T. KAYSI

ATILIM UNIVERSITY 2024

A SURVEY ON REPRODUCING KERNEL HILBERT SPACES

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ATILIM UNIVERSITY

TUBA KAYSI

A MASTER OF SCIENCE THESIS
IN
THE DEPARTMENT OF MATHEMATICS

JUNE 2024

A SURVEY ON REPRODUCING KERNEL HILBERT SPACES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ATILIM UNIVERSITY

BY

TUBA KAYSI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
MATHEMATICS

JUNE 2024

Approval of the Graduate School of Natural and Applied Sciences, Atilim University.

Prof. Dr. Ender KESKİNKILIÇ
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of **Master of Science in Mathematics Department, Atilim University**.

Prof. Dr. Ayhan AYDIN
Head of Department

This is to certify that we have read the thesis A SURVEY ON REPRODUCING KERNEL HILBERT SPACES submitted by TUBA KAYSI and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Prof. Dr. Ferihe ATALAN
Co-Supervisor

Asst. Prof. Dr. Serdar AY
Supervisor

Examining Committee Members:

Prof. Dr. Aurelian B.N. GHEONDEA E.
Mathematics, Bilkent University

Asst. Prof. Dr. Serdar AY
Mathematics, Atılım University

Asst. Prof. Dr. Emel YILDIRIM KAVGACI
Mathematics, Atılım University

Date: June 24, 2024

I declare and guarantee that all data, knowledge and information in this document has been obtained, processed and presented in accordance with academic rules and ethical conduct. Based on these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : TUBA KAYSI

Signature :

ABSTRACT

A SURVEY ON REPRODUCING KERNEL HILBERT SPACES

KAYSI, Tuba

M.S., Department of Mathematics

Supervisor : Asst. Prof. Dr. Serdar AY

Co-Supervisor : Prof. Dr. Ferihe ATALAN

June 2024, 50 pages

The content of this thesis consists of general information about reproducing kernel Hilbert spaces, which are widely used in various fields such as Mathematics, Statistics, and machine learning. In this study, we first introduced the concept of a reproducing kernel Hilbert space (shortly RKHS) and provided the definition of a reproducing kernel. We also discussed the characteristic property of the reproducing kernels, gave the statement and a brief proof of the Moore-Aronszajn Theorem, which is one of the classical theorems in the theory of reproducing kernel Hilbert spaces. Next, we explored how to construct a reproducing kernel Hilbert space from a given kernel function in some concrete cases. Finally, we briefly discussed several applications of reproducing kernel Hilbert spaces, including their use in interpolation-approximation theory, statistics, and machine learning.

Keywords: RKHS, reproducing kernel, positive semidefinite kernel, Moore-Aronsazjn Theorem, interpolation-approximation, machine learning.

ÖZ

DOĞURAN ÇEKİRDEKLİ HİLBERT UZAYLARI ÜZERİNE BİR İNCELEME

KAYSI, Tuba

Yüksek Lisans, Matematik

Tez Yöneticisi : Dr. Öğr. Üyesi Serdar AY

Ortak Tez Yöneticisi : Prof. Dr. Ferihe ATALAN

Haziran 2024, 50 sayfa

Bu tezin içeriği, Matematik, İstatistik ve makine öğrenmesi gibi pek çok alanda önemli bir araç olarak kullanılan doğuran çekirdekli Hilbert uzayları ile ilgili genel bilgilerden oluşmaktadır. Bu çalışmada ilk olarak doğuran çekirdekli Hilbert uzayı (kısaca DÇHU) ve doğuran çekirdek tanımı verildi ve birkaç DÇHU örneğinden bahsedildi. Doğuran çekirdeğin temel karakteristik özelliğine, doğuran çekirdekli Hilbert uzayları teorisinin klasik teoremlerinden biri olan Moore-Aronszajn Teoremi'nin ifadesine ve kısaca ispatına değinildi. Sonrasında bir çekirdek fonksiyonu verildiğinde nasıl doğuran çekirdekli bir Hilbert uzayı inşa edildiğine bakıldı. Son olarak, doğuran çekirdekli Hilbert uzaylarının bazı uygulamaları tartışıldı. Bunlardan ilki interpolasyon ve yaklaşım teorisi üzerine diğeri ise İstatistik ve makine öğrenmesi üzerine uygulamalarıdır.

Anahtar Kelimeler: DÇHU, doğuran çekirdek, pozitif yarıtanımlı çekirdek, Moore-Aronszajn Teoremi, interpolasyon-yaklaşım, makine öğrenmesi.

To my family; my mother Hacer, my father Hüsamettin, and my siblings Sümeyye, Ramazan, and Hilal.

ACKNOWLEDGMENTS

I would like to express my greatest thank and appreciation to my supervisor Asst. Prof. Dr. Serdar Ay for his constant guidance, support, and invaluable patience. I could not complete my thank to him without mentioning his assistance at every stage of this journey. I am equally thankful to my co-supervisor Prof. Dr. Ferihe Atalan for her continuous support, encouragement, and to be in every meeting with insightful comments.

In addition to my supervisor, I would extend my appreciation to the members of my thesis committee: Prof. Dr. Aurelian B.N. Gheondea E., and Asst. Prof. Dr. Emel Yıldırım Kavgacı for their valuable feedback.

I shall also thank all instructors and staff in our department for their cheerful faces. A special thank is to Prof. Dr. Sofiya Ostrovska for her belief in me since the first day of this journey. My sincere gratitude goes to our head of department Prof. Dr. Ayhan Aydın for his aid in getting a full scholarship from the college.

Further acknowledgment is made for the financial assistance provided by The Scientific and Technical Research Council of Turkey (TÜBİTAK) (BİDEB 2210-A).

Lastly, my deepest gratitude goes to my family: to my mom Hacer who always makes me feel confident with her prayers, to my father Hüsamettin, and my siblings Sümeyye, Ramazan, and Hilal for their endless love, consistent encouragement, and steadfast belief. Their faith in me has kept my spirit and motivation high. Without their support, I could not start anything. My biggest supporters, I am truly grateful to them.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The theory of reproducing kernel Hilbert space is quite remarkable and has many applications in various areas such as complex analysis, interpolation-approximation, theory of integral operators, statistics and machine learning. The term reproducing kernel Hilbert space refers to a Hilbert space of functions with a reproducing kernel.

Introduction of the kernels dates back to beginning of the twentieth century. In 1904 D. Hilbert [1] defined the definite kernel. Later, the reproducing kernels are seen in Zeramba's paper in 1909 [2]. In the mean time, Hilbert space was introduced [3]. The arise of theory of the reproducing kernel Hilbert space was in 1920's with the studies of S. Szegö in 1921 [4] and S. Bergman in 1922 [5] which were mostly about the Bergman and Szegö kernels in Complex Analysis. N. Aronszajn's work [6] in 1950 is a milestone in development of the theory. His paper constructs the general theory of the reproducing kernel Hilbert spaces. About fifteen years later L. Schwartz in his work [7] also notably improved the theory as given in [8].

This thesis is based on Paulsen and Raghupathi's 2006 book [9]. In the first chapter, we define basic definitions of the theory of RKHS and discuss some examples of RKHS. In the second part, we look at the construction of RKHS from a given kernel. We state and prove the Moore-Aronszajn Theorem and consider some concrete examples of the theorem. Last two chapters are devoted to applications of RKHS to interpolation-approximation problems, and basic statistics and machine learning concepts.

1.1 Fundamental Definitions

Reproducing kernel Hilbert space is a Hilbert space which consists of only functions defined on a set. Formally, it can be defined as in Definition 1.1.1 without referring directly to the reproducing property. But before going to the definition, note that for a nonempty set X and the scalar field $\mathbb{F}$ the set $\mathcal{F}(X,\mathbb{F}) \coloneqq \{f \in \mathbb{F} \mid f : X \to \mathbb{F}\}$ is a vector space over the field $\mathbb{F}$ which is $\mathbb{C}$ or $\mathbb{R}$ as it is nonempty, and closed under addition and scalar multiplication.

Definition 1.1.1 *Let X be a nonempty set and let $\mathcal{H} \subseteq \mathcal{F}(X,\mathbb{F})$ be a subset of the function space. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space, shortly written RKHS, if the following conditions are satisfied:*
(i) $\mathcal{H}$ is a vector subspace of $\mathcal{F}(X,\mathbb{F})$.
(ii) $\mathcal{H}$ has an inner product, $\langle\cdot,\cdot\rangle$, and is complete in the inner product.
(iii) for every $x \in X$, the linear evaluation functional

$$E_x : \mathcal{H} \to \mathbb{F}$$
$$f \mapsto f(x) = E_x(f)$$

is bounded.

Moreover, for every bounded linear functional in the RHKS the Riesz representation theorem plays a key role to give them an inner product representation. For each $x \in X$, there exits a unique function $k_x \in \mathcal{H}$ such that

$$f(x) = E_x(f) = \langle f, k_x\rangle, \qquad \forall f \in \mathcal{H}$$

k_x is called the reproducing kernel for the point x. The equation $f(x) = \langle f, k_x\rangle$ is often called the reproducing property.

Definition 1.1.2 *Reproducing kernel for $\mathcal{H}$ is a function $K : X \times X \to \mathbb{F}$ defined by $K(x,y) \coloneqq k_y(x)$. The unique function $k_y : X \to \mathbb{F} \in \mathcal{H}$ is called the reproducing kernel for the point y.*

Reproducing kernel will frequently briefly be written RK. RKs for the points are special choice of a function $f = k_y$ in $\mathcal{H}$. An RK for $\mathcal{H}$, $K(x,y)$ is conjugate symmetric in

$\mathbb{C}$ and symmetric in $\mathbb{R}$ by the inner product representation of the evaluation functional

$$K(x, y) = k_y(x) = \langle k_y, k_x \rangle = \overline{\langle k_x, k_y \rangle} = \overline{K(y, x)}$$

In addition, for each $y \in X$, $K(y, y)$ is positive by

$$0 \leq \|E_y\|^2 = \|k_y\|^2 = \langle k_y, k_y \rangle = K(y, y).$$

Now, to became more familiar with the theory, we will look at some classical examples of the reproducing kernel Hilbert space.

1.2 Examples of RKHS

1.2.1 $\mathbb{C}^n$

Consider the set of all n-tuples of complex numbers $\mathbb{C}^n$. Let $X = \{1, 2, \ldots, n\}$ and let $\nu : X \to \mathbb{C}$ be a function defined by $j \mapsto \nu(j) \coloneqq \nu_j$. Then the vectors in $\mathbb{C}^n$ are identified as $(\nu(1), \ldots, \nu(n))$, and it can be seen that $\mathbb{C}^n$ is a space of functions. Clearly, $\mathbb{C}^n$ is a vector space and $\mathcal{F}(X, \mathbb{C}) = \mathbb{C}^n$. For any $\nu, \omega \in \mathbb{C}^n$ with $\nu \coloneqq (\nu_1, \ldots, \nu_n,), \omega \coloneqq (\omega_1, \ldots, \omega_n)$, the usual inner product on $\mathbb{C}^n$ is

$$\langle \nu, \omega \rangle = \sum_{j=1}^{n} \nu_j \overline{\omega_j}$$

and it is complete in this inner product since $\mathbb{C}^n$ is finite dimensional. So $\mathbb{C}^n$ is a complex Hilbert space. Now we define the evaluation functional $E_j : \mathbb{C}^n \to \mathbb{C}$ by $E_j(\nu) \coloneqq \nu_j$. By the Riesz representation theorem ν has inner product form

$$E_j = \nu_j = \nu(j) = \langle \nu, k_j \rangle \tag{1.1}$$

for some $k_j \in \mathbb{C}^n$. Then by Schwarz inequality

$$|E_j(\nu)| = |\langle \nu, k_j \rangle| \leq \|\nu\| \|k_j\|$$

the evaluation functional is bounded. So, $\mathbb{C}^n$ is an RKHS.

For the reproducing kernel of the point and $\mathcal{H}$, consider the inner product representation 1.1 (reproducing property) of any $\nu \in \mathbb{C}^n$. First find the reproducing kernels k_j, $j = 1, \ldots, n$, here k_j denotes the *j-th* column vector and so $k_j(i)$ will denote the *i-th*

entry of the *j-th* column vector. To do this, make the choice $v = (1, 0, \dots, 0)$ and take the inner product of v and k_1

$$1 = v_1 = \langle v, k_1 \rangle = \overline{k_1(1)} + 0 + \cdots + 0 = \overline{k_1(1)}$$

so $k_1(1) = 1$, and then make another choice of $v = (0, 1, \dots, 0)$ and take the inner product of v and k_1, that is $0 = v_1 = \langle v, k_1 \rangle = 0 + \overline{k_1(2)} + 0 + \cdots + 0 = \overline{k_1(2)}$ so $k_1(2) = 0$, keeping going of the choices of v in this manner we find k_1. Similarly we can find $k_j, j = 1, \dots, n$ and then the reproducing kernel for $\mathbb{C}^n$. The reproducing kernel is just the $n \times n$ identity matrix, which is

$$K(i, j) = k_j(i) = e_j(i) = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

The notion here can be generalized if the set is $X = \mathbb{N} = \{1, 2, \dots\}$. Consider the space $l^2(X) \coloneqq \{v : X \to \mathbb{C} \mid \sum_{j=1}^{\infty} |v|^2 < \infty\}$. Again, each component of a sequence in l^2 can be identified with the function defined as above that turns it to a function space. This function space is a vector subspace, that is, $l^2(\mathbb{N}) \subseteq \mathcal{F}(X, \mathbb{C})$. For any $u, v \in l^2$, the inner product on l^2 is

$$\langle u, v \rangle = \sum_{j=1}^{\infty} u(j)\overline{v(j)}.$$

It is well known that l^2 is a Hilbert space. By 1.1 and Schwarz inequality every linear evaluation map is bounded, hence it is an RKHS.

By a similar discussion it can be seen that the reproducing kernel for H is the "Schauder basis" (see Appendix *A*.0.1)

$$k_j(i) = \langle k_j, k_i \rangle = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

1.2.2 Nonexample L^2

Let $X = [0, 1]$. Consider the space $C([0, 1]) = \{f : [0, 1] \to \mathbb{R} \mid \text{f is continuous}\}$ in the space of functions $\mathcal{F}([0, 1], \mathbb{R})$. $C([0, 1])$ is obviously a subspace of $\mathcal{F}(X, \mathbb{R})$. On this subspace, for any $f, g \in C([0, 1])$ define the inner product form

$$\langle f, g \rangle \coloneqq \int_0^1 f(t)\overline{g(t)}dt$$

which induces the usual two-norm

$$\|f\|^2 = \int_0^1 |f(t)|^2 dt.$$

By adding the limits of all Cauchy sequences to the space of all continuous functions one obtains the completion space $L^2[0,1]$ in this inner product, so a Hilbert space is obtained. Now it can be checked that whether every evaluation functional in the Hilbert space $L^2[0,1]$ is bounded or not. Unfortunately, there is a problem here. $L^2[0,1]$ cannot be an RKHS, to show why recall that a continuous linear operator on a dense set of a normed space can be continuously extended to the whole space by the bounded linear extension theorem in Appendix A.0.7. Consider the linear evaluation map $E_t(f) = f(t)$ in $C([0,1])$. Since it has only one image at every point $x \in [0,1]$ it is well-defined on $C([0,1])$. However, it cannot have an extension on all of $L^2[0,1]$ which is bounded and linear. To see this consider the following. Let $x \in (0,1)$ be fixed and

$$E_t(f_n) = f_n(t) := \begin{cases} (\frac{t}{x})^n & \text{if } 0 \le t \le x, \\ (\frac{1-t}{1-x})^n & \text{if } x < t \le 1. \end{cases}$$

The evaluation map is defined on the dense space $C([0,1])$ since f_n is in $C([0,1])$ for all $n \ge 1$. Notice that $\|f_n\|_{L^2[0,1]} = \|f_n\|_{C([0,1])}$ by completion theorem. At $t = x$, $f_n(x) = 1 = \lim_{n\to\infty} f_n(x)$ then $|E_x(f_n)| = |f_n(x)| = 1\ \forall n \ge 1$. For any fixed x, an easy integration shows

$$\lim_{n\to\infty} \|f_n\|_{L^2[0,1]} = \lim_{n\to\infty} \Big(\int_0^x |f_n(t)|^2 dt \Big)^{1/2} = \lim_{n\to\infty} \frac{x^{1/2}}{(2n+1)^{1/2}} = 0,$$

and

$$\lim_{n\to\infty} \|f_n\|_{L^2[0,1]} = \lim_{n\to\infty} \Big(\int_x^1 |f_n(t)|^2 dt \Big)^{1/2} = \lim_{n\to\infty} \frac{(1-x)^{1/2}}{(2n+1)^{1/2}} = 0.$$

We try to show that at $t = x$ the evaluation functional is unbounded in $C([0,1])$ i.e. there exists no constant $c \in \mathbb{R}$ such that $|E_x(f_n)| = |f_n(x)| \le c\|f_n\|$, $\forall n \ge 1$. Clearly, the left side of such an inequality is 1 and the right side is 0 for all $n \ge 1$, a contradiction. Thus there is no $c \ge 0$ satisfying the inequality. The evaluation map at $t = x$ is not bounded in $C([0,1])$ hence it cannot have a bounded extension on all of $L^2[0,1]$ at $t = x$.

