State of Libya
Sirt University
Office of Graduate Studies
Faculty of Sciences
Department of Mathematics

LEVERAGING POSITIVE DEFINITE KERNELS for IMPROVED PERFORMANCE in MACHINE LEARNING MODELS

This thesis is submitted in fulfillment of the requirements for a higher degree (Master's) in Mathematics

Prepared by:

AYA ZIDAN MOHAMMED ALZARQIU

(22026102)

Supervised by:

Dr. SOUAD AHMED ABUMARYAM

2024-2025

دولة ليبيا
جامعة سرت
مكتب الدراسات العليا
كلية العلوم
قسم الرياضيات

الاستفادة من النواة الإيجابية المحددة لتحسين الأداء في نماذج التعلم الآلي

قدمت هذه الرسالة استكمالاً لمتطلبات الإجازة العالية (الماجستير) في الرياضيات

إعداد
أية زيدان محمد الزرقي
(22026102)
تحت إشراف
د. سعاد أحمد أبو مريم

2024-2025

Declaration

I am Aya Zidan Mohammed Alzarqiu, and I confirm that the work in this thesis, unless otherwise referenced is researcher`s own work, and has not been previously submitted to meet requirements of an award at this University or any other education or research institution, I furthermore, cede copyright of this thesis in favour of Sirt University.

Student name: Aya zidan Mohammed Alzarqiu.

Signature:

Date: 27/05/2025.

Dedication

To the souls of those who left an everlasting mark on my journey, Dr. Abdulhamid Al-Hamad, Dr. Emtair Emharhar and Ms. Joud Makhzom, you remain in our hearts forever, their memory will never fade away. May they be forever remembered in my heart. May God have mercy on you. To my parents, my first teachers and my forever greatest heroes, your sacrifices paved the way I walk today. To my siblings, my refuge and strength, thank you for always standing by me.

To my little princess, Doha, my dream is to one day see you holding your own PhD thesis. To my professors, colleagues, and every kind soul who crossed my path, thank you for being part of this journey. Finally, to everyone who inspired me to be better, through a word, a gesture, or even silent presence, these pages carry your imprint.

Acknowledgments

First and foremost, I praise Allah, the Most Gracious, the Most Merciful, for granting me the strength, wisdom, and perseverance to complete this work. To my esteemed supervisor, Dr. Souad Ahmed Abumaryam, words cannot express my gratitude for your unwavering guidance, patience, and invaluable feedback. Your expertise and encouragement were the compass that led me through every challenge. To the faculty members of my department and college, thank you for enriching my academic journey with your knowledge, inspiring lectures, and continuous support. This achievement is a reflection of your dedication to education. To my beloved university you were the soil where my academic roots grew. Thank you for the resources, challenges, and community that molded me. Finally, to every person who contributed to this journey whether through a word of advice, a helping hand, or silent encouragement you have my deepest gratitude.

الملخص

على الرغم من التقدم الكبير في اســـتخدام النوى موجبة التحديد في تعلم الآلة، لا تزال هناك تحديات جوهرية في فهم العلاقة بين الخصـائص الرياضـية للنوى وقدرتها العملية على تحسـين أداء النماذج في التطبيقات الواقعية.

تتناول هذه الرســـالة دراســـة شـــاملة بين التحليل النظري والتجريبي حول النواة المحددة الموجبة Positive definite kernel في مجـال التعلم الآلي. في الجـانـب النظري، تم عرض أنواع النوى وخصـــائصـــها الرياضـــية بالإضـــافة إلى النظريات والتعاريف المرتبطة بها. يبرز البحث دور فضـــاء هيلبرت للنواة المولـدة(Reproducing Kernel Hilbert space (RKHS في تحويـل البيـانـات المعقدة إلى تمثيلات خطية قابلة للفصل، مع تقديم شرح دقيق لشروط ميرسر Mercer's condition باعتبارها إطاراً أساسيًا لاختيار النواة الأمثل.

أما في الجانب العملي، تم اختيار ثلاثة أنواع من النوى لتحليل أدائها، وهي النواة الخطية Linear kernel ونواة متعدد الحدود polynomial kernel والنواة Radial Basic Function (RBF) وقد أُجريت مقارنة بين هذه الأنواع باستخدام مجموعة بيانات تمثيلية مع تسليط الضوء على تأثير كل نواة في تحسين أداء النموذج.

تم تنفيذ التحليل باستخدام لغة البرمجة بايثون Python، مما يعزز الجانب التطبيقي للدراسة ويبرز فاعلية النواة المحددة الموجبة في معالجة البيانات المعقدة. تهدف هذه الرسالة إلى تقديم رؤية متكاملة تجمع بين التحليل النظري العميق والتنفيذ العملي لتوفير إطار عمل يساعد في تحسين أداء النماذج في مجال التعلم الآلي من خلال اختيار النواة الأمثل لكل تطبيق.

الكلمات المفتاحية: طريقة النواة، النواة الإيجابية المحددة، آلة دعم المتجهات، فضاء هيلبرت، فضاء هيلبرت المولد للنواة، دالة الأساس الشعاعي، التعلم الآلي.

Abstract

Despite significant advances in using positive definite kernels in machine learning, fundamental challenges remain in understanding the relationship between the mathematical properties of kernels and their practical ability to enhance model performance in real-world applications.

This Master's thesis presents a comprehensive study combining theoretical analysis and empirical investigation of positive definite kernels in machine learning. On the theoretical side, the research examines various kernel types and their mathematical properties, along with relevant theorems and definitions. The study highlights the role of Reproducing Kernel Hilbert Space (RKHS) in transforming complex data into linearly separable representations, while providing a rigorous explanation of Mercer's conditions as a fundamental framework for optimal kernel selection.

On the practical side, three kernel types were selected for performance analysis: the Linear kernel, Polynomial kernel, and Radial Basis Function (RBF) kernel. A comparative study of these kernels was conducted using a representative dataset, with particular emphasis on each kernel's impact on model performance enhancement.

The analysis was implemented using the Python programming language, strengthening the applied aspect of the study and demonstrating the effectiveness of positive definite kernels in processing complex data. This thesis aims to provide an integrated vision that combines deep theoretical analysis with practical implementation, offering a framework to improve model performance in machine learning through optimal kernel selection for each application.

Key words: Kernel Methods, positive definite kernel, Support Vector Machine (SVM), Hilbert Space, Reproducing Kernel Hilbert Space (RKHS), Radial Basis Function (RBF), Machine learning.

Contents

List of Figures

List of tables

List of symbols

$\boldsymbol{k}$	Kernel
$\boldsymbol{\mathcal{H}}$	Hilbert space
$\boldsymbol{\mathcal{T}}$	Integral operator
$\boldsymbol{G}$ or $\boldsymbol{K}$	Gram matrix
$\boldsymbol{l^2}$	Sequence space
$\boldsymbol{L^2}$	Lebesgue space
$\boldsymbol{L_2}$	Square integrable functions
$\mathbb{R}$	The set of real number
$\mathbb{N}$	The set of natural number
$\mathbb{C}$	The set of complex number

Chapter one

Introduction

Chapter 1

Introduction

1.1 Introduction

Nowadays, there have been significant improvements in machine learning, particularly regarding kernel methods. In machine learning, a kernel is a function that examines how similar two data points are to another. Specifically, a kernel function is a mathematical tool that calculates the inner product between two feature vectors in a high-dimensional space without explicitly computing the feature vectors themselves.

The representational strength and performance of these models are greatly affected by the kernel function choice. The choice of kernel function is an important design decision that can have a considerable impact on the performance of a kernel-based model. However, this is not a limitation of the method, but rather an indication of the necessity of determining the appropriate inductive bias for a given situation. With a suitable kernel, kernel methods can provide innovative performance in a wide range of applications. kernel methods have shown to be among the most effective and widely employed machine learning techniques.

A particular kind of kernels known as positive definite kernels fulfills specific mathematical requirements, which makes them especially helpful for a range of machine learning applications and provide the stability and validity of the underlying optimization problems. Positive kernel can be specified on any set, which is an important advantage. To apply linear techniques in high-dimensional feature spaces, one needs understand fundamental concepts in machine learning, such as kernel functions, kernel methods and positive definite kernels. It is essential for understanding these ideas in order to implement kernel-based.

Positive definite kernels are crucial in machine learning, particularly within kernel methods. A positive definite kernel is a function that satisfies certain mathematical properties, which make it a valuable tool for various applications. They define a method for quantifying similarity among data points and play a pivotal role in algorithms such as Support Vector Machines (SVMs), Gaussian Processes, kernel ridge regression and Kernel Principal Component Analysis (KPCA). These kernels enable effective learning and generalization

from data in various machine learning tasks. Positive definite kernel functions, which implicitly send the input data into a high-dimensional feature space, which are essential to the effectiveness of kernel methods.

Kernel methods are classification algorithms that use a kernel function K to map data points from input space V to a higher dimensional feature space V'. This allows for clearer separability between classes of data (Mengoni & Di Pierro, (2019).

Reproducing Kernel Hilbert Space (RKHS) theory was developed by Aronszajn (1950) and Mercer (1962), introducing foundational concepts for kernel methods. This theoretical advancement paved the way for subsequent applications across computational domains (Schölkopf, 2002).

Kernel methods aim to identify linear patterns within a transformed, high-dimensional feature space, where input items are compared using dot products of their representations. This feature space provides a vector-based description of the data defined by a specific set of features. However, kernel methods can avoid direct access to this feature space. Instead, they utilize a kernel function, a symmetric positive semi-definite function that computes the similarity between pairs of items directly in their original input space. This approach allows for effective analysis without the need to explicitly map the data into the high-dimensional feature space.

One of the reasons of this success is the use of kernels. Positive definiteness has to be checked for kernels to be suitable for most of these methods. kernels can also be understood from the perspective of functional analysis, as each kernel k defined on a set X induces a unique Hilbert space $\mathcal{H}_k$ of real-valued functions on X. The term “reproducing” indicates that for every f in $\mathcal{H}$, the values $f(x)$ can be obtained from the inner product in $\mathcal{H}$. Through a standard construction, involving the set X and the kernel K, one can derive a Hilbert space $\mathcal{H}$ of functions defined on X, known as reproducing kernel Hilbert space (RKHS). Machine learning algorithms are built on assumptions derived from scientific models, incorporating principles from calculus, statistics, and probability. The term "cybernetics," coined by mathematician and philosopher Norbert Wiener, refers to the scientific study of control and communication among animals, humans, and machines. This notion that such processes can

be studied within the same framework across biological and artificial systems marks a conceptual revolution.

Machine learning is grounded in various interdisciplinary fields, including linear algebra, statistical learning theory, neural networks, pattern recognition, and artificial intelligence.

Kernel machines are algorithms that select functions with desirable properties from a pre-defined reproducing kernel Hilbert space (RKHS) based on sample data. Each kernel estimation procedure first establishes a criterion that combines multiple properties. The optimal element f from the RKHS, which meet this criterion, is then selected through an optimization procedure.

Mercer (1909) introduced kernel trick and Mercer's theorem, which provides conditions under which a function can be represented as a sum of products of functions, forming the basis for the concept of positive definite kernels. This paved the way for the development of the idea of positive definite kernels related to the integral equations. Schoenberg (1938) provided foundational insights into which kernels can be used to compute distances in feature spaces. which play a crucial role in the foundation for understanding the positive definite kernels, which are essential for kernel methods in machine learning and functional analysis. Kolmogorov (1941) Studied methods for representing kernels in linear spaces specifically for countable input domains. Aronszajn (1950) Developed methods for representing kernels in linear spaces for the general case.

Dunford and Schwartz (1963) Showed that Mercer's theorem, which concerns positive definite kernels, holds true for general compact spaces. Aizerman, Braverman, and Rozonoer (1964) Introduced the use of Mercer's theorem in machine learning, specifically interpreting kernels as inner products in feature spaces, which became foundational for kernel methods. Cover (1965) Justified the application of a non-linear transformation followed by a linear transformation in machine learning contexts, influencing the development of non-linear kernels. Poggio (1975) Introduced the idea of polynomial kernels, expanding the types of kernels used in machine learning. Berg, Christensen, and Ressel (1984) Published a significant study on the theory of kernels, contributing to the theoretical understanding of kernel methods. Micchelli (1986) Explored closure properties related to kernels, further deepening the mathematical foundations of kernel methods. The concept of a positive-

definite kernel further developed to become a central concept in Machine Learning. It was not until 1995 that the theory of positive definite kernels was applied in practical contexts within the field of machine learning.

The practical application of positive definite kernels crystallized with Vapnik's (1995) introduction of the Support Vector Machine (SVM) algorithm, a culmination of his work on Statistical Learning Theory between 1960–1990 Vapnik (1998). Statistical Learning Theory is framework formalized key machine learning concepts including regularization, loss functions, and kernel-based feature mappings providing theoretical guarantees for model generalization.

Sabri and colleagues (2005) characterized conditionally positive definite kernels. In particular, conditionally positive definite kernels have been shown to be suitable for the SVM algorithm. they presented a novel kernel in the framework of SVM, which they termed the Log kernel, and it seems to perform particularly well in their image recognition testing. Marco et al. (2010) explain the kernel methods that are for lazy people but should not say that. They offer solutions for performing a wide range of tasks on various datasets with greater ease and efficiency.

Badillo (2020) emphasized that machine learning algorithms including kernel methods are grounded in mathematical principles from calculus, statistics, and probability theory, necessitating rigorous validation of their assumptions. Harald (2022) proposed reproducing kernel Hilbert spaces and made a first construction of an RKHS out of a Hermitian positive definite kernel conditionally positive definite kernels, a generalization of positive definite kernels, and create an RKHS with regard to them. Maryam (2024) present two distinct approaches were described for developing new stable bases for conditionally positive definite kernel-based spaces. Both of these methods rely on working with a reproducing kernel from the corresponding Native Hilbert Space of CPD kernels. Other bases with distinct properties were obtained by factorizing the kernel matrix differently.

The research addresses several critical issues in the study and application of positive definite kernels in machine learning. A significant theoretical gap exists, as many researchers and practitioners lack a comprehensive understanding of the mathematical properties underlying these kernels, despite their importance in model performance and generalization. Practical

challenges also arise when applying these kernels to real-world problems, as effective implementation requires careful selection and customization tasks often hindered by limited familiarity and insufficient guidance.

Current research also suffers from a narrow focus on a restricted subset of kernel types, overlooking the potential benefits that a more diverse range of positive definite kernels could offer across different applications. Additionally, there is a notable absence of systematic empirical comparisons between models utilizing positive definite kernels and traditional kernel methods, leaving their relative advantages and trade-offs poorly understood.

To address these gaps, there is a pressing need for a comprehensive study that bridges theoretical insights with practical applications while expanding empirical investigations into the broader utility of positive definite kernels. This thesis aims to explore the theoretical and practical applications of positive definite kernels in machine learning, with the goal of enhancing model accuracy, flexibility, and generalization.

The study seeks to establish a rigorous mathematical foundation for kernel-based learning, emphasizing the role of advanced mathematical concepts in improving machine learning techniques. Specifically, it will investigate the theoretical properties of positive definite kernels and their relationship with reproducing kernel Hilbert spaces (RKHS), examining how this connection can be exploited to enhance kernel-based learning algorithms. Additionally, the research will implement and integrate these kernels into widely used machine learning models, such as support vector machines (SVMs) and Gaussian processes, to assess their effectiveness in real-world applications. Another key objective is to analyze the interplay between positive definite kernels and Hilbert spaces, elucidating their theoretical implications for kernel methods.

By bridging theoretical insights with practical implementations, this study aims to contribute to a deeper understanding of kernel methods while expanding their applicability in machine learning.

