For any given function $h \in H$, we can decompose it in $h = \sum a_i k(\cdot, x_i) + h_\perp = h_1 + h_\perp$, where $h_\perp \in M^\perp$. By the reproducing property of the kernel k we have that $h(x_j) = \langle h, k(\cdot, x_j) \rangle = \langle h_1, k(\cdot, x_j) \rangle + \langle h_\perp, k(\cdot, x_j) \rangle = \langle h_1, k(\cdot, x_j) \rangle = h_1(x_j)$. Thus the evaluation of the loss function is the same on h and

h_1, i.e.

$$L((x_1, y_1, h(x_1)), \cdots, (x_m, y_m, h(x_m))) = L((x_1, y_1, h_1(x_1)), \cdots, (x_m, y_m, h_1(x_m))).$$

By the orthogonality of h_1 and $h_\perp$ we have that $\|h\|_H = \sqrt{\|h_1\|_H^2 + \|h_\perp\|_H^2}$, and since G is non-decreasing the next inequalities follow

$$G(\|h_1\|_H) \leq G(\sqrt{\|h_1\|_H^2 + \|h_\perp\|_H^2}) = G(\|h\|_H),$$

so if h is a solution to the optimization problem, its orthogonal projection h_1 is also a solution. If G is assumed to be strictly increasing, then the last inequality is also strict, showing that any solution must be such that $h_\perp = 0$. □

So the search for the solution can be found in the finite dimensional space M spanned by the functions $k(\cdot, x_i)$. And the dual form (3.7) allows us to calculate it independently of the dimension of the Hilbert space H, since we only need the evaluations $k(x_i, x_j)$.

An aspect we want to recover from this result is the representation of a solution through linear combinations of $k(\cdot, x_i)$. As we will see, this exact same result will not be achieved every time. What we will do is show that solutions can still be obtained by working on a finite dimensional space.

Going from solutions expressed through affine functions to the solutions expressed through kernel functions has some drawbacks. A solution to the regularized problem has the form $\sum_i \alpha_i k(x_i, \cdot)$, those x_i whose coefficients α_i are not 0 are called support vectors. In theory these can be any number $n \leq m$. But in practice, it is found that without specific regularizers most if not all samples are support vectors. One may find that some of these coefficients are much smaller than the rest, which means that when computing $f(x) = \sum_i \alpha_i k(x_i, x)$, they will contribute very little. This brings storage and computation difficulties for little gain, since this means that we will use every support vector in the training and the evaluation of the solution [46]. Solving a regularized problem with the ℓ^1 norm has been shown to reduce the amount of support vectors without a dramatic impact on its prediction capabilities [46][44]. Another possible issue is that the input data could have a meaningful metric defined a priori. This could be incorporated in the choice of reproducing kernel, but not every metric space can be embedded into a Hilbert space.

3.2 Representer theorem for RKBS

The Representer theorem for RKHS shows that by using a positive definite function (i.e. a kernel function for the RKHS), we can solve optimization problems, including the hyperplane separation problem from the beginning of the chapter [see 35, example 4.6] in a higher dimensional space. If a coefficient α_{j_0} is not zero, the evaluation must take it into account , whether it is significant or not for the solution. One way that has been found to mitigate this problem is to use regularization with respect to the ℓ^1 norm applied to the coefficients. This approach has been shown to yield solutions with fewer support vectors [44] [10]. There exist other choices of regularization terms but for any $p \neq 2$, the problem lies outside the scope of the Representer theorem for RKHS, see Figure 3.5. Actually, a necessary and sufficient condition for a

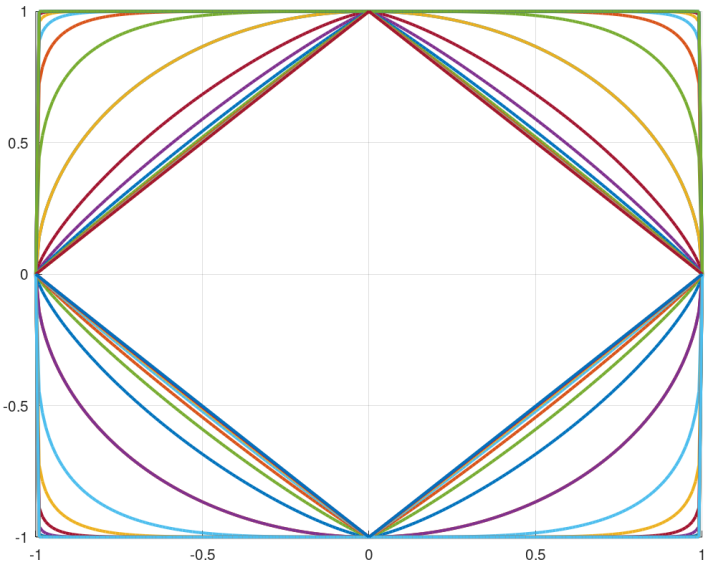

FIGURE 3.5: Coefficient based regularization, the curve-levels of the equations $\|f\|_p = 1$, for $p = \frac{2^9}{2^9-1}, \frac{2^8}{2^8-1}, \cdots, 2^8, 2^9$, illustrate that a regularizer of the form $\Omega_1 = \|\cdot\|_p$ cannot be thought as a regularizer f the form $\Omega(\|\cdot\|_q)$ for $p \neq q$, therefore the Representer theorem does not apply to this case, in particular for $p = 1$.

regularization problem to have a solution that can be expressed as $\sum_1^m \alpha_i k(\cdot, x_j)$, is for the regularizer to be of the form $\phi(\|\cdot\|_H)$ [3]. Therefore it is in our interests to extend the results obtained for RKHS to Banach spaces to expand this result to regularizers that depend on other non-Hilbertian norms.

The Representer Theorem for Hilbert spaces uses the geometric properties of Hilbert spaces to show the existence of solution in the finite dimensional subspace M. To show that any solution must be of the same form, we also used the hypothesis of strict monotony, but it essentially came from the orthogonality to the space spanned by the functions $k(\cdot, x_j)$. This means that if we can give algebraic or geometric conditions similar to those given by an inner product we could potentially prove the same result.

It turns out that an useful problem related to the regularized problem is *minimum norm interpolation*. This is to find the minimum of

$$\|f\|_{\mathscr{B}}$$

over all functions which $f(x_i) = y_i$. This problem can be obtained from regularization problems by making $\lambda \to 0$ [3]. Furthermore we will see that the solutions to a minimum interpolation problem (if any exists) can be used to justify the existence of a solution to the regularization problem.

The reason to consider minimum norm interpolation problems is that their solutions are related to the solutions of regularized problems. This is because the assumptions on the loss functions make them compatible with evaluation functionals. We say this in the sense that they keep the weak convergence with respect to the evaluation functionals, i.e. if $eval_{x_i}(f_v) = f_v(x_i) \to f(x_i) = eval_{x_i}(f)$ then

$L(\mathbf{x}, \mathbf{y}, f_v) \to L(\mathbf{x}, \mathbf{y}, f)$. Thus they serve as base for solutions to regularized problems.

Since a minimum interpolation problem and a regularization problem are different a priori, we will separate their respective results unless proven to be equivalent.

Definition 3.2.1. Remember the Construction 1.2.1.1 of an adjoint pair of RKBS. With the same notation and hypothesis, and given an adjoint pair of RKBS B_1, B_2 induced by the spaces W_1 and W_2, we define the following sets for vectors , $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ and an arbitrary subset $A \subseteq W_1$:

$$\begin{aligned} S_{\mathbf{x},\mathbf{y}} &:= \{f \in B_1 : f(x_i) = y_i,\ i = 1, \cdots, m\}, \\ V_{\mathbf{x},\mathbf{y}} &:= \{u \in W_2 : \langle \Phi_1(x_i), u\rangle = y_i,\ i = 1, \cdots, m\} \\ S^{\mathbf{x}} &:= \operatorname{Span}\{k(x_j, \cdot) : j = 1, \cdots, m\}, \\ A^{\vdash} &:= \{g \in W_2 : \langle a, g\rangle_{W_1 \times W_2} = 0\ \forall a \in A\}. \end{aligned}$$

We begin with a lemma that gives conditions for $V_{\mathbf{x},\mathbf{t}}$ to be non-empty given the samples $\mathbf{x} := (x_1, \cdots, x_m)$.