1.2.3 Sobolev Spaces on $[0, 1]$ as RKHS

Sobolev spaces are used as an important tool by ensuring a nice domain for problems of partial differential equations. It is introduced by Sergey Sobolev in 1930's. At that time it was realized that the space $C^m(X)$ where X is an open subset of $\mathbb{R}^n$ and m is a nonnegative integer containing m-th order continuously differentiable functions is not sufficient for the solutions of the partial differential equations. In the absence of this property the space of functions $L^p(X)$ having weak derivatives of functions up to the m-th order is proposed so that weak derivatives are one the main characteristics of the Sobolev spaces. For more historical background on the subject see the book [10].

A Sobolev space, generally denoted by $W^{m,p}$ where m is a nonnegative integer and $1 \leq p \leq \infty$, is an example of a Banach space. In particular when $p = 2$ it is a Hilbert space. In this thesis we will not define what a Sobolev space is. We will only consider a specific Sobolev space and show that it is an RKHS, corresponding to the case $p = 2$ and $m = 1$.

Before starting recall the definition of absolute continuity. A function $f : [0, 1] \to \mathbb{R}$ is said to be absolutely continuous for any given $\epsilon > 0$ there exists $\delta > 0$ such that for any disjoint intervals (x_i, y_i), $i = 1, \ldots, n$ with $\sum_{i=1}^{n} |x_i - y_i| < \delta$ we have $\sum_{i=1}^{n} |f(x_i) - f(y_i)| < \epsilon$. By Fundamental Theorem of Lebesgue Calculus, the following definitions of absolute continuity are equivalent:

i) f is absolutely continuous;

ii) the three conditions are satisfied: f admits derivative f' almost everywhere, f' is integrable, and $f(x) = f(a) + \int_a^x f'(t)dt$ for all $x \in [a, b]$;

iii) for a Lebesgue integrable function g on [0,1], $f(x) = f(a) + \int_a^x g(t)dt$ for all $x \in [a, b]$.

Let us define a set

$$\mathcal{H} := \left\{ f \mid f \text{ is absolutely continuous, } \int_0^1 |f'(t)|^2 dt < +\infty, \text{ and } f(0) = f(1) = 0 \right\}$$

here $f : [0, 1] \to \mathbb{R}$. It can be shown that $\mathcal{H}$ is a Sobolev space. We will show that $\mathcal{H}$ is an RHKS by showing that $\mathcal{H}$ is a subspace of the function space, is endowed

with an inner product, complete in the inner product defined on $\mathcal{H}$, and all evaluation functions are bounded.

$\mathcal{H}$ is clearly a vector subspace of $\mathcal{F}([0,1],\mathbb{R})$ since;
i) $f = 0 \in \mathcal{H}$ and for absolutely continuous functions f, g on $[0,1]$, for all $x \in X$ $(\alpha f + \beta g)(x) = \alpha f(x) + \beta g(x)$ is absolutely continuous,
ii) $\left(\int_0^1 |(f'+g')(t)|^2 dt\right)^{1/2} \leq \left(\int_0^1 |f'(t)|^2 dt\right)^{1/2} + \left(\int_0^1 |g'(t)|^2 dt\right)^{1/2} < \infty$ by Minkowski inequality, and
iii) $(f+g)(0) = f(0) + g(0) = 0 = f(1) + g(1) = (f+g)(1)$.

Define a sesquilinear form on $\mathcal{H}$ as

$$\langle f, g\rangle := \int_0^1 f'(t)g'(t)dt.$$

Note that for the norm one has $\|f\|^2 = \int_0^1 f'(t)^2 dt \geq 0$. By the Fundamental Theorem of Lebesgue Calculus, it is obtained that

$$f(x) = f(0) + \int_0^x f'(t)dt = \int_0^1 f'(t)\chi_{[0,x]}(t)dt.$$

$$|f(x)| = \left|\int_0^1 f'(t)\chi_{[0,x]}dt\right| \leq \left(\int_0^1 |f'(t)|^2 dt\right)^{1/2}\left(\int_0^1 \chi_{[0,x]}(t)dt\right)^{1/2} = \|f\|\sqrt{x} \qquad (1.2)$$

which shows that $f = 0 \iff \langle f, f\rangle = 0$ by the Schwarz inequality. Since $E_x(f) = f(x)$ it also ensures the boundedness of evaluation function for any $x \in [0,1]$. Moreover, taking supremum of E_x over all f of norm one it can be seen that $\|E_x\| \leq \sqrt{x}, \forall x \in [0,1]$. Thus only the completeness of the inner product space $\mathcal{H}$ remains to be shown.

Suppose $\{f_n\}_{n\geq 1} \in \mathcal{H}$ is a Cauchy sequence. For any $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $\|f_n - f_m\|_{\mathcal{H}} < \epsilon$ whenever $m, n > N$. Then

$$\|f_n - f_m\|_{\mathcal{H}} = \left(\int_0^1 |f_n'(t) - f_m'(t)|^2 dt\right)^{1/2} = \|f_n' - f_m'\|_{L^2[0,1]} < \epsilon \qquad (1.3)$$

that is the sequence $\{f_n'\}_{n\geq 1}$ is Cauchy in $L^2[0,1]$. Since $L^2[0,1]$ is complete there exists a function $g \in L^2[0,1]$ such that $\{f_n'\}_{n\geq 1}$ converges to g in L^2 sense that is

$$\lim_{n\to\infty}\|f_n' - g\|^2_{L^2[0,1]} = \lim_{n\to\infty}\int_0^1 |f_n'(t) - g(t)|^2 dt = 0.$$

By the Schwarz inequality one also has

$$\int_0^x |f_n'(t) - g(t)|dt \leq \left(\int_0^x |f_n'(t) - g(t)|^2 dt\right)^{1/2}\left(\int_0^x 1^2 dt\right)^{1/2} \to 0$$

as $n \to \infty$ then $\lim_{n\to\infty} \int_0^1 |f_n'(t) - g(t)|dt = 0 \ \forall x \in [0,1]$. In addition by (1.3) above, for each $x \in [0,1]$, $|f_n(x) - f_m(x)| \leq \|f_n - f_m\| \sqrt{x}$ by (1.2) which means $\{f_n\}_{n\geq 1}$ must be pointwise Cauchy (hence pointwise convergent) so one has $f(x) = \lim_{n\to\infty} f_n(x)$ for some function f on $[0,1]$. Then

$$f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \int_0^x f_n'(t)dt = \lim_{n\to\infty} \int_0^x (f_n'(t) - g(t))dt + \int_0^x g(t)dt = \int_0^x g(t)dt.$$

Since g is Lebesgue integrable f is absolutely continuous, by the Fundamental Theorem of Lebesgue Calculus taking derivatives of both sides one gets $f'(x) = g(x)$ a.e.. Then $f' \in L^2[0,1]$ since g is independent of particular choice of its class and so one has $\langle f, f\rangle = \int_0^1 f'(t)^2 dt < \infty$. Also $f(0) = \lim_{n\to\infty} f_n(0) = 0 = \lim_{n\to\infty} f_n(1) = f(1)$, so $f \in \mathcal{H}$. Therefore $\mathcal{H}$ is complete and hence $\mathcal{H}$ is an RKHS.

To find the kernel function of the RKHS $\mathcal{H}$, start by considering the reproducing property. First attempt is to consider the Lebesgue integral of form $f(x) = \int_0^1 f'(t)\chi_{[0,x]}(t)dt$. Since $\mathcal{H}$ is an RKHS one has $f(x) = \langle f, g\rangle = \int_0^1 f'(t)\chi_{[0,x]}(t)dt$ where $g = k_x$ and $g' = \chi_{[0,x]}$. Thus, since $g \in \mathcal{H}$ one obtains the boundary value problem

$$g'(t) = \chi_{[0,x]}(t),$$
$$g(0) = g(1) = 0.$$

Taking the integral of both sides from 0 to x, $g(x) = \int_0^x \chi_{[0,x]}(t)dt = x$ and considering boundary values one easily sees that g is not continuous; that is, the problem has no continuous solution. It is known by the Riesz representation that such a $g = k_x$ exists in $\mathcal{H}$ and must be absolutely continuous (and is unique).

In the following, we will compute k_x and K by formally computing the Green's function for a boundary value problem. Since we know such a function exists, start by assuming the existence of k_x in the representation

$$f(x) = \langle f, k_x\rangle = \int_0^1 f'(t)k_x'(t)dt.$$

Then applying integration by parts

$$f(x) = k_x' f(1) - k_x'' f(0) - \int_0^1 f(t)k_x''(t)dt = -\int_0^1 f(t)k_x''(t)dt.$$

Letting δ_x Dirac's delta function one also has $f(x) = \int_0^1 f(t)\delta_x(t)dt$, then one tries to find k_x by constructing the new boundary value problem

$$-k_x''(t) = \delta_x(t)$$

$$k_x(0) = k_x(1) = 0.$$

The solution to the boundary value problem is the Green's function G. The corresponding Green's function satisfies $G_{tt}(t,x) = -k_x''(t) = 0, t \neq x$, G_{tt} refers to second derivative of G with respect to t, the boundary conditions $G(0,x) = 0 = G(1,x)$, and will be of the form

$$G(t,x) = \begin{cases} c_0 t + c_1 & \text{if } t \leq x \\ c_2(t-1) + c_4 & \text{if } t \geq x. \end{cases}$$

Let $G_1(t,x) = c_0 t + c_1, G_2(t,x) = c_2(t-1) + c_4$, for any fixed x. On the one hand we have the system of equations $G_1(t,x) = c_0 t + c_1, G_1(0,x) = G(0,x) = 0$ after solving one gets $G_1(t,x) = c_0 t$, on the other hand we have the system of equations $G_2(t,x) = c_2(t-1) + c_4, G_2(1,x) = G(1,x) = 0$ again by solving one obtains that $G_2(t,x) = c_2(t-1)$. G is continuous at $t = x, G_1(x,x) = G_2(x,x)$, i.e. $c_0 x - c_2(x-1) = 0$. In addition, G_t the first derivative of G, has a jump discontinuity at $t = x$; that is, we have that $\lim_{t\to x^+}(G_2)_t - \lim_{t\to x^-}(G_1)_t = 1$. Thus we get the system of equations $c_0 x - c_2(x-1) = 0, c_2 - c_0 = 1$. By solving this system, one finds $c_0 = (x-1), c_2 = x$.

Hence the solution of the boundary value problem is the Green's function constructed to be

$$K(t,x) = k_x(t) = G(t,x) = \begin{cases} (1-x)t & \text{if } t \leq x \\ (1-t)x & \text{if } t \geq x. \end{cases}$$

We can check that $k_x(t)$ is really a member of $\mathcal{H}$. Except at $t = x$, k_x is differentiable

$$k_x'(t) = \begin{cases} (1-x) & \text{if } t < x \\ -x & \text{if } t > x \end{cases}$$

and equal to the integral of $k_x'(t)$. Therefore we get k_x is absolutely continuous, $\int_0^1 k_x'(t)^2 dt = \int_0^x (1-x)^2 dt + \int_x^1 (-x)^2 dt < \infty$, and $k_x(0) = k_x(1) = 0$. Thus $k_x \in \mathcal{H}$. Finally, check that for any $f \in \mathcal{H}$, k_x is a kernel function (reproducing property).

$$\begin{aligned} \langle f, k_x \rangle = \int_0^1 f'(t) k_x'(t) dt &= \int_0^x f'(t)(1-x)^2 dt + \int_x^1 f'(t)(-x)^2 dt \\ &= (1-x)(f(x) - f(0)) - x(f(1) - f(x)) = f(x). \end{aligned}$$

Hence, K meets all the requirements to serve as the reproducing kernel for $\mathcal{H}$.

CHAPTER 2

CONSTRUCTION OF THE RKHS FROM A KERNEL

2.1 Characterization of RK

In this part, we will look at a necessary and sufficient condition for a kernel function to be an RK.

Definition 2.1.1 (Kernel Function) *Let $X \neq \emptyset$. Let $K : X \times X \to \mathbb{C}$ be a function. If for any $n \geq 1$, any finite set of distinct points $\{x_1, \ldots, x_n\}$ in X, and any $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$, the matrix $(K(x_i, x_j))_{i,j=1}^n$ is positive semidefinite i.e*

$$\sum_{i,j=1}^{n} \alpha_j \overline{\alpha_i} K(x_i, x_j) \geq 0$$

then the function K will be called a kernel function. This is denoted by $K \geq 0$.

There can be various terminologies for the definition above but we prefer to call it as kernel function or positive semidefinite kernel. Therefore we will be writing K is a kernel function or a positive semidefinite kernel. Also, by writing p.s.d. we will be referring to the terminology positive semidefinite.

In some cases, definiteness of a kernel function may not be sufficient. In such cases strictly positive kernels arise.

Definition 2.1.2 (Strictly Positive Kernel Function) *With the same assumptions as in Definition* 2.1.1*, if the matrix $(K(x_i, x_j))_{i,j=1}^n$ is strictly positive i.e*

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

then the function K will be called a strictly positive kernel function or a strictly positive definite kernel. This is denoted by $K > 0$.

Note that if a kernel is strictly p.s.d. then the matrices $Q = (K(x_i, x_j))_{i,j=1}^n > 0$ are invertible.

2.2 The Moore-Aronszajn Theorem

There is a relation between RK for an RKHS and kernel function, and it is in the following proposition which can be regarded as the converse of the Moore-Aronszajn theorem, that we will state and prove below.

Proposition 2.2.1 *Let* $\mathcal{H}$ *be an RKHS on any set* $X \neq \emptyset$. *Assume* $K : X \times X \to \mathbb{C}$ *is its RK. Then K is a kernel function.*

Proof. Consider any choice of distinct points $\{x_1, \ldots, x_n\} \subseteq X \; \forall n \geq 1$. Then for any $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$, by definition of RK and the reproducing property we had $K(x, y) = k_y(x) = \langle k_y, k_x \rangle, \forall x, y \in X$,

$$\sum_{i,j=1}^{n} \overline{\alpha_i} \alpha_j K(x_i, x_j) = \sum_{i,j=1}^{n} \overline{\alpha_i} \alpha_j \langle k_{x_j}, k_{x_i} \rangle = \left\langle \sum_{j=1}^{n} \alpha_j k_{x_j}, \sum_{i=1}^{n} \alpha_i k_{x_i} \right\rangle = \| \sum_{i=1}^{n} \alpha_i k_{x_i} \|^2 \geq 0.$$

Therefore the reproducing kernel K for $\mathcal{H}$ is always a kernel function. □

We see that if we are given an RKHS, then the RKHS has always a positive semidefinite kernel.

Note that the reproducing kernels for the points $\{x_1, \ldots, x_n\}$ will be the kernel functions $k_{x_1}, \ldots, k_{x_n}$, and all linear combinations of kernel functions form a dense set in an RKHS.

Proposition 2.2.2 *Let* $\mathcal{H}$ *be an RKHS on X with the reproducing kernel K. Let* $M = \{k_y(\cdot) =: K(\cdot, y) \mid y \in X\}$ *be the set of kernel functions in* $\mathcal{H}$. *Then span M is dense in* $\mathcal{H}$.

Proof. Let $f \in \mathcal{H}$. Any element in *span M* is of the form $\sum_{j=1}^{n} \alpha_j k_{y_j}, \alpha_1, \ldots, \alpha_n \in \mathbb{C}$. $f \perp$ *span M* ($f \in$ *span* $M^{\perp}$) $\iff \langle f, M \rangle = 0$. Since f is orthogonal to each element

of M, $\langle f, g\rangle = \langle f, \sum_{j=1}^{n} \alpha_j k_{y_j}\rangle = 0 \iff \langle f, k_y\rangle = f(y) = 0, \forall y \in X \iff f = 0$. Therefore *span* M is dense in $\mathcal{H}$, since *span* $M^{\perp} = \{0\}$, by a well-known Hilbert space fact. □

The following theorem, one of the fundamentals of the theory of RKHS, ensures that for a given p.s.d. kernel K there can always be constructed RKHS.