Positive definite kernels represent a fundamental component in many machines learning algorithms, offering powerful capabilities for data transformation and effective modeling. In light of this, the present study proposes a structured plan to explore the benefits and applications of positive definite kernels, aiming to enhance model performance and

contribute to both theoretical and practical advancements in the field. To achieve this, the study is divided into the following chapters:

Chapter 1: **Introduction**

Provides a comprehensive introduction to the mathematical foundation of machine learning and kernels which include functional analysis, linear algebra, and statistical theory.

Chapter 2: **Kernel method**

Introduce the concept of the kernel method and kernel function types and its properties.

Chapter 3: **Positive definite kernel**

Introduce the theoretical foundation of positive definite kernel, their properties and reproducing kernel Hilbert space. Discuss the relation between them.

Chapter 4: **Implementation and Results**

Applying the positive definite kernel function in machine learning.

Chapter 5: Conclusion and Discussion

Summarizes the main findings and their implications, and provides practical or theoretical recommendations based on the results.

1.2 Some Basic Concept in Function Analysis

Definition 1.2.1 (Metric space): Let X be a set and function $d: X \times X \to \mathbb{R}$ is the distance between two elements such that for any $x, y \in X$:

1- $d(x,y) \geq 0$
2- $d(x,y) = 0 \iff x = y$
3- $d(x,y) = d(y,x)$
4- $d(x,z) + d(z,y) \geq d(x,y)$

Definition 1.2.2 (vector space): Let a set V is vector over field $\mathcal{F}$ called scalars then v is vector space $(V, +, .)$ if we have two operations:

1- $\forall\, \alpha, \beta \in V \implies \alpha + \beta \in V$.
2- $\forall\, w \in V$ and $\alpha \in \mathcal{F} \implies \alpha w \in V$.

Definition 1.2.3 (Inner product space): The inner product space (Pre-Hilbert space) is a vector space V over $\mathbb{R}$ with function $(\langle .,. \rangle, V)$ such that for any $x, y, z \in V$ and $\alpha \in \mathbb{R}$:

1- $\langle x, y \rangle \geq \langle y, x \rangle$
2- $\langle x, x \rangle \geq 0 \; and \; \langle x, x \rangle = 0 \Leftrightarrow x = 0$
3- $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$

Definition 1.2.4 (Normed vector space) (Bloehdorn, 2008)**:** A vector space V on function $\|.\|: V \to \mathbb{R}$ is called normed space $(V, \|.\|)$. For any $x, y \in V$ and $\alpha \in \mathbb{R}$, the norm satisfies:

1- $\|x\| \geq 0$ and $\|x\| = 0 \Leftrightarrow x = 0$.
2- $\|x + y\| \leq \|x\| + \|y\|$.
3- $\|\alpha x\| = |\alpha| \|x\|$

Where $\|x\| = \sqrt{\langle x, x \rangle}$ and $d(x, y) = \|x - y\|$.

Definition 1.2.5 (Cauchy sequence): Let a sequence $\{x_i\}_{i=1}^{\infty}$ in a normed space v, a Cauchy sequence is for every $\varepsilon > 0$, there exists $N \in \mathbb{N}$, such that

$$|x_m - x_n| < \varepsilon \quad \forall \, n, m > N$$

Where a Cauchy sequence is converging to point $x \in V$ if $\|x_n - x\| \to 0$ as $n \to \infty$.

Definition 1.2.6 (Complete space): A space is complete if every Cauchy sequence in the space is convergent.

Definition 1.2.7 (Hilbert space) (Krebs, 2007)**:** A Hilbert space is a complete inner product space.

Definition 1.2.8 (Hilbert space of function) (Schölkopf, 2002)**:** Let $C[a, b]$ be the real-valued continuous function over the interval $[a, b]$. For all $f, g \in C[a, b]$, such that

$$\langle f, g \rangle = \int_a^b f(x) g(x) dx$$

Definition 1.2.9 (Banach space): A Banach space is a complete normed space.

Theorem 1.2.1 (Cauchy-Schwarz inequality): Let V be an inner product space and $x, y \in v$. Then we have: $|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$

Theorem 1.2.2 (Holder inequality) (Wigren, 2015)**:** For any $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ and $b = (b_1, \dots, b_n) \in \mathbb{R}^n$ satisfies Holder's inequality if:

$$\sum_{k=1}^{n} |a_k b_k| \leq \left(\sum_{k=1}^{n} |a_k|^p \right)^{\frac{1}{p}} \left(\sum_{k=1}^{n} |a_k|^q \right)^{\frac{1}{q}}$$

Where $p, q \in (1, \infty)$ and $\frac{1}{p} + \frac{1}{q} = 1$.

Definition 1.2.10 (Fredholm integral equation): A Fredholm integral equation is of the type for all $x \in [a, b]$:

$$h(x)u(x) = f(x) + \int_a^b k(x, s)f(s)ds \quad (1.2.1)$$

where if $h(x) = b$ then we say the Fredholm integral equation lower and upper limits are constant.

If $h(x) = 0$ in (1.2.1) then this equation called Fredholm integral equation of first kind such that:

$$-f(x) = \int_a^b k(x, s)f(s)ds \quad (1.2.2)$$

If $h(x) = 1$ in (1.2.2) then this equation is called Fredholm integral equation of second kind such that:

$$u(x) = f(x) + \int_a^b k(x, s)f(s)ds$$

Definition 1.2.11 (Hadamard product): Consider A and B be $m \times n$ matrices then the Hadmard product of A and B such that:

$$[A \circ B]_{ij} = [A]_{ij}[B]_{ij}$$

Definition 1.2.12 (Hermitian matrices): Consider A be matrix on $\mathbb{C}$, then A is called Hermitian, if $A^* = A$, where $A^* = \overline{A^T}$ the conjugate transpose.

Definition 1.2.13 (Orthonormal Basis) (Deisenroth, Faisal, & Ong, 2020)**:** Let we have an n-dimensional vector space V, a basis $\{b_1, b_2, \dots, b_n\}$ of V. For all $i, j = 1, \dots, n$, such that:

1- $\langle b_i, b_j \rangle = 0 \; for \; i \neq j.$

2- $\langle b_i, b_j \rangle = 1 \; for \; i = j.$

Then the basis is called an orthonormal basis.

Definition 1.2.14 (Eigenvalues and Eigenvector) (Bloehdorn, 2008)**:** Consider an eigenvalue of an $n \times n$ matrix A is the real number λ and the vector $x \in \mathbb{R}^n$ and eigenvector, if $Ax = \lambda x$.

Theorem 1.2.3 (Pythagoras): Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be orthonormal vectors in space X, and let $x \in X$ be nonempty set. Then:

$$\|x\|^2 = \sum_{i=1}^{n} \langle x, \lambda_i \rangle^2 + \left\| x - \sum_{i=1}^{n} \langle x, \lambda_i \rangle \lambda_i \right\|^2$$

Definition 1.2.15 (Evaluation functional): Let X be an arbitrary set and $\mathcal{H}$ be a Hibert space consisting of real valued function defined on X. For each point $x \in X$, the evaluation function:

$$L_x : f \to f(x) \; \forall \; f \in \mathcal{H} \; L_x(f) = f(x)$$

Evaluation functional L_x is linear functional of $\mathcal{H}$, for any $f, g \in \mathcal{H}, a \in \mathbb{R}$, which satisfies:

$$L_x(af + g) = (af + g)(x) = af(x) + g(x) = aL_x(f) + L_x(g)$$

1.3 Some Basic Concept in Machine Learning

Machine learning ML is a scientific discipline that allows computers to learn and make decisions autonomously, without the need for explicit programming. By analyzing vast amounts of data, ML algorithms can identify patterns, make predictions, and solve complex problems. It's a subset of artificial intelligence, a broader field focused on developing intelligent agents that can perceive, reason, learn, and interact with the world. AI aims to create machines capable of human-like cognitive abilities, such as understanding language, recognizing objects, and making strategic choices. Three fundamental elements form the core of any machine learning system: data, models, and learning processes, highlighting the importance of models performing effectively on unseen data. Deep learning, a subset of

machine learning, involves neural networks with multiple hidden layers. While machine learning is one avenue of artificial intelligence, there are other approaches as well. (Deisenroth et al., 2020).

Definition 1.3.1 (Machine learning) (Colliot, 2023): A computer program is considered to learn from experience E with respect to a task T and some performance measure P if its performance on task T, as evaluated by P, improves with experience E.

1.3.1 Machine learning Process

Step 1. Data Collection**:** Data online from different sources are brought together and tabulated at which quantity and quality are directly which is approximately correct depending on fidelity of spin model.

Step 2. Data preparation: The preprocessed collected are and missing value of data is collected and made ready for the model.

Step 3. For many of the machine learning algorithms that need to be used, suitable models must be selected and can be expressed according to the demand, which provides suitable answers to the presented query.

Step 4. Train the Model: Training the large set of data based on the selected model from Step 3 The new data is verified against a model is still general, e.g., model for specific task prediction or classification etc.

Step 5. Prediction: This step is performed to predict or classify the new object or data.

Step 6. Evaluation: The results that obtained from the model are compared with the results of testing so that the error and accuracy of the results could be checked.(Patil, Jadhav, Talekar & Bag, 2023)

1.3.2 Types of machine learning

1- Supervised learning: predict output given input- classification and regression see also image classification, forecast. Labelled data goal is to «predict».

2- Unsupervised learning: learn useful patterns out of different input based on no goal. Use case dimensionality reduction. Unlabeled data goal is to «explore».
3- Self-supervised: learning–supervised learning without human-given labels examples auto encoders.
4- Reinforcement learning: where an agent learns to take actions to maximize reward. examples Self-driving car, robotics. (Khaliq, 2022)

1.3.3 Machine learning algorithm

These are the models used to implement the various types of machine learning:

1- Neural Networks.
2- Decision Trees.
3- Support Vector Machines (SVM).
4- Ridge Regression.
5- Naive Bayes.
6- Random Forest. (Khaliq, 2022)

1.4 Some Basic Concept in Probability Theory and Statistical

Statistics and probability theory serve as the cornerstone of machine learning, offering a robust framework for modeling uncertainty, optimizing learning algorithms, and ensuring their ability to generalize effectively. This foundational role is particularly evident in kernel-based methods, where positive definite kernels play a critical role. Their spectral properties, as formalized by Mercer's theorem, enable efficient embeddings into high-dimensional feature spaces, thereby facilitating the development of powerful and theoretically sound learning algorithms. (Gopala Rao & Sankara Rao, 2024)

Definition 1.4.1 (σ −Field or σ −Algebra) (Schölkopf, 2022)**:** Let Ω be non empty set, the collection $\mathcal{F} = \{A: A \subset \Omega\}$ is called σ −Algebra, also known as sigma field, if it satisfies:

1- $\phi, \Omega \in \mathcal{F}$.
2- It is closed under complementation, meaning if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.
3- It is closed under countable unions, meaning if $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Definition 1.4.2 (measure space): A measure space $(\Omega, \mathcal{F}, \mu)$ or measure μ defined on $\mathcal{F}$ is σ −Field if Ω can be decomposed into a countable union of measurable sets, each with finite measure that is $\Omega = \bigcup_{i \leq \infty} \Omega_k, \Omega_i \cap \Omega_j = \phi, i \neq j$ such that $\mu(\Omega_k) < \infty, \Omega_k \in \mathcal{F}$.

Definition 1.4.3 (probability measure): Let a measure $\mu: \mathcal{F} \rightarrow [0,1]$ defined on measurable space $(\Omega, \mathcal{F})$ such that $\mu(\Omega) = 1$.

Theorem 1.4.1 (Bayes' theorem)

Bayes' Theorem is a fundamental concept in probability theory and plays a pivotal role in binary classification problems. It establishes a relationship between the conditional probabilities of two events, A and B, and is expressed as follows:

$$P(B/A) = \frac{P(B) \times P(A/B)}{P(A)}$$

Definition 1.4.4 (Expected value): The expected value of a function $f: \mathbb{R} \rightarrow \mathbb{R}$ which applied to univariate continuous random variable X with probability density function $p(x)$ is defined as:

$$\mathbb{E}_X[f(x)] = \int f(x)p(x)dx$$

And the expected value of a function f of a discrete random variable presents as:

$$\mathbb{E}_X[f(x)] = \sum_{x \in X} f(x)p(x)$$

where X denotes the set of all possible values (the target space) that the random variable X can take.

Definition 1.4.5 (Mean): The mean of a random variable X with states $x \in \mathbb{R}^n$ is an average and is given by:

$$\mathbb{E}_X[x] = \begin{bmatrix} \mathbb{E}_{x_1}[x_1] \\ \vdots \\ \mathbb{E}_{x_d}[x_d] \end{bmatrix} \in \mathbb{R}^n$$

Where: $\mathbb{E}_{x_d}[x_d] = \begin{cases} \int x_d p(x_d) dx_d \; if \; X \; continous \; variable \\ \sum_{x_i \in X} x_i p(x_d = x_i) \; if \; X \; discrete \; variable \end{cases}$.

Definition 1.4.6 (covariance): The covariance between two univariate random variables X and Y is the expected value of the product of their deviation from their respective means. Mathematically, it is defined as:

$$Cov_{X,Y}[x, y] = \mathbb{E}_{X,Y}[(x - \mathbb{E}_X[x])(y - \mathbb{E}_Y[y])]$$

Remark 1.4.1: The covariance of a variable with itself $Cov[x, x]$ is called variance and denoted by $\mathbb{V}_X[x]$.

Definition 1.4.7 (Variance): The variance of d-dimensional random variable X with states X in $\mathbb{R}^n$ with mean vector μ in $\mathbb{R}^n$ is defined as the expected squared Euclidean distance between X and μ (its mean). Mathematically its expressed as:

$$\mathbb{V}_X[x] = Cov_X[x, x]$$

$$= \mathbb{E}_X[(x - \mu)(x - \mu)^T]$$

$$= \mathbb{E}_X[xx^T] - \mathbb{E}_X[x]\mathbb{E}_X[x]^T$$

$$= \begin{bmatrix} Cov[x_1, x_1] & \cdots & Cov[x_1, x_d] \\ \vdots & \ddots & \vdots \\ Cov[x_d, x_1] & \cdots & Cov[x_d, x_d] \end{bmatrix}$$

The matrix known as the Covariance matrix of the multivariate random variable X is symmetric and positive semidefinite, providing insights into the spread of the data.

Definition 1.4.8 (Empirical Mean and Covariance) (Brinker, 2004)**:** The empirical mean vector is statistical measure that represents the average value of each variable in a given dataset. It is calculated by taking the arithmetic mean of the observed values for each variable Mathematically: For a dataset with n observation, the empirical mean vector $\hat{\mu}$ is defined as:

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n; \quad x_n \in \mathbb{R}^n$$

And the empirical Covariance matrix is $n \times n$ matrix defined as:

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T$$

Definition 1.4.9 (Loss function): A loss function $L: X \times Y \times F \rightarrow [0, \infty)$ is mapping from the space of input- output pairs (x, y) and prediction function f to the non-negative real number such that $L(x, y, f) = 0$ for all couples of labeled data $(x, y) \in X \times Y$ is called loss function if:

$$L(x, y, f) = \frac{1}{2}|f(x) - y|$$

Chapter two

Kernel method

Chapter 2

Kernel method

Introduction

Kernel methods are a diverse family of algorithms within the machine-learning domain. These methods excel at analyzing data residing in high-dimensional spaces and are especially well suited for regression and classification tasks. This chapter presents some key points about kernel method and their properties, and the theoretical foundations of kernel methods.