Lemma 3.2.1. The set $V_{\bar{x},\bar{t}}$ is non-empty for all $\mathrm{t} \in \mathbb{R}^m$ if and only if $\{\Phi_1(x_i)\}_{I=1,\cdots,m}$ is a linearly independent set in W_1.

Proof. Since the sample set $\{x_j\}$ is arbitrary but fixed, we denote by $\mathbf{f}_u$ for any given $u \in W_2$ the vector $(f_u(x_1), \cdots, f_u(x_m))$. Suppose that the vector $\mathbf{c} \in \mathbb{R}^m$ is such that $g_c = \sum_{j=1}^{m} c_j \Phi_1(x_j) = 0$. Then for an arbitrary $u \in W_2$ we have that

$$\langle g_c, u\rangle = \langle \sum_{j=1}^{m} c_j \Phi_1(x_j), u\rangle = 0.$$

Due to the bilinear form, this last expression is equivalent to

$$\sum_{j=1}^{m} c_j \langle \Phi_1(x_j), u\rangle = \sum_{j=1}^{m} c_j f_u(x_j) = 0$$

for all $u \in W_2$. But this is the inner product in $\mathbb{R}^m$, therefore the set $\{\Phi_1(x_i)\}$ is linearly dependent if and only if the the span of $\{\mathbf{f}_u : u \in W_2\}$ is not $\mathbb{R}^m$. And this last condition if equivalent to $V_{\bar{x},\bar{t}} = \emptyset$ for $\mathbf{t} = \mathbf{c}$. □

At the end of the previous section, we brought up some problems that appear when training SVM with kernels. We proposed ℓ^1 regularization as a way to address one of them. And we also went over why the Representer theorem for RKHS cannot cover this regularization. So the rest of the chapter will be devoted to unify these tools to some degree.

3.2.1 Representer theorem for uniformly convex and Gateaux differentiable spaces.

We will work first with uniformly convex and Gateaux differentiable spaces. The reason is that their geometry allows for orthogonality arguments, even if we do not explicitly use the semi-inner product, as seen in Section 2.2. We now define the Gateaux differential of a point.

Definition 3.2.2. [6] Let V be a uniformly convex and Gateaux differentiable space. The Gateaux differential at f, $\mathcal{G}(f)$ is the unique linear functional defined by the limit

$$\mathcal{G}(f)(g) = \lim_{t\to 0}\left(\frac{\|f+tg\|_V - \|f\|_V}{t}\right).$$

The limit is called the differential at f in the direction of g.

Remark. The Gateaux differential at a point f coincides with the semi-inner product with f in its second argument, as seen in Section 2.2.

The Representer theorem for Hilbert spaces relied on the following:

- Strict monotony to show that any solution can be projected onto the space we want.
- Orthogonality to show that any solution must be of the same form.

From Section 2.2 we know that with some extra assumptions, the notion of orthogonality can be translated to the Banach spaces by the inequality 2.2.4. Thus the following results can be proven by adapting the ideas, considering the appropriate modifications.

The following lemmas will help when showing the existence of solutions to the problem of minimal norm interpolation.

Lemma 3.2.2. [29] Every convex and closed non-empty subset C of a reflexive and strictly convex space B has the best approximation property. This means that for every $x \in B$, there exists a unique $c \in C$ such that

$$\|x-c\| = \min_{y\in C}\|x-y\|.$$

Lemma 3.2.3. In a Gateaux differentiable Banach space B, x is orthogonal to y if and only if $\langle y, \mathcal{G}(x)\rangle_B = \mathcal{G}(x)(y) = 0$.

Proof. This is a rephrasing of Theorem 2.2.4. □

We first show that the minimal norm interpolation can be covered by the Representer theorem.

Theorem 3.2.4. (Representer theorem for minimal norm interpolation in strictly convex and Gateaux differentiable spaces) Assume W_1, W_2, $\langle\cdot,\cdot\rangle_{W_1\times W_2}$ are as in Construction 1.2.1.1. Assume further that W_2 is a strictly convex and Gateaux differentiable space. Then the problem of finding a function that interpolates the data $\{y_i, i = 1...m\}$ of minimum norm, i.e.

$$\min_{f\in S_{\mathbf{x},\mathbf{y}}}\|f\|_B$$

has a solution $f \in W_2$ with the following property:

$$\mathcal{G}(f) \in \left(\Phi_1(X)^{\vdash}\right)^{\perp}.$$

Proof. Note that for a fixed vector $\mathbf{t}$ the set $V_{\mathbf{x},\mathbf{t}} = \bigcap \delta_{x_i}^{-1}(\{t_i\})$ is closed due to the continuity of the evaluation functionals. It also is a convex set, therefore by Lemma 3.2.2 we have that there exists an element with minimum norm, we denote it by v_0.

We will show next that this element is orthogonal to $V_{x,0}$ in the sense of Theorem 2.2.4. It is easy to see that the following statement is true:

$$V_{\mathbf{x},\mathbf{t}} = v_0 + V_{\mathbf{x},\mathbf{0}}.$$

So the inequality

$$\|v_0\| \leq \|v_0 + \nu\|$$

is trivially true for every $\nu \in V_{\mathbf{x},\mathbf{0}}$, showing that our claim is true. Finally, from Lemma 3.2.3, we know that this is equivalent to having that $\langle \nu, \mathcal{G}(v_0)\rangle = 0$, which is to say that $\mathcal{G}(v_0)$ lies in the annihilator of $V_{\mathbf{x},\mathbf{0}}$, and this last one if equal to $(\Phi_1(X))^{\vdash}$. □

If we assume that $W_2 = W_1^*$ then we will have that $A^{\perp} \subseteq A^{\vdash}$, so the following corollary is a direct consequence of the theorem above.

Corollary 3.2.4.1. Assume the same hypothesis as in the previous theorem, and further let $W_2 = W_1^*$. Then for a set $A \subset W_1$ we have that $A^{\vdash} = A^{\perp}$, therefore $(\Phi_1(X))^{\vdash} = (\Phi_1(X))^{\perp}$ and $\mathcal{G}(v_0) \in \operatorname{Span}\Phi_1(\mathbf{x})$.

Finally to show the Representer theorem for regularized problems we need a lemma to make sure that there is at least one solution, even if it may not lie in the finite dimensional space $S^{\mathbf{x}}$. This result is a particular case of the Generalized Weierstrass Theorem [24].

Lemma 3.2.5. Let B be a reflexive Banach space and $F : B \longrightarrow \mathbb{R} \cup \{\infty\}$ be a lower semi-continuous convex function. If for some $M > 0$ the set $\{x \in B : F(x) \leq M\}$ is non-empty and bounded, then F attains its minimum in B [27].