Theorem 2.2.3 (Moore-Aronszajn) *Let $X \neq \emptyset$ and let $K : X \times X \to \mathbb{C}$ be a kernel function. Then there is an RKHS $\mathcal{H}$ admitting the kernel function K to be reproducing kernel for $\mathcal{H}$.*

Proof. The proof will be shown in parts:
part a) Defining a function and dense set W to be able to construct an RKHS on it.
part b) Definition of a sesquilinear form $B(f, g)$ on W then showing this sesquilinear form is well defined and is indeed an inner product.
part c) Completion of the inner product space W which is the Hilbert space $\mathcal{H}$.
part d) Identification of any element in $\mathcal{H}$ with a function on X constructing the space $\widehat{\mathcal{H}}$ uniquely such that $\mathcal{H}$ is isomorphic to a subspace $\widehat{\mathcal{H}}$ of $\mathcal{F}(X, \mathbb{C})$.

a) Let $k_y : X \to \mathbb{C}$ defined by $k_y(x) \coloneqq K(x, y)$. Consider the set of all linear combinations of k_y, $y \in X$, say W ($W \subseteq \mathcal{F}(X, \mathbb{C})$). We consider this kind of set W because by previous Proposition 2.2.2, W is a nice candidate to construct an RKHS on it as being a dense set in an RKHS.

b) To see W as an inner product space, define a sesquilinear form on W. Let $B : W \times W \to \mathbb{C}$ be a sesquilinear form given by

$$B(f, g) \coloneqq \sum_{i,j=1}^{n,m} \alpha_j \overline{\beta_i} K(y_i, y_j)$$

where $f = \sum_{j=1}^{n} \alpha_j k_{y_j}$, $g = \sum_{i=1}^{m} \beta_i k_{y_i} \in W$ and $\alpha_j, \beta_i, i, j = 1, \ldots, n, m$ are scalars. Here it is required to discuss whether B is well-defined or not since notice that any function $f \in W$ might be written in more than one form, in other words

$$f = \sum_{j=1}^{n} \alpha_j k_{y_j} = \sum_{l=1}^{p} \alpha_l k_{y_l} = \sum_{a=1}^{q} \alpha_a k_{y_a}.$$

However, B must be independent of representer choice of a function. To show it is independent, use the classical idea that $f \equiv 0$ (f is the zero function on X) if and only

if $B(f,g) = B(g,f) = 0$ for all $g \in W$ but it is enough to show just for $k_y \in W$ since W is the linear span of k_y's. So letting $f(x) = \sum_{j=1}^{n} \alpha_j k_{y_j}(x) = 0 \; \forall x \in X$, we see that

$$B(f,k_x) = \sum_{j=1}^{n} \alpha_j K(x,y_j) = \sum_{j=1}^{n} \alpha_j k_{y_j}(x) = f(x) = 0$$

$$B(k_x,f) = \sum_{j=1}^{n} \overline{\alpha_j} K(y_j,x) = \sum_{j=1}^{n} \overline{\alpha_j K(x,y_j)} = \overline{f(x)} = 0.$$

On the other hand, letting $B(f,g) = B(g,f) = 0$ for all $g \in W$ in particular $B(f,k_x) = B(k_x,f) = 0$ one obtains $\overline{f(x)} = B(k_x,f) = 0 = B(f,k_x) = f(x)$. Thus B is well-defined on W. Here notice that we always have $f(x) = B(f,k_x)$ for all $f \in W$.

It is clear that B is linear in the first component, conjugate linear in the second one and $B(f,g) = \overline{B(g,f)}$.

An observation for the positivity of B discussed in Chapter 1 was $K(y,y) \geq 0 \; \forall y \in X$, so $B(f,f) = \sum_{j=1}^{n} |\alpha_j|^2 K(y_j,y_j) \geq 0$ for all $y_j \in X$. The definiteness of B can be seen by $B(f,f) = 0 \iff B(f,g) = B(g,f) = 0 \iff B(f,k_x) = f(x) = 0 \iff f = 0$, for all $x \in X$, where we deduce the first logical equivalence by an application of the Schwarz inequality to the positive semidefinite sesquilinear form B. Therefore B is an inner product on W hence W is an inner product space.

c) Any inner product space can be completed in the given inner product. Let $\mathcal{H}$ denote the completion of the inner product space W. Instead of directly showing how W is completed to the space consisting of only functions and then how the evaluation functional is bounded at every point, we can study $\mathcal{H}$ as follows.

d) For a given $h \in \mathcal{H}$ let us define a function as

$$\hat{h}(x) \coloneqq \langle h, k_x \rangle_{\mathcal{H}}$$

and denote the set of such functions $\widehat{\mathcal{H}} \coloneqq \{\hat{h} : X \to \mathbb{C} \mid \hat{h}(x) \coloneqq \langle h, k_x \rangle_{\mathcal{H}}, \forall h \in \mathcal{H}\}$. Also set a map $L : \mathcal{H} \to \widehat{\mathcal{H}}$ by

$$L(h) \coloneqq \hat{h}.$$

L is oclearly linear since for any $\alpha \in \mathbb{C}$, $h_1, h_2 \in \mathcal{H}$ and $\forall x \in X$, closed under addition

$$\begin{aligned} L((h_1+h_2)(x)) = \widehat{(h_1+h_2)}(x) &= \langle h_1+h_2, k_x \rangle_{\mathcal{H}} = \alpha \langle h_1, k_x \rangle_{\mathcal{H}} + \langle h_2, k_x \rangle_{\mathcal{H}} \\ &= \hat{h_1}(x) + \hat{h_2}(x) = L(h_1(x)) + L(h_2(x)), \end{aligned}$$

and scalar multiplication

$$L(\alpha h_1(x)) = \widehat{\alpha h_1}(x) = \langle \alpha h_1, k_x \rangle_{\mathcal{H}} = \alpha \langle h_1, k_x \rangle_{\mathcal{H}} = \alpha \hat{h_1}(x) = \alpha L(h_1(x)).$$

So $\widehat{\mathcal{H}}$ is a vector subspace of $\mathcal{F}(X, \mathbb{C})$.

Also since B is an inner product on W and $f : X \to \mathbb{C}$ we have $f(x) = B(f, k_x) = \langle f, k_x \rangle_W = \hat{f}(x)$. It follows that $f(x) = \hat{f}(x)$ for all $f \in W$ ($W \subseteq \widehat{\mathcal{H}}$).

Then we show that L is one-to-one by the idea that $ker(L) = \{0\}$ if and only if f is one-to-one, where $ker(L) = \{h \in \mathcal{H} \mid L(h(x)) = \hat{h}(x) = \langle h, k_x \rangle_{\mathcal{H}} = 0, \forall h \in \mathcal{H}\}$. To show that $h = 0$, we use a classical idea in Hilbert space used also in Proposition 2.2.2: The span of a nonempty set is dense in a Hilbert space if and only if the orthogonal companion of the span set contains only 0. We have $\hat{h}(x) = \langle h, k_x \rangle_{\mathcal{H}} = 0$ if and only if $h \perp k_x$ at every $x \in X$ if and only if $h \perp W$, i.e. $h \in W^{\perp}$, if and only if $h \in W^{\perp} = \{0\}$ since W is dense in the Hilbert space $\mathcal{H}$, so $h = 0$. Therefore, L is $1-1$ and onto from $\mathcal{H}$ to $\widehat{\mathcal{H}}$.

Defining an inner product on $\widehat{\mathcal{H}}$ by

$$\langle \hat{h_1}, \hat{h_2} \rangle_{\widehat{\mathcal{H}}} \coloneqq \langle h_1, h_2 \rangle_{\mathcal{H}} \tag{2.1}$$

shows that $L : \mathcal{H} \to \widehat{\mathcal{H}}$ is an isometry. L was one-to-one, onto and linear and so L is an inner product preserving isomorphism, hence $\widehat{\mathcal{H}}$ and $\mathcal{H}$ isomorphic. Hence $\widehat{\mathcal{H}}$ is a Hilbert space of functions on X.

Finally, every point evaluation in $\widehat{\mathcal{H}}$ is bounded since at every $x \in X$

$$E_x(\hat{h}) = \hat{h}(x) = \langle h, k_x \rangle_{\mathcal{H}} =: \langle \hat{h}, \hat{k_x} \rangle_{\widehat{\mathcal{H}}}$$

$\forall \hat{h} \in \widehat{\mathcal{H}}$ then $|E_x(\hat{h})| = |\langle h, k_x \rangle_{\mathcal{H}}| \leq \|\hat{h}\| \|\hat{k_x}\|$ for all $x \in X$. Thus $\widehat{\mathcal{H}}$ is an RKHS on W hence W is an RKHS on X.

By Proposition 2.2.1, we know $\widehat{\mathcal{H}}$ has the kernel function. Thus we see that $k_x = \hat{k_x}$, $\forall x \in X$ so $\hat{k_x}$ is the reproducing kernel for the point x. It follows that $\forall x, y \in X$

$$\hat{k_y}(x) = k_y(x) = K(x, y)$$

is the reproducing kernel for the $\widehat{\mathcal{H}}$. Therefore we see that $\widehat{\mathcal{H}}$ is indeed the space $\mathcal{H}$. Hence when we are given kernel function it produces only one RKHS. □

Note that this theorem, along with Proposition 2.2.1, establishes a one-to-one relationship between RKHS defined on a set and kernel functions associated with that set. Hereafter, we will be using the notation $\mathcal{H}(K)$ for the unique RKHS generated by the p.s.d. kernel K.

2.3 Some Consequences About Kernel Function

To find out continuity of a function in an RKHS, it is enough to look at the continuity of the kernel function.

Theorem 2.3.1 *Let X be a topological space, and consider $X \times X$ the product topology. Let $K : X \times X \to \mathbb{C}$ be a kernel function. If K is continuous then any function $f \in \mathcal{H}(K)$ is continuous.*

Proof. We show that for any given $\epsilon > 0$ and a fixed point y_0, there exists a neighborhood of y_0, $U \subset X$ such that for any $y \in U$, $|f(y) - f(y_0)| < \epsilon$ where $f \in \mathcal{H}(K)$.

Let $f : X \to \mathbb{C} \in \mathcal{H}(K)$ and let $y_0 \in X$ be fixed. Since K is continuous on $X \times X$ there exists a neighborhood, $V \subseteq X \times X$ of the fixed point (y_0, y_0) such that for any $(x, y) \in V$

$$|K(x, y) - K(y_0, y_0)| < \frac{\epsilon^2}{3(\|f\|^2 + 1)}.$$

Because $X \times X$ has the product topology, selecting a neighborhood $U \subseteq X$ of y_0 such that $U \times U \subseteq V$ is possible. By the reproducing property and the Schwarz inequality

$$|f(y) - f(y_0)|^2 = |\langle f, k_y\rangle - \langle f, k_{y_0}\rangle|^2 = |\langle f, k_y - k_{y_0}\rangle|^2 \leq \|f\|^2 \|k_y - k_{y_0}\|^2.$$

Then one can make the following discussion by using $K(x, y) = k_y(x) = \langle k_y, k_x\rangle$,

$$\begin{aligned}
\|k_y - k_{y_0}\|^2 =& \langle k_y - k_{y_0}, k_y - k_{y_0}\rangle \\
=& \langle k_y, k_y\rangle - \langle k_y, k_{y_0}\rangle - \langle k_{y_0}, k_y\rangle + \langle k_{y_0}, k_{y_0}\rangle \\
=& K(y, y) - K(y, y_0) - K(y_0, y) + K(y_0, y_0) - K(y_0, y_0) + K(y_0, y_0) \\
=& [K(y, y) - K(y_0, y_0)] - [K(y, y_0) - K(y_0, y_0)] - [K(y_0, y) - K(y_0, y_0)] \\
<& \frac{\epsilon^2}{\|f\|^2 + 1}
\end{aligned}$$

for any $y \in U$. Hence we obtain that

$$|f(y) - f(y_0)| < \epsilon$$

($|f(y) - f(y_0)|^2 < \frac{\|f\|^2\epsilon}{\|f\|^2+1} < \epsilon^2$) that is any function $f \in \mathcal{H}(K)$ is continuous since f was arbitrary. □

There is a relation between a kernel function and its conjugate. In addition, the RKHS induced by a positive semidefinite kernel and the RKHS generated by the conjugate of the given p.s.d. kernel are connected.

Proposition 2.3.2 *Let $K : X \times X \to \mathbb{C}$ be a positive semidefinite kernel, and let $\mathcal{H}(K)$ be the RKHS produced by K. Then the conjugate $\overline{K} : X \times X \to \mathbb{C}$ of K is a positive semidefinite kernel and the corresponding RKHS is $\mathcal{H}(\overline{K}) = \{\overline{f} \mid f \in \mathcal{H}(K)\}$. Furthermore, if $C : \mathcal{H}(K) \to \mathcal{H}(\overline{K})$ is a map given by $C(f) := \overline{f}$ then C is a surjective conjugate-linear isometry.*

Proof. For any $n \geq 1$, let $\{x_1, \ldots, x_n\} \subseteq X$ be given and for any complex scalars $\alpha_1, \ldots, \alpha_n$

$$\begin{aligned}
\sum_{i,j=1}^{n} \overline{\alpha_i}\alpha_j\overline{K(x_i, x_j)} &= \overline{\alpha_1}[\alpha_1\overline{K(x_1, x_1)} + \cdots + \alpha_n\overline{K(x_1, x_n)}] \\
&\quad + \cdots + \overline{\alpha_n}[\alpha_1\overline{K(x_n, x_1)} + \cdots + \alpha_n\overline{K(x_n, x_n)}] \\
&= \overline{\alpha_1[\overline{\alpha_1}K(x_1, x_1) + \cdots + \overline{\alpha_n}K(x_1, x_n)]} \\
&\quad + \cdots + \overline{\alpha_n[\overline{\alpha_1}K(x_n, x_1) + \cdots + \overline{\alpha_n}K(x_n, x_n)]} \\
&= \overline{\sum_{i,j=1}^{n} \alpha_i\overline{\alpha_j}K(x_i, x_j)} \geq 0,
\end{aligned}$$

thus $\overline{K}$ is a positive semidefinite kernel since K is a p.s.d. kernel. $\overline{K}$ is hermitian since $\overline{K}(x, y) = \overline{k_y}(x) = \langle\overline{k_y}, \overline{k_x}\rangle = \overline{\langle k_y, k_x\rangle} = \langle k_x, k_y\rangle = K(y, x)$ for any $x, y \in X$. From here one can have that $\forall x, y \in X, \overline{K}(x, y) = \overline{K(x, y)}$.