2.1 kernel Function

Kernel methods are classification algorithms that employ a kernel function k to map data points from a lower-dimensional input space V into a higher-dimensional feature space V'. This transformation often makes the separation between data classes more distinct. A kernel function calculates the dot product of vectors in the feature space without explicitly mapping the original data points. This allows many algorithms to operate solely on kernel values rather than the raw data. The kernel method is a highly accurate and flexible approach to data processing. Various kernel functions are designed to capture different relationships and structures within datasets. Each kernel comes with its own mathematical formulation, advantages, and specific use cases, making it crucial to select the right kernel based on the dataset's characteristics and the problem being addressed. (Campbell, 2001) (Cristianini, 2001) (Rasmussen & Williams, 2006)

Definition 2.1.1 (Feature Map): A feature map $\varphi: X \to \mathcal{H}$ that maps elements from an input space X to a Hilbert space $\mathcal{H}$.

Definition 2.1.2 (Feature Space) (Herbrich, 2001)**:** The feature space $\mathcal{H}$ associated with a feature map $\varphi: X \to \mathcal{H}$ is the closure of the linear span of the mapped data: $\mathcal{H} = \overline{span\{\varphi(x) | \, x \in X\}}$.

Definition 2.1.3 (Kernel Function): Let X be a non-empty set. A function $k: X \times X \to \mathbb{R}$ is a kernel function if there exists a feature map $\varphi: X \to \mathbb{R}^n$ mapping into a real Hilbert space $\mathcal{H}$(feature space) such that for all $x, z \in X$,

$$k(x,z) = \langle \varphi(x), \varphi(z) \rangle$$

Where $\mathbb{R}^n$ feature space and φ feature map.

2.1.1 Some Types of Kernel Function

In this overview, we will examine several common types of kernel functions, including linear, polynomial, Laplacian, Gaussian, radial basis function (RBF), sigmoid, exponential, and translation invariant kernels. Each of these kernels possesses unique properties that render them suitable for different data types and tasks, empowering practitioners to improve the performance of their machine learning models:

1- **Linear kernel:** For $x, z \in \mathbb{R}^n$ the linear kernel defined as

$$k(x,z) = \langle x, z \rangle = x^T z$$

In this case, the feature space of the linear kernel is identical to the input space, representing the straightforward dot product of the input vectors. Linear kernel is best to apply on linear separable data.

2- **Polynomial kernel:** The derived polynomial kernel is defined as

$$k_d(x,z) = P\big(K(x,z)\big) = (x^T z + c)^d ; c \geq 0 , d \in N$$

Where $P(.)$ is any polynomial with positive coefficients, c is constant and d is the degree of the polynomial called non-homogeneous kernel. The polynomial kernel is called homogeneous kernel if $c = 0$. Polynomial kernel can distinguish curved or nonlinear input space.

3- **Laplacian kernel:** The Laplacian kernel is expressed as

$$k(x,z) = \exp\left(-\gamma \|x - z\|_1\right)$$

Where

$$\|x - z\|_1 = \sum_{i=1}^{n} |x_i - z_i|$$

Where $\|x - z\|_1$is the L_1norm Manhattan distance between vectors *x* and z, $\gamma > 0$.

4- **Gaussian kernel or Radial basis function kernels (RBF):** for $\gamma > 0$

$$k(x, z) = exp(-\gamma \|x - z\|^2)$$

Where $\gamma = \frac{1}{2\sigma^2}$ and σ is coefficient of variation. This kernel is popular for its ability to handle non-linear relationship. The Gaussian kernel is the most commonly used RBF kernel.

5- **Sigmoid kernels:** The kernel is defined as

$$k(x, z) = tanh\ [\alpha x^T z + c]$$

Where x, z are input data and α is parameter. (Often used in networks).

6- **Exponential kernel:** Let $\beta > 0$ and $x, z \in \mathbb{R}^n$ then

$$k_n(x, z) = 1 + \beta xz + \frac{\beta^2}{2}(x.z)^2 + \cdots + \frac{\beta^n}{n!}(x.z)^n; n \geq 1$$

And

$$k_\infty(x, z) = \exp(\beta(x.z)) = \lim_{n \to \infty} k_n(x.z)$$

Where This kernel is similar to the RBF kernel but uses the Euclidean norm directly instead of its square. It is also a translation-invariant kernel.

7- **Translation invariant kernel:** A kernel k is called translation-invariant if it can be expressed as $k(x, z) = T(x - z)$, where $T: \mathbb{R}^n \to \mathbb{R}$ is differentiable function.

Where this kernel depends only on the difference between the input vectors $(x - y)$. It is commonly used in applications where the absolute position of the data is irrelevant, such as image processing. (Shawe-Taylor, 2004).

Now, we will show some examples of kernel function

Example 2.1.1: Consider a two-dimensional vector space $x \in \mathbb{R}^2$, the feature map:

$$\varphi: x = (x_1, x_2) \rightarrow \varphi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2) \in \mathbb{R}^3$$

The kernel function is given by

$$k(x,z) = \langle \varphi(x), \varphi(z) \rangle = \langle (x_1^2, x_2^2, \sqrt{2}, x_1x_2).(z_1^2, z_2^2, \sqrt{2}, z_1z_2) \rangle$$

$$= x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 = (x_1z_1 + x_2z_2)^2 = \langle x, z \rangle^2.$$

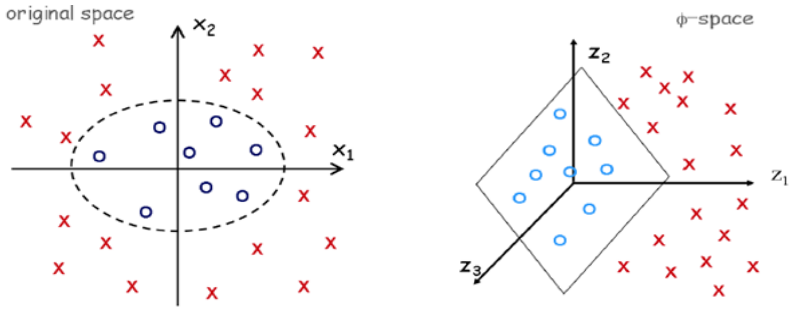

Figure 2.1 (Boughorbel & Boujemaa, 2005) The polynomial kernel transforms the original 2D data into a 3D space, allowing a linear hyperplane to separate the classes effectively.

Example 2.1.2: if the input space $x = (x_1, x_2)$ and feature space $\varphi(x) = (x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, 1) \in \mathbb{R}^6$

Then the kernel $k(x,z) = \langle \varphi(x), \varphi(z) \rangle$

$$= x_1^2z_1^2 + x_2^2z_2^2 + 2x_1z_1 + 2x_2z_2 + 2x_1z_1x_2z_2 + 1$$

$$= (x_1z_1 + x_2z_2 + 1)^2 = (\langle x, z \rangle + 1)^2$$

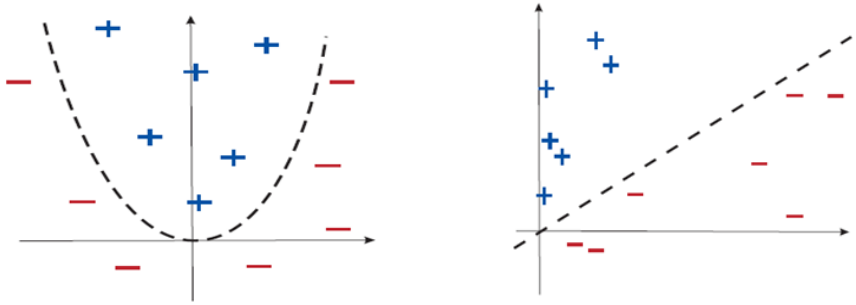

Figure 2.2 Transforming a complex, non-linear model into a simpler, linear model that can be partitioned by a hyper plane for efficient data processing.

Now we will show example of polynomial kernel in general case.

Theorem 2.1.1: The polynomial kernel with d-degree, $d \in \mathbb{N}$ defined as

$$k(x,z) = (1 + \langle x,z\rangle)^d$$

$$= \left(\sum_{j=0}^{m} x_j z_j\right).\left(\sum_{j=0}^{m} x_j z_j\right)....\left(\sum_{j=0}^{m} x_j z_j\right)$$

$$= \sum_{j=0}^{m}\prod_{i=1}^{n} x_{j_i} z_{j_i}$$

$$= \sum_{j=0}^{m}\prod_{i=1}^{n} x_{j_i}\prod_{i=1}^{n} z_{j_i}$$

$$= \langle \varphi(x), \varphi(z)\rangle$$

Where $c = 1$ is kernel function and $\varphi(x) = \left\{\prod_{i=1}^{n} x_{j_i}\right\}_{j\in[0:m]}$. We see that the complexity of implementing k is $O(m)$, the feature space is dimensional $(\mathrm{m} + 1)^d$.

Example 2.1.3: Consider the kernel polynomial family $\{k_i(x,z) = \alpha_i(\langle x,z\rangle + c_i)^{d_i}; d_i \in \mathbb{N}, \alpha_i > 0, c_i \geq 0\}$ for any $n > 0, 0 \in \mathbb{N} \Rightarrow \sum_{i=0}^{n} k_i(x,z)$ is a kernel.

If $\alpha_i = \frac{\alpha^i}{i!}; \alpha > 0,\ c_i = 0, d_i = i\ and\ n \to \infty,$ Where $\varphi(z) = \sum_{i=0}^{\infty}\sqrt{\frac{\alpha^i}{i!}} z^i$ then $\sum_{i=0}^{\infty} k_i(x,z)$ is a kernel because

$$\sum_{i=0}^{\infty} k_i(x,z) = \sum_{i=0}^{\infty}\sqrt{\frac{\alpha^i}{i!}}(\langle x,z\rangle)^i = e^{\alpha\langle x,z\rangle} = \langle \varphi(x), \varphi(z)\rangle.$$

Example 2.1.4 (Gaussian kernel): Let $X = \mathbb{R}$ and consider the mapping given by $\varphi(x) = \left\{\frac{1}{\sqrt{n!}} e^{\frac{-x^2}{2}} x^n\right\}_{n=0}^{\infty}$. Then we have

$$\langle \varphi(x), \varphi(z)\rangle = \sum_{n=0}^{\infty}\left(\frac{1}{\sqrt{n!}} e^{\frac{-x^2}{2}} x^n\right)\left(\frac{1}{\sqrt{n!}} e^{\frac{-z^2}{2}} z^n\right)$$

$$= e^{\frac{-x^2+z^2}{2}} \sum_{n=0}^{\infty} \left(\frac{(xz)^n}{n!}\right)$$

$$= e^{\frac{-(x-z)^2}{2}} = k(x,z)$$

Where the evaluating the Gaussian kernel is very simple while in sharp contrast the feature space is of infinite dimension. Note that since (x) includes all the monomial terms, using the Gaussian kernel we can learn polynomial predictor of any degree over the original space. It's evident that evaluating this kernel function is computationally straightforward. However, it's crucial to note that the corresponding feature space associated with this kernel is infinite-dimensional. This is because the mapping function $\varphi(x)$ implicitly encompasses all possible monomial terms, effectively enabling the model to learn polynomial predictors of any degree within the original input space. We observe that the Gaussian kernel is computationally more efficient than the polynomial kernel. We will assess the effectiveness of the Gaussian kernel in this thesis and present the results in Chapter 4.

Definition 2.1.4 (Kernel Section) (Dinuzzo, 2011)**:** Let k be a kernel, then the kernel section centered on $\bar{x} \in X$ is a function $K_{\bar{x}}: X \to \mathbb{R}$ defined by: $K_{\bar{x}}(x) = K(\bar{x}, x)$.

Definition 2.1.5 (All-subset embedding) (Shawe-Taylor, 2004)**:** All-subset kernel is defined as $\varphi: X \to \varphi_B(x)_{B\subseteq\{1,2,\ldots,n\}}$ with the corresponding kernel and $k \subset (x,z)$ such that

$$k \subset (x,z) = \langle \varphi(x), \varphi(z) \rangle$$

$$= \sum_{B\subseteq\{1,2,\ldots n\}} \varphi_B(x)\varphi_B(z) = \sum_{B\subseteq\{1,2,\ldots n\}} \prod_{i\in B} x_i z_i = \prod_{i=1}^{n} (1 + x_i z_i)$$

For n-dimensional input vectors.

Remark 2.1.1: The polynomial and all-subsets kernels exhibit limited flexibility in terms of feature selection and weighting. The polynomial kernel, by design, incorporates all monomials up to degree d, with their weights determined by a single parameter R. Similarly, the all-subsets kernel is inherently restricted to incorporating all monomials derived from

every possible combination of the n input features. We now introduce a method that offers greater control in defining the set of monomials.

Definition 2.1.6 (ANOVA kernel): The ANOVA kernel of d-degree defines as:

$$\varphi_d : X \to (\varphi_B(X))_{|d|}$$

Where, the feature is defined for every subset B by

$$\varphi_B(X) = \prod_{i \epsilon B} x_i = X^{i_B}$$

where i_B denotes the indicator function for the set B. The dimension of the resulting embedding is explicitly $\binom{n}{d}$, corresponding to the number of possible subsets. The inner product in this space is determined by the specific kernel function used such that:

$$k_d(x,z) = \langle \varphi_d(x), \varphi_d(z) \rangle = \sum_{|B|=d} \varphi_B(x)\varphi_B(z)$$

$$= \sum_{i \le i_1 \le i_2 \le \cdots \le n} (x_{i_i} z_{i_i})(x_{i_2} z_{i_2}) \dots (x_{i_d} z_{i_d}) = \sum_{i \le i_1 \le i_2 \le \cdots \le n} \prod_{j=1}^{d} x_{i_j} z_{i_j}.$$

Lemma 2.1.1: The ANOVA kernel recursion is given by:

$$k_0^m(x,z) = 1 \ \ if \ m \ge 0$$

$$k_s^m(x,z) = 0 \ if \ m < s$$

$$k_s^m(x,z) = (x_m z_m) k_{s-1}^{m-1}(x,z) + k_s^{m-1}(x,z)$$

The next is an example illustrating the ANOVA kernel recursion.

Example 2.1.5: Consider $x = (x_1, x_2, x_3), z = (z_1, z_2, z_3)$ we computing $k_1^2(x,z)$ using the recursion. For $m = 2, s = 1$, we have

$$k_1^2(x,z) = (x_2, z_2) k_0^1(x,z) + k_1^1(x,z)$$

Where we know $k_0^1(x,z) = 1$, therefore

$$k_1^1(x,z) = (x_1, z_1) k_0^0(x,z) + k_1^0(x,z)$$

$$k_1^1(x,z) = (x_1, z_1) * 1 + 1 = x_1 z_1 + 1$$

Where $k_0^0(x,z) = 1$ and $k_1^0(x,z) = 1$, then

$$k_1^2(x,z) = (x_2, z_2).1 + (x_1 z_1 + 1) = x_2 z_2 + x_1 z_1 + 1.$$

Definition 2.1.7 (Alignment kernel) (Shawe-Taylor, 2004)**:** The alignment kernel measures the similarity between two kernel matrices K and K_2 is defined as

$$A(K_1, K_2) = \frac{\langle K_1, K_2 \rangle}{\sqrt{\langle K_1, K_1 \rangle \langle K_2, K_2 \rangle}}$$

We will now present definition of positive semi-definite, some properties, examples, and theories.