Now we set up what we need for the regularized problem. We fix a set of samples $\{x_i\}$ and their labels $\{y_i\}$, a continuous and convex loss function $L_{\mathbf{y}}(f(\mathbf{x}) := L(\mathbf{y}, f(\mathbf{x}))$, a continuous, convex, strictly increasing and unbounded function ϕ. For a function f in $\mathscr{B}_1$ and $\lambda \in \mathbb{R}^+$ we define

$$\mathcal{E}_{\mathbf{z},\lambda}(f) := L_{\mathbf{y}}(f) + \lambda\phi(\|f\|_{\mathscr{B}_1}),$$

where $\mathbf{z} = \{(x_i, y_i)_{i=1,\cdots,n}.\}$

Theorem 3.2.6. (Representer theorem for regularized problems in strictly convex and Gateaux differentiable spaces) With the same hypothesis as Theorem 3.2.4, the problem

$$\inf_{f\in\mathscr{B}_1} \mathcal{E}_{\mathbf{z},\lambda}(f) \tag{3.11}$$

has a solution f_{v_0}, where $v_0 \in W_2$ is such that $\mathcal{G}(v_0) \in \left((\Phi_1(X))^{\vdash}\right)^{\perp}$

Proof. We first show the uniqueness of the solution. Assume $f_1, f_2 \in \mathscr{B}_1$ are two different solutions and let $f_3 = \frac{f_1+f_2}{2}$. From Theorem 1.2.2 we know that $\mathscr{B}_1$ is reflexive, Gateaux differentiable and strictly convex space. Therefore, due to the strict convexity of the space, $\|(f_1 + f_2)/2\|_{\mathscr{B}_1} < (\|f_1\|_{\mathscr{B}_1} + \|f_2\|_{\mathscr{B}_1})/2$ and

$$\mathcal{E}_{\mathbf{z},\lambda}(f_3) = L_y(\frac{f_1+f_2}{2}) + \lambda\phi(\|\frac{f_1+f_2}{2}\|_{\mathscr{B}_1}) < L_y(\frac{f_1+f_2}{2}) + \lambda\phi(\frac{\|f_1\|_{\mathscr{B}_1} + \|f_2\|_{\mathscr{B}_1}}{2})$$

$$\leq \frac{L_Y(f_1)}{2} + \frac{L_y(f_2)}{2} + \lambda\left(\frac{\phi(\|f_1\|_{\mathscr{B}_1})}{2} + \frac{\phi(\|f_2\|_{\mathscr{B}_1})}{2}\right).$$

Since f_1 and f_2 minimize (3.11), then the last member is $\mathcal{E}_{\mathbf{z},\lambda}(f_1)$, so we conclude

$$\mathcal{E}_{\mathbf{z},\lambda}(f_3) < \mathcal{E}_{\mathbf{z},\lambda}(f_1),$$

which contradicts the choice of f_1. For the existence, given any $f \in \mathscr{B}_1$ such that $\|f\|_{\mathscr{B}_1} > \phi^{-1}(\frac{\mathcal{E}_{\mathbf{z},\lambda}(0)}{\lambda})$ we have

$$\mathcal{E}_{\mathbf{z},\cdot}(f) \geq \lambda\phi(\|f\|_{\mathscr{B}_1}) > \mathcal{E}_{z,\lambda}(0).$$

Then the minimum over the whole space is the same if we restrict it to the set

$$\{f \in \mathscr{B}_1 : \|f\|_{\mathscr{B}_1} \leq \phi^{-1}(\frac{\mathcal{E}_{\mathbf{z},\lambda}(0)}{\lambda})\} = \{f \in \mathscr{B}_1 : \mathcal{E}_{z,\lambda}(f) \leq \mathcal{E}_{z,\lambda}(0)\}.$$

This set is thus non-empty and $\mathcal{E}_{z,\lambda}$ is convex and continuous, so Lemma 3.2.5 assures that there the minimum is attained in $\mathscr{B}_1$. Let $f_v \in \mathscr{B}_1$ be the solution, with $v \in W_2$. Let $y_i = f_v(x_i)$, by Theorem 3.2.4 there exists a $v_0 \in W_2$ such *that* f_{v_0} interpolates the data $\{y_i\}$ and is solution to the problem of minimum norm, in other words

$$\|f_{v_0}\| \leq \|f_v\|.$$

So f_{v_0} is a solution to the regularized problem, and by uniqueness $f_v = f_{v_0}$. □

We remember a bit about the duality mapping from Chapter 1. We know that for a Gateaux differentiable space, the Gateaux differential defines a bijective mapping from B to its dual. In the case of a Hilbert space, this mapping ends up being the identity map [11]. So the Representer theorem for RKHS hides the use of the duality map. In the general case, the duality map is not even linear, not to mention it is not the identity. Therefore these theorems are a step shy of the actual solution. What they give instead is $J(f)$, where J is the unique duality map induced by the Gateaux differential [17] and f is the actual solution. To read more about duality mappings, we refer the reader to [19] [11] [29].

One approach to solve the ℓ^1 regularization problem posed in the previous section is to first solve the problem for $p_n \to 1$ regularization. This gives solutions lying in reflexive, uniformly convex and Gateaux differentiable spaces. Under some hypothesis these solutions converge to the solution to the ℓ^1 regularization problem.

Theorem 3.2.7. [43] Consider a set of pairwise distinct samples $\{x_1, \cdots, x_m\} \subset X$ and associated labels $\{y_1, \cdots, y_n\} \subset \mathbb{R}$. Let $\mathscr{B}_1, \mathscr{B}_2$ be an adjoint pair of RKBS with a generalized Mercer kernel k induced by expansion sets which satisfy property *A1* and the set $\{k(x,\cdot)\}_{x\in X}$ is linearly independent. We define

$$\mathcal{F}_{\mathbf{z},\lambda}(f) := \frac{1}{N}\sum_{j=1}^{n} L(x_j, y_j, f(x_j)) + R(\|f\|_{\mathscr{B}_1}).$$

If furthermore R is like in Theorem 3.2.6 and $t \mapsto L(x,y,t)$ is convex for fixed $(x,y) \in X \times \mathbb{R}$, then the problem of minimizing $\mathcal{F}$ has a global solution $s_1 \in \mathscr{B}_1$ and there exists a sequence $s_{p_m} \in \mathscr{B}_1^p$, where s_{p_m} are the solutions to the regularized problem (3.11) posed in $\mathscr{B}_1^{p_m}$, such that

$$\lim_{p_m \to 1} s_{p_m}(x) = s_1(x)$$

for every $x \in X$ and

$$\mathcal{F}(s_{p_m}) \to \mathcal{F}(s_1).$$

In [16], the authors developed a RKBS with the intent to show the viability of working with RKBS for learning tasks. What they did is work with the spaces constructed in Section 2.1.1 but considering the bilinear form

$$\langle f, g \rangle_{\mathscr{B}_\Psi^p \times \mathscr{B}_\Psi^q} := \int_{\mathbb{R}^d} \hat{f}\overline{\hat{g}} d\mu.$$

This yields the reproducing kernel $k(\mathbf{x}, \mathbf{y}) := \Phi(\mathbf{x} - \mathbf{y})$. Their version of the Representer theorem is a particular case to this one, but it gives an even more explicit form given the known duality map of $L^p(\mathbb{R}^d, \mu)$ we see in (2.2).

Theorem 3.2.8. Assume the same hypothesis as Theorem 3.2.4, with the RKBS constructed in Theorem 2.1.5 with the bilinear form and kernel defined above. Then the regularized problem

$$\min_{f \in \mathscr{B}_\Psi^p} \frac{1}{m} \sum_{i=1}^{m} L(\mathbf{x}_i, \mathbf{y}_i, f(\mathbf{x}_i)) + R(\|f\|_{\mathscr{B}_\Psi^p})$$

has the form

$$f^*(\mathbf{x}) = \left(\frac{1}{2\pi}\right)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \hat{\Phi}(y)^{p-1} \sum_{k=1}^{m} c_k e^{i(\mathbf{x}-\mathbf{x_k})} \left| \sum_{j=1}^{m} c_j e^{i(\mathbf{x}-\mathbf{x_j})} \right| d\mathbf{y}.$$

If p is an even integer this can be further simplified to [16]:

$$f^*(\mathbf{x}) = \sum_{j_1, \cdots, j_{p-1}=1}^{m} c_{j_1} \overline{c_{j_2}} \cdots c_{j_{p-1}} \Phi(\mathbf{x} - \mathbf{x}_{j_1} + \mathbf{x}_{j_2} - \cdots - \mathbf{x}_{j_{p-1}}).$$

The spaces used in this subsection were all isometric to l^p spaces as seen in Section 2.3. But the case for $p = 1$ is not directly treated here, since this space is not uniformly convex [29]. This means that the dual mapping cannot be defined by Gateaux differentials or semi-inner products in a natural way. Instead we will be imposing conditions on the reproducing kernel to find solutions to the problem of minimal norm interpolation.