Now assume $C : \mathcal{H}(K) \to \mathcal{H}(\overline{K})$ is a map given by $f \mapsto \overline{f}$. Let W be a subset of $\mathcal{H}(K)$ spanned by the kernel functions $k_{y_1}, \ldots, k_{y_n}$, and let $Z = span\{\overline{k_{y_1}}, \ldots, \overline{k_{y_n}}\} \subseteq \mathcal{H}(\overline{K})$. We know that W is dense in $\mathcal{H}(K)$ and Z is dense in $\mathcal{H}(\overline{K})$ by Proposition 2.2.2. Thus for any number of points chosen $y_1 \ldots, y_n \in X$ and for any complex scalars

α_j, $j = 1, \dots, n$ we may set $\hat{C} : W \to Z$ by

$$\hat{C}\Big(\sum_{j=1}^{n} \alpha_j k_{y_j}\Big) = \sum_{j=1}^{n} \overline{\alpha_j} \overline{k_{y_j}}.$$

In this case one needs to discuss the well-definedness of the map $\hat{C}$ since as in proof of the Moore-Aronszajn Theorem 2.2.3, f may be written in more than one form. To show that $\hat{C}$ is well-defined, again by previous discussions it is enough to show that $\hat{C}$ is isometry.

$$\begin{aligned}
\|\hat{C}(\sum_{j=1}^{n} \overline{\alpha_j} k_{y_j})\|^2_{\mathcal{H}(\overline{K})} &= \|\sum_{j=1}^{n} \alpha_j \overline{k_{y_j}}\|^2_{\mathcal{H}(\overline{K})} = \Big\langle \sum_{j=1}^{n} \alpha_j \overline{k_{y_j}}, \sum_{i=1}^{n} \alpha_i \overline{k_{y_i}} \Big\rangle_{\mathcal{H}(\overline{K})} \\
&= \overline{\alpha_1}[\alpha_1 \langle \overline{k_{y_1}}, \overline{k_{y_1}} \rangle + \cdots + \alpha_n \langle \overline{k_{y_1}}, \overline{k_{y_n}} \rangle] \\
&+ \cdots + \overline{\alpha_n}[\alpha_1 \langle \overline{k_{y_n}}, \overline{k_{y_1}} \rangle + \cdots + \alpha_n \langle \overline{k_{y_n}}, \overline{k_{y_n}} \rangle] \\
&= \sum_{i,j=1}^{n} \alpha_j \overline{\alpha_i} \langle \overline{k_{y_i}, k_{y_j}} \rangle_{\mathcal{H}(\overline{K})} = \sum_{i,j=1}^{n} \alpha_j \overline{\alpha_i} \overline{K}(y_j, y_i) \\
&= \overline{\sum_{i,j=1}^{n} \overline{\alpha_j} \alpha_i K(y_j, y_i)} = \overline{\sum_{i,j=1}^{n} \overline{\alpha_j} \alpha_i \langle k_{y_i}, k_{y_j} \rangle_{\mathcal{H}(K)}} \\
&= \overline{\Big\langle \sum_{j=1}^{n} \overline{\alpha_j} k_{y_j}, \sum_{j=1}^{n} \overline{\alpha_j} k_{y_j} \Big\rangle_{\mathcal{H}(K)}} = \overline{\|\sum_{j=1}^{n} \overline{\alpha_j} k_{y_j}\|_{\mathcal{H}(K)}} \\
&= \|\sum_{j=1}^{n} \overline{\alpha_j} k_{y_j}\|_{\mathcal{H}(K)}
\end{aligned}$$

Therefore $\hat{C}$ is isometry and hence $\hat{C}$ is well-defined on the dense spaces W and Z. On the dense spaces $\hat{C}$ is obviously conjugate linear since

$$\begin{aligned}
\hat{C}(a \sum_{j=1}^{n} \alpha_j k_{y_j} + b \sum_{i=1}^{m} \alpha_j k_{y_i}) &= \overline{a \sum_{j=1}^{n} \alpha_j k_{y_j} + b \sum_{i=1}^{m} \alpha_j k_{y_i}} = \overline{a} \overline{\sum_{j=1}^{n} \alpha_j k_{y_j}} + \overline{a} \overline{\sum_{i=1}^{m} \alpha_j k_{y_i}} \\
&= \overline{a} C(\sum_{j=1}^{n} \alpha_j k_{y_j}) + \overline{a} C(\sum_{i=1}^{m} \alpha_j k_{y_i}).
\end{aligned}$$

Since dense spaces on $\mathcal{H}(K)$ and $\mathcal{H}(\overline{K})$ are normed spaces (indeed inner product spaces again by 2.2.3) and $\hat{C}$ is bounded (on finite dimensional space W), we can extend uniquely the map $\hat{C}$ to $C : \mathcal{H}(K) \to \mathcal{H}(\overline{K})$ which is isometric and conjugate linear, by bounded extension theorem in Appendix A. In addition to these, C will map the space $\mathcal{H}(K)$ onto $\mathcal{H}(\overline{K})$ since C is isometric, and there is a dense subspace in its range. Notice that C takes a function to its complex conjugate on W, thus $\mathcal{H}(\overline{K}) = C(\mathcal{H}(K)) = \{\overline{f} \mid f \in \mathcal{H}(K)\}$. □

2.4 From Kernels to RKHSs

This section will cover the concrete creation of RKHSs by certain fundamental kernels.

2.4.1 One Dimensional RKHS

Proposition 2.4.1 *Let X be a nonempty set. Let $f : X \to \mathbb{C}$ be a nonzero function. Consider the function $K(x, y) = f(x)\overline{f(y)}$ on X. Then K is a kernel function, produces a one dimensional RKHS, $\mathcal{H}(K)$ and $\|f\| = 1$.*

Proof. For any $n \geq 1$, any points $\{x_1, \ldots, x_n\} \subseteq X$ and for any complex scalars $\alpha_1, \ldots, \alpha_n$

$$\sum_{i,j=1}^{n} \alpha_j \overline{\alpha_i} K(x_i, x_j) = \sum_{i,j=1}^{n} \alpha_j \overline{\alpha_i} f(x_i)\overline{f(x_j)} = \left| \sum_{j=1}^{n} \overline{\alpha_j} f(x_j) \right|^2 \geq 0,$$

K is a positive semidefinite kernel.

Thus by the Moore-Aronszajn Theorem there exists an RKHS, $\mathcal{H}(K)$ whose RK for $\mathcal{H}(K)$ is the kernel function $K(x, y) = f(x)\overline{f(y)}$ and reproducing kernel for the fixed point y is $k_y = \overline{f(y)}f$. So if we let $W = span\{k_y \mid y \in X\} = \{\overline{f(y)}f \mid y \in X\}$ we see that W is a just one-dimensional space. Since W is finite dimensional, it is closed. Therefore $W = \mathcal{H}(K)$, that is the $\mathcal{H}(K)$ is a one dimensional space spanned by f. Moreover, for any fixed y with $f(y) \neq 0$,

$$|f(y)|^2 = f(y)\overline{f(y)} = K(y, y) = k_y(y) = \langle k_y, k_y \rangle = \|k_y\|^2 = \|\overline{f(y)}f\|^2 = |\overline{f(y)}|^2 \|f\|^2$$

hence $\|f\| = 1$. □

2.4.2 Min Function

Consider the function $K : X \times X \to \mathbb{R}$, $X = [0, +\infty)$, defined by $K(x, y) := min\{x, y\}$. We will show that K is a positive semidefinite kernel and look at some properties of the RKHS generated by it.

Proposition 2.4.2 *The map* $K : [0, +\infty) \times [0, +\infty) \to \mathbb{R}$ *given by* $K(x, y) := min\{x, y\}$ *is a kernel function.*

Proof. First, it is required to show the statement in linear algebra that the $n \times n$ matrix $J_n = [\xi_{i,j}]_{i,j=1}^n$ whose all entries are 1 is positive semidefinite and has the eigenvalues $\lambda = n$ with multiplicity one and $\lambda = 0$ with multiplicity $(n - 1)$.

To show that J_n is positive semidefinite use an equivalent definition of positive semidefiniteness of a matrix. Let $v = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{C}^n$. Then

$$\langle Jv, v\rangle = \sum_{i,j=1}^{n} \alpha_j \overline{\alpha_i} \xi_{i,j} = \Big(\sum_{i=1}^{n} \alpha_i \Big) \Big(\sum_{j=1}^{n} \overline{\alpha_j} \Big) \geq 0.$$

To find the eigenvalues of J_n consider the eigenvalue-eigenvector equation $J_n v = \lambda v$ for any nonzero vector v. It is easy to see that $J_n v_1 = n v_1$, so n is an eigenvalue of J_n. If for a square matrix A, there exists a square matrix S such that $A = S^{-1} D S$, here D denotes the diagonal matrix whose diagonal entries are the eigenvalues of J_n with their multiplicity, we have $tr(A) = tr(S^{-1} D S) = tr(D S S^{-1}) = tr(D)$.

By directly multiplying J_n again with itself one also has $J_n^2 = n J_n$. Then

$$\lambda^2 \omega = J_n^2 \omega = n J_n \omega = n \lambda \omega$$

for any $\omega \neq 0$. Thus $\lambda^2 = n\lambda$ which shows that $\lambda = 0$ or $\lambda = n$. Since $tr(J_n) = n$, the trace of the diagonal matrix consisting of the eigenvalues of J_n must be n, hence all the other eigenvalues of J_n must be $\lambda = 0$ with the multiplicity $(n - 1)$.

Second, let $\{x_1, \ldots, x_n\} \subseteq X$ be points, and let $B = (K(x_i, x_j))_{i,j=1}^n = (min\{x_i, x_j\})_{i,j=1}^n$ be an $n \times n$ matrix. Note that if we put the points $\{x_1, \ldots, x_n\}$ in an order, we get the $n \times n$ matrix Q obtained by conjugating the matrix B by a permutation matrix P. Also note that B is unitary matrix, so we have, in particular $B^* = B^{-1}$. In this case, the matrix $Q = PBP^{-1}$ is p.s.d. if and only if B is a p.s.d. matrix by. To see this in a particular case, let $X = \{x_1, x_2, x_3\}$ and assume that the points are given in the order $x_2 \leq x_3 \leq x_1$. Then we have a permutation (231) in the permutation group S_3 and the corresponding permutation matrix is

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

Note that $P^* = P^{-1} = P^T$. Then if we conjugate the matrix

$$B = \begin{bmatrix} x_1 & x_2 & x_3 \\ x_2 & x_2 & x_2 \\ x_3 & x_2 & x_3 \end{bmatrix}$$

by the permutation unitary P, (PBP^{-1}) we get

$$Q = PBP^{-1} = \begin{bmatrix} x_2 & x_2 & x_2 \\ x_2 & x_3 & x_3 \\ x_2 & x_3 & x_1 \end{bmatrix}.$$

The matrix B is positive if and only if the matrix PBP^* is positive, where the expression PBP^* is sometimes called P sandwich.

Now we are able to show that K is a positive semidefinite kernel. Let $\{x_1, \ldots, x_n\} \in [0, +\infty)$ be any finite number of set of points. Without loss of generality assume that $0 \leq x_1 \leq x_2 \leq \cdots \leq x_n$. Show that K is p.s.d. by induction on the number of points. Let $\alpha_1, \ldots, \alpha_n$ be any complex scalars. When $n = 1$, $K(x_1, x_1) = min\{x_1, x_1\} = x_1 \geq 0$, so it is trivially positive $\sum_{i,j=1}^{1} \alpha_j \overline{\alpha_i} x_1 = |\alpha_1|^2 x_1 \geq 0$. For $(n-1)$, suppose the matrix

$$\begin{bmatrix} x_1 & x_1 & \cdots & x_1 \\ x_1 & x_2 & \cdots & x_2 \\ \vdots & & \ddots & \\ x_1 & x_2 & \cdots & x_{n-1} \end{bmatrix}$$

is p.s.d. that is $\sum_{i,j=1}^{(n-1)} \alpha_j \overline{\alpha_i} (min\{x_i, x_j\}) \geq 0$. When n points are given the matrix $B = (K(x_i, x_j))_{i,j=1}^{n} = (min\{x_i, x_j\})_{i,j=1}^{n}$ can be written in the form

$$Q = \begin{bmatrix} x_1 & x_1 & \cdots & x_1 \\ x_1 & x_2 & \cdots & x_2 \\ \vdots & & \ddots & \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} = x_1 \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & & \ddots & \\ 1 & 1 & \cdots & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & x_2 - x_1 & \cdots & x_2 - x_1 \\ \vdots & & \ddots & \\ 0 & x_2 - x_1 & \cdots & x_n - x_1 \end{bmatrix}.$$

The $(n-1) \times (n-1)$ matrix on the right in the matrix addition is also in ordering since only x_1 is substracted and it is p.s.d. since $\sum_{i,j=2}^{n} \alpha_j \overline{\alpha_i} min\{x_i - x_1, x_j\} \geq 0$ by the assumption. It was showed that J_n is p.s.d.. Since sum of two p.s.d. matrices is p.s.d. the matrix Q is also positive semidefinite. □

Hence, by the Moore-Aronszajn theorem there is an RKHS, $\mathcal{H}_{\mathbb{R}}(K)$ such that the kernel function K is RK for $\mathcal{H}_{\mathbb{R}}(K)$. Note that it is very easy to see that $(K(x_i, x_j))_{i,j=1}^n = (min\{x_i, x_j\})_{i,j=1}^n = (min\{x_j, x_i\})_{i,j=1}^n = (K(x_j, x_i))_{i,j=1}^n$.

Some properties of $\mathcal{H}_{\mathbb{R}}(K)$ can be deduced. First observation is about continuity of functions $f \in \mathcal{H}_{\mathbb{R}}(K)$. The kernel function is continuous since for any given $\epsilon > 0$ there exists a neighborhood $V \subset [0, +\infty) \times [0, +\infty)$ of (x_0, y_0), $|(x_0, y_0) - (x, y)| < \delta$, such that for any point (x, y) in V, where without loss of generality $x_0 < y_0, x < y$, $|K(x_0, y_0) - K(x, y)| = |min\{x_0, y_0\} - min\{x, y\}| = |x_0 - x| \leq [(x_0 - x)^2 + (y_0 - y)^2]^{1/2} = ||(x_0 - x), (y_0 - y)|| = ||(x_0, y_0) - (x, y)|| < \delta = \epsilon$. So by Theorem 2.3.1, each function $f \in \mathcal{H}_{\mathbb{R}}(K)$ is continuous on $[0, +\infty)$.

Kernel functions, letting $x < y$, are $x = min\{x, y\} = K(x, y) \coloneqq k_y(x)$.

We skip a complete description of a function in the $\mathcal{H}_{\mathbb{R}}(K)$, and only describe the appearance of a typical function in the dense space $W_{\mathbb{R}}$. Take some points in the order $y_1 < y_2 < \cdots < y_n \in [0, +\infty)$ and scalars $a_1, \ldots, a_n \in \mathbb{R}$. Any function in $W_{\mathbb{R}}$ is a linear combination of kernel functions k_{y_j}. A typical function is of the form

$$\sum_{i=1}^{n} a_i k_{y_j}(x) = \begin{cases} (a_1 + \cdots + a_n)x \text{ if } 0 \leq x < y_1 \\ a_1 y_1 + (a_2 + \cdots + a_n)x \text{ if } y_1 \leq x < y_2 \\ \vdots \\ (a_1 y_1 + \cdots + a_{n-1} y_{n-1}) + a_n x \text{ if } y_{n-1} \leq x < y_n \\ (a_1 y_1 + \cdots + a_n y_n) \text{ if } y_n \leq x. \end{cases}$$

Thus it can be seen from here that any $f \in W_{\mathbb{R}}$ is continuous, piece-wise linear, 0 at 0, and eventually constant. Such functions can be called "sawtooth". Conversely, one can prove that every such function is in $W_{\mathbb{R}}$.

2.4.3 Positive Semidefinite Matrices

A positive semidefinite matrix naturally builts an RKHS.

Let $P = (p_{i,j})_{i,j=1}^n$ be a complex positive semidefinite matrix. By letting $X = \{1, \ldots, n\}$ one can set a function $K : X \times X \to \mathbb{C}$ by $K(i, j) = p_{i,j}$. By definitions it can be easily

seen that K is a kernel function if and only if P is positive semidefinite. Thus there exits an RKHS associated to K, say $\mathcal{H}(K)$. We now investigate the characteristics of the space $\mathcal{H}(K)$.

First observation is about kernel functions. We have that $K(i,j) = k_j(i) = p_{i,j}$ here $k_j = p_j$ is the *j-th* column of the matrix P. The set of all linear combinations of k_j, W, is closed since it is a finite dimensional space and so $W = \overline{W} = \mathcal{H}(K)$. Second, notice that P defines an operator, say P and its range is $\mathcal{R}(P) \subseteq \mathbb{C}^n$. Since $\mathcal{H}(K)$ is spanned by $k_j = p_j$, $j = 1, \ldots, n$, we have that $\mathcal{R}(P) = \mathcal{H}(K) \subseteq \mathbb{C}^n$. Another observation is that the inner product on $\mathcal{H}(K)$ might be different from the usual inner product on $\mathbb{C}^n$ and depends on whether the matrix P is invertible or not. We find inner product on $\mathcal{H}(K)$ in two cases.

Suppose the matrix P is invertible (p_j' s are linearly independent). In this case, $\mathcal{R}(P) = \mathcal{H}(K) = \mathbb{C}^n$. Consider P as a linear operator. The matrix P is positive semidefinite, therefore the operator P is positive, in particular, bounded and self-adjoint ($P = P^* = \overline{P^T}$). Hence it has the positive square root operator $P^{1/2}$. The matrix $P^{1/2}$ is also invertible since $0 \neq det(P) = det(P^{(1/2)2}) = det(P^{(1/2)})^2$. Then the columns of $P^{1/2}, P^{1/2}e_j$ are linearly independent and so $P^{1/2}e_j$, $j = 1, \ldots, n$ is a basis for $\mathbb{C}^n$ as e_j, $j = 1, \ldots, n$ is the standard orthonormal basis for $\mathbb{C}^n$.

Now define a map $A : \mathbb{C}^n \to \mathcal{H}(K)$ by $A(P^{1/2}e_j) = k_j$, i.e. a map matching the *j-th* columns of the matrices $P^{1/2}$ and P. A is clearly well defined since every element in $\mathbb{C}^n$ can be uniquely written by the vectors $P^{1/2}e_j$. Take two elements $v = \sum_{j=1}^n \alpha_j P^{1/2}e_j$, $\omega = \sum_{i=1}^n \beta_i P^{1/2}e_i \in \mathbb{C}^n$. Then

$$\begin{aligned}
\langle v, \omega \rangle =& \Big\langle \sum_{j=1}^n \alpha_j P^{1/2}e_j, \sum_{i=1}^n \beta_i P^{1/2}e_i \Big\rangle \\
=& \sum_{i,j=1}^n \alpha_j \overline{\beta_i} \langle P^{1/2}e_j, P^{1/2}e_i \rangle \\
=& \sum_{i,j=1}^n \alpha_j \overline{\beta_i} \langle P^{(1/2)2}e_j, e_i \rangle = \sum_{i,j=1}^n \alpha_j \overline{\beta_i} \langle Pe_j, e_i \rangle \\
=& \sum_{i,j=1}^n \alpha_j \overline{\beta_i} p_{i,j} = \sum_{i,j=1}^n \alpha_j \overline{\beta_i} k_j(i) = \sum_{i,j=1}^n \alpha_j \overline{\beta_i} \langle k_j, k_i \rangle_{\mathcal{H}} \\
=& \Big\langle \sum_{j=1}^n \alpha_j k_j, \sum_{i=1}^n \beta_i k_i \Big\rangle_{\mathcal{H}} = \langle Av, A\omega \rangle_{\mathcal{H}}
\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product on $\mathbb{C}^n$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product on $\mathcal{H}(K)$. Thus, A represents a map between the Hilbert spaces $\mathbb{C}^n$ and $\mathcal{H}(K)$ that maintains the product structure while being bijective.