2.2 positive semi-Definite kernel

Definition 2.2.1 (Positive Semi-Definite) (Fukumizu, 2010)**:** Let X be a non-empty set. A function $k: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called positive semi-definite kernel if it satisfies the following condition:

1- Symmetric: $k(x,z) = k(z,x)\ \forall\ x, z \in \mathbb{R}^n$

2- Positive semi-definiteness: for any set $x_1, x_2, \dots, x_n \in \mathbb{R}^n$ and any real numbers $c_1, \dots, c_m \in \mathbb{R}$, the kernel matrix defined by:

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} > 0$$

Definition 2.2.2 (Gram Matrix) (Rossmann, 2023)**:** Let V be a vector space over the field $\mathcal{F}$, and $S = \{x_1, x_2, \dots, x_n\}$, be a set of vectors in V. The Gram matrix G is an $n \times n$ matrix defined by:

$$G_{i,j} = \langle x_i, x_j \rangle, \quad \forall\ i, j \in \{1,2, \dots, n\}$$

Where $\langle .,. \rangle$ denotes the inner product on V. In matrix form:

$$G = \begin{bmatrix} \langle x_1, x_1 \rangle & \cdots & \langle x_1, x_n \rangle \\ \vdots & \ddots & \vdots \\ \langle x_n, x_1 \rangle & \cdots & \langle x_n, x_n \rangle \end{bmatrix}$$

If a kernel k compute inner products in a feature space induced by a mapping function φ, such that:

$$k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$$

Then the associated kernel matrix (or Gram matrix in feature space) is given by:

$$G_{i,j} = k(x_i, x_j)$$

And the matrix form:

$$G = \begin{bmatrix} k(x_1, x_2) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix}$$

This matrix is called kernel matrix.

Definition 2.2.3 (Positive Definite Matrix) (Schölkopf et al. 2001)**:** A real-valued $\boldsymbol{m \times n}$ matrix K is called positive definite if for all finite sequences $\boldsymbol{\alpha_1, \alpha_2, \dots, \alpha_n} \in \mathbb{R}$, the following inequality holds:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \boldsymbol{\alpha_i \alpha_j K_{ij}} \geq \mathbf{0}$$

Lemma 2.2.1 (Rossmann, 2023)**:** Let $F = span\{\varphi(x_1), \dots, \varphi(x_n)\}$ and $\varphi(x_i) = \Lambda^{1/2} u_i \in \mathbb{R}^n$ for $i = 1, .., n$, then the Gram matrix G is defined as $G_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$

Proof: The inner product between $\varphi(x_i)$ and $\varphi(x_j)$ in the space F is given by:

$$\langle \varphi(x_i), \varphi(x_j) \rangle_F = \left(\Lambda^{\frac{1}{2}} u_i \right)^T \Lambda^{\frac{1}{2}} u_j$$

$$= u_i^T \Lambda u_j = G_{ij}$$

Definition 2.2.4 (Distance between a point and a set): Let $x \in X$ is any point and $s = \{x_1, x_2, .., x_n\}$ be a finite set in X then the distance between x and s defined as:

$$d_k(x, s) = \|\phi(x) - m\|$$

Where $m = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$ such that the distance by kernel trick given by:

$$d_k(x, s) = \sqrt{K(x, x) - \frac{2}{n} \sum_{i=1}^{n} K(x, x_i) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K(x_i, x_j)}$$

Lemma 2.2.2 (Centered Gram matrix): Let $s = \{x_1, x_2, \dots, x_n\}$ associated with kernel k and let G be Gram matrix $n \times n$ the computation centered Gram matrix for $0 \leq i, j \leq n$:

$$G^{c}_{i,j=}\langle \phi(x_i) - m, \phi\big(x_j\big) - m\rangle$$

$$= \langle \phi(x_i), \phi\big(x_j\big)\rangle - \langle m, \phi(x_i) + \phi\big(x_j\big)\rangle + \langle m, m\rangle$$

$$= G_{i,j} - \frac{1}{n}\sum_{k=1}^{n}\big(G_{i,k}, G_{j,k}\big) + \frac{1}{n^2}\sum_{k,l=1}^{n} G_{k,l}$$

Where $m = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$ and $G_{i,j} = K(x_i, x_j)$.

2.2.1 Properties of positive semi-definite kernel

Property 1(Belanche, 2014)**:** A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semi-definite kernel if and only if $x^T Ax \geq 0, \forall x \in \mathbb{R}^n$.

Example 2.2.1: To show that the matrix $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ is positive semi definite, consider any vector $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$. The product Ax is:

$$Ax = \begin{bmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 - x_3 \\ -x_2 + 2x_3 \end{bmatrix}$$

Then:

$$x^T Ax = [x_1 \quad x_2 \quad x_3]Ax$$

$$= 2x_1^2 - x_1x_2 - x_1x_2 + 2x_2^2 - x_2x_3 - x_3x_2 + 2x_3^2$$

$$= (x_1 - x_2)^2 + x_1^2 + (x_2 - x_3)^2 + x_3^2$$

Since each term in this expression is a squared term, then $x^T Ax \geq 0$.

Property 2 (Belanche, 2014)**:** A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semi-definite kernel if and only if $\lambda v = Av\,; v \neq 0$ where for all λ eigenvalues of A are non-negative.

Example 2.2.2: The symmetric matrix $A = \begin{bmatrix} 3 & -1 & 1 \\ -1 & 5 & -1 \\ 1 & -1 & 3 \end{bmatrix}$, is positive semi-definite, firstly examine its characteristic polynomial, which is given by

$$\lambda^3 - 11\lambda^2 + 36\lambda - 36 = 0$$

The roots of this polynomial, which correspond to the eigenvalues of the matrix A, are $\lambda_1 = 2, \lambda_2 = 3$ and $\lambda_3 = 6$ all the eigenvalues are non-negative. Since all eigenvalues of A are non-negative, we affirm that the matrix A is indeed positive semi-definite.

Property 3 (Belanche, 2014)**:** A symmetric matrix A is positive semi-definite kernel if all its upper left sub matrixes have non-negative determinants.

Example 2.2.3: The symmetric matrix $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ is positive semi- definite,

The upper left determines of degree 1 is:

$$|2| > 0$$

The upper left determines of degree 2 is:

$$\begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} = 3 > 0$$

The upper left determines of degree 3 is:

$$\begin{vmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{vmatrix} = 4 > 0$$

Since all the leading principal minors are positive, conclude that the matrix A is positive semi-definite.

Property 4 (Belanche, 2014)**:** A matrix A is positive semi-definite kernel if and only if there exists B is positive semi-definite matrix such that $BB^T = A$.

Example 2.2.4: The matrix $A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}$, is positive semi-definite, first compute the product $A^T A$:

$$A^T A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 5 \\ 5 & 6 \end{bmatrix}$$

This matrix is symmetric. To confirm that it is positive semi-definite by property 3.

Theorem 2.2.1 (Shawe-Taylor, 2004)**:** Let A be symmetric then A is positive semi-definite if and only if $\langle A, B\rangle \geq 0, \forall$ for all positive semi-definite kernels B is positive semi-definite kernel B.

Proof: Let $\alpha_1, \dots, \alpha_n$ be eigenvalues of A which corresponding with eigenvectors $\beta_1, \dots, \beta_n$. Then

$$\langle A, B\rangle = \langle \sum_{i=1}^{n} \alpha_i \beta_i \beta_i^T, B\rangle = \sum_{i=1}^{n} \alpha_i \langle \beta_i \beta_i^T, B\rangle = \sum_{i=1}^{n} \alpha_i \beta_i^T B \beta_i$$

Where $\beta_i^T B \beta_i \geq 0$ because B is positive semi-definite.

Theorem 2.2.2: Let G be a Gram matrix and k be kernel such that, there exists $\varphi: k(x, z) = \langle \varphi(x), \varphi(z)\rangle$. Then G is symmetric positive semi- definite.

Proof: Since $G = G^T$ and $\forall\, \alpha \in \mathbb{R}^n$ we have

$$\alpha^T K \alpha = \sum_{i,j=1}^{n} \alpha_i \alpha_j K_{ij} = \langle \sum_{i=1}^{n} \alpha_i \varphi(x_i), \sum_{j=1}^{n} \alpha_j \varphi(x_j)\rangle$$

Therefore

$$= \left\| \sum_{i=1}^{n} \alpha_i \varphi(x_i) \right\|^2 \geq 0.$$

This shows G is positive semi definite and symmetric.

Theorem 2.2.3: If matrix A is positive semi-definite kernel matrix, then the function $k: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by $k(x, z) = x^T A z$ is a kernel

Proof: Since A is positive semi-definite kernel, by property (4) we can express $A = BB^T$.

$\forall\, n \in \mathbb{N}$, there exist $x_1, \dots, x_n \in \mathbb{R}^n$ and construct the matrix $K = k_{ij}$ with entries

$$K_{ij} = k(x_i, x_j) = x_i^T A x_j$$

Then for every $\alpha \in \mathbb{R}^n$, we have:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j x_i^T A x_j = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j (B^T x_i)^T (B^T x_j)$$

Thus

$$= \left\| \sum_{i=1}^{n} \alpha_i (B^T x_i) \right\|^2 \geq 0$$

Where $\varphi(z) = B^T z$.

2.3 Valid kernel

Definition 2.3.1 (Valid Kernel) (Dan San Martino, 2009)**:** Let X be a non-empty set. A function $k: X \times X \to \mathbb{R}$ is valid kernel if satisfies the following condition:

1- symmetric: $k(x, z) = k(z, x)\ \forall\ x, z \in X$

2- positive semi-definite: for any $\{x_1, x_2, \dots, x_n\} \subset X$ and $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{R}$, the kernel function satisfies:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\alpha_j k(x_i, x_j)$$

This is equivalent to stating the Gram matrix K, is defined by $K_{i,j} = k\langle x_i, x_j\rangle$ is positive semi definite for any finite subset of X.

Now, we have the following properties of valid function.

2.3.1 Properties of valid kernel

Let $c > 0$ and k_1 and k_2 be a valid kernel over X and K_1 and K_2 be their associated Gram matrices, and let $f(.)$ and $p(.)$ be a polynomial with non-negative coefficient, exp $(.)$ be the exponential function. Then the following are valid kernels:

1- Multiplication by a Positive Constant:

$$k(x, z) = c k_1(x, z) \qquad (2.3.1)$$

Proof: If k_1 is valid kernel, then it satisfies $k_1(x, z) = k_1(z, x)$. This implies the following

$$k(z, x) = c k_1(z, x) = c k_1(x, z) = k(x, z)$$

Given that K is symmetric, for any arbitrary vector $V \in \mathbb{R}^n$,the following hold

$$V^\top k(x,z)V = V^\top c k_1(x,z)V \geq 0$$

Given that $c > 0$ and $V^\top k_1(x,z)V \geq 0$, is non-negative, hence $V^\top k(x,z)V \geq 0$. Hence, $k(z,x)$ is positive semi-definite, ensuring it is a valid kernel.

2- Sum of Valid Kernels:

$$k(x,z) = k_1(x,z) + k_2(x,z) \qquad (2.3.2)$$

Proof: If k_1 and k_2 are valid kernels, then $k_1(x,z) = k_1(z,x)$ and $k_2(x,z) = k_2(z,x)$. Thus, we have:

$$k(z,x) = k_1(z,x) + k_2(z,x) = k_1(x,z) + k_2(x,z) = k(x,z).$$

Given that K is symmetric, for any arbitrary vector $V \in \mathbb{R}^n$,the following hold

$$V^\top k(x,z)V = V^\top (k_1(x,z) + k_2(x,z))V = V^\top k_1(x,z)V + V^\top k_2(x,z)V \geq 0$$

Given that $V^\top k_1(x,z)v \geq 0$ and $V^\top k_2(x,z)v \geq 0$, we conclude that $V^\top k(x,z)v \geq 0$. hence $k(x,z)$ is positive semi-definite.

3- Product of Valid Kernels:

$$k(x,z) = k_1(x,z)k_2(x,z) \qquad (2.3.3)$$

Proof: If k_1 and k_2 are valid kernels, then by definition, it must satisfy $k_1(x,z) = k_1(z,x)$ and $k_2(x,z) = k_2(z,x)$. Thus, we have:

$$k(z,x) = k_1(z,x)k_2(z,x) = k_1(x,z)k_2(x,z) = k(x,z).$$

Given that K is symmetric, for any arbitrary vector $V \in \mathbb{R}^n$ also let $K_1 = B^\top \Lambda B$ and $K_2 = A^\top \Lambda A$ using the spectral decompositions

$(K_1)_{ij} = b_{iq}\lambda_{qq}b_{jq}$ and $(K_2)_{ij} = a_{ik}\omega_{kk}a_{jk}$, then

$$V^\top KV = V^\top (k_1(x,z)k_2(x,z))V = V^\top \big(B^\top \Lambda B\big)\big(A^\top \Lambda A\big)V$$

$$\sum_{i,j,k,q=1}^{n} v_i v_j b_{iq}\lambda_{qq}b_{jq}a_{ik}\omega_{kk}a_{jk} = \sum_{i,k,q=1}^{n} \omega_{kk}\lambda_{qq}(v_i b_{iq}a_{ik})^2 \geq 0.$$

Given that the eigenvalues of K_1 and K_2 are non-negative, it follows that K is positive semi-definite. The symmetry and positive semi-definiteness of k ensure that it is a valid kernel.

4- Weighted Valid Kernel:

$$k(x,z) = f(x)k_1(x,z)f(z) \tag{2.3.4}$$

Proof: If k_1 is valid kernel, then by definition, it must satisfy $k_1(x,z) = k_1(z,x)$. That gives

$$k(z,x) = f(x)k_1(x,z)f(z)$$

$$\sum_{i,j=1}^{n} f(x_i)f(x_j)k_{1_{ij}} = \sum_{i,j=1}^{n} f(x_j)f(x_i)k_{1_{ij}}$$

$$f(z)k_1(x,z)f(x) = f(z)k_1(z,x)f(x) = k(x,z).$$

Given that K is symmetric, for any arbitrary vector $V \in \mathbb{R}^n$, the following hold

$$V^\top KV = V^\top f(x)K(x,z)f(z)V = \sum_{i,j=1}^{n} v_i v_j f(x_i)f(x_j)\langle\varphi(x_i),\varphi(x_j)\rangle$$

which

$$= \langle \sum_{i=1}^{n} v_i f(x_i)\varphi(x_i) , \sum_{j=1}^{n} v_j f(x_j)\varphi(x_j)\rangle$$

This lead to

$$\left\| \sum_{i=1}^{n} v_i f(x_i)\varphi(x_i) \right\|^2 \geq 0.$$

It follows that $k(x,z)$ is positive semi-definite. The symmetry and positive semi-definiteness of k ensure that it is a valid kernel.

5- Polynomial of a Valid Kernel

$$k(x,z) = p(k_1(x,z)) \tag{2.3.5}$$

Proof: Suppose $\alpha_m \in \mathbb{R}$ holds for every $\alpha_m \geq 0$ for all $m \in \mathbb{N}$. Then

$$k(x,z) = q(k_1(x,z))$$

$$= \alpha_0 + \alpha_1 k_1(x,z) + \alpha_2 (k_1(x,z))^2 + \cdots + \alpha_m \big(k_1(x,z)\big)^m = \sum_{n=0}^{m} \alpha_n (k_1(x,z))^n$$

In this case, $k_2(x,z) = (k_1(x,z))^n$ is a kernel which follows from applying (2.3.3) repeatedly n times. Then

$$k(x,z) = \sum_{n=0}^{m} \alpha_n k_2(x,z)$$

When $\alpha_n > 0$, $k_3(x,z) = \alpha_n k_2(x,z)$ is a kernel, as stated in (2.3.1). If $\alpha_n = 0$, we have $V^\top KV = 0$. Finally:

$$k(x,z) = \sum_{n=0}^{m} k_3(x,z)$$

Repeated application of (2.3.2) shows that k is a valid kernel.