3.2.2 Representer theorem for minimum norm interpolation in spaces with ℓ^1 norm.

The following approach we treat here is to solve the problem in the space with ℓ^1 norm. This one imposes conditions on the kernel function to make up for the lack of smoothness and uniform convexity. As a RKBS constructed in Section 2.3 will not be reflexive, it will not be uniformly convex either. This is the reason why the previous approach does not work.

In this case minimal norm interpolation satisfying the Representer theorem is equivalent to the regularized problem satisfying it.

Definition 3.2.3. Let $k : \Omega \times \Omega \longrightarrow \mathbb{R}$ be a kernel function. For a set of examples $x_1, \cdots, x_m$ we define the following the matrix

$$K[\mathbf{x}] := \left(k(x_i, x_j)\right)_{i,j=1,\cdots,m},$$

and the vectorial functions

$$k^{\mathbf{x}}(y) := (k(x_i, y))_{i=1,\cdots,m},$$

$$k_{\mathbf{x}}(y) := (k(y, x_i))^T_{i=1,\cdots,m}.$$

Remark. We do not assume any symmetry of the kernel function, so $k[\mathbf{x}]$ is not necessarily symmetric, and $k_{\mathbf{x}}$ is not necessarily the transpose of $k^{\mathbf{x}}$. But $k[\mathbf{x}]$ can be see as the matrix made of columns $k_{\mathbf{x}}(x_j)$ or rows $k^{\mathbf{x}}(x_j)$.

For the remainder of the section, we will assume that for the regularized problem with sample set $\{x_1, \cdots, x_m\}$

$$\min_{f \in \mathscr{B}_1} L(f(\mathbf{x})) + \lambda\phi(f)$$

both L and ϕ are continuous, L is everywhere finite, ϕ is non-decreasing and

$$\lim_{t\to\infty} \phi(t) = +\infty.$$

Lemma 3.2.9. A RKBS $\mathscr{B}_1$, constructed as in 2.3.5, satisfies the Representer theorem for regularized problems if and only if it satisfies the Representer theorem for minimum norm interpolation.

Proof. Let V, ϕ, λ be as the assumption above. Assume first that the minimum interpolation has a solution for any $\mathbf{y} \in \mathbb{R}^m$. Choose an arbitrary $f \in \mathscr{B}_1$. Then by hypothesis, we can find a solution $f_0 = \sum_{j=1}^{m} a_j k(x_j, \cdot)$ to

$$\inf_{g \in \mathcal{I}_{\mathbf{x}}(f(\mathbf{x}))} \|g\|_{\mathscr{B}_1}.$$

Since it interpolates $f(\mathbf{x})$ we have that $L(f(\mathbf{x})) = L(f_0(\mathbf{x}))$, and it also satisfies $\lambda\phi(\|f\|_{\mathscr{B}_1}) \geq \lambda\phi(\|f_0\|_{\mathscr{B}_1})$ because f_0 has the minimum norm from those functions that interpolate $f(\mathbf{x})$, and the regularizer is non-decreasing. So the next equation follows

$$\inf_{f \in \mathscr{B}_1} L(f(X)) + \lambda\phi(\|f\|_{\mathscr{B}_1}) = \inf_{f \in S^{\mathbf{x}}} L(f(X)) + \lambda\phi(\|f\|_{\mathscr{B}_1}).$$

Now since $\phi(x) \xrightarrow[x\to\infty]{} \infty$, we can restrict the problem to $\{f \in \mathscr{B}_1 : \|f\|_{\mathscr{B}_1} \leq M\}$ for some $M > 0$, thus the following is true

$$\inf_{f \in S^{\mathbf{x}}} L(f(X)) + \lambda\phi(\|f\|_{\mathscr{B}_1}) = \inf_{f \in S^{\mathbf{x}}, \|f\|_{\mathscr{B}_1} \leq M} L(f(X)) + \lambda\phi(\|f\|_{\mathscr{B}_1}).$$

Since $S^{\mathbf{x}}$ is finite dimensional, the set over which the problem is taken is compact and the minimum is reached there.

Now assume that for every regularization problem there is a solution lying in $S^{\mathbf{x}}$. Choose a minimizer $f_{0,\lambda}$ for the regularization problem

$$\inf_{f \in \mathscr{B}_1} \|f(\mathbf{x}) - \mathbf{y}\|_2^2 + \lambda\|f\|_{\mathscr{B}_1}$$

where $\|\cdot\|_2$ is the usual euclidean norm for $\mathbb{R}^n$. The form of each $f_{0,\lambda}$ implies that there exists a set $\mathbf{c}_\lambda \subset \mathbb{R}^m$ such that $f_{0,\lambda}(\cdot) = k^{\mathbf{x}}(\cdot)\mathbf{c}_\lambda$. We now show that the set of $\mathbf{c}_\lambda$ is bounded. Indeed,

$$\|k[\mathbf{x}]\mathbf{c}_\lambda - \mathbf{y}\|_2^2 = \|f_{0,\lambda}(\mathbf{x}) - \mathbf{y}\|_2^2 \leq L(f_{0,\lambda}) + \lambda\|f_{0,\lambda}\|_{\mathscr{B}_1} \leq V(0) + 0 = \|\mathbf{y}\|_2^2.$$

So the triangle inequality implies that

$$\|k[\mathbf{x}]\mathbf{c}_\lambda\|_2 \leq 2\|\mathbf{y}\|_2,$$

thus $\{\mathbf{c}_\lambda\}$ must be bounded in $\mathbb{R}^m$. We choose a sequence of $\mathbf{c}_{\lambda_n}$ as $\lambda_n \to 0$ which converges to a point $\mathbf{c}_0 \in \mathbb{R}^n$ and let f_0 be its associated function. By construction of f_0 and f_{λ_n} we have that

$$\lim_{n\to\infty} \|f_0 - f_{\lambda_n}\|_{\mathscr{B}_1} = \lim_{n\to\infty} \|\mathbf{c}_{\lambda_n} - \mathbf{c}_0\|_1 = 0. \tag{3.12}$$

Let g be an arbitrary interpolant of $\mathbf{y}$. This means that

$$g(\mathbf{x}) = \mathbf{y}.$$

The choice of f_{0,λ_n} and g makes it so that $\lambda_n\|g\|_{\mathscr{B}_1} = \|g(\mathbf{x}) - \mathbf{y}\|_2^2 + \lambda_n\|g\|_{\mathscr{B}_1} \geq \|f_{0,\lambda_n}(\mathbf{x}) - \mathbf{y}\|_2^2 + \lambda_n\|f_{0,\lambda_n}\|_{\mathscr{B}_1}$. By continuity of the evaluation functionals, we have $f_{0,\lambda_n}(x_j) \to f_0(x_j)$, and coupled with the continuity of the chosen V and ϕ we obtain that $f_{0,\lambda_n}(\mathbf{x}) = \mathbf{y}$ by letting λ_n tend to 0. This shows that $f_0 \in \mathcal{I}_\mathbf{x}(\mathbf{y})$ and is the solution to the problem of minimum norm interpolant.

□

This means that we can focus on solving only the minimal norm interpolation problem. One of the first results we need for this is related to assumption A4.

Assumption. *A4* - For any finite subset of pairwise distinct points $x_1, \cdots, x_{n+1} \subset X$ we define the column vector $K_\mathbf{x}(x_{n+1}) := (k(x_{n+1}, x_j))_{j=1,\cdots,n}$. Then

$$\|(K[\mathbf{x}])^{-1} K_\mathbf{x}(x_{n+1})\|_{\ell^1} \leq 1.$$

The following result gives a condition that is equivalent to assumption *A4*. This condition is for the solution of the minimal interpolation problem formulated in $S^\mathbf{x}$ to stay the same even if we add a finite number of dimensions.