Since $k_j = Pe_j$ one has $AP^{1/2}e_j = k_j = Pe_j = P^{(1/2)2}e_j$ then $(A - P^{1/2})P^{1/2}e_j = 0$, since $P^{1/2}e_j \neq 0$ we get $A = P^{1/2} \iff A^{-1} = P^{-1/2}$. For any $v, \omega \in \mathcal{H}(K) = \mathcal{R}(P) = \mathbb{C}^n$

$$\langle v, \omega \rangle_{\mathcal{H}} = \langle A^{-1}v, A^{-1}\omega \rangle = \langle P^{-1/2}v, P^{-1/2}\omega \rangle = \langle P^{(-1/2)2}, \omega \rangle = \langle P^{-1}, \omega \rangle.$$

Therefore, if an invertible p.s.d. matrix P is given then the inner product on the RKHS produced by columns of P can be determined by the inverse of P.

When P is not invertible ($p'_j s$ are linearly dependent) we have, for the null space of P, $\mathcal{N}(P) = \{\alpha \in \mathbb{C}^n \mid P\alpha = 0\} \neq \emptyset$. Notice that we have $\mathcal{N}(P)^{\perp} = \mathcal{R}(P) = \mathcal{H}(K) \subset \mathbb{C}^n$. $P^{1/2}$ still exits, and is p.s.d.. $P^{1/2}$ is also noninvertible and so the columns, $P^{1/2}e_j$ are linearly dependent. However a map $A : \mathcal{R}(P^{1/2}) \to \mathcal{H}(K)$ set by multiplication by the matrix $P^{1/2}$ is again well-defined, linear, inner product preserving. By a similar discussion to above one can find that for any $v, \omega \in \mathcal{H}(K)$

$$\langle v, \omega \rangle_{\mathcal{H}} = \langle P^{\dagger}v, \omega \rangle$$

where $P^{\dagger} : \mathcal{N}(P)^{\perp} \to \mathcal{N}(P)^{\perp}$ is defined by $P^{\dagger}v = \omega$ if and only if $v = P\omega$.

2.4.4 Inner Product of a Hilbert Space

Inner product of a Hilbert space can be viewed as a kernel function. In this regard, the RKHS constructed by the inner product of a Hilbert space is the dual space of the Hilbert space.

Definition 2.4.3 *Let $\mathcal{L}$ be a Hilbert space and let $h_1, \ldots, h_n$ be a finite collection of elements of $\mathcal{L}$. Then the $n \times n$ matrix G whose entries are inner product of elements of $\mathcal{L}$, $G = (\langle h_i, h_j \rangle)_{i,j=1}^n$, is called the Gram or Grammian matrix.*

Proposition 2.4.4 *A Gram matrix G is positive semidefinite. Furthermore, G is a positive definite matrix if and only if the elements $h_1, \ldots, h_n$ taken from a Hilbert space are linearly independent.*

Proof. Use an equivalent definition of positive semidefiniteness. Let $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{C}^n$ and let $h_1, \ldots, h_n$ in a Hilbert space. It follows that

$$\begin{aligned}
\langle G\alpha, \alpha\rangle =& \left\langle \begin{bmatrix} \langle h_1, h_1\rangle & \ldots & \langle h_1, h_n\rangle \\ \vdots & \ddots & \vdots \\ \langle h_n, h_1\rangle & \ldots & \langle h_n, h_n\rangle \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}, \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \right\rangle \\
=& \left\langle \begin{bmatrix} \alpha_1\langle h_1, h_1\rangle & \ldots & \alpha_n\langle h_1, h_n\rangle \\ \vdots & \ddots & \vdots \\ \alpha_1\langle h_n, h_1\rangle & \ldots & \alpha_n\langle h_n, h_n\rangle \end{bmatrix}, \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \right\rangle \\
=& \overline{\alpha_1}[\alpha_1\langle h_1, h_1\rangle + \cdots + \alpha_n\langle h_1, h_n\rangle] + \cdots + \overline{\alpha_n}[\alpha_1\langle h_n, h_1\rangle + \cdots + \alpha_n\langle h_n, h_n\rangle] \\
=& [\langle \overline{\alpha_1}h_1, \overline{\alpha_1}h_1\rangle + \cdots + \langle \overline{\alpha_1}h_1, \overline{\alpha_n}h_n\rangle] + \cdots + [\langle \overline{\alpha_n}h_n, \overline{\alpha_1}h_1\rangle + \cdots + \langle \overline{\alpha_n}h_n, \overline{\alpha_n}h_n\rangle] \\
=& \langle \overline{\alpha_1}h_1, (\overline{\alpha_1}h_1 + \cdots + \overline{\alpha_n}h_n)\rangle + \cdots + \langle \overline{\alpha_n}h_n, (\overline{\alpha_1}h_1 + \cdots + \overline{\alpha_n}h_n)\rangle \\
=& \langle (\overline{\alpha_1}h_1 + \cdots + \overline{\alpha_n}h_n), (\overline{\alpha_1}h_1 + \cdots + \overline{\alpha_n}h_n)\rangle \\
=& \|\sum_{j=1}^{n} \overline{\alpha_j}h_j\|^2 \geq 0.
\end{aligned}$$

Therefore, any Gram matrix G is positive semidefinite.

Moreover, for any $\alpha \in \mathbb{C}^n$, $\langle G\alpha, \alpha\rangle = \|\sum_{j=1}^{n} \overline{\alpha_j}h_j\|^2 = 0 \iff \overline{\alpha_1}h_1 + \cdots + \overline{\alpha_n}h_n = 0$ which means the elements $h_1, \ldots, h_n$ are linearly dependent in this case. So $\langle G\alpha, \alpha\rangle \neq 0$ if and only if $\overline{\alpha_1}h_1 + \cdots + \overline{\alpha_n}h_n \neq 0$ for all $\alpha \neq 0 \in \mathbb{C}^n$ in other words $\langle G\alpha, \alpha\rangle > 0 \iff [\,\overline{\alpha_1}h_1 + \cdots + \overline{\alpha_n}h_n = 0 \iff \alpha = 0]$.

□

Proposition 2.4.5 *For a Hilbert space $\mathcal{L}$ with the inner product $\langle\cdot,\cdot\rangle$, the kernel function $K : \mathcal{L} \times \mathcal{L} \to \mathbb{C}$ given by*

$$K(x, y) := \langle x, y\rangle$$

is a kernel function and the corresponding RKHS $\mathcal{H}(K)$ is the vector space of bounded linear functionals f_ω on $\mathcal{L}$ where $\omega \in \mathcal{L}$ with the identification given by the Riesz Representation Theorem. Moreover,

$$\|f_\omega\|_{\mathcal{H}} = \|\omega\|_{\mathcal{L}}$$

where the norm $\|\cdot\|_{\mathcal{H}}$ is induced by the inner product on $\mathcal{H}(K)$ and $\|\cdot\|_{\mathcal{L}}$ is induced by the inner product on $\mathcal{L}$.

Proof. Notice that K is a positive semidefinite kernel by Proposition 2.4.4 hence $\mathcal{H}(K)$ exists by the Moore-Aronszajn Theorem 2.2.3.

To show that $\mathcal{H}(K)$ consists of all linear combinations of bounded linear functionals on the Hilbert space one can make the following discussion. First of all we have that $k_y(u) = K(u, y) = \langle u, y\rangle$ for all $u \in \mathcal{L}$. By the Schwarz inequality k_y is clearly a linear and bounded functional. Any addition and multiplication by a scalar of the functions k_y's are again a bounded linear functional on $\mathcal{L}$. We will show that the space of all bounded linear functionals is indeed the space $\mathcal{H}(K)$.

By the Riesz Representation Theorem any bounded linear functional f on $\mathcal{L}$ is given by

$$f(x) = \langle x, \omega\rangle$$

for all $x \in \mathcal{L}$ and for some unique $\omega \in \mathcal{L}$. With the identification $f = f_\omega$ we may define a set $\mathcal{H} := \{f_\omega \mid f_\omega : \mathcal{L} \to \mathbb{C}, \omega \in \mathcal{L}\}$. It is not difficult to see that $\mathcal{H}$ is an inner product space with the inner product $\langle f_\omega, f_v\rangle_{\mathcal{H}} := \langle v, \omega\rangle_{\mathcal{L}}$.

Now consider a map $C : \mathcal{H} \to \mathcal{L}$ defined by $Cf_\omega = \omega$. For any $\alpha, \beta \in \mathbb{C}$ and $f_\omega, f_v \in \mathcal{H}$, note that $\forall x \in \mathcal{L}$,

$$\begin{aligned}(\alpha f_\omega + \beta f_v)(x) = \alpha f_\omega(x) + \beta f_v(x) =& \alpha\langle x, \omega\rangle + \beta\langle x, v\rangle = \langle x, \overline{\alpha}\omega\rangle + \langle x, \overline{\beta}v\rangle \\ =& \langle x, \overline{\alpha}\omega + \overline{\beta}v\rangle = f_{\overline{\alpha}\omega+\overline{\beta}v}(x),\end{aligned}$$

it follows that

$$C(\alpha f_\omega + \beta f_v)(x) = f_{\overline{\alpha}\omega+\overline{\beta}v}(x) = \overline{\alpha}\omega(x) + \overline{\beta}v(x) = \overline{\alpha}Cf_\omega(x) + \overline{\beta}Cf_v(x)$$

C is a conjugate linear map for any $x \in \mathcal{L}$ since inner product is conjugate linear in the second variable. Because $\|\omega\|_{\mathcal{L}} = \|f_\omega\|_{\mathcal{H}}$,

$$\|Cf_\omega\|_{\mathcal{L}} = \|\omega\|_{\mathcal{L}} = \|f_\omega\|_{\mathcal{H}}$$

that is C is isometric hence it is one to one. Also, by the Riesz Representation Theorem C is onto. We can see this by considering the idea in the definition of the k_y, i.e. since for any vector in $\mathcal{L}$ defines bounded linear functional on $\mathcal{L}$. Hence $\mathcal{L}$ is conjugate linear isomorphic to $\mathcal{H}$. Therefore $\mathcal{H}$ is a Hilbert space in the inner product

$$\langle f_\omega, f_v\rangle_{\mathcal{H}} = \langle v, \omega\rangle_{\mathcal{L}}.$$

For each $x \in \mathcal{L}$, the evaluation functional $E_x(f_\omega) = f_\omega(x) = \langle x, \omega \rangle_{\mathcal{L}} = \langle f_\omega, f_x \rangle_{\mathcal{H}}$

$$|E_x(f_\omega)| = |\langle x, \omega \rangle| = ||x||||\omega|| = ||x||||f_\omega||$$

is bounded at any $x \in \mathcal{L}$. Thus $\mathcal{H}$ is an RKHS. Also $E_x(f_\omega) = \langle f_\omega, k_x \rangle_{\mathcal{H}}$ since the evaluation functional E_x is bounded. Comparing the two forms of E_x, $\langle f_\omega, f_x \rangle = \langle f_\omega, k_x \rangle$ for any $f_\omega \in \mathcal{H}$, we obtain that the kernel function for the point x is $k_x = f_x$. It follows that the kernel function for $\mathcal{H}$ is

$$K_{\mathcal{H}}(x, y) = k_y(x) = f_y(x) = \langle x, y \rangle$$

and so $K = K_{\mathcal{H}}$. Therefore by the uniqueness of the RKHS we see that $\mathcal{H}$ is the RKHS $\mathcal{H}(K)$. □

CHAPTER 3

APPLICATIONS OF RKHS IN INTERPOLATION

In this chapter we will discuss applications of RKHS is on interpolation and approximation problems. Unlike approximation, interpolation means that the given data and values are exact matches. Intuitively, interpolation is a method of making an estimation to a function by looking the range of the given points.

There are two major problems in this concept and we will consider these problems in an RKHS. One problem is the interpolation problem which is to find a function that exactly fits the points. The other is the approximation problem which arises when the interpolant is not in the RKHS. In this case our main aim is to find a function that is the "most similar" to the original one.

3.1 Interpolation in RKHS

We will discuss the existence and uniqueness of an interpolant in an RKHS.

Definition 3.1.1 (Interpolation) *Let X, Y be two nonempty sets. For a given set of distinct points $\{x_1, \ldots, x_n\} \subseteq X$ and the values $\{\lambda_1, \ldots, \lambda_n\} \subseteq Y$, a function $g : X \to Y$ is said to be an interpolant if $g(x_i) = \lambda_i$, for all $i = 1, \ldots, n$. The subset $\{x_1, \ldots, x_n\} \subseteq X$ is often called the set of data points.*

Theorem 3.1.2 (Uniqueness) *Let X and $\mathcal{H}$ be an RKHS on X. Let $E = \{x_1, \ldots, x_n\} \subseteq X$ be a finite collection of distinct points and $\{\lambda_1, \ldots, \lambda_n\} \subseteq \mathbb{C}$ be values. Assume $\exists g \in \mathcal{H}$ such that $g(x_i) = \lambda_i$, $\forall i = 1, \ldots, n$. Then $P_E(g)$ is the unique interpolant*

of minimum norm where $P_E : \mathcal{H} \to \mathcal{H}_E$ is the orthogonal projection onto $\mathcal{H}_E \coloneqq span\{k_{x_1}, \dots k_{x_n}\}$.

Proof. Let $\mathcal{H}_E \coloneqq \text{span}\{k_{x_1}, \dots k_{x_n}\}$. Observe that $\mathcal{H}_E \subseteq \mathcal{H}$ and $\mathcal{H}_E$ is a closed subspace of $\mathcal{H}$ since $\mathcal{H}_E$ is finite dimensional. Then $\mathcal{H}$ has the form $\mathcal{H} = \mathcal{H}_E \bigoplus \mathcal{H}_E^{\perp}$. Hence defining such a projection operator is meaningful. Let $P_E : \mathcal{H} \to \mathcal{H}_E$ be the orthogonal projection onto $\mathcal{H}_E$. Let $u \in \mathcal{H}_E^{\perp}$. By the reproducing property and $u \perp k_{x_i}$, $u(x_i) = \langle u, k_{x_i} \rangle = 0 \ \forall i = 1, \dots, n$. Conversely, if $0 = u(x_i) = \langle u, k_{x_i} \rangle \ \forall i = 1, \dots, n$ then $u \in \mathcal{H}_E^{\perp}$ because $k_{x_i} \in \mathcal{H}_E \ \forall i = 1, \dots, n$. Therefore, $i = 1, \dots, n$

$$P_E(f)(x_i) = f(x_i)$$

for any $f \in \mathcal{H}$.

Now assume $g_1, g_2 \in \mathcal{H}$ such that $g_1(x_i) = \lambda_i$ and $g_2(x_j) = \lambda_j \ \forall i, j = 1, \dots, n$. By the observation above $\forall i, j = 1, \dots, n$

$$g_1(x_i) - g_2(x_j) = \langle g_1, k_{x_i} \rangle - \langle g_2, k_{x_i} \rangle = \langle g_1 - g_2, k_{x_i} \rangle = 0$$

the function $(g_1 - g_2)$ is in the null space of P_E ($(g_1 - g_2) \in \mathcal{H}_E^{\perp}$), that is there can be more than one interpolant such that $g_1(x_i) = g_2(x_j) \ \forall i, j = 1, \dots, n$. Thus, for $h \in \mathcal{H}_E^{\perp}$ the functions in the set of solutions of the interpolation problem are in the form $g(x_i) = (g + h)(x_i) \ \forall i = 1, \dots, n$. Notice that $P_E(g)$ is also in the set of solutions. Then $\forall i = 1, \dots, n$

$$P_E(g)(x_i) = g(x_i) = (g + h)(x_i).$$

Since $P_E(\mathcal{H}) \neq \{0\}$ and the projection operator is bounded by norm 1

$$||P_E(g)|| = ||P_E(g + h)|| \leq ||g + h||.$$

Therefore $P_E(g)$ has the minimum norm among all the other solutions.