6- Exponential of a Valid Kernel

$$k(x,z) = \exp\,(k_1(x,z)) \qquad (2.3.6)$$

Proof: Using the Taylor series expansion $\sum_{n=0}^{\infty} \frac{x^n}{n!}, n \in \mathbb{N}$, we obtain:

$$k(x,z) = \exp\,(k_1(x,z))$$

$$= \sum_{n=0}^{\infty} \frac{(k_1(x,z))^n}{n!}$$

By applying n times, $k_2(x,z) = (k_1(x,z))^n$ is a valid kernel by applying (2.3.3). Letting $c = \frac{1}{n!}$ where $c > 0$, $k_3(x,z) = ck_2(x,z)$ is also kernel by (2.3.1). Consequently, we have:

$$k(x,z) = \lim_{a \to \infty} \sum_{n=0}^{\infty} k_3(x,z)$$

Applying (2.3.2) n times, we conclude that $k(x,z)$ is a kernel in the limit.

The following are some examples of valid kernels.

Example 2.3.1: (Gaussian / Radial Basis Function (RBF) kernel) Let $k: X \times X \to \mathbb{R}$ be a kernel function with mapping $\varphi(x)$ then $k(x,z)$ is valid kernel.

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

Starting with

$$-\|x-z\|^2 = -\langle x,x\rangle + 2\langle x,z\rangle - \langle z,z\rangle$$

Where $k_2(x,z) = 2(x,z)$ is a kernel by $k(x,z) = ck_1(x,z)$. Then we have

$$\exp\left(\frac{-\|x-z\|^2}{2\sigma^2}\right) = \exp\left(\frac{-\|x\|^2 - \|z\|^2 + 2\langle x,z\rangle}{2\sigma^2}\right)$$

$$= \exp\left(\frac{-\|x\|^2}{2\sigma^2}\right)\exp\left(\frac{-\|z\|^2}{2\sigma^2}\right)\exp\left(\frac{2\langle x,z\rangle}{2\sigma^2}\right)$$

Each term is kernel, the sum of a finite number of kernels is a kernel. Letting $c = \frac{1}{2\sigma^2}$ and $c > 0, \sigma^2 > 0$, the following hold:

$$k_3(x,z) = ck_2(x,z) = \frac{2\langle x,z\rangle}{2\sigma^2}$$

is kernel by $k(x,z) = ck_1(x,z)$. Therefore:

$$k_4(x,z) = \exp\left(k_3(x,z)\right)$$

is also kernel by $k(x,z) = \exp\left(k_1(x,z)\right)$. Suppose $f(x) = \exp\left(-\frac{\langle x,x\rangle}{2\sigma^2}\right)$ and $f(z) = \exp\left(-\frac{\langle z,z\rangle}{2\sigma^2}\right)$, then we have:

$$k(x,z) = f(x)k_4(x,z)f(z)$$

is kernel since $k(x,z) = f(x)k_1(x,z)f(z)$.

Example 2.3.2: The function $k(x,z) = \exp(\gamma\|x-z\|^2)\,, \gamma > 0$ is not valid kernel.

A function $k(x,z)$ is considered a valid kernel function. For any set $x_1, x_2, \ldots, x_n$, the corresponding Gram matrix K defined by $K_{ij} = k(x_i, x_j)$ be positive semi-definite.

Let $x_1, x_2 \in \Re^2$, $x_1 = (0,0)^\top$ and $x_2 = (1,0)^\top$. Then

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_2, x_1) & k(x_2, x_2) \end{bmatrix} = \begin{bmatrix} 1 & \exp(\gamma) \\ \exp(\gamma) & 1 \end{bmatrix}$$

The matrix determinant is calculated as:

$$\det(K) = 1 - \exp(2\gamma)$$

Since $\gamma > 0$ and $\exp(2\gamma) > 1$, it follows that $\det(K) < 0$. This means that, the matrix K is not positive semi-definite. Consequently, the function $k(x, z)$ cannot be a valid kernel function, as a valid kernel must always produce a positive semi-definite Gram matrix for any set of inputs.

2.4 Application of kernel method

Kernel methods are a sophisticated class of algorithms in machine learning and artificial intelligence that enable the analysis of nonlinear data by implicitly mapping it to a higher-dimensional space. These methods rely on semi-definite positive kernel functions to assess the similarity between data points without explicitly performing a transformation. Some of these applications are as follows:

2.4.1 Support Vector Machines (SVMs) (Bishop & Nasrabadi, 2006)**:** (SVMs) and Bayes Point Machines (BPMs) are prominent linear learning algorithms that can be significantly enhanced through the employ of kernel methods. Support Vector Machines (SVMs) have demonstrated remarkable effectiveness in tackling various machine learning challenges, particularly those involving high-dimensional data and complex structures. They perform exceptionally well in situations where the number of features significantly surpasses the number of observations, making them especially valuable in areas like natural language processing and bioinformatics.

Support Vector Machines employ a dual representation defined by the following optimization problem:

$$min_\alpha \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_i \alpha_i$$

with $\sum_i \alpha_i y_i = 0$ and $\alpha_i \geq 0$ for all i. They operate in a kernel induced feature space such as the decision function is

$$f(x) = \sum_i \alpha_i y_i \langle \phi(x_i), \phi(x) \rangle + b$$

and the classification is determined by:

$$h(x) = singn\big(f(x)\big)$$

where is the function linear in implicit feature space by K.

2.4.2 Kernel Ridge Regression

2.4.2.1 Linear regression: The linear ridge regression model takes the primal form

$$y = w^T x + \varepsilon$$

Where $x \in \mathbb{R}^n$ is input feature vector and $w = (XX^T + \lambda I)^{-1} Xy$. In the dual representation

$$y = x^T X\alpha$$

where $\alpha = (X^T X + \lambda I)^{-1} y; \alpha \in \mathbb{R}^n, X^T X \in \mathbb{R}^{n \times n}$ is the matrix and $y \in \mathbb{R}^n$. w is linear combination of training data points for $w \in \mathbb{R}^n$. The model can be expressed as:

$$w = \sum_{i=1}^{n} \alpha_i x_i = X\alpha$$

From this, we can write it as:

$$w = (XX^T + \lambda I)^{-1} Xy = X\alpha$$

$$\Leftrightarrow (XX^T + \lambda I) X\alpha = Xy$$

$$\Leftrightarrow (XX^T X + \lambda X)\alpha = Xy$$

$$\Leftrightarrow X(X^T X + \lambda I)\alpha = Xy$$

$$\Leftrightarrow X^T X(X^T X + \lambda I)\alpha = X^T Xy$$

$$\Leftrightarrow (K + \lambda I)\alpha = y$$

$$\Leftrightarrow \alpha = (K + \lambda I)^{-1} y$$

Thus, we can express the kernel ridge regression model in primal form as:

$y = w^T \Phi(x) + \varepsilon$ (primal)

or in dual form as:

$y = K_x \alpha$ where $\alpha = (K + \lambda I)^{-1} y$ (Dual) and K is the matrix of mutual scalar products such that

$$XX^T = K = \begin{bmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \cdots & \langle x_1, x_n \rangle \\ \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & & \langle x_2, x_n \rangle \\ \vdots & & \ddots & \vdots \\ \langle x_n, x_1 \rangle & \langle x_n, x_2 \rangle & \cdots & \langle x_n, x_n \rangle \end{bmatrix}$$

The prediction function can then be generalized using the kernel trick:

$$\hat{y}(x') = k(x')(K + \lambda I)^{-1} y$$

Where $k(x') = \begin{bmatrix} k(x', x_1) \\ \vdots \\ k(x', x_n) \end{bmatrix}^T$. (Rasmussen & Williams, 2006)

2.4.2.2 Non-linear regression: For a mapping $\varphi: X \to S$, where S is feature space or some vector space associated with inner product $\langle ., . \rangle$. The prediction of ridge regression on a signal x can be presented as:

$$\gamma_{RR} = y'(aI + K)^{-1} k$$

Where K is defined as

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & & k(x_2, x_n) \\ \vdots & & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{bmatrix}$$

Where the function $k: X \times X \to \mathbb{R}$ is given by $k(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle$. (Kalnishkan, 2009)

2.4.3 Signal processing: Kernel methods have shown effectiveness in addressing various signal processing and communications challenges. They have been applied in tasks such as detection, estimation, and classification, often leveraging the standard Support Vector Machine (SVM) algorithm. To enhance performance and incorporate domain-specific

knowledge, researchers have explored techniques like virtual training samples and custom kernel design. Applications span diverse fields, including speech and audio processing (e.g., recognition, identification, feature extraction, segmentation), image processing (e.g., face detection, recognition, coding), and communications (e.g., channel equalization, non-stationary modeling, multi-user detection). (Schölkopf et al. 2001)

2.4.4 Principal Component Analysis (PCA): Principal component analysis (PCA) is a fundamental technique in unsupervised learning, aimed at transforming correlated variables into a set of uncorrelated features called principal components. This transformation identifies the directions of maximum variance within the data. A key strength of PCA is that a small subset number of principal components can often capture most of the underlying structure in the original dataset. However, a key limitation of PCA is in its inability to uncover nonlinear relationships inherent in the original variables.

Let assume the mean of data is $\sum_{i=1}^{n} x_i = 0$ then, the standard PCA is

$$sv_i = \lambda_i v_i$$

Where $s = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$

For a non-linear transformation $\Phi(x_i)$ where $\sum_{i=1}^{n} \Phi(x_i) = 0$, we can define C as:

$$C = \frac{1}{n}\sum_{i=1}^{n} \Phi(x_i)\Phi(x_i)^T \quad (2.4.1)$$

From $Cu_i = \lambda u_i$ substitute in (2.4.1), the following can be derived:

$$\frac{1}{n}\sum_{i=1}^{n} \Phi(x_i)\Phi(x_i)^T u_i = \lambda u_i$$

This means $u_i = \sum_{i=1}^{n} a_{in}\Phi(x_n)$ where a_{in} values

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \Phi(x_i)\Phi(x_i)^T \sum_{j=1}^{n} a_{ij}\Phi(x_j) = \lambda_i \sum_{i=1}^{n} a_{in}\Phi(x_n) \quad (2.4.2)$$

Multiplying both sides in (2.4.2) by $\Phi(x_l)$ gives

$$\frac{1}{n}\sum_{i=1}^{n} k(x_l, x_i)\sum_{j=1}^{n} a_{ij}k(x_i, x_j) = \lambda_i \sum_{i=1}^{n} a_{in}k(x_l, x_n)$$

Where $k(x_l, x_n) = \Phi(x_l)^T \Phi(x_n)$ kernel function then, it can be found the projection of the image of x onto a given PCA as:

$$\Phi(x)^T u_i = \sum_{i=1}^{n} a_{in}\Phi(x)^T \Phi(x_i) = \sum_{i=1}^{n} a_{in}k(x, x_i).$$

In this thesis, will utilize kernel methods in support vector machines with a Gaussian kernel, and the results will be presented in Chapter Four. (Karatzoglou, 2006)

Chapter Three

Positive definite kernel

Chapter 3

Positive definite kernel

Introduction

Learning methods employing positive definite kernels which have gained significant popularity in machine learning over the past few decades. Their strong mathematical foundation has attracted considerable interest from statisticians and mathematicians. Traditionally, these experts have developed theories and algorithms within the linear framework of machine learning and statistics. However, due to the inherent limitations of linear models in capturing complex relationships within real-world data, the need for nonlinear approaches has become evident. To achieve successful predictions, it is essential to model the dependent relationships present in data, necessitating the exploration of nonlinear methods.

In this chapter, comprehensive theoretical overview of positive definite kernels will be provided, highlighting their key properties and significance in machine learning. The concept of Reproducing Kernel Hilbert Space (RKHS) will be introduced, a framework that elegantly combines the ideas of functional analysis and statistical learning. Furthermore, we will explore some of the fundamental properties and theorems associated with RKHS will be explored, illustrating how they facilitate the development of effective learning algorithms and enhance our understanding of the relationships within data. Through this exploration, a solid foundation will be established for elucidating and understanding how positive definite kernels derive advancement in machine learning methodologies.

3.1. Positive definite Kernel

In machine learning and statistics, positive definite kernels often called kernels are vital functions utilized in various algorithms. They are particularly significant in support vector machines, kernel principal component analysis, and Gaussian processes. Below is a concise definition and explanation:

Definition 3.1.1 (positive definite kernel) (Yang, 2016)**:** A positive definite kernel on a set $\boldsymbol{X}$ is a symmetric function $\boldsymbol{k}: \boldsymbol{X} \times \boldsymbol{X} \rightarrow \mathbb{R}$ that satisfies the property:

For all $\boldsymbol{x_1, x_2, \dots, x_n \in X}$ and $\boldsymbol{\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{R}, n \in \mathbb{N}}$, the kernel satisfies:

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

It can be generalized to complex case where $k: X \times X \to \mathbb{C}$ and $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{C}$ such that:

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

Where $\bar{\alpha}$ denotes the complex conjugate of α where $k(x, z) = \overline{k(z, x)}$.

Symmetry ensures that the relationship measured by the kernel between two points remains unchanged regardless of their order. For example, the similarity between points x and y is the same as that between y and x. Positive definiteness guarantees that the kernel can be considered an inner product in a feature space, allowing for the analysis or transformation of data while preserving the concepts of distance and similarity.

Definition 3.1.2 (Integrally positive definite kernel) (Fasshauer, 2011)**:** Let k be a symmetric kernel, then k is called integrally positive definite on $\varphi \times \varphi$ for all $f \in L_2(\varphi)$ such that:

$$\int\int k(x, z)f(x)f(z)dxdz \geq 0.$$

An integrally positive definite kernel $k(x, z)$ serves as the kernel of a positive integral operator $\mathcal{T}$**.** The positivity of $\mathcal{T}$ is expressed by the condition:

$$\langle \mathcal{T}f, f\rangle_{\mathcal{H}} = \int\int k(x, z)f(x)f(z)dxdz$$

Thus, an integrally positive definite kernel is the kernel associated with a positive integral operator. This property is fundamental in the study of kernel methods, functional analysis, and operator theory.

In the following sections, we will explore various properties of positive definite kernels. These properties are essential for understanding their behavior and applications in functional analysis and statistics.

3.1.1 Some Properties of Positive Definite Kernels

Property 1 (Suzuki, 2022)**:** Let X be a nonempty set and k_1, k_2 are positive definite kernel on $X \times X$. Then $ak_1 + bk_2$ is positive definite kernel, a, b are nonnegative numbers.

Proof: Since k_1 and k_2 are positive definite kernel, such that

$$k_1 = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \overline{\alpha}_j k_1\left(x_i, x_j\right) \geq 0$$

and

$$k_2 = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \overline{\alpha}_j k_1\left(x_i, x_j\right) \geq 0$$

Then

$$\begin{aligned} k_1 + k_2 &= \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \overline{\alpha}_j (a k_1(x_i, x_j) + b k_2(x_i, x_j)) \\ &= a\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \overline{\alpha}_j \, k_1(x_i, x_j) + b\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \overline{\alpha}_j \, k_2(x_i, x_j) \\ &= a\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \overline{\alpha}_j \, k_1(x_i, x_j) + b\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \overline{\alpha}_j \, k_2(x_i, x_j) \geq 0 \end{aligned}$$

Since both $k_1(x_i, x_j)$ and $k_2(x_i, x_j)$ are positive definite kernel, then $\alpha k_1 + \beta k_2$ is positive definite kernel.

The sum of two kernels is always a kernel, maintaining the positive definiteness needed for developing more complex models in machine learning. However, the difference of two kernels does not necessarily result in a kernel, as it may violate the non-negativity condition for inner products. This distinction highlights the limitations of using subtraction in kernel-based methods.