Lemma 3.2.10. Let k be a kernel function which satisfies assumptions *A1* to *A3*. Then condition A4 is equivalent to the solution to the minimum norm interpolation in $S^\mathbf{x}$ being the same as the one formulated in $S^{\overline{\mathbf{x}}}$ where $\overline{x} = (x_1, \cdots, x_m, x_{m+1})$ and $x_{m+1} \neq x_i i = 1, \cdots, m$.

Proof. The set $\mathcal{I}_\mathbf{x}(y) \cap S^\mathbf{x}$ consists solely of the function $f = k^\mathbf{x}(\cdot)\,(k[\mathbf{x}])^{-1}\,\mathbf{y}$. A function $g \in \mathcal{I}_\mathbf{x}(y) \cap S^{\overline{\mathbf{x}}}$ is completely determined by where it sends $x_m + 1$. We label $g(x_{m+1}) = c$ and let $\overline{\mathbf{y}} := \left(y^T, c\right)^\mathbb{T}$. Moreover, the function g has the explicit form $g = k^{\overline{\mathbf{x}}}(\cdot)k[\overline{\mathbf{x}}]\overline{\mathbf{y}}$ so we proceed to bound its norm. We consider the matrix

$$k[\overline{\mathbf{x}}]^{-1} = \begin{bmatrix} k[\mathbf{x}] & k_\mathbf{x}(x_{m+1}) \\ k^\mathbf{x}(x_{m+1}) & k(x_{m+1}, x_{m+1}) \end{bmatrix}^{-1}.$$

This can be calculated by blocks [7] which yields the following:

$$k[\overline{\mathbf{x}}]^{-1}\overline{\mathbf{y}} = \begin{pmatrix} k[\mathbf{x}]^{-1}\mathbf{y} + \frac{q}{p}k[\mathbf{x}]^{-1}k_\mathbf{x}(x_{m+1}) \\ -\frac{q}{p} \end{pmatrix}$$

where $p := k(x_{m+1}) - k^\mathbf{x}(x_{m+1})k[\mathbf{x}]^{-1}k_\mathbf{x}(x_{m+1})$ and $q := k^\mathbf{x}(x_{m+1})k[\mathbf{x}]^{-1}\mathbf{y} - c$. The norm for a function in $\mathscr{B}_0$ was defined in Theorem 2.3.2. Also from its definition it

can be shown that

$$\|k[\overline{\mathbf{x}}]^{-1}\overline{\mathbf{y}}\|_{\mathscr{B}_0} = \|k[\mathbf{x}]^{-1}\mathbf{y} + \frac{q}{p}k[\mathbf{x}]^{-1}k_{\mathbf{x}}(x_{m+1})\| + |\frac{q}{p}|.$$

We now estimate $\|g\|_{\mathscr{B}_1} = \|g\|_{\mathscr{B}_0}$:

$$\|g\|_{\mathscr{B}_0} = \|k[\mathbf{x}]^{-1}\mathbf{y} + \frac{q}{p}k[\mathbf{x}]^{-1}k_{\mathbf{x}}(x_{m+1})\|_{\mathscr{B}_0} + |\frac{q}{p}| \geq$$

$$\|k[\mathbf{x}]^{-1}\mathbf{y}\|_{\mathscr{B}_0} - |\frac{q}{p}|\|k[\mathbf{x}]^{-1}k_{\mathbf{x}}(x_{m+1})\|_{\mathscr{B}_0}| + \frac{q}{p}|.$$

If k satisfies 3.2.2 then $|\frac{q}{p}| - |\frac{q}{p}|\left(\|k[\mathbf{x}]^{-1}k_{\mathbf{x}}(x_{m+1})\|_{\mathscr{B}_0}\right) \geq 0$ thus

$$\|g\|_{\mathscr{B}_1} \geq \|k[\mathbf{x}]^{-1}\mathbf{y}\|_{\mathscr{B}_0} = \|f\|_{\mathscr{B}_1}.$$

This means that

$$\min_{f\in\mathcal{I}_{\mathbf{x}}(y)\cap S^{\overline{\mathbf{x}}}}\|f\|_{\mathscr{B}} \geq \min_{f\in\mathcal{I}_{\mathbf{x}}(y)\cap S^{\overline{\mathbf{x}}}}\|f\|_{\mathscr{B}},$$

so we can conclude these are equal.

Conversely, suppose that

$$\min_{f\in\mathcal{I}_{\mathbf{x}}(y)\cap S^{\overline{\mathbf{x}}}}\|f\|_{\mathscr{B}} = \min_{f\in\mathcal{I}_{\mathbf{x}}(y)\cap S^{\overline{\mathbf{x}}}}\|f\|_{\mathscr{B}}$$

for any vector $\mathbf{y} \in \mathbb{R}^m$. Then we choose $\mathbf{y} = k_{\mathbf{x}}(x_{m+1})$ and $c = k^{\mathbf{x}}(x_{m+1})k[\mathbf{x}]^{-1}k_{\mathbf{x}}^T(x_{m+1})$. By doing this we get the following norms:

$$\|k[\overline{\mathbf{x}}]^{-1}\overline{\mathbf{y}}\|_{\mathscr{B}_1} = \|\begin{pmatrix}0\\1\end{pmatrix}\|_1 = 1 \text{ and } \|k[\mathbf{x}]^{-1}\mathbf{y}\|_{\mathscr{B}_1} = \|\left(k[\mathbf{x}]\right)^{-1}k_{\mathbf{x}}(x_{m+1})\|_{\mathscr{B}_1}. \tag{3.13}$$

And finally by the property of the being the function of minimum norm we have

$$\|\left(k[\mathbf{x}]\right)^{-1}k_{\mathbf{x}}(x_{m+1})\|_{\mathscr{B}_1} \leq 1.$$

□

This theorem can be applied repeatedly to show that adding a finite amount of does not change the solution.

Theorem 3.2.11. (Representer theorem for spaces with ℓ^1 norm.) Every solution to the minimum norm interpolation posed on a space constructed as in Section 2.3.5 can be represented as $\sum_i \alpha_i k(x_i, \cdot)$, if and only if, k satisfies assumption A4.

Proof. Assume we have condition A4. We will show that $\|g\|_{\mathscr{B}_1} \geq \min_{f\in\mathcal{I}_{\mathbf{x}}(y)\cap S^{\mathbf{x}}}\|f\|_{\mathscr{B}_1}$ for $g \in \mathscr{B}_0$. If we show that, then for an arbitrary $g \in \mathscr{B}_1$ we can choose a sequence $g_j \in \mathscr{B}_0$ which converges to it. Then by the continuity of the norm and the previous inequality we have that

$$\|g\|_{\mathscr{B}_1} = \lim_{n\to\infty}\|g_n\|_{\mathscr{B}_1} \geq \min_{f\in\mathcal{I}_{\mathbf{x}}(y)\cap S^{\mathbf{x}}}\|f\|_{\mathscr{B}_1}.$$

From where we deduce that

$$\min_{f\in\mathcal{I}_{\mathbf{x}}(y)\cap S^{\mathbf{x}}}\|f\|_{\mathscr{B}_1} = \min_{f\in\mathcal{I}_{\mathbf{x}}(y)}\|f\|_{\mathscr{B}_1}.$$

Let $g \in \mathscr{B}_0 \cap \mathcal{I}_{\mathbf{x}}(y)$ be arbitrary. By construction of $\mathscr{B}_0$, by possibly adding zero coefficients along with extra sample points x_j, g has an expression of the form $\sum_{j=1}^{l} \alpha_j k(x_j, \cdot)$ for one $l \geq m$. We define for $1 \leq i \leq l$ the vector $\mathbf{u}_i := (g(x_n) : 1 \leq n \leq i)$ and the set $\mathbf{v}_i := \{x_n : 1 \leq n \leq i\}$. With this we see that $g \in \mathcal{I}_{\mathbf{v}_l} \cap S^{\mathbf{u}_l}$, therefore the following is true:

$$\|g\|_{\mathscr{B}_1} \geq \min_{f \in \mathcal{I}_{\mathbf{v}_l}(\mathbf{u}_l) \cap S^{\mathbf{u}_l}} \|f\|_{\mathscr{B}_1} \geq \min_{f \in \mathcal{I}_{\mathbf{v}_{l-1}}(\mathbf{u}_{l-1}) \cap S^{\mathbf{u}_l}} \|f\|_{\mathscr{B}_1}.$$