Uniqueness of the interpolant follows immediately from the uniqueness of the projection operator. □

Now we will give an important theorem which turns the interpolation problem to a linear algebra problem in an RKHS.

Theorem 3.1.3 (Existence of Interpolation in an RKHS) *Let $\mathcal{H}$ be an RKHS on X with RK K. Given a set $E = \{x_1, \ldots, x_n\} \subseteq X$ of distinct points and $\{\lambda_1, \ldots, \lambda_n\} \subseteq \mathbb{C}$, there exists an interpolant $g \in \mathcal{H}$ if and only if the vector $v = (\lambda_1, \ldots, \lambda_n)^T$ lies in the range of the matrix $(K(x_i, x_j))_{i,j=1}^n$.*

In addition, if $\omega = (\alpha_1, \ldots, \alpha_n)^T$ is a vector satisfying the equation $(K(x_i, x_j))_{i,j=1}^n \omega = v$ then the interpolant given by the formula

$$h = \sum_{i=1}^{n} \alpha_i k_{x_i}$$

is of minimum norm in $\mathcal{H}$. Moreover, $\|h\|^2 = \langle v, \omega \rangle$.

Proof. ($\Rightarrow$:) Assume that a function $g \in \mathcal{H}$ that interpolates the values exists. Then by the previous Theorem 3.1.2 $P_E(g)$ is the solution having minimal norm and it is

$$P_E(g) = \sum_{i=1}^{n} \beta_i k_{x_i}$$

for some scalars $\beta_1, \ldots, \beta_n$. Then for all $i = 1, \ldots, n$

$$\lambda_i = g(x_i) = P_E(g)(x_i) = \sum_{j=1}^{n} \beta_j k_{x_j}(x_i).$$

Therefore

$$v = \left(\sum_{j=1}^{n} \beta_j k_{x_j}(x_i) \right)_{i=1}^{n} = \begin{bmatrix} k_{x_1}(x_1) & k_{x_2}(x_1) & \cdots & k_{x_n}(x_1) \\ k_{x_1}(x_2) & k_{x_2}(x_2) & \cdots & k_{x_n}(x_2) \\ \vdots & & \ddots & \vdots \\ k_{x_1}(x_n) & k_{x_2}(x_n) & \cdots & k_{x_n}(x_n) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} = (K(x_i, x_j))_{i,j=1}^n \omega_1$$

where $\omega_1 = (\beta_1, \ldots, \beta_n)^T$ is the solution of the linear algebra problem

$$(K(x_i, x_j))_{i,j=1}^n \omega = v. \tag{3.1}$$

($\Leftarrow$:) Now assume that v is in the range of the kernel matrix $(K(x_i, x_j))_{i,j=1}^n$, that is there exists a solution $\omega_1 = (\beta_1, \ldots, \beta_n)^T$ of the matrix-vector equation $v = (K(x_i, x_j))_{i,j=1}^n \omega$. Reversing the discussion above by assuming the vector $\omega = (\alpha_1, \ldots, \alpha_n)^T$ satisfies the equation $(K(x_i, x_j))_{i,j=1}^n \omega = v$ and defining a function $h = \sum_{j=1}^n \alpha_j k_{x_j}$, one can easily see that h is in $\mathcal{H}$ and an interpolation function there.

In addition, considering $(K(x_i, x_j))_{i,j=1}^n \omega = v = (K(x_i, x_j))_{i,j=1}^n \omega_1$ one sees that

$$(K(x_i, x_j))_{i,j=1}^n \omega - (K(x_i, x_j))_{i,j=1}^n \omega_1 = (K(x_i, x_j))_{i,j=1}^n (\omega - \omega_1) = 0$$

therefore $\omega - \omega_1 \in ker(K(x_i, x_j))_{i,j=1}^n$. Then $\sum_{j=1}^n (\alpha_j - \beta_j) k_{x_j} = 0$ whence

$$\sum_{j=1}^n \alpha_j k_{x_j} = \sum_{j=1}^n \beta_j k_{x_j} = P_E(g).$$

Thus by calling $\sum_{j=1}^n \alpha_j k_{x_j} = h$, h will be the interpolation function of minimum norm in $\mathcal{H}$.

Moreover, once the solution of the linear algebra problem (3.1) is found, the norm of h can easily be found from

$$\begin{aligned} \|h\|^2 = \left\langle \sum_{j=1}^n \alpha_j k_{x_j}, \sum_{i=1}^n \alpha_i k_{x_i} \right\rangle &= \sum_{j=1}^n \alpha_j \overline{\alpha_i} \langle k_{x_j}, k_{x_i} \rangle \\ &= \sum_{j=1}^n \alpha_j \overline{\alpha_i} K(x_i, x_j) = \left\langle (K(x_i, x_j))_{i,j=1}^n \omega, \omega \right\rangle = \langle v, \omega \rangle. \end{aligned}$$

□

Note that in the case of the kernel matrix $(K(x_i, x_j))_{i,j=1}^n$ is invertible, the solution h of the interpolation problem is readily found by $h = \sum_{i=1}^n \alpha_i k_{x_i}$ where $(\alpha_1, \ldots, \alpha_n)^T = \omega = ((K(x_i, x_j))_{i,j=1}^n)^{-1} v$ which is the solution of the linear algebra problem.

3.1.1 Fully Interpolating RKHS

Recall that a matrix is strictly positive if and only if it is positive semidefinite and invertible. In this part, we will see without giving any proof that when an RK is strictly positive there is an easy way to find the interpolating function having minimum norm.

Theorem 3.1.4 *For given a nonempty set X and a kernel function $K : X \times X \to \mathbb{C}$ the following statements are equivalent.*

I. $K > 0$.

II. *Let $\forall n$, $\{x_1, \ldots, x_n\} \subseteq X$ be any set of distinct points. Then the kernel functions $k_{x_1}, \ldots, k_{x_n}$ are linearly independent.*

III. *Let $\forall n$, $\{x_1, \ldots, x_n\} \subseteq X$ be any set of distinct points, and $\{\alpha_1, \ldots, \alpha_n\} \subseteq \mathbb{C}$ be any set which is not all zero. Then $\exists f \in \mathcal{H}$, where $\mathcal{H}$ is the RKHS of K resulting from theorem* 2.2.3 *satisfying $\alpha_1 f(x_1) + \cdots + \alpha_n f(x_n) \neq 0$.*

IV. *Let $\forall n$, $\{x_1, \ldots, x_n\} \subseteq X$ be any set of distinct points. Then $\exists g_1, \ldots, g_n \in \mathcal{H}$ with*

$$g_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

A collection of functions that meets the condition specified in last statement of the theorem above is commonly referred to as a partition of unity for the set $\{x_1, \ldots, x_n\}$.

Definition 3.1.5 *An RKHS is said to be fully interpolating if one of the equivalent conditions of Theorem* 3.1.4 *is satisfied.*

The way of finding such a partition of unity is to consider the set $\{x_1, \ldots, x_n\}$ and the invertible RK $P = (K(x_i, x_j))_{i,j=1}^n$. By letting $P^{-1} = (b_{i,j})_{i,j=1}^n = B$ with the columns of B ω_j, $j = 1, \ldots, n$ and taking the standard basis e_j, $j = 1, \ldots, n$ for $\mathbb{C}$ we see that for every $j = 1, \ldots, n$, the columns of B are the unique vectors ω_j, which are solutions to $e_j = P\omega_j$. Suppose we set

$$g_j = \sum_{j=1}^{n} b_{i,j} k_{x_j}$$

then $g_j(x_i) = \sum_{j=1}^n b_{i,j} k_{x_j}(x_i) = \sum_{j=1}^n b_{i,j} \langle k_{x_j}, k_{x_i} \rangle = \langle \sum_{j=1}^n b_{i,j} k_{x_j}, k_{x_i} \rangle = \delta_{i,j}$ where $\delta_{i,j}$ refers to the Dirac's delta function. Thus g_j's constitute a partition of unity for given points x_j, $j = 1, \ldots, n$. Hence once a partition of unity for the points $\{x_1, \ldots, x_n\}$ is obtained then the interpolation function f in the RKHS is unique with the minimum norm and can be written easily in the form

$$f = \sum_{j=1}^{n} \lambda_j g_j.$$

Note that this is a specific partition of unity and called canonical partition of unity.

3.2 Best Least Squares Approximant

When the vector $v = (\lambda_1, \ldots, \lambda_n)^T$ is not in the range of the kernel matrix $(K(x_i, x_j))_{i,j=1}^n$ e.g. in the case when the matrix $(K(x_i, x_j))_{i,j=1}^n$ is not invertible, the solution of the in-

terpolation problem may not exist in the RKHS. However, the problem of finding a function that "best resembles" to the interpolant can still be turned to a linear algebra problem and still be found in the RKHS.

Definition 3.2.1 (Best Least Squares Approximant) *A function closest to the interpolant of minimum norm and making the least square error, $J(f) = \sum_{i=1}^{n} |f(x_i) - \lambda_i|^2$, minimum is said to be the best least squares approximant.*

The function $J(f) = \sum_{i=1}^{n} |f(x_i) - \lambda_i|^2$ is also often called loss. The following theorem is frequently called the existence and uniqueness of the best least squares approximant. It shows how the problem of finding a best least squares approximant reduces to a linear algebra problem and that the problem has a simple form of solution. In the following $\mathcal{N}(Q) = \{u \in \mathbb{C}^n \mid Qu = 0\}$ denotes the null space of Q and $\mathcal{R}(Q) = \{Q\omega \in \mathbb{C}^n \mid \omega \in \mathbb{C}^n\}$ refers to the range of Q.

Theorem 3.2.2 *Given an RKHS $\mathcal{H}$ on X with RK K. Let $E = \{x_1, \ldots, x_n\} \subseteq X$ be a set of distinct points, $\{\lambda_1, \ldots, \lambda_n\} \subseteq \mathbb{C}$ and let $Q = (K(x_i, x_j))_{i,j=1}^{n}$. Then a vector $\omega = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{C}^n$ exists satisfying that $(v - Q\omega) \in \mathcal{N}(Q)$ where $v = (\lambda_1, \ldots, \lambda_n)^T$. Here $Q\omega$ is the orthogonal projection P of v onto the range of Q. Moreover, in this case the function g given by the formula*

$$g = \sum_{i=1}^{n} \alpha_i k_{x_i}$$

makes the least square error minimum and is unique with the minimum norm amongst all least square error-minimizing functions in $\mathcal{H}$.

Proof. First notice that there may exist some matrices Q such that $\mathcal{N}(Q) \neq \{0\}$. Let $v = (\lambda_1, \ldots, \lambda_n)^T$. By Theorem 3.1.3 for any $f \in \mathcal{H}$ there exists vector $\omega = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{C}^n$ such that $Q\omega = f$; i.e. we have

$$Q\omega = \begin{bmatrix} k_{x_1}(x_1) & \cdots & k_{x_n}(x_1) \\ k_{x_1}(x_2) & \cdots & k_{x_n}(x_2) \\ \vdots & \ddots & \vdots \\ k_{x_1}(x_n) & \cdots & k_{x_n}(x_n) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} \alpha_1 k_{x_1}(x_1) + \cdots + \alpha_n k_{x_n}(x_1) \\ \alpha_1 k_{x_1}(x_2) + \cdots + \alpha_n k_{x_n}(x_2) \\ \vdots \\ \alpha_1 k_{x_1}(x_n) + \cdots + \alpha_n k_{x_n}(x_n) \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix}.$$

Thus one gets $J(f) = \sum_{j=1}^{n} |f(x_j) - \lambda_j|^2 = \|Q\omega - v\|^2$. The norm $\|Q\omega - v\|$ which is the distance from the point v to the set $\mathcal{R}(Q)$ is the smallest if ω is chosen as in the statement of the theorem, since $\mathcal{R}(Q)$ is complete and convex since it is a finite dimensional vector space in $\mathbb{C}^n$. It follows that by the projection theorem, see Appendix A.0.5, there exists a vector $\omega = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{C}^n$ such that $P_{\mathcal{R}(Q)}v = Q\omega$ where $P_{\mathcal{R}(Q)}$ is the orthogonal projection of $\mathcal{H}$ onto $\mathcal{R}(Q)$. Therefore we see that $\mathbb{C}^n = \mathcal{R}(Q) \bigoplus \mathcal{R}(Q)^{\perp}$ then $v = Q\omega \bigoplus (v - Q\omega)$, hence $Q(v - Q\omega) = 0$, i.e. $(v - Q\omega) \in \mathcal{N}(Q)$.

Assume there is another vector $\omega' = (\alpha'_1, \ldots, \alpha'_n)^T \in \mathbb{C}^n$ such that $P_{\mathcal{R}(Q)}v = Q\omega'$. So $Q\omega = Q\omega'$ which implies that $Q(\omega - \omega') = 0$, $\omega - \omega' \in \mathcal{N}(Q)$. Hence

$$\begin{bmatrix} \alpha_1 k_{x_1}(x_1) + \cdots + \alpha_n k_{x_n}(x_1) \\ \alpha_1 k_{x_1}(x_2) + \cdots + \alpha_n k_{x_n}(x_2) \\ \vdots \\ \alpha_1 k_{x_1}(x_n) + \cdots + \alpha_n k_{x_n}(x_n) \end{bmatrix} = \begin{bmatrix} \alpha'_1 k_{x_1}(x_1) + \cdots + \alpha'_n k_{x_n}(x_1) \\ \alpha'_1 k_{x_1}(x_2) + \cdots + \alpha'_n k_{x_n}(x_2) \\ \vdots \\ \alpha'_1 k_{x_1}(x_n) + \cdots + \alpha'_n k_{x_n}(x_n) \end{bmatrix} = \begin{bmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_n) \end{bmatrix} = P_{\mathcal{R}(Q)}v$$

and g is the unique solution with the minimum norm since here indeed f is projected onto the span of the kernel functions $k_{x_1}, \ldots, k_{x_n}$ which keeps the value of f at $x_1, \ldots, x_n$ and does not increase the norm of f, so the norm of g is minimum. □

Now a natural question is how to find such a best least squares approximant. To find it we can give a method by summarizing this chapter. Let $v = (\lambda_1, \ldots, \lambda_n)^T \in \mathbb{C}^n$ and distinct points $x_1, \ldots, x_n \in X \neq \emptyset$ be given. The best least squares approximant is the solution of the approximation problem of finding a function f that is nearest to the interpolant. In this sense, the method is to find the function f such that the least square error

$$J(f) = \sum_{i=1}^{n} |f(x_i) - \lambda_i|^2$$

is minimized. The method naturally offers the optimization problem

$$\min_{f \in \mathcal{H}} J(f) = \min_{f \in \mathcal{H}} \sum_{i=1}^{n} |f(x_i) - \lambda_i|^2$$

here $\mathcal{H}$ refers to an RKHS, and is called the method of the least square errors. The method of the least squares is among the most traditional statistical techniques. This minimization problem is over an RKHS $\mathcal{H}$, and the solution function $f \in \mathcal{H}$ will be in the form

$$f = \sum_{j=1}^{n} \alpha_j k_{x_j} \tag{3.2}$$

where $k_{x_1}, \ldots, k_{x_n}$ are kernel functions and $\omega = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{C}^n$. In this form, the unknowns are α_j, $j = 1, \ldots, n$. As shown in Theorem 3.2.2, the vector ω is obtained by solving the linear algebra problem $Pv = Q\omega$ where $P = P_{\mathcal{R}(Q)}$, $Q = (K(x_i, x_j)_{i,j=1}^n)$ and $\omega \perp \mathcal{N}(Q)$. Hence the problem of finding the best least squares approximant eventually becomes a linear algebra problem, in other words the method is reduced to solving a linear algebra problem in an RKHS. Furthermore if the matrix Q is invertible then the least square error is 0 and interpolant is in the form (3.2). To find the interpolant, we solve the problem $v = Q\omega$. When Q is not invertible the solution will be again in the form (3.2) but in this case we find the best least squares approximant by solving the problem $Pv = Q\omega$.

CHAPTER 4

APPLICATIONS OF RKHS TO STATISTICS AND MACHINE LEARNING

4.1 The Kernel Trick

Kernel trick which is also called kernelization is a very useful method to turn nonlinear functions in the existence space into linear functions in a space of higher dimension. As we will see, the kernel trick enables one to treat nonlinear functions in $\mathbb{R}^p$ as linear functions in $\mathbb{R}^{p+m}$, where $p, m \geq 1$. Thus some nonlinear problems reduces to linear problems. This kernel trick is just a particular example of pull-back which is applied by a map called the feature map ϕ. The notion of pull-back is described briefly in Appendix B.1.