Property 2 (Fukumizu, 2010)**:** Let X be a nonempty set, and the kernels k_1, k_2 are positive definite kernel on $X \times X$. Then the product $k_1 k_2$ is also positive definite kernel.

Proof: Since k_1, k_2 is positive definite kernel such that

$$k_1 = \sum_{i}\sum_{j} k_1(x_i, x_j)\alpha_i \overline{\alpha}_j \geq 0$$

and

$$k_2 = \sum_{i}\sum_{j} k_2(x_i, x_j)\alpha_i \overline{\alpha}_j \geq 0$$

Then

$$k_1 k_2 = \sum_i \sum_j \alpha_i \overline{\alpha}_j k_1(x_i, x_j) k_2(x_i, x_j)$$

By the Cauchy-Schwarz inequality (Hadmard product). Since both $k_1(x_i, x_j)$ and $k_2(x_i, x_j)$ is positive definite kernel, then $k_1 k_2 \geq 0$ are positive definite kernel.

Property 3 (Fukumizu, 2008)**:** Let X be a nonempty set, k is positive definite kernel on X. Then $k(x, x) \geq 0, \ \forall x \in X$.

Proof: The proof of this property is straightforward because

$$k(x, x) = \|x\|^2 \geq 0.$$

Property 4 (Fukumizu, 2008)**:** Let X be a nonempty set and k is positive definite kernel on X such that $k: X \times X \to \mathbb{C}$ then for any x, y in X:

$$|k(x, z)|^2 \leq k(x, x) k(z, z).$$

Proof: By using the property $k(x, z) = \overline{k(z, x)}$, the definition of positive definiteness implies that the eigenvalue of the Hermitian matrix is:

$$K = \begin{bmatrix} k(x, x) & \overline{k(x, z)} \\ k(x, z) & (z, z) \end{bmatrix} \geq 0$$

The determinate of matrix is $k(x, x) k(z, z) - |k(x, z)|^2 \geq 0$, then:

$$|k(x, z)|^2 \leq k(x, x) k(z, z).$$

is a consequence of the Cauchy-Schwarz inequality applied to the inner product defined by the kernel k.

Property 5 (Fukumizu, 2010)**:** Let k_α be a generalized sequence of positive definite kernels, where sequence k_α converges uniformly to a limit kernel k such that:

$$\lim_r k_r(x, z) = k(x, z) \ \ \forall x, z \in X$$

Then k is a positive definite kernel.

Proof: Assume the above statement is true, and let

$$B_\infty = \sum_i \sum_j k_r(x_i, x_j) \alpha_i \alpha_j = -\varepsilon.$$

For $x_1, x_2, \dots, x_n \in X, \alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{R}$ with $\epsilon > 0$, we have

$$B_r = \sum_i \sum_j k_r(x_i, x_j) \alpha_i \alpha_j \geq 0$$

when $r \to \infty$ then $B_\infty \to 0$ which is a contradiction. Thus $B_\infty \geq 0$ implies that our assumption is false.

Proposition 3.1.1 (Manton, J. H., et al. 2015)**:** Let $k: X \times X \to \mathbb{R}$ be a symmetric function. Then k is positive definite if and only if k is a kernel.

Proof: ($\Longrightarrow$) Since k is positive definite kernel, we can write in the form $k = UDU^T$

Where U is an orthogonal matrix, D is a diagonal matrix containing the (non-negative) eigenvalues of K. We can define

$$\varphi(x) = D^{1/2}U^T 1_x$$

Where 1_x is the indicator vector for x. The inner product of $\varphi(x)$ and $\varphi(z)$ is:

$$\langle \varphi(x), \varphi(z) \rangle = \langle D^{1/2}U^T 1_x, D^{1/2}U^T 1_z \rangle = 1_x^T UDU^T 1_z$$

Since $k = UDU^T$, Then:

$$\langle \varphi(x), \varphi(z) \rangle = 1_x k 1_z = k(x, z)$$

Thus, k is a kernel.

($\Longleftarrow$) Assume k is a kernel, then there exists a feature map $\varphi: X \to \mathcal{H}$, such that:

$$k(x, z) = \langle \varphi(x), \varphi(z) \rangle_{\mathcal{H}}$$

For any points $x_1, x_2, \dots, x_n \in X$ and any vectors $u_1, u_2, \dots, u_n \in \mathbb{R}^n$, we have:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} u_i u_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} u_i u_j \langle \varphi(x_i), \varphi(x_j) \rangle$$

$$= \langle \sum_{i=1}^{n} u_i \varphi(x_i), \sum_{j=1}^{n} u_j \varphi(x_j) \rangle = \left\| \sum_{i=1}^{n} u_i \varphi(x_i) \right\|^2 \geq 0$$

This implies that k is positive definite.

Proposition 3.1.2 (Rachev, Klebanov & Fabozzi, 2013)**:** Suppose (X, Ω) be a measurable space, and let μ be a σ −field. Let $k: X \times X \to \mathbb{C}$ is a positive definite kernel on X^2, that is integrable and measurable in respect with $\mu \times \mu$. Then, the double integral of k satisfies:

$$\int \int k(x, z) d\mu(x) d\mu(z) \geq 0$$

Proof: Since k is a measurable function of two variables with respect to the dot product of $\sigma-$field $\varphi \times \varphi$, the function $k(t,t)$ of one variable is also measurable. Therefore, there exists a set $x_0 \in \Omega$ such that $\mu(x_0) < \infty$ and the function $k(t,t)$ is bounded on x_0. The positive definiteness of k indicates that for any $n \geq 2$ and any points $t_1, t_2, \dots, t_n \in X$, the following inequality holds:

$$\sum_{i=1}^{n} k(t_i, t_i) + \sum_{i \neq j} k(t_i, t_j) \geq 0 \quad (3.1.1)$$

Integrating both sides (3.1.1) over the set x_0 with respect to the n-times product measure $\mu \times \dots \times \mu$ we have:

$$n\big(\mu(x_0)\big)^{n-1} \int_{x_0} k(t,t)d\mu(t) + n(n-1)\big(\mu(x_0)\big)^{n-2} \int_{x_0}\int_{x_0} k(x,z)d\mu(x)d\mu(z) \geq 0$$

Since the inequality holds for arbitrary $n \geq 2$, we can take the limit as n grows large. The dominant term in the inequality is the double integral term, which must therefore satisfy:

$$\int_{x_0}\int_{x_0} k(x,z)d\mu(x)d\mu(z) \geq 0$$

Theorem 3.1.1 (Fasshuer, 2014)**:** If k is strictly positive definite, then the point evaluation functionals δ_x in the reproducing kernel Hilbert space $\mathcal{H}_k$ are linearly independent.

Proof: Assume the opposite, that the point evaluation functionals $\delta_{x_1}, \delta_{x_2}, \dots, \delta_{x_n}$ are linear dependent then there exist points $x_1, \dots, x_n \in X$ and the coefficient $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{C}$ not all zero such that:

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

Thus

$$\sum_{i=1}^{n} \alpha_i k(.\,, x_i) = 0$$

For any $f \in \mathcal{H}_k$ arbitrary function and utilizing the reproducing property of k where $\mathcal{H}_k$ Hilbert space

$$\langle f, \sum_{i=1}^{n} \alpha_i k(.\,, x_i)\rangle = 0$$

Which simplifies to

$$\sum_{i=1}^{n} \alpha_i \langle f, k(.\, x_i) \rangle = 0$$

Then

$$\sum_{i=1}^{n} \alpha_i f(x_i) = \sum_{i=1}^{n} \alpha_i \delta_{x_i}(f)$$

Where $\delta_{x_i}(f) = f(x_i); i = 1, \dots, n$, this implies that the linear combination of the point evaluation functionals δ_{x_i} vanishes for all $f \in \mathcal{H}_k$.
Now, we will introduce a range of examples of positive definite kernels.

Example 3.1.1: Let X be any set and $A \subseteq X$, the kernel:

$$k(x, z) = \begin{cases} 1; & x \in A \text{ and } z \in A \\ 0 & \text{otherwise} \end{cases}$$

Is positive definite kernel because if all points $x_1, x_2, \dots, x_n \in A$, then by definition k we have $k(x_i, x_j) = 1$, and

$$\sum_{i=1}^{n} \sum_{j=1}^{m} k(x_i, x_j) \alpha_i \alpha_j = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \alpha_j = \left(\sum_{i=1}^{n} \alpha_i \right)^2 \geq 0$$

Also if $x_1, x_2, \dots, x_n \notin A$, then by definition k we have $k(x_i, x_j) = 0$, and

$$\sum_{i=1}^{n} \sum_{j=1}^{m} k(x_i, x_j) \alpha_i \alpha_j = 0$$

Example 3.1.2: For any $a, x, y \in \mathbb{R}$, define a set:

$$U_a(x) = \begin{cases} 1 \; for \; x \leq a \\ 0 \; for \; x \geq a \end{cases} \; ; \; a \in \mathbb{R}$$

Let F be a bounded non-decreasing function on $\mathbb{R}$, define the kernel:

$$k(x, z) = \int_{-\infty}^{\infty} U_a(x \vee z) dF(a)$$

Where the $(x \vee z)$ is the maximum of x and z. For all $x_1, x_2, \dots, x_n \in \mathbb{R} \; and \; \alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{R}$, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j) \alpha_i \alpha_j = \int_{-\infty}^{\infty} \left(\sum_{i=1}^{n} U_a(x_i) \alpha_i \right)^2 dF(a) \geq 0$$

Then k is positive definite kernel.

Example 3.1.3: If x, z are elements of the Hilbert space $\mathcal{H}$, the kernel

$$k(x,z) = \exp\{-\|x-z\|^2\}$$

is positive definite kernel because $\forall\, x_1, x_2, \dots, x_n \in \mathcal{H}$ and $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{C}$ we have:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} k(x,z)\alpha_i\overline{\alpha}_j$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\overline{\alpha}_j e\{-\|x_i - x_j\|^2\}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\overline{\alpha}_j \exp\{-\|x_i\|^2\}.\exp\left\{-\|x_j\|^2\right\}.\exp\{2Re\langle x_i, x_j\rangle\}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i^{\mathrm{T}}\overline{\alpha_j^{\mathrm{T}}} \exp\{2Re\langle x_i, x_j\rangle\} \geq 0$$

Where $\alpha_i = \exp\{\|x_i\|^2\}, \alpha_i^{\mathrm{T}} = \exp\{-\|x_i\|^2\}$.

Example 3.1.4: The kernel $k(x,z) = e^{i(x-z)t}$ is positive definite on $\mathbb{R}^n \times \mathbb{R}^n$ for a fixed $t \in \mathbb{R}^n$. This can be shown by rewriting the kernel as:

$$k(x,z) = e^{ix^{\mathrm{T}}t} * e^{-iz^{\mathrm{T}}t}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} k(x_i, x_j)\alpha_i\,\alpha_j$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\,\alpha_j(e^{ix_i^{\mathrm{T}}t} * e^{-ix_jt})$$

$$\left|\sum_{i=1}^{N} \alpha_i\varphi(x_i)\right|^2 \geq 0$$

Since the squared modulus is always non-negative, the kernel $k(x,z)$is positive definite.

Example 3.1.5: For any $x, z \in \mathbb{R}$, the cosine function defines to a positive definite kernel on $\mathbb{R} \times \mathbb{R}$, such that $k(x,z) = \cos(x-z)t$

From pervious example (3.1.4) $k_1 = e^{i(x-z)t}$ is positive definite on $\mathbb{R}^n \times \mathbb{R}^n$ for any fixed $t \in \mathbb{R}^n, k_1 = e^{i(x-z)t}$ and $k_2 = e^{-i(x-z)t}$ are both positive definite on $\mathbb{R} \times \mathbb{R}$. Using Euler's formula, the cosine kernel can be presented as:

$$\cos(x-z)t = \frac{1}{2}(e^{i(x-z)} + e^{-i(x-z)})$$

Now, consider the sum

$$\sum_{i=1}^{n}\sum_{j=1}^{n} k(x,z)\alpha_i\alpha_j = \sum_{i=1}^{n}\sum_{j=1}^{n} \cos(x-z)\,\alpha_i\alpha_j$$

This can be written as

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j \frac{1}{2}(k_1+k_2) == \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j\left(e^{i(x_i-x_j)t} + e^{-i(x_i-x_j)t}\right)$$

Which can be simplifies

$$= \frac{1}{2}\left(\sum_i\sum_j \alpha_i\alpha_j e^{i(x_i-x_j)t} + \sum_i\sum_j \alpha_i\alpha_j\, e^{-i(x_i-x_j)t}\right)$$

$$= \frac{1}{2}\left[\left(\sum_i \alpha_i e^{ix_it}.\sum_j \alpha_j e^{-ix_jt}\right) + \left(\sum_j \alpha_j e^{ix_jt}.\sum_i \alpha_i e^{-ix_it}\right)\right]$$

By rearranging, we get

$$= \frac{1}{2}\left[\left|\sum_i \alpha_i e^{ix_it}\right|^2 + \left|\sum_j \alpha_j e^{ix_jt}\right|^2\right] \geq 0$$

This demonstrates that the expression is non-negative, thereby affirming the positive definiteness of the kernel.

3.2 Hilbert- Schmidt

The Hilbert-Schmidt theorem plays a fundamental role in functional analysis, mathematical physics, and kernel-based methods, ensuring that a wide range of functions can be expressed as infinite series in terms of the eigenfunctions of Hilbert-Schmidt operators.

Definition 3.2.1 (Hilbert- Schmidt operator) (Fasshauer, 2014): Let $\mathcal{H}$ be a Hilbert space, and let $\mathcal{T}: \mathcal{H} \to \mathcal{H}$ be bonded linear operator. The operator $\mathcal{T}$ is called Hilbert-Schmidt operator if there exists an orthonormal basis $\{\lambda_n\}$ in $\mathcal{H}$, where

$$\sum_{n=1}^{\infty} \|\mathcal{T}\lambda_n\|^2 < \infty\,,$$

where $\|.\|$ denotes the norm in $\mathcal{H}$ induced by the inner product $\langle .,. \rangle$. The norm of $\mathcal{T}$

$$\|\mathcal{T}\|_{HS} = \sqrt{\sum_{n=1}^{\infty} \|\mathcal{T}\lambda_n\|^2}$$

is called Hilbert-Schmidt.

The Hilbert- Schmidt norm generalizes the Frobenius norm to infinite dimensional space in other words, if $\mathcal{T}$ is presented as a matrix and $\mathcal{H}$ is finite dimensional then the Hilbert-Schmidt norm is equivalent to the Frobenius norm. This norm of Frobenius of a matrix $A = [a_{ij}]_{m\times n}$ is defined as:

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} |a_i a_j|^2}$$

In this context, the Hilbert-Schmidt norm can be calculated using the singular values of the matrix, leading to the following relationship:

$$\|\mathcal{T}\|_{HS} = \|A\|_F$$

This equivalence highlights the connection between these norms in different contexts. The Hilbert-Schmidt norm applies in infinite-dimensional settings, while the Frobenius norm is relevant in finite-dimensional settings.

3.2.1 The eigenvalue and eigenfunction of Hilbert-Schmidt

This eigenvalue problem for a Hilbert-Schmidt operator is equivalent to solving the second kind homogeneous Fredholm integral equation. This means that for certain values of λ (eigenvalues) and their corresponding functions φ (eigenfunctions). Consider the homogeneous Fredholm equation:

$$y(x) = \lambda \int_a^b k(x,z)y(z)dz \quad (3.2.1)$$

where the kernel $k(x,z)$ is symmetric so that $k(x,z) = k(z,x)$.

Theorem 3.2.1 (Pipkin, 1991)**:** Eigenfunctions corresponding with distinct eigenvalues of a symmetric kernel are orthogonal over the interval $[a,b]$.