Last inequality is justified because $\mathcal{I}_{\mathbf{v}_l}(\mathbf{u}_l) \subseteq \mathcal{I}_{\mathbf{v}_{l-1}}(\mathbf{u}_l - 1)$. Now by applying Lemma 3.2.10 we get that

$$\|g\|_{\mathscr{B}_1} \geq \min_{f \in \mathcal{I}_{\mathbf{v}_{l-1}}(\mathbf{u}_{l-1}) \cap S^{\mathbf{u}_l}} \|f\|_{\mathscr{B}_1} = \min_{f \in \mathcal{I}_{\mathbf{v}_{l-1}}(\mathbf{u}_{l-1}) \cap S^{\mathbf{u}_{l-1}}} \|f\|_{\mathscr{B}_1}.$$

We repeat the same argument until we reach u_m, so we get the following result:

$$\|g\|_{\mathscr{B}_1} \geq \min f \in \mathcal{I}_{v_m}(u_m) \cap S^{v_m} \|f\|_{\mathscr{B}_1} = \min_{f \in \mathcal{I}_{\mathbf{x}}(y) \cap S^{\mathbf{x}}} \|f\|_{\mathscr{B}_1}.$$

Thus it can be concluded that

$$\min_{f \in \mathcal{I}_{\mathbf{x}}(y) \cap S^{\mathbf{x}}} \|f\|_{\mathscr{B}_1} = \min_{f \in \mathcal{I}_{\mathbf{x}}(y)} \|f\|_{\mathscr{B}_1}$$

by using properties of the min function. For the other implication, note that every minimal norm interpolant being in $S^{\mathbf{x}}$ is equivalent to

$$\min_{f \in \mathcal{I}_{\mathbf{x}}(y) \cap S^{\mathbf{x}}} \|f\|_{\mathscr{B}_1} = \min_{f \in \mathcal{I}_{\mathbf{x}}(y)} \|f\|_{\mathscr{B}_1}.$$

And since $\mathcal{I}_{\mathbf{x}}(\mathbf{y}) \cap S^{\mathbf{x}} \subseteq \mathcal{I}_{\mathbf{x}}(\mathbf{y}) \cap S^{\overline{\mathbf{x}}} \subset \mathcal{I}_{\mathbf{x}}$ for every $\mathbf{y} \in \mathbb{R}^m$, if we apply the minimum of the norms over these sets, we obtain the assumption A4 by Lemma 3.2.10. □

As examples of kernels which satisfy conditions *A1* through *A4*, the authors of [38] show that the **Brownian bridge kernel**

$$k(x, y) := min\{x, y\} - xy \ , x, y \in (0, 1)$$

and the **exponential kernel**

$$k(x, y) := e^{-|x-y|} \ , x, y \in \mathbb{R}$$

both satisfy all the requirements. They also point out that other known kernels like the Gaussian kernel do not verify assumption *A4*.

Appendix A

Appendix A

A.1 Fourier transform and Positive definite functions

In this section we give a summary of results about the Fourier Transform and positive definite functions that we need to develop the theory of RKBS.

First are the space of Schwarz and its dual.

Definition A.1.1. We denote by $C^\infty(\mathbb{R}^m)$ the space of infinitely differentiable functions.

Definition A.1.2. [22] The Schwartz space $\mathcal{S}$ or $\mathcal{S}(\mathbb{R}^m)$ is the subspace of $\mathcal{C}^\infty(\mathbb{R}^m)$ of functions f such that

$$\sup_{x\in\mathbb{R}^m} |x^\beta \partial^\alpha f(x)| < \infty$$

where α, β are any pair of multi-indices. Its topology is defined by the seminorms

$$|x^\beta \partial^\alpha f(x)|.$$

Its continuous dual space $\mathcal{S}'$ will be referred to as the space of temperate distributions.

We follow with the definition of the Fourier transform and its inverse on a function in $L_1(\mathbb{R}^d)$.

Definition A.1.3. [41] Let $f \in L_1(\mathbb{R}^d)$ with respect to Lebesgue measure. We define its Fourier transform by:

$$\hat{f}(\omega) := (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} f(x) e^{-i\langle\omega,x\rangle} dx$$

and its inverse

$$\check{f}(\omega) := (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} f(x) e^{i\langle\omega,x\rangle} dx.$$

If μ is a tempered distribution, then we define its distributional Fourier transform by the equation [22] :

$$\hat{\mu}(\phi) := \mu(\hat{\phi}).$$

This tool is necessary for establishing some properties of positive definite functions. It was also used extensively for the RKBS defined in section 2.1.1.

Theorem A.1.1. [22] The distributional Fourier transform is an isomorphism of $\mathcal{S}'$ and we have that Fourier's inversion formula $\hat{\hat{\phi}} = \phi$.

Definition A.1.4. A function $\Phi : \mathbb{R}^d \longrightarrow \mathbb{C}$ is called a positive definite function if it is continuous and for any $x_1, \cdots, x_j \in \mathbb{R}^d$, and $\alpha_1, \cdots, \alpha_j \in \mathbb{C}$ the following is always

true

$$\sum_{i,j=1}^{n} \alpha_i \overline{\alpha_j} \Phi(x_i - x_j) > 0.$$

The following results are characterizations of positive definite functions in terms of the Fourier transform.

Theorem A.1.2. *(Bochner's Characterization of positive definite functions)* A continuous function $\Phi : \mathbb{R}^d \to \mathbb{C}$ is a positive definite function if and only if it is the Fourier transform of a finite non-negative Borel measure on $\mathbb{R}^d$ [41].

Theorem A.1.3. Assume $\Phi : \mathbb{R}^d \to \mathbb{C}$ is absolutely integrable and continuous. Then it is positive definite if and only if it is bounded and its Fourier transform is non-negative and non-vanishing everywhere.[41]

This last theorem was what allowed us to define the $\mathscr{B}^p_\Phi(\mathbb{R}^d)$ spaces. It also allowed us to construct the isomorphism to $L^q(\mathbb{R}^d, \mu)$.

A.2 Banach spaces and Optimization

In this section, the field $\mathbb{K}$ represents either $\mathbb{R}$ or $\mathbb{C}$.

Definition A.2.1. A norm on a Vector spaces B is a function defined on $\|\| : B \longmapsto \mathbb{R}$ with the following properties:

- $\|x\| \geq 0 \; \forall x \in B$ and $\|x\| = 0 \iff x = 0$.
- $\|\lambda x\| = |\lambda| \|x\|, \; \forall x \in B, \lambda \in \mathbb{K}$.
- $\|x + y\| \leq \|x\| + \|y\|$.

Definition A.2.2. A $\mathbb{K}$ vector space B is called a Banach space if it is endowed with a norm $\|\cdot\|_B : B \longrightarrow \mathbb{R}$ such that every Cauchy sequence converges with respect to the norm.

One of the most important aspects of this approach to RKBS is the emphasis on the natural duality between a Banach space and its continuous dual.

Definition A.2.3. The dual space B^* of a Banach space B is defined as the space of all the continuous linear functionals. It is also a Banach space with the norm

$$\|f\|_{B'} := \sup_{\|x\|=1} |f(x)|.$$

From Definition 1.2.4 and the discussion following shortly after 1.2.2 we know that Construction 1.2.1.1 asks for an embedding of Banach spaces. It follows easily from this definition and the non-degeneracy of the bilinear form.