Definition 4.1.1 *A function $\phi : X \to \mathcal{L}$ is called a feature map where X is a nonempty set and $\mathcal{L}$ is a Hilbert space.*

4.2 Finding the Best Least Squares Approximant in RKHS

In this section, we find the best least squares approximant among all linear functionals, all affine functions on $\mathbb{R}^p$, and all polynomials on $\mathbb{R}$.

4.2.1 Over All Linear Functionals on a Hilbert Space

Consider the space all linear functionals on $\mathbb{R}^p$. In Proposition 2.4.5, it is obtained that all bounded linear functionals on a Hilbert space $\mathcal{L}$ construct an RKHS $\mathcal{H}$ on $X = \mathcal{L}$, whose kernel function is the inner product on $\mathcal{L}$ (i.e. $\mathcal{H} = \mathcal{L}'$, $\mathcal{L}'$ is the dual of $\mathcal{L}$). Since $X = \mathcal{L} = \mathbb{R}^p$, we have $\mathcal{H} = \mathcal{L}' = \mathbb{R}^p$.

Let $g \in \mathcal{H}$ be the best least squares approximant of minimum norm of the problem

$$\min_{f\in\mathcal{H}} J(f) = \min_{f\in\mathcal{H}} \sum_{i=1}^{n} |f(x_i) - \lambda_i|^2.$$

Recall that any $f \in \mathcal{H}$ had the form $f(x) = \langle x, u\rangle$, $\forall x \in \mathbb{R}^p$ where $u \in \mathbb{R}^p$ (reproducing property). By the proof of Proposition 2.4.5, $\mathcal{H}$ has the kernel function for the point x, $k_x = f_x$ so the solution $g \in \mathcal{H}$ is of the form

$$g = \sum_{i=1}^{n} \alpha_i k_{x_i} = \sum_{i=1}^{n} \alpha_i f_{x_i}.$$

The only unknowns here are α_i's $i = 1, \ldots, n$. To find them use the linear algebra problem $Pv = Q\omega$ where P is orthogonal projection onto the $\mathcal{R}(Q)$, $v = (\lambda_1, \ldots, \lambda_n)^T$, Q is the kernel matrix of K and ω is orthogonal to the $\mathcal{N}(Q)$ as in Theorem 3.2.2. Suppose the solution of $Pv = Q\omega$ is found to be $z = (\alpha'_1, \ldots, \alpha'_n)^T$ with $z \perp \mathcal{N}(Q)$. It follows that $\forall x \in \mathbb{R}^p$,

$$g(x) = \sum_{i=1}^{n} \alpha'_i f_{x_i}(x) = \sum_{i=1}^{n} \alpha'_i \langle x, x_i\rangle = \left\langle x, \sum_{i=1}^{n} \alpha'_i x_i \right\rangle = \langle x, u\rangle.$$

Thus, $u = \sum_{i=1}^{n} \alpha'_i x_i$ and hence the best least squares approximant g is found by just solving $Pv = Q\omega$.

4.2.2 Over All Affine Functions in an RKHS

Let $X = \mathbb{R}^p$ and let $f : X \to \mathbb{R}$ be an affine function. Note that any affine function on X can be written as $f(x) = \langle x, \omega\rangle + a$ where $\omega \in \mathbb{R}^p$, $a > 0$. Consider the space of all affine functions f on $\mathbb{R}^p$, say $\mathcal{A}$. Let us identify the affine function $f \in \mathcal{A}$ to be $f := f_{\omega,a}$.

Let $\phi : X \to Y$ where $Y = \mathbb{R}^p \bigoplus \mathbb{R}$ be an embedding map defined by $\phi(x) = (x, 1)$. Then for any $x \in X$

$$f_{\omega,a}(x) = \langle x, \omega \rangle_{\mathbb{R}^p} + a = \langle (x, 1), (\omega, a) \rangle_{\mathbb{R}^{p+1}} = \langle \phi(x), (\omega, a) \rangle_{\mathbb{R}^{p+1}} = g_{\omega,a}(\phi(x)) \quad (4.1)$$

where g is in the space of all bounded linear functionals on Y restricted to $\phi(\mathbb{R}^p) \subset \mathbb{R}^{p+1}$.

We know that on Y there is an RKHS $(\mathbb{R}^{p+1})'$ by Proposition 2.4.5. Assume that $\mathcal{H}_\phi$ is the space of all bounded linear functionals on $\mathbb{R}^{p+1}$ restricted to $\phi(\mathbb{R}^p) \subset \mathbb{R}^{p+1}$. By pull-back of $K'(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathbb{R}^{p+1}} = \langle x, y \rangle_{\mathbb{R}^p} + 1$, see Appendix B.1, the space $\mathcal{A}$ is the RKHS on $\mathbb{R}^{p+1}$ restricted to $\phi(\mathbb{R}^p) \subset \mathbb{R}^{p+1}$. Therefore, the problem of minimization of $J(f)$ over all affine functions on $\mathbb{R}^p$ (over $\mathcal{A}$) reduces to the minimization problem of the loss function over all linear functionals on $\phi(\mathbb{R}^p) \subset \mathbb{R}^{p+1}$ (over $\mathcal{H}_\phi$).

The problem

$$\min_{f \in \mathcal{A}} J(f) = \min_{f \in \mathcal{A}} \sum_{i=1}^{n} |f(x_i) - \lambda_i|^2 \quad (4.2)$$

is equivalent to the problem

$$\min_{g \in \mathcal{H}_\phi} J(f) = \min_{g \in \mathcal{H}_\phi} \sum_{j=1}^{n} |g(\phi(x_j)) - \lambda_j|^2. \quad (4.3)$$

Considering the problem (4.3) we see that we have turned to the previous case of linear functional that is the nonlinear least square error problem (4.2) turns to a linear least square error problem (4.3). Thus solution of (4.3) is obtained by the linear algebra problem $Pv = Q\omega$ as in Section 4.2.1 where Q is the (kernel) matrix corresponding to the points $x_1, \ldots, x_n$, i.e. $Q = (K'(x_i, x_j))_{i,j=1}^n = (\langle x_i, x_j \rangle + 1)_{i,j=1}^n$. Assume the solution of this problem is found to be $z = (\alpha_1, \ldots, \alpha_n)^T$. Then any affine function $f_{\omega,a} \in \mathcal{A}$ can be written by the kernel functions

$$f_{\omega,a} = \sum_{j=1}^{n} \alpha_j k_{\phi(x_j)}.$$

It follows that

$$f_{\omega,a}(x) = \langle \phi(x), (\omega, a) \rangle = \sum_{j=1}^{n} \alpha_j k_{\phi(x_j)}(\phi(x)) = \left\langle (\phi(x)), \sum_{j=1}^{n} \alpha_j \phi(x_j) \right\rangle$$

and one gets $(\omega, a) = \sum_{j=1}^n \alpha_j \phi(x_j)$. Therefore $\omega = \sum_{j=1}^n \alpha_j x_j$ and $a = \sum_{j=1}^n \alpha_j$ since $\phi(x) = (x, 1)$. Hence the problem of optimization over nonlinear affine functions is

solved by the kernel trick which allows to considers $\mathcal{A}$ as $\mathcal{H}_\phi$ and thus by solving the corresponding linear algebra problem.

4.2.3 Over All Polynomials in an RKHS

Consider $E = \{x_1, \dots, x_n\} \subset X = \mathbb{R}$ be data points, $\{\lambda_1, \dots, \lambda_n\} \subset \mathbb{R}$ be values and let d be a fixed integer. In this case the objective is to find a polynomial $f : X \to \mathbb{R}$ that makes the loss $J(f) = \sum_{i=1}^{n} |f(x_i) - \lambda_i|^2$ minimum over an RKHS of all polynomials of degree at most d. Consider the space of all polynomials of degree d on $\mathbb{R}$. Assume $d + 1 < n$ since there can always be found a function that makes the loss 0 when $d \geq n - 1$ for any $n \geq 1$.

To be able to see the space of all polynomials as an RKHS of linear functionals, we apply the kernel trick as before in the affine case. Let $\phi : X \to \mathbb{R}^{d+1}$ be the embedding map given by $\phi(x) = (1, x, \dots, x^d)$. For a vector $\omega = (\omega_0, \dots, \omega_d) \in \mathbb{R}^{d+1}$, any polynomial f identified as $f \coloneqq f_\omega$ on X can be written

$$f_\omega(x) = \omega_0 + \cdots + \omega_d x^d = \langle (1, \dots, x^d), (\omega_0, \dots, \omega_d) \rangle = \langle \phi(x), (\omega_0, \dots, \omega_d) \rangle = g(\phi(x)) \tag{4.4}$$

$\forall x \in X$. Here $g : \mathbb{R}^{d+1} \to \mathbb{R}$ is the linear functional of inner product against $(\omega_0, \dots, \omega_d)$. This shows that by the feature map ϕ the space of all polynomials of degree at most d on X is equal to the space of all linear functionals on $\mathbb{R}^{d+1}$ restricted to $\phi(X)$.

By the Corollary in Appendix B.1, the space of all polynomials of degree at most d is the pull-back of the RKHS of all bounded linear functionals on $\mathbb{R}^{d+1}$ along ϕ. Thus we are able to see the space of nonlinear functions as the space of linear functions by (4.4), that is the space of all polynomials of degree at most d on X is the RKHS on X. The kernel function of the RKHS is $K(x, y) = K'(\phi(x), \phi(y)) = \langle \phi(x), \phi(y) \rangle_{\mathbb{R}^{d+1}}$ for all $x, y \in \mathbb{R}$ and so the kernel function of the RKHS of all polynomials of degree at most d on $\mathbb{R}$ is given by $K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathbb{R}^{d+1}} = 1 + xy + \cdots + (xy)^d$.

Now we can consider the problem of finding the best least squares approximant over all linear functionals on $\mathbb{R}^{d+1}$ restricted to $\phi(X)$ instead of polynomials of degree at

most d on $\mathbb{R}$. The problem of minimization of the loss becomes

$$J(f) = \sum_{j=1}^{n} |g(\phi(x_j)) - \lambda_j|^2$$

over all linear functionals on $\phi(X) \subset \mathbb{R}^{d+1}$ hence it reduces to linear least squares problem. Hence the solution can be found by the linear algebra problem $Pv = Q\omega$ as considered in the previous Sections of this chapter with the only difference the $n \times n$ matrix $Q = (K(\phi(x_i), \phi(x_j)))_{i,j=1}^{n} = (\langle \phi(x_i), \phi(x_j) \rangle)_{i,j=1}^{n}$.

The polynomial solution of the minimization problem is

$$f_\omega = \sum_{j=1}^{n} \alpha_j k_{\phi(x_j)}$$

where

$$\begin{aligned} \omega = \sum_{j=1}^{n} \alpha_j \phi(x_j) =& \alpha_1 \phi(x_1) + \cdots + \alpha_n \phi(x_n) \\ =& \alpha_1 (1, x_1, \ldots, x_1^d) + \cdots + \alpha_n (1, x_n, \ldots, x_n^d) \\ =& ((\alpha_1 + \cdots + \alpha_n), (\alpha_1 x_1 + \cdots + \alpha_n x_n), \ldots, (\alpha_1 x_1^d + \cdots + \alpha_n x_n^d)) \end{aligned}$$

and $z = (\alpha_1, \ldots, \alpha_n)$ is the solution the linear algebra problem $Pv = Q\omega$ with $z \perp ker(Q)$.

Note that all these three cases considered when the matrix Q is not invertible, i.e. the vector $(v - Q\omega) \in \mathcal{N}(Q)$. When the matrix is invertible there will be a function in the RKHS such that the loss is 0. Also, for the map $\phi : X \to Y$ where $Y = \mathcal{L}$ is a Hilbert space, the kernel function $K_Y(\phi(x), \phi(y))$ is on $\phi(X) \subseteq \mathcal{L}$ and $K(x, y)$ is on X and the relation between the two positive semidefinite kernels is $K(x, y) = K_Y(\phi(x), \phi(y))$ by the Proposition B.1.1 in the Appendix. It follows that

$$K(x, y) = K_Y(\phi(x), \phi(y)) = \langle \phi(x), \phi(y) \rangle_{\mathcal{L}}$$

since there occurs an RKHS on $Y = \mathcal{L}$ with the kernel $K_Y(u, v) = \langle u, v \rangle_{\mathcal{L}}$.

4.3 The Representer Theorem

We have seen that the loss function is of the form $J(f) = \sum_{i=1}^{n} |f(x_i) - \lambda_i|^2$. However it can be considered in a more general form, and solution of the general form under

mild assumptions is again in an RKHS and given by a linear combination of kernel functions. The application of a loss or penalty function and functional approximation are common in statistical results.

Let $E = \{x_1, \ldots, x_n\}$ be a subset of $X \neq \emptyset$. Consider the optimization problem

$$min_{f \in H} W(\|f\|_H^2) + L(f(x_1), \ldots, f(x_2)) \tag{4.5}$$

where H is a Hilbert space, W is a monotonically increasing function and L is a loss function. We will specify the assumptions on the functions L later. Indeed the term $W(\|f\|_H^2)$ is added to overcome the problem of overfitting the data to the problem of minimization of the loss. A function overfits data if it makes a good prediction on the existing data and is not good while classifying the new data.

We will show that the existence and uniqueness of a form the Representer Theorem. but first recall the definition of a convex function. Letting $x, y \in S \neq \emptyset$, if for any $0 < \alpha < 1$ we have $\alpha x + (1 - \alpha)y \in S$, S is called a convex set. A real valued function L is said to be convex if for any $0 < \alpha < 1$, the inequality $L(\alpha x + (1 - \alpha)y) \leq \alpha L(x) + (1 - \alpha)L(y)$ holds for all $x, y \in S$.

Theorem 4.3.1 *If L is a convex function then the solution of the problem*

$$min_{f \in H} J(f) = \|f\|_H^2 + L(f(x_1), \ldots, f(x_n))$$

exists and unique.

Proof. To show that the solution is unique, we use a classical idea in the theory of Hilbert space and convexity of L. Let L be convex on a Hilbert space H. To make the notation simpler, we will write $L(f) = L(f(x_1), \ldots, f(x_2))$. Let $f, g \in H$ be solutions i.e. $f, g \in H$ makes $J(f)$ minimum. Then we have $\|f\|_H^2 + L(f) \leq \|\frac{f+g}{2}\|_H^2 + L(\frac{f+g}{2})$ and by the same idea $\|g\|_H^2 + L(g) \leq \|\frac{f+g}{2}\|_H^2 + L(\frac{f+g}{2})$. By adding these two inequalities and dividing by 2 we get

$$\frac{\|f\|_H^2 + \|g\|_H^2}{2} + \frac{L(f) + L(g)}{2} \leq \|\frac{f+g}{2}\|_H^2 + L\Big(\frac{f+g}{2}\Big).$$

Rearranging the inequality

$$\frac{1}{2}\|f\|_H^2 + \frac{1}{2}\|g\|_H^2 \leq \frac{1}{2}L(f) + \frac{1}{2}L(g).$$

Recalling the parallelogram law ($\|f+g\|^2 + \|f-g\|^2 = 2(\|f\|^2 + \|g\|^2)$) one gets

$$0 \leq \|\frac{1}{2}(f-g)\|_H^2 \leq \frac{1}{2}L(f) + \frac{1}{2}L(g) - \frac{1}{2}L(f) - \frac{1}{2}L(g) = 0$$

it follows that $f - g = 0 \iff f = g$. □

Note that the result is still true when the norm of f in Theorem 4.3.1 is replaced by $c\|f\|_H^2$ for some constant c.

By using the representer theorem, we can work directly with the kernel function instead of direct computation of the feature map. Because under reasonable assumptions of the theorem below the solution to the problem in (4.5) above is the linear combination of kernel functions $k_{x_1}, \ldots, k_{x_n}$. Thus the general problem (4.5), possibly infinite dimensional by nature becomes finite dimensional.

Theorem 4.3.2 (A Form of the Representer Theorem) *Let $\mathcal{H}$ be an RKHS on X. Let the function $W : \mathbb{R} \to \mathbb{R}$ is monotonically increasing, the function $L : \mathbb{R}^n \to \mathbb{R}$ is continuous. If there exists a function $f^* = inf_{f\in\mathcal{H}} J(f)$ where*

$$J(f) = W(\|f\|_{\mathcal{H}}^2) + L(f(x_1), \ldots, f(x_n))$$

then f^* is in the span of kernel functions of H, i.e.

$$f^* = \sum_{j=1}^{n} \alpha_i k_{x_i}$$

for some constants $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$.