Proof: Let $\lambda_n$and λ_m be two distinct eigenvalues of (3.2.1) with corresponding eigenfunctions $y_n(x)$ and $y_m(x)$ respectively. Then

$$y_m(x) = \lambda_m \int_0^b k(x,z) y_m(z) dz \quad (3.2.2)$$

and

$$y_n(x) = \lambda_n \int_0^b k(x,z) y_n(z) dz \quad (3.2.3)$$

Multiplying both sides of (3.2.2) by $y_n(x)$, integrating over $[a,b]$ and applying Fubini's theorem for swapping integrals

$$\int_a^b y_m(x) y_n(x) dx = \lambda_m \int_a^b y_n(x) \left[\int_a^b k(x,z) y_m(z) dz \right] dx$$

Swap integration order

$$= \lambda_m \int_a^b y_m(x) \left[\int_a^b k(x,z) y_n(z)\, dx \right] dz$$

By symmetry of $K(x,z)$, we have

$$\lambda_m \int_a^b y_m(z) \left[\int_a^b k(z,x) y_n(x)\, dx \right] dz \quad (3.2.4)$$

From the above, we have

$$y_n(z) = \lambda_n \int_a^b k(z,x) y_n(x) dx$$

Hence

$$\int_a^b k(z,x) y_n(x) dx = \frac{y_n(z)}{\lambda_n} \quad (3.2.5)$$

Substitute this value for the inner integral in (3.2.4), we get

$$\int_a^b \lambda_m(x) y_n(x) dx = \lambda_m \int \frac{y_m(x).y_n(x)}{\lambda_n} dx$$

Simplify to

$$\lambda_n \int_a^b y_m(x) y_n(x) dx = \lambda_m \int_a^b y_m(x) y_n(x) dx$$

By rearranging, we have

$$(\lambda_n - \lambda_m) \int_a^b y_m(x) y_n(x) dx = 0 \quad (3.2.6)$$

Since λ_n and $\lambda_m$are different then $\lambda_n - \lambda_m \neq 0$, that is mean:

$$\int_a^b y_m(x) y_n(x) dx = 0$$

Therefore, $y_n(x)$ and $y_m(x)$ are orthogonal over (a, b). Thus, eigenfunctions corresponding to distinct eigenvalues are orthogonal.

Theorem 3.2.2 (Pipkin, 1991)**:** All eigenvalues of a symmetric kernel are real.

Proof: Consider equation (3.2.6). By setting $\lambda_n = \overline{\lambda_m}$ and $y_n(x) = \overline{y_m(x)}$, we have

$$(\overline{\lambda_m} - \lambda_m) \int_a^b y_m(x) y_m(x) dx = 0 \quad (3.2.7)$$

Letting $\lambda_m = \alpha_m + i\beta_m$ and $y_m(x) = f_m(x) + ig_m(x) \Longrightarrow \overline{\lambda_m} = \alpha_m - i\beta_m$ and $\overline{y_m(x)} = f_m(x) - ig_m(x)$ then Substituting into the equation (3.2.7) becomes:

$$-2\beta_m \int_a^b [f_m(x) + ig_m(x)][f_m(x) - ig_m(x)] dx = 0$$

This simplifies

$$\beta_m \int_a^b [f_m^2(x) + g_m^2(x)] dx$$

Since $f_m^2(x) + g_m^2(x) \neq 0$ and $y_m(x) \neq 0$, it follows that $\beta_m = 0$. Consequently $\lambda_m = \alpha_m$ is demonstrating that λ_m are real. Thus, all eigenvalues of a symmetric kernel are real. We can prove in other way:

Assuming $\lambda_m \in \mathbb{C}$ and the corresponding eigenfunction $y_m(x)$. By the complex conjugate of (3.2.2)

$$\overline{y_m} = \overline{\lambda_m} \int_a^b \overline{K(x,z)}\, \overline{y_m(z)}\, dz$$

And multiplying by $y_m(x)$, integrates and use the symmetric $\overline{k(x,z)} = k(x,z)$:

$$\int_a^b |y_m(x)|^2\, dx = \overline{\lambda_m} \int_a^b y_m(z) \left[\int_a^b k(x,z)\, \overline{y_m(x)} \right] dz$$

Using Fubini's theorem, swap the order of integration:

$$\int_a^b |y_m(x)|^2\, dx = \overline{\lambda_m} \int_a^b \overline{y_m(x)} \left[\int_a^b k(x,z) y_m(z) dz \right] \quad (3.2.8)$$

From the equation (3.2.2), we have

$$\int_a^b k(x,z) y_m(z) dz = \frac{1}{\lambda_m} y_m(x)$$

Substitute this into the equation (3.2.8)

$$\int_a^b |y_m(x)|^2\, dx = \frac{\overline{\lambda_m}}{\lambda_m} \int_a^b |y_m(z)|^2 dz$$

Since $\int_a^b |y_m(x)|^2\, dx = \int_a^b |y_m(z)|^2\, dz$, we can write

$$\left(\frac{\overline{\lambda_m}}{\lambda_m} - 1 \right) \int_a^b |y_m(z)|^2 dz = 0$$

Since $\int_a^b |y_m(z)|^2 dz \neq 0$, then $\left(\frac{\overline{\lambda_m}}{\lambda_m} - 1 \right) = 0 \implies \frac{\overline{\lambda_m}}{\lambda_m} = 1$. This implies $\lambda_m = \overline{\lambda_m}$ which means λ_m is a real number.

Proposition 3.2.1: Let k be a kernel in L_2 where $\mathcal{H} = L_2(\varphi, \rho)$ be a Hilbert space and ρ is a weight function such that $\int \rho(x) dx = 1$

And

$$\int \int |k(x,z)|^2 \rho(x) \rho(z) dx dz < \infty$$

Then the operator $\mathcal{T}$ defined as:

$$(\mathcal{T}f)(x) = \int k(x,z)f(z)\rho(z)dz \quad f \in L_2(\varphi)$$

is Hilbert-Schmidt operator.

Moreover, every Hilbert-Schmidt operator on the space L_2 can be uniquely represented by a kernel k that also belongs to L_2. To better understand this idea, we will now present an example of such a Hilbert-Schmidt operator.

Example 3.2.1: Consider the kernel defined by

$$k(x,z) = \min(x,z) = \begin{cases} x, & x \leq z \\ z, & x > z \end{cases}$$

on the domain $\Omega = [0,1]$. The operator $\mathcal{T}$ satisfies the equation

$$\mathcal{T}\varphi = \lambda\varphi \Leftrightarrow \int k(x,z)\varphi(z)\rho(z)dz = \lambda\varphi(x)$$

We have $\Omega = [0,1]$, $\rho = 1$ and $k(x,z) = \min(x,z)$ such that:

$$\int_0^1 \min(x,z)\,\varphi(z)dz = \lambda\varphi(x)$$

This can be expressed

$$\int_0^x z\varphi(z)dz + \int_x^1 x\varphi(z)dz = \lambda\varphi(x) \qquad (3.2.9)$$

we can rewrite the kernel as:

$$k(x,z) = \frac{1}{2}(x + z - |x - z|) = \begin{cases} {}^1\!/_2\,(x + z - (z - x) = x\,, x \leq z \\ {}^1\!/_2\,(x + z - (x - z) = z\,, x > z \end{cases}$$

Applying the operator $\mathcal{T} = \frac{d^2}{dx^2}$ on (3.2.9), we have

$$\frac{d^2}{dx^2}\left[\int_0^x z\varphi(z)dz + \int_x^1 x\varphi(z)dz\right] = \frac{d^2}{dx^2}[\lambda\varphi(x)]$$

By calculating the first derivative, we have

$$\frac{d}{dx}\left[\int_0^x z\varphi(z)dz + \int_x^1 x\varphi(z)dz\right] = x\varphi(x) - \int_1^x \varphi(z)dz - x\varphi(x) = -\int_1^x \varphi(z)dz$$

Also, the second derivative

$$\frac{d}{dx}\left[-\int_1^x \varphi(z)dz\right] = -\varphi(x)$$

Hence, we can write it as

$$\frac{d^2}{dx^2}\left[\int_0^x z\varphi(z)dz - x\int_x^1 \varphi(z)dz\right] = \lambda\varphi''(x)$$

This equation simplifies to

$$-\varphi(x) = \lambda\varphi''(x) \Leftrightarrow -\varphi''(x) = \frac{1}{\lambda}\varphi(x).$$

3.2.2 Extended Hilbert-Schmidt operator

The Extended Hilbert-Schmidt operator space, denoted as $HS_X(\mathcal{H})$ is defined as:

$$HS_X(\mathcal{H}) = \{A + \lambda I : A \in HS(\mathcal{H}), \lambda \in \mathbb{R}\}$$

Where A is a Hilbert-Schmidt operator on a Hilbert space $\mathcal{H}$, λ is a real scalar and I the identity operator. This set of such operator forms a Hilbert space with the respect to the extended Hilbert–Schmidt inner product, defined as:

$$\langle A + \lambda I, B + \mu I\rangle_{HS_X} = \langle A, B\rangle_{HS} + \lambda\mu$$

Where $\langle .,. \rangle_{HS}$ is the standard Hilbert-Schmidt inner product. The associated extended Hilbert–Schmidt norm is given by:

$$\|A + \lambda I\|^2_{HS_X} = \|A\|^2_{HS} + \lambda^2$$

Where $\|I\|_{HS_X} = 1$, in contrast to $\|I\|_{HS} = \infty$.

Reproducing kernel Hilbert space provides a powerful for understanding kernel methods in machine learning. Reproducing kernel Hilbert space, a specific type of Hilbert space that provides a unique perspective on kernel methods, will now be introduced.

3.3 Reproducing kernel Hilbert Space (RKHS)

kernel can be viewed from the functional analysis viewpoint. Kernel methodology involves transforming data into a high-dimensional feature space defined by Reproducing Kernel Hilbert Spaces (RKHS). Reproducing Kernel Hilbert Spaces (RKHS) establish a connection between the task of selecting an appropriate hypothesis space for learning functional relationships and the concept of a positive semidefinite kernel. In this context, "hypothesis space" refers to the set of all potential functions that the learning algorithm might consider as solutions to the problem. By employing a positive definite kernel, optimal performance can be achieved in both linear and nonlinear scenarios. To construct an RKHS, a feature map

is needed to transform the input space into a specific inner product known as the feature space. (Dinuzzo, 2011)

Definition 3.3.1 (**Reproducing kernel Hilbert space)** (Krbes, 2007)**:** Consider $\mathcal{H}$ be a Hilbert space consisting of function $f: X \to \mathbb{R}$ defined on a non-empty set X. A function $k: X \times X \to \mathbb{R}$ is termed as a reproducing kernel of $\mathcal{H}$ if:

1- For every $x \in X$, $k_x = k(.,x) \in \mathcal{H}$.

2- For every $f \in \mathcal{H}$, $x \in X$, $f(x) = \langle f, k(.,x)\rangle_{\mathcal{H}}$ (the reproducing property)

Where if reproducing kernel exists, then $\mathcal{H}$ called RKHS. A space $\mathcal{H}$ admitting a reproducing kernel is called a reproducing kernel Hilbert space (RKHS) and denoted by $\mathcal{H}_K(X)$.

Remark 3.3.1: Most Hilbert spaces are Reproducing Kernel Hilbert spaces (RKHS). Spaces (with their respective inner product) such as $\mathbb{R}$, $\mathbb{R}^n$, $l^2$and L^2 are examples of RKHS. The terms of reproducing refer to that every f in $\mathcal{H}$, the values $f(x)$ can be reconstructed from the inner product in the Hilbert space $\mathcal{H}$, using standard construction starting with X and k. This leads to the formation of the Hilbert space $\mathcal{H}$ of functions on X, knowing as reproducing kernel Hilbert space (RKHS). This space is determined by the pair (X, k). (Martorell, 2023)

Lemma 3.3.1 (Tsunekawa, 2023)**: (Riesz's representation)** For any φ from Hilbert space $\mathcal{H}$, all bounded linear functional are of the form $\varphi(f) = \langle g, f\rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$ and $g \in \mathcal{H}$. In addition, g satisfies$\|\varphi\|_{\mathcal{H}} = \|g\|$.

Theorem 3.3.1 (Bllakcori, 2022)**:** Let $(\mathcal{H}, \langle .,. \rangle)$ be a reproducing kernel Hilbert space on φ, and k be the corresponding reproducing kernel. Then the following holds.

1- $k(x,z) = \overline{k(z,x)}\ \forall x, z \in \mathcal{H}$.

Proof: For all $\forall x, z \in \varphi$ the reproducing kernel k satisfies:

$$k(x,z) = (k_z, k_x) = \overline{(k_x, k_z)} = \overline{k(z,x)}$$

2- $\sum_i \sum_j k(x_i, x_j)\alpha_i \bar{\alpha}_j \geq 0\ \forall\ i, j \in \mathbb{N}, x_1, \dots, x_n \in \varphi, \alpha_1, \dots, \alpha_n \epsilon \mathbb{C}$.

Proof: For all $x_1, \dots, x_n \in \varphi$ and $\alpha_1, \dots, \alpha_n \in \mathbb{C}$, consider the sum

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \bar{\alpha}_i \alpha_j (k_{x_i}, k_{x_j})$$

Using the reproducing property of k, this can be written as:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{\alpha_i}\alpha_j\,\langle k_{x_i}, k_{x_j}\rangle$$

$$=\langle\sum_{i=1}^{n}\alpha_i k_{x_i}, \sum_{j=1}^{n}\alpha_j k_{x_j}\rangle$$

Since the inner product is non-negative, we have:

$$\langle\sum_{i=1}^{n}\alpha_i k_{x_i}, \sum_{j=1}^{n}\alpha_j k_{x_j}\rangle \geq 0.$$

3- The functions $k_x \in \mathcal{H}$ are dense: $\{k_x : x \in \varphi\}$ is dense in $\mathcal{H}$

Proof: Let $\mathcal{H}_0 = span\{\, k(.,x):\ x \in X\}$ be the subspace spanned by the kernel function. For any $f, g \in \mathcal{H}_0$, we can write:

$$f = \sum_{i=1}^{n} a_i k(.,x_i),\ \ g = \sum_{j=1}^{m} b_j k(.,z_j)$$

The inner product of f and g is given by:

$$\langle f, g\rangle\ = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j k(x_i, z_j)$$

This expression is invariant under the specific choice of f and g. Using the reproducing property, we also have:

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

The norm of f is given by:

$$\|f\|^2 = \sum_{i,j=1}^{n} a_i \bar{a_j} k(x_i, x_j) \geq 0$$

If $\|f\| = 0$, then by Cauchy-Schwarz inequality

$$|f(x)| = |\langle f, k(.,x)\rangle| \leq \|f\|\|k(.,x)\| = 0$$

This implies $f(x) = 0$. (Fukumizu, 2008)

3.3.1 Properties of Reproducing Kernel Hilbert Spaces

There is a one-to-one correspondence relationship between kernels and reproducing kernel Hilbert spaces (RKHS). This connection allows us to study the properties of functions in an RKHS by leveraging the characteristics of the reproducing kernel. Conversely, we can infer properties of the reproducing kernel, such as boundedness, measurability, and integrability, by analyzing the corresponding properties of all functions in the associated RKHS.

Property 1 (Bounded) (Davies, 2022): Let $k: X \times X \to \mathbb{R}$ be a kernel on X with a corresponding reproducing kernel Hilbert space $\mathcal{H}$ and the feature map $\varphi: X \to \mathcal{H}$. Then k is bounded if and only if the feature map φ is bounded, $\|k\|_\infty = sup_{x \in X}\sqrt{K(x,x)} < \infty$.