Definition A.2.4. A bilinear form defined on a pair (V, W) of vector spaces is a function $\langle \cdot, \cdot \rangle : V \times W \longmapsto \mathbb{K}$ such that for any vectors $v_1, v_2 \in V$, $w_1, w_2 \in W$ and scalars $\alpha, \beta \in \mathbb{K}$:

- $\langle v_1 + \alpha v_2, w_1 \rangle = \langle v_1, w_1 \rangle + \alpha \langle v_2, w_1 \rangle$.
- $\langle v_1, w_1 + \beta w_2 \rangle = \langle v_1, w_1 \rangle + \beta \langle v_1, w_2 \rangle$.

A class of RKBS that gets used often due to their geometric properties is the class of Reflexive Banach spaces.

Definition A.2.5. A Banach space B is said to be reflexive if it is isometrically isomorphic to its double dual $(B^*)^*$.

We generalized many arguments from their Hilbert space version. One of the tools to translate orthogonality was the annihilator of a subset.

Definition A.2.6. Let V and W be Banach spaces and $\langle \cdot, \cdot \rangle_{V \times W}$ a bilinear form such that V and W are dense with respect to it. For subspaces $M \subset V$ $N \subset W$ we define their annihilators $M^\perp$ and ${}^\perp N$ as [1]:

$$M^\perp := \{g \in W : \langle f, g \rangle_{V \times W} = 0, \ \forall f \in V\},$$

$${}^\perp N := \{f \in V : \langle f, g \rangle_{V \times W} = 0, \ \forall g \in W\}.$$

Theorem A.2.1. If V and W are Banach spaces with a bilinear map $\langle \cdot, \cdot \rangle : v \times W \longmapsto \mathbb{K}$ and $M_1, M_2 \subset V$, the following properties can be established:

- If $M_1 \subset M_2$ then $M_2^\perp \subset M_1^\perp$.
- $M \subset M^{\perp\perp}$.
- $M^\perp = M^{\perp\perp\perp}$.

A set which has this last property is said to be orthogonally closed with respect to the bilinear form.[29] Furthermore, if W is the dual space of V and M is a subspace, then the following properties hold.

- There exists an isometric isomorphism that identifies the dual space M^* with $V^*/M^\perp$ such that an element $x^* \in M^*$ identified with $x^* + M^\perp$ has the following action on the elements m of M:

$$(x + M^*)(m) = x^*(m) = \langle m, x^* \rangle.$$

- If M is closed then there exists an isometric isomorphism that identifies $(X/M)^*$ with $M^\perp$ such that an element in $(X/M)^*$ identified with $x^* \in M^\perp$ has the following action on the elements $m + M$ of X/M:

$$x^*(m + M) = x^*(m) = \langle m, x^* \rangle.$$

An important property of Banach spaces is that their dual space has always "enoguh" functionals to separate different points.

Theorem A.2.2. Hahn-Banach Let V be a Banach space over $\mathbb{K}$ and f a bounded linear functional defined over a subspace $Z \subseteq V$. Then there exists a bounded linear functional $F \in V^*$ such that

$$f(z) = F(z), \ \forall z \in Z$$

and

$$\|f\|_{Z^*} = \|F\|_{V^*}.$$

Another prominent tool is the semi-inner product.

Definition A.2.7. Let $\mathscr{B}$ be a Banach space. A function $[\cdot,\cdot] : \mathscr{B} \times \mathscr{B} \to \mathbb{K}$ is called a **semi-inner product** (SIP) if it has the next properties:

- $[x,x] \geq 0 \; \forall x \in \mathscr{B}$ and $[x,x] = 0 \iff x = 0$
- $[\lambda x, y] = \lambda [x,y]$ and $[x, \lambda y] = \overline{\lambda}[x,y] \; \forall x, y \in \mathscr{B}, \lambda \in \mathbb{K}$.
- $[x+z, y] = [x,y] + [z,y] \; \forall x, y, z \in \mathscr{B}$.
- $|[x,y]| \leq [x,x]^2 [y,y]^2$

  A semi-inner product induces a norm by defining

$$\|f\|_{[\cdot,\cdot]} := ([f,f])^{1/2}.$$

As a consequence of the Hahn-Banach theorem, there always exists a semi-inner product that induces the norm.

Definition A.2.8. Let $\mathscr{B}$ be a Banach space. A function $\langle \cdot,\cdot \rangle : \mathscr{B} \times \mathscr{B} \to \mathbb{K}$ is called an **inner product** on B if it has the following properties:

- $\langle x,x \rangle \geq 0 \; \forall x \in \mathscr{B}$ and $\langle x,x \rangle = 0 \iff x = 0$
- $\langle \lambda x, y \rangle = \lambda \langle x,y \rangle$.
- $\langle x+z, y \rangle = \langle x,y \rangle + \langle z,y \rangle \; \forall x, y, z \in \mathscr{B}$.
- $\langle x,y \rangle = \overline{\langle y,x \rangle}$ for all $x, y \in \mathscr{B}$.

An inner product defines a norm $\|\cdot\|_{\langle \cdot,\cdot \rangle}$ on B by the formula

$$\|x\|^2_{\langle \cdot,\cdot \rangle} := \langle x,x \rangle.$$

A Banach space with an inner product which induces its norm will be called a Hilbert space.

The only difference between these two concepts is the linearity of the second argument. This has some geometric implications.

Theorem A.2.3. A semi-inner product on a $\mathbb{K}$ vector space is an inner product if and only if it is linear on its second argument [45]. Equivalently, a semi-inner product is an inner product if and only if its induced norm verifies the parallelogram law [28] :

$$2\|x\|^2 + 2\|y\|^2 = \|x+y\|^2 + \|x-y\|^2.$$

Since a semi-inner product lacks structure to replicate every property of an inner product, we need to make more use of the norm to define orthogonality.

Definition A.2.9. Let B be a Banach space. We say that $x \in B$ is orthogonal to $y \in B$ if for any scalar quantity λ we have

$$\|x\|_B \leq \|x + \lambda y\|_B.$$

If B is a Hilbert space then this orthogonality is equivalent to the following definition.

Definition A.2.10. Let H be a Hilbert space. We say that $x \in H$ is orthogonal to $y \in H$ if we have

$$\langle x,y \rangle = 0.$$

We defined smoothness in terms of functionals (2.2.2, and later mentioned how it is equivalent to a condition on the duality map and to differentiability of the norm. Thus we can expect that the geometric structure of the Banach space affects the geometric structure of its dual.

Definition A.2.11. [29] A Banach space is said to be strictly convex if $\|tx_1 + (1-t)x_2\| < 1$ for any $0 \leq t \leq 1$ and any unitary x_1, x_2.

Definition A.2.12. A Banach space B is **uniformly convex** if for every two sequences $\{x_n\}, \{y_n\} \subset B$ such that $\|x_n\|, \|y_n\| \leq 1$ and $\|x_n + y_n\| \longrightarrow 2$, we have that $\lim_{n\longrightarrow\infty}\|x_n - y_n\| = 0$.

Theorem A.2.4. A Banach space is uniformly smooth if and only if its dual is strictly convex and it is strictly convex if and only if its dual is uniformly smooth [6].

Theorem A.2.5. A Banach space is smooth if and only if the Gateaux differential exists at every point in any direction [23].

Theorem A.2.6. Milman-Pettis Every strictly convex space is reflexive [29].

In the last sections we dealt with functions from the Banach space to $\mathbb{R}$. To ensure the existence of solutions or convergence to solutions we needed the convexity and weak lower-semicontinuity.

Definition A.2.13. Let B be a Banach space. A function $f : B \to \mathbb{R} \cup \{\infty\}$ is a :

- Convex function if for $0 \leq t \leq 1$ and $x, y \in B$:

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

- Lower semi-continuous function if

$$\{x \in B : f(x) > c\}$$

  is open for every $c \in \mathbb{R}$.

A result in analysis says that any continuous function from a compact to $\mathbb{R} \cup \{\infty\}$ meets its infimum. The following result is a generalization to lower semi-continuous functions.