Proof. Let $x_1 \ldots, x_n \subset X$. Suppose $S = span\{k_{x_1}, \ldots, k_{x_n}\}$. Let $f^* = g + h \in \mathcal{H}$, where $g \in S$ and $h \in S^\perp$. We show that $h = 0$. By the reproducing property

$$h(x_j) = \langle h, k_{x_j} \rangle = 0$$

for all $j = 1, \ldots, n$ since $h \in S^\perp$ and $k_{x_j} \in S$. Then since g and h are orthogonal

$$\begin{aligned} J(f^*) = J(g+h) =& W(\|g+h\|_{\mathcal{H}}^2) + L((g+h)(x_1), \ldots, (g+h)(x_n)) \\ =& W(\|g\|_{\mathcal{H}}^2 + \|h\|_{\mathcal{H}}^2) + L((g+h)(x_1), \ldots, (g+h)(x_n)). \end{aligned}$$

Furthermore, one has $J(g) = W(\|g\|_{\mathcal{H}}^2) + L(g(x_1), \ldots, g(x_n))$. Since W is monotonically increasing we get $J(f^*) \geq J(g)$. Assume $J(f^*) > J(g)$. But this contradicts with the

assumption that $f^* = inf_{f \in \mathcal{H}} J(f)$. Hence $J(f^*) = J(g)$, therefore $\|h\|^2 = 0 \iff h = 0$. □

The theorem shows that when there is a solution f of the general form of the optimization problem over the possibly infinite dimensional RKHS, the solution is again in the finite dimensional space spanned by the kernel functions of $x_1 \ldots, x_n$ i.e.

$$f = \sum_{j=1}^{n} \alpha_j k_{x_j}.$$

4.4 The Kernel Method

The kernel method can be described as a composition of ideas demonstrated in this last chapter. The concepts of feature maps, prediction and optimization in the theory connect the ideas. Feature maps are a key to bring the problems to an RKHS. By a feature map $\phi : X \to \mathcal{L}$ where $\mathcal{L}$ is a Hilbert space, a kernel on X is induced via $K(x, y) = \langle \phi(x), \phi(y) \rangle$ so by the pull-back construction, an RKHS is formed on X. In this sense, the elements of that RKHS is said to be predictors. The advantages of this method are that some problems are linearized and some optimization problems over possibly infinite dimensional RKHS of predictors are frequently turned to finite dimensional linear algebra problems including the matrix $Q = (K(x_i, x_j))_{i,j=1}^n$.

A prediction is the result of an algorithm that was trained using historical data. Prediction problems are the problems of finding a function $f : X \to Y$ that performs well on the available data and makes accurate predictions for new data. Many statistical models include prediction problems. Y might be an interval or be a finite set of points. In the case of Y is an interval the problem of prediction is called the regression problem, when $Y = \{y_1, \ldots, y_m\}$ it is said to be the classification problem. In classification problems, the values within Y may constitute different collections called classes. General way of making prediction is that the predictor f assigns a y_i value to the new data x and then it is predicted that x belongs to the *i-th* class of Y, that is we determine the class of the data not in the given data. This study concerns the classification problems where Y has two finite classes subset of $\mathbb{R}$.

Optimization is the process of determining the maximum or minimum value of a

given real-valued function. Some prediction or classification problems reduce to optimization problems, as we have seen as minimization of the loss, over an RKHS of functionals on a nonempty set X.

4.5 The Problems of Classification and Geometric Separation

In this part, we will mention shape recognition, one of the most basic problems in machine learning and it introduces the concept of linear separation. In this thesis we will consider a concrete case.

Let $S \subseteq \mathbb{R}^2$. Consider S to be a fixed shape which is not known such as ellipse or more generally a conic section. Let $\{x_1, \ldots, x_n\}$ be points in $\mathbb{R}^2$ and assume that x_i's are inside the shape S and others are outside S. Our purpose is to give a considerably nice guess for the shape of S. To do this we first consider the boundary of S as graph of a function g belonging to a particular set of functions. Set $S = \{x = (\alpha, \beta) \in \mathbb{R}^2 \mid g(x) < 0\}$ for some g of the form $g(x) = a + b\alpha + c\beta + d\alpha^2 + e\beta^2$ where $x = (\alpha, \beta)$. Define a feature map $\phi : X = \mathbb{R}^2 \to \mathbb{R}^5$ by $\phi(x) = (1, \alpha, \beta, \alpha^2, \beta^2)$ where $x = (\alpha, \beta)$ and so for the vector $u = (a, b, c, d, e) \in \mathbb{R}^5$, the predictor function can be written by

$$g(x) = \langle \phi(x), u \rangle.$$

This is similar to the problem we had with least squares problems before. We had obtained a linearized problem by getting an RKHS created by a pull-back embedding ϕ on $\mathbb{R}^p$ with the kernel $K(x, y) = \langle \phi(x), \phi(y) \rangle$. In this sense, if we take $p = 2$ we get an RKHS of predictors on X, namely $\mathcal{H} = \{g : X \to \mathbb{R} \mid g(x) = \langle \phi(x), u \rangle\}$. $\mathcal{H}$ contains all circles, ellipses in $X = \mathbb{R}^2$ and many other functions. For example if our problem was to find a circle which resembles the most to S, then a constraint in the form $u_4 = u_5, u_2^2 + u_3^2 - 4u_1^2 > 0$, which increases the complexity, would have added to the optimization problem.

Instead, to predict g try to get information using S. Choose some points $\{x_1, \ldots, x_n\} \subset \mathbb{R}^2$ and then ask an *oracle* whether or not the points are inside or outside S. Then label data points to be $\lambda_i = -1$ if $x_i \in S$ and $\lambda_i = 1$ if $x_i \notin S$, indeed one has $\lambda_i = sign(g(x))$ because S is defined when $x \in S, g(x) < 0$.

Let $\mathcal{X}_- = \{x_i \in \mathbb{R}^2 \mid x_i \in S\}$ and $\mathcal{X}_+ = \{x_i \in \mathbb{R}^2 \mid x_i \notin S\}$. Now we are in search of a function f such that $sign(f(x)) < 0$ at x_i with $\lambda_i = -1$ and $sign(f(x)) > 0$ at the remaining data x_i with $\lambda_i = 1$. Thus the problem of finding the shape of S f was reduced to finding a vector $u \in \mathbb{R}^5$ such that $f(x) = \langle \phi(x), u \rangle$ by the feature map and divided into two parts

$$f(x_i) = \langle \phi(x_i), u \rangle < 0$$

on $\mathcal{X}_-$ and on $\mathcal{X}_+$

$$f(x_i) = \langle \phi(x_i), u \rangle > 0$$

i.e. two sets $\phi(\mathcal{X}_-)$ and $\phi(\mathcal{X}_+)$ are separated by a hyperplane determined by u. Hence the problem boils down to finding the point at which a hyperplane in a Hilbert space separates two sets of points, and when this occurs. If there is such a hyperplane it can be shown that there are infinitely many.

REFERENCES

[1] D. Hilbert, "Grundzüge einer allgemeinen theorie der linearen integralrechnungen.(zweite mitteilung)," *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, vol. 1904, pp. 213–260, 1904.

[2] S. Zaremba, *Léquation Biharmonique et une Classe Remarquable de Fonctions Fondamentales Harmoniques*. Imprimerie de L'Universite, 1907.

[3] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Reproducing kernel hilbert space, mercer's theorem, eigenfunctions, nyström method, and use of kernels in machine learning: Tutorial and survey," *ArXiv Preprint ArXiv:2106.08443*, 2021.

[4] G. Szegő, "Über orthogonale polynome, die zu einer gegebenen kurve der komplexen ebene gehören," *Mathematische Zeitschrift*, vol. 9, no. 3, pp. 218–270, 1921.

[5] M. Bergmann, H. Schotte, and W. Lechinsky, "Über die ungesättigten reduktionsprodukte der zuckerarten und ihre umwandlungen, iii.: Über 2-desoxyglucose (gluco-desose)," *Berichte Der Deutschen Chemischen Gesellschaft (A and B Series)*, vol. 55, no. 1, pp. 158–172, 1922.

[6] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, no. 3, pp. 337–404, 1950.

[7] L. Schwartz, "Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants)," *Journal D'analyse Mathématique*, vol. 13, pp. 115–256, 1964.

[8] S. Saitoh, Y. Sawano *et al.*, *Theory of Reproducing Kernels and Applications*. Springer, 2016.

[9] V.I. Paulsen and M. Raghupathi, *An Introduction to The Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016, vol. 152.

[10] L. Tartar, *An Introduction to Sobolev Spaces and Interpolation Spaces*. Springer Science & Business Media, 2007, vol. 3.

[11] E. Kreyszig, "Introductory functional analysis with applications. jonh wiley & sons," *Inc., New York*, 1978.

Appendix A

Hilbert Space

Definition A.0.1 (Schauder Basis) *Let X be a normed space. A sequence $(e_n)_{n\geq 1} \in X$ with the property for every $x \in X$ there exists a unique sequence of numbers $(\alpha_n)_{n\geq 1}$ such that as $n \to \infty$*

$$\|x - \alpha_1 e_1 + \cdots + \alpha_n e_n\| \to 0$$

is called a Schauder basis for X.

In addition, any $x \in X$ can be written as $x = \sum_{n=1}^{\infty} \alpha_n e_n$ and the expression $\sum_{n=1}^{\infty} \alpha_n e_n$ is said to be the expansion of x with respect to $(e_n)_{n\geq 1}$.

Theorem A.0.2 *Let X be a normed space and $Y \subseteq X$ be a finite dimensional subspace. Then Y is closed.*

Theorem A.0.3 *For a normed space X and a finite dimensional subspace Y of X, Y is complete. Specifically, all normed spaces with finite dimension are complete.*

Theorem A.0.4 *Let X be an inner product space and $\emptyset \neq M \subset X$. If M is convex, complete in the metric induced by the inner product then for every given $x \in X$ there exists a unique vector $y \in M$ such that*

$$inf_{\tilde{y}\in M}\|x - \tilde{y}\| = \|x - y\|.$$

Theorem A.0.5 *Let H be a Hilbert space and Y be a closed subspace of H. Then $H = Y \bigoplus Y^{\perp}$ where $Y^{\perp}$ is the orthogonal complement of Y.*

Definition A.0.6 *For a Hilbert space H and its closed subspace Y the linear operator $P : H \to Y$ given by $x \mapsto y = Px$ is called (orthogonal) projection of H onto P.*

Theorem A.0.7 (Bounded Linear Extension) *Let $T : \mathcal{D}(T) \to Y$ be a bounded linear operator where $\mathcal{D}(T) \subseteq X$ is a vector space, X is a normed space and Y is a Banach space. Then T has an extension $\tilde{T}$ from the closure of the domain, $\overline{\mathcal{D}(T)}$ to Y which is bounded linear and the norms of T and $\tilde{T}$ are the same i.e. $\|T\| = \|\tilde{T}\|$.*

For further information and detailed explanations, the book [11] can be revised.

Appendix B

Pull-Back

The following theorem given without proof states the characterization of the elements of an RKHS.

Theorem B.0.1 (Characterization of Elements of $\mathcal{H}(K)$ by K) *For a given RKHS $\mathcal{H}$ on X with RK K and a function $f : X \to \mathbb{C}$ the following conditions are equivalent:*

I. f *in* $\mathcal{H}$*;*

II. $\exists c \geq 0$*, such that for any* $E = x_1, \ldots, x_n \subseteq X$ *there exists a function* $h \in \mathcal{H}$ *having the norm* $\|h\| \leq c$ *and* h *equals to* f *at any* x_j*,* $j = 1 \ldots, n$ $f(x_j) = h(x_j)$*;*

III. $\exists c \geq 0$*, such that the function* $c^2K(x, y) - f(x)\overline{f(y)} \geq 0$ *i.e. is a positive semidefinite kernel.*

In addition, when $f \in \mathcal{H}$*,* $\|f\|$ *is the least* c *satisfying both* $\|h\| \leq c$ *and* $c^2K(x, y) - f(x)\overline{f(y)} \geq 0$*.*

B.1 The Pull-Back

Let $X \neq \emptyset$ with the subset $S \subseteq X$ and let $K : X \times X \to \mathbb{C}$ be a kernel function. Then the restriction of K on S $K|_{S\times S} : S \times S \to \mathbb{C}$, often denoted as $K|_S$, is also a kernel function. As expected, there is a relation between the K and $K|_S$ and that will be shown for a more general case in the following by Proposition $B.1.1$ below.

Let $\varphi : S \to X$ be a function. Then $K \circ \varphi : S \times S \to \mathbb{C}$ is another function defined by

$K \circ \varphi(s,t) \coloneqq K(\varphi(s), \varphi(t))$, here the function $K \circ \varphi$ is actually of the form $K \circ (\varphi \times \varphi)$. Then $K(\varphi(s), \varphi(t)) = K(x,y)$ for some $x, y \in X$.

We wonder if $K \circ \varphi$ is a kernel function for any φ. We will now consider this in two cases. When φ is one to one, by identifying each x_i with $\varphi(s_i)$ $\forall i = 1, \ldots, n$ we see that $\{x_1, \ldots, x_n\} = \{\varphi(s_1), \ldots, \varphi(s_n)\} \subseteq X$ and so the definition of a positive semidefinite kernel is automatically satisfied by the positive semidefiniteness of K on X. If φ is any function, $K \circ \varphi$ is also a positive semidefinite kernel which will be shown in the following proposition.

Proposition B.1.1 *Let $X \neq \emptyset$, $S \neq \emptyset$ be two nonempty sets and let $K : X \times X \to \mathbb{C}$ be a kernel function. Then for any function $\varphi : S \to X$, the function $K \circ \varphi : S \times S \to \mathbb{C}$ is a kernel.*

Proof. Let $n \geq 1$, $\{s_1, \ldots, s_n\} \subseteq S$ be a subset and complex scalars $(\alpha_i)_{i=1}^n$ be given. It is needed to be shown that for any $n \geq 1$, choice of $\{s_1, \ldots, s_n\} \subseteq S$, and for any $(\alpha_i)_{i=1}^n \in \mathbb{C}$, $\sum_{i,j=1}^n \alpha_j \overline{\alpha_i} K(\varphi(s_i), \varphi(s_j)) \geq 0$. Start with identifying the points in X and the images of φ as $\{x_1, \ldots, x_p\} = \{\varphi(s_1), \ldots, \varphi(s_n)\}$ where $n \geq p$. Then define

$$A_k \coloneqq \{i \mid \varphi(s_i) = x_k\}, \qquad \beta_k \coloneqq \sum_{i \in A_k} \alpha_i.$$

Considering $\{x_1, \ldots, x_p\} = \{\varphi(s_1), \ldots, \varphi(s_n)\}$ and the definition of A_k one obtains

$$\sum_{i,j=1}^n \alpha_j \overline{\alpha_i} K(\varphi(s_i), \varphi(s_j)) = \sum_{k,l=1}^n \sum_{i \in A_k} \sum_{j \in A_l} \alpha_j \overline{\alpha_i} K(x_i, x_j) = \sum_{k,l=1}^n \beta_l \overline{\beta_k} K(x_i, x_j) \geq 0.$$

Therefore, desired result is obtained since K is p.s.d on X. □

The following theorem, given without proof explains the connection between the RKHSs on X and S generated by the kernel functions K and $K \circ \varphi$, respectively.

Theorem B.1.2 (Pull-Back) *For two given nonempty sets S, X, a given function $\varphi : S \to X$ and a given kernel function $K : X \times X \to \mathbb{C}$, any element in $\mathcal{H}(K \circ \varphi)$ is of the form $f \circ \varphi$ where $f \in \mathcal{H}$ and the norm of any $u = f \circ \varphi \in \mathcal{H}(K \circ \varphi)$ which can be written by different $f \in \mathcal{H}$ is the minimum norm of such f's, that is $\|u\|_{\mathcal{H}(K \circ \varphi)} = min\{\|f\|_{\mathcal{H}} \mid u = f \circ \varphi\}$.*

Corollary B.1.3 (Restriction) *Let $X \neq \emptyset$ with $S \subset X$. Let K be a kernel function on X. If $K|_S$ denotes to restriction of K on S then $K|_S$ is a positive semidefinite kernel on S. A function $h \in \mathcal{H}(K|_S)$ if and only if $h = f|_S$ for $f \in \mathcal{H}(K)$. Furthermore, for any $u \in \mathcal{H}(K|_S)$, $\|u\|_{\mathcal{H}(K|_S)} = min\{\|f\|_{\mathcal{H}(K)} \mid u = f|_S\}$.*

Definition B.1.4 (Pull-Back) *Let X and S be nonempty sets, let $\varphi : S \to X$ be any function and let $K : X \times X \to \mathbb{C}$ be a kernel function. Then the RKHS, $\mathcal{H}(K \circ \varphi)$, produced by $K \circ \varphi$ is said to be pull-back of $\mathcal{H}(K)$ along φ. In addition, the linear map $C_\varphi : \mathcal{H}(K) \to \mathcal{H}(K \circ \varphi)$ is called the pull-back map.*