Theorem A.2.7. (Generalized Weierstrass Theorem) Let B be a reflexive Banach space and $f : A \subseteq B \to \mathbb{R} \cup \{\infty\}$ a weakly lower semicontinuous function, where A is a bounded and weakly sequentially closed subset. Then f attains its minimum in A [24]

Bibliography

[1] C.D. Aliprantis and K.C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2007.

[2] J.P Antoine and K Gustafson. "Partial inner product spaces and semi-inner product spaces". In: *Advances in Mathematics* 41.3 (1981), pp. 281–300.

[3] Andreas Argyriou, C. Micchelli, and M. Pontil. "When is there a representer theorem? Vector versus matrix regularizers". In: *J. Mach. Learn. Res.* 10 (2009), pp. 2507–2529.

[4] Andreas Argyriou, Charles A. Micchelli, and Massimiliano Pontil. "When Is There a Representer Theorem? Vector Versus Matrix Regularizers". In: *J. Mach. Learn. Res.* 10 (Dec. 2009), pp. 2507–2529.

[5] Francesca Bartolucci et al. *Understanding neural networks with reproducing kernel Banach spaces*. 2021. arXiv: 2109.09710 [stat.ML].

[6] B. Beauzamy. *Introduction to Banach Spaces and their Geometry*. Information Research and Resource Reports. North-Holland, 1985.

[7] D.S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas (Second Edition)*. Princeton reference. Princeton University Press, 2009.

[8] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. "A Training Algorithm for Optimal Margin Classifiers". In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. COLT '92. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 1992, pp. 144–152.

[9] Liangzhi Chen and Haizhang Zhang. "Margin Error Bounds for Support Vector Machines on Reproducing Kernel Banach Spaces". In: *Neural Computation* 29.11 (Nov. 2017), pp. 3078–3093. eprint: https://direct.mit.edu/neco/article-pdf/29/11/3078/1024162/neco_a_01013.pdf.

[10] Liangzhi Chen and Haizhang Zhang. "Margin Error Bounds for Support Vector Machines on Reproducing Kernel Banach Spaces". In: *Neural Computation* 29.11 (2017), pp. 3078–3093.

[11] C. Chidume. *Geometric Properties of Banach Spaces and Nonlinear Iterations*. Lecture Notes in Mathematics. Springer London, 2009.

[12] I. Cioranescu. *Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems*. Mathematics and Its Applications. Springer Netherlands, 2012.

[13] J.B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer New York, 2019.

[14] N. Cristianini et al. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[15] Frank Deutsch. "Linear selections for the metric projection". In: *Journal of Functional Analysis* 49.3 (1982), pp. 269–292.

[16] Gregory E Fasshauer, Fred J Hickernell, and Qi Ye. "Solving support vector machines in reproducing kernel Banach spaces with positive definite functions". In: *Applied and Computational Harmonic Analysis* 38.1 (2015), pp. 115–139.

[17] Francisco Garcia-Pacheco, Alejandro Miralles, and Daniele Puglisi. "Selectors of the duality mapping". In: *Mathematical Proceedings of the Royal Irish Academy* 116A (Jan. 2016), p. 105.

[18] PANDO G GEORGIEV, LUIS SÁNCHEZ-GONZÁLEZ, and PANOS M PARDALOS. "REPRODUCING KERNEL BANACH SPACES". In: ().

[19] J. Giles. "Classes of semi-inner-product spaces". In: *Transactions of the American Mathematical Society* 129 (1967), pp. 436–446.

[20] H. Haghshenas, A. Assadi, and T. D. Narang. "A look at proximinal and Chebyshev sets in Banach spaces". In: *Le Matematiche* 69 (2014), pp. 71–87.

[21] M. Hein, Olivier Bousquet, and Bernhard Schölkopf. "Maximal Margin Classification for Metric Spaces". In: *Journal of Computer and System Sciences, v.71, 333-359 (2005)* 71 (Jan. 2003).

[22] L. Hörmander. *The Analysis of Linear Partial Differential Operators I: Distribution Theory and Fourier Analysis*. Classics in Mathematics. Springer Berlin Heidelberg, 2015.

[23] G. Köthe. *Topological Vector Spaces*. Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen n.º 1. Springer-Verlag, 1969.

[24] A.J. Kurdila and M. Zabarankin. *Convex Functional Analysis*. Systems & Control: Foundations & Applications. Birkhäuser Basel, 2006.

[25] L. Li. *Selected Applications of Convex Optimization*. Springer Optimization and Its Applications. Springer Berlin Heidelberg, 2015.

[26] Zheng Li, Yuesheng Xu, and Qi Ye. "Sparse Support Vector Machines in Reproducing Kernel Banach Spaces". In: *Contemporary Computational Mathematics - A Celebration of the 80th Birthday of Ian Sloan*. Ed. by Josef Dick, Frances Y. Kuo, and Henryk Woźniakowski. Springer International Publishing, 2018, pp. 869–887.

[27] Rongrong Lin, Haizhang Zhang, and Jun Zhang. *On Reproducing Kernel Banach Spaces: Generic Definitions and Unified Framework of Constructions*. 2019. arXiv: 1901.01002 [cs.LG].

[28] G. Lumer. "SEMI-INNER-PRODUCT SPACES". In: *Transactions of the American Mathematical Society* 100 (1961), pp. 29–43.

[29] Robert E Megginson. *An Introduction to Banach Space Theory*. Vol. 183. Springer Science & Business Media, 1998.

[30] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning, second edition*. Adaptive Computation and Machine Learning series. MIT Press, 2018.

[31] Houman Owhadi and Clint Scovel. "Separability of reproducing kernel spaces". In: *Proceedings of the American Mathematical Society* 145.5 (2017), pp. 2131–2138.

[32] W. Rudin. *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1991.

[33] Saburou Saitoh and Yoshihiro Sawano. *Theory of reproducing kernels and applications*. Springer, 2016.

[34] Kevin Schlegel. "When is there a representer theorem?" In: *Journal of Global Optimization* 74.2 (2019), pp. 401–415.

[35] B. Schölkopf et al. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive computation and machine learning. MIT Press, 2002.

[36] Thomas Schuster et al. *Regularization Methods in Banach Spaces:* De Gruyter, 2012.

[37] Guohui Song and Haizhang Zhang. "Reproducing kernel Banach spaces with the ℓ^1 norm II: Error analysis for regularized least square regression". In: *Neural computation* 23.10 (2011), pp. 2713–2729.

[38] Guohui Song, Haizhang Zhang, and Fred J. Hickernell. "Reproducing kernel Banach spaces with the ℓ^1 norm". In: *Applied and Computational Harmonic Analysis* 34.1 (2013), pp. 96–116.

[39] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[40] Vladimir Vapnik and Alexey Chervonenkis. *Theory of pattern recognition*. 1974.

[41] Holger Wendland. *Scattered data approximation*. Vol. 17. Cambridge university press, 2004.

[42] Matthew A. Wright and Joseph E. Gonzalez. *Transformers are Deep Infinite-Dimensional Non-Mercer Binary Kernel Machines*. 2021. arXiv: 2106.01506 [cs.LG].

[43] Yuesheng Xu and Qi Ye. *Generalized Mercer kernels and reproducing kernel Banach spaces*. Vol. 258. 1243. American Mathematical Society, 2019.

[44] Qi Ye Ying Lin Rongrong Lin. "Sparse regularized learning in the reproducing kernel banach spaces with the ℓ^1 norm". In: *Mathematical Foundations of Computing* 3.3 (2020), pp. 205–218.

[45] Haizhang Zhang, Yuesheng Xu, and Jun Zhang. "Reproducing Kernel Banach Spaces for Machine Learning." In: *Journal of Machine Learning Research* 10.12 (2009).

[46] Shenglong Zhou. "Sparse SVM for Sufficient Data Reduction". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Apr. 2021), pp. 1–11.