Proof: By reproducing property of $\mathcal{H}$ and definition of the kernel function that leads to $\forall\, x, z \in X$:

$$k(x,z) = \langle \varphi(z), \varphi(x) \rangle = \langle k(.,x), k(.,z) \rangle$$

Taking the supremum over x, we have:

$$|k(x,z)|^2 = |\langle k(.,x), K(.,z) \rangle|^2 \leq \|k(.,x)\|_{\mathcal{H}}^2 . \|k(.,z)\|_{\mathcal{H}}^2 = k(x,x).k(z,z)$$

By using properties of the supremum and we get that:

$$\|k\|_\infty = sup_{(x,z)\in X\times X}|k(x,z)| \leq sup_{x,z\in X}\sqrt{k(x,x)}.\sqrt{k(z,z)} =$$

$$= sup_{x\in X}\sqrt{k(x,x)}.sup_{z\in X}\sqrt{k(z,z)} = sup_{x\in X}k(x,x)$$

On the other hand, we have:

$$sup_{(x,z)\in X\times X}|k(x,z)| \geq sup_{(x,x)\in X\times X}k(x,x)$$

since $\{k(x,x), x \in X\} \subseteq \{k(x,z), x, z \in X\}$ if $z = x$. Therefore, k is bounded if and only if

$$\|k\|_b = sup_{(x,z)\in X\times X}|k(x,z)| = sup_{x\in X}k(x,x) < \infty$$

By the monotonicity of the square root and $k(x,x) \geq 0$. Then

$$\|k\|_\infty = sup_{x\in X}\sqrt{k(x,x)} < \infty.$$

Property 2 (Bounded on feature map) (Davies, 2022)**:** Let $k: X \times X \to \mathbb{R}$ be a kernel on X with a reproducing kernel Hilbert space $\mathcal{H}$, a feature space $\mathcal{H}_0$ and feature map $\varphi: X \to \mathcal{H}_0$. Then k is bounded if and only if φ is bounded.

Proof: Since $\varphi: X \to \mathcal{H}_0$ is a feature map of k and by using reproducing property, we get

$$\|\varphi(x)\|_{\mathcal{H}_0}^2 = \langle \varphi(x), \varphi(x) \rangle_{\mathcal{H}_0} = \langle k(.,x), k(.,x) \rangle_{\mathcal{H}} = m(x,x)$$

Taking the supremum on both sides, we obtain:

$$\|\varphi\|_b^2 = \|k\|_\infty^2$$

Therefore

$$\|\varphi\|_b^2 < \infty \iff \|k\|_\infty^2 < \infty.$$

Property 3 (Bounded on feature space) (Davies, 2022)**:** Let $k: X \times X \to \mathbb{R}$ be a kernel on X associated with a reproducing kernel Hilbert space $\mathcal{H}$. Then, k is bounded if and only if every function $f \in \mathcal{H}$ is bounded. And the inclusion map $(id): \mathcal{H} \to l^\infty(X)$ is continuous, with $\|(id): \mathcal{H} \to l^\infty(X)\| = sup_{x \in X}\sqrt{\langle x, x\rangle} = \|k\|_\infty$.

Proof: ($\Longrightarrow$): Assume k is bounded, such that $\|k\|_\infty = sup_{x \in X}\sqrt{\langle x, x\rangle}$. For any $f \in \mathcal{H}$ and $x \in X$, the reproducing property of $\mathcal{H}$ is $f(x) = < f, K(.,x) >_{\mathcal{H}}$

Since $\|k(.,x)\|_{\mathcal{H}}^2 = \langle k(.,x), k(.,x)\rangle_{\mathcal{H}} = k(x,x)$, we have

$$|f(x)|^2 = |< f, k(.,x) >_{\mathcal{H}}|^2 \leq \|f\|_{\mathcal{H}}^2 k(x,x)$$

Taking the supremum over X:

$$\|f\|_\infty \leq \|f\|_{\mathcal{H}} \|k\|_\infty$$

Since $\|f\|_{\mathcal{H}} < \infty$ and $\|k\|_\infty < \infty$, it follows that $\|f\|_\infty < \infty$. Then f is bounded. Moreover, the inclusion map $(id): \mathcal{H} \to l^\infty(X)$ is well define and satisfies:

$$\|(id): \mathcal{H} \to l^\infty(X)\| \leq \|k\|_\infty \quad (3.3.1)$$

($\Longleftarrow$): Assume every $f \in \mathcal{H}$ is bounded. Then, the inclusion map $(id): \mathcal{H} \to l^\infty(X)$, is well define and linear as:

$$(id)(\alpha f + g)(x) = (\alpha f + g)(x) = \alpha f(x) + g(x)$$
$$= \alpha(id)(f)(x) + (id)(g)(x).$$

For any $f, g \in \mathcal{H}$ and $\alpha \in \mathbb{R}$.To prove (id) is bounded, we use the closed graph theorem, which asserts that a linear operator between Banach spaces is continuous if its graph is closed. Let $(f_n)_{n=1}^\infty$ be a sequence of function in $\mathcal{H}$ such that $f_n \to f$ in $\mathcal{H}$ and $f_n \to g$ in $l^\infty(X)$ as $n \to \infty$. Then $\lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}} \to 0$ and $\lim_{n\to\infty} \|f_n - g\|_\infty \to 0$. Convergence in $\mathcal{H}$ implies pointwise convergence, thus $f_n \to g$ for all $x \in X$. Since

$$|f_n(x) - g(x)| \leq \|f_n - g\|_\infty \Longrightarrow \lim_{n\to\infty} \|f_n - g\|_\infty = 0$$

This implies $f(x) = g(x)$ for all $x \in X$, which means $f = (id)(f) = g$. Therefore the graph of id is closed and by the closed graph theorem, id is continuous.

For any $x \in X$, $f \in \mathcal{H}$ and $g \in l^{\infty}(X)$, we use the reproducing property of $\mathcal{H}$:

$$\lim_{n\to\infty} |f_n(x) - f(x)|^2 = \lim_{n\to\infty} |\langle f_n, k(.,x)\rangle_{\mathcal{H}} - \langle f, k(.,x)\rangle_{\mathcal{H}}|^2$$

By the properties of inner product:

$$\lim_{n\to\infty} |\langle f_n - f, k(.,x)\rangle_{\mathcal{H}}|^2$$

Applying Cauchy-Schwarz inequality:

$$\leq \lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}}^2 \|k(.,x)\|_{\mathcal{H}}^2 = 0$$

Then $(id)\colon \mathcal{H} \to l^{\infty}(X)$ has a closed graph and hence is bounded. Since (id) is linear, it is also continuous. Finally, $\forall\, x \in X$ we have:

$$|k(x,x)| \leq \|k(.,x)\|_{\infty} \leq \|(id)\colon \mathcal{H} \to l^{\infty}(X)\| \|k(.,x)\|_{\mathcal{H}}$$
$$= \|(id)\colon \mathcal{H} \to l^{\infty}(X)\| \sqrt{k(x,x)}$$

This implies

$$\sqrt{k(x,x)} \leq \|(id)\colon \mathcal{H} \to l^{\infty}(X)\|$$

Taking the supremum on both sides:

$$\|k\|_{\infty} \leq \|(id)\colon \mathcal{H} \to l^{\infty}(X)\| \quad (3.3.2)$$

By (3.3.1) and (3.3.1), we conclude:

$$\|(id)\colon \mathcal{H} \to l^{\infty}(X)\| = \|k\|_{\infty}.$$

Lemma 3.3.2: (Measurability) Let (X, μ) be a measurable space and let k be a kernel on X with associated reproducing kernel Hilbert space $\mathcal{H}$. Then every function $f \in \mathcal{H}$ is μ −measurable if and only if the kernel function $k(.,x_0)\colon X \to \mathbb{R}$ is μ −measurable for all $x_0 \in X$.

Proof: ($\Longrightarrow$) Assume that every function $f \in \mathcal{H}$ is μ −measurable, by the definition of a reproducing kernel Hilbert space, we have that for any $x_0 \in X$ the function $K(.,x_0) \in \mathcal{H}$ and since every function in $\mathcal{H}$ is measurable then $k(.,x_0)$ is μ −measurable for all $x_0 \in X$.

($\Longleftarrow$): Assume the kernel function $k(.,x_0)$ is μ −measurable. Let $f \in \mathcal{H}$, by the reproducing property of RKHS, for any $x \in X$, we have

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

This means that $f(x)$ can be expressed as an inner product of f with the measurable function $k(.,x)$. Since $k(.,x)$ is μ −measurable for all $x \in X$, and the inner product is a

continuous linear functional, $f(x)$ is a pointwise limit of measurable functions. Then the function $f(x)$ can be approximated by:

$$f(x) = \lim_{n\to\infty} \langle f, k(.,x_n)\rangle_{\mathcal{H}}$$

Since $k(.,x_0)$, then each term $\langle f, k(.,x)\rangle_{\mathcal{H}}$ is measurable and $f(x)$ is $\mu-$measurable because the pointwise limit of measurable functions is also measurable.

Property 4 (Integrability) (Davies, 2022)**:** Let (X,μ) be a $\sigma-$finite measure space and $\mathcal{H}$ be a separable reproducing kernel Hilbert space RKHS over X with measurable $k: X \times X \to \mathbb{R}$. Suppose the kernel k satisfies the following integrability condition:

$$\|k\|_{L^P} = (\int_X k(x,x)^{P/2}\, d\mu(x))^{1/P} < \infty$$

For some $1 \leq P \leq \infty$.Then:

1- Every $f \in \mathcal{H}$ lies in $L^P(\mu)$ integrable function

2- The inclusion map $(id): \mathcal{H} \to L^P(\mu)$ is continuous

3- The adjoint operator $S_k: L^P \to \mathcal{H}$ is given by:

$$S_k g(x) = \int_X k(x,z) g(z) d\mu(z)$$

where $g \in L^{P^T}, x \in X$ and $\frac{1}{P} + \frac{1}{P^T} = 1$.

Proof: For $f \in \mathcal{H}$, and by $\|k(.x)\|_{\mathcal{H}} = \sqrt{k(x,x)}$ we get:

$$\|f\|_{L^P}^P = \int_X |f(x)|^P d\mu(x)$$

$$= \int_X |< f, k(.x) >|^P d\mu(x)$$

$$\leq \|f\|^P \int_X k(x,x)^{P/2} d\mu(x)$$

$$= \|f\|^P \|k\|^P$$

Then $(id): \mathcal{H} \to L^P(\mu)$ is continuous with $\|(id): \mathcal{H} \to L^P(\mu)\| \leq \|k\|_{L^P}$ where $f \in L^P(\mu)$.Now we prove the ad joint of the inclusion map exists, For $g \in L^{P^T}$ by the Schwartz-inequality we have:

$$\int_X |k(x,z)g(z)|d\mu(z) \leq \sqrt{k(x,x)} \int_X \sqrt{k(z,z)}|g(z)|d\mu(z)$$

By Holder inequality, we have:

$$|S_k g(x)| \leq \sqrt{k(x,x)}\|K\|_{L^P}\|g\|_{L^{P^T}}$$

The integrable is exist in $\mathbb{R}$ of $S_k(x)$ $\forall\, x \in X$. Then we have:

$$S_k g(x) = \int_X \langle \Phi(z), \Phi(x)\rangle_{\mathcal{H}} g(z)d\mu(z)$$

$$= \langle \int_X g(z)\Phi(z)d\mu(z), \Phi(x)\rangle_{\mathcal{H}}$$

Where $S_k g = \bar{g} = \int g(z)\Phi(z)d\mu(z) \in \mathcal{H}$.

3.3.2 Build an Reproducing kernel Hilbert space

Let k be a kernel and $\varphi: X \to \mathbb{R}^n$ reproducing kernel feature map is $\varphi(x) = k(.,x)$, and for any $x_i \in X$ and $\alpha_i \in \mathbb{R}$ feature vector space given by:

$$span(\{\varphi(x): x \in X\}) = f(.) = \sum_{i=1}^{n} \alpha_i k(.,x_i)$$

Then we define:

$$\langle f, g\rangle = \sum_i \sum_j \alpha_i \beta_j k(u_i, v_j)$$

Where $f = \sum_i \alpha_i k(.,u_i)$ and $g = \sum_i \beta_i k(.,v_i)$ such that:

$$\langle f, k(.,x)\rangle = \sum_i \alpha_i k(x,u_i) = f(x)$$

Where k has the reproducing property. Note that $\langle f, g\rangle$ is an inner product by the condition:

1- Symmetry:

$$\langle f, g\rangle = \sum_i \sum_j \alpha_i \beta_j k(u_i, v_j) = \sum_j \sum_i \beta_j \alpha_i k(v_j, u_i) = \langle g, f\rangle$$

2- Linearity:

$$\langle f, g\rangle = \sum_i \alpha_i g(u_i) = \sum_j \beta_j f(v_j)$$

3- Positive definiteness:

$$\langle f, f\rangle = \alpha^T k \alpha \geq 0$$

We have defined an inner product space. Complete it to give the Hilbert space

Now will introduce some examples of reproducing kernel Hilbert space.

Example 3.3.1: Let us determine the reproducing kernel Hilbert space associated with the polynomial kernel of degree 2:

$$k(x,z) = \langle x,z\rangle_{\mathbb{R}^n}^2 = \left(x^\top z\right)^2 \; \forall \; x,z \in \mathbb{R}^n$$

First, we will demonstrate the inner product, where the Frobenius norm for matrices $\mathbb{R}^{n\times n}$ denoted by F such that:

$$k(x,z) = trac(x^\top z x^\top z) = trac(z^\top x x^\top z)$$
$$= trac(x x^\top z z^\top) = \langle x x^\top, z z^\top\rangle_F$$

k is positive definite kernel. In the second step, we suggest a potential reproducing kernel Hilbert space where:

$$f(x) = \sum_i \alpha_i k(x_i,x) = \sum_i \alpha_i \langle x_i {x_i}^\top, x x^\top\rangle_F$$
$$\langle \sum_i \alpha_i x_i x_i^\top, x x^\top\rangle_F$$

Any symmetric matrix in $S \in \mathbb{R}^{n\times n}$ can be written as a linear combination of outer products of vectors, $\sum_i \alpha_i x_i x_i^\top$. Thus, we define the candidate RKHS, $\mathcal{H}$ as the space of all quadratic functions

$$f_S(x) = \langle S, x x^\top\rangle_F = x^\top S_x$$

Where $S \in S^{n\times n}$ the set of symmetric matrices in $\mathbb{R}^{n\times n}$and the inner product in $\mathcal{H}$ given by $\langle f_{S_1}, f_{S_2}\rangle_{\mathcal{H}} = \langle S_1, S_2\rangle_F$. The candidate space $\mathcal{H}$ is a Hilbert space because, it is isomorphic to $S^{n\times n}$ via the mapping $\varphi: S \to f_S$. Since $S \in S^{n\times n}$ is Euclidean space, $\mathcal{H}$ inherits its completeness and Hilbert space properties.

To confirm that $\mathcal{H}$ is the RKHS associated with kernel, we verify two conditions:

1- $\mathcal{H}$ contains all functions $k_x: y \to K(x,y) = \langle x x^\top, y y^\top\rangle_F$.

2- For all $f_S \in \mathcal{H}$ and $x \in \mathbb{R}^n$ the reproducing property hold:

$$f_S(x) = \langle S, x x^\top\rangle_F = \langle f_S, f_{x x^\top}\rangle_{\mathcal{H}} = \langle f_s, K_x\rangle_{\mathcal{H}}.$$

Thus, the space $\mathcal{H}$ of quadratic function $f_S(x) = x^\top S x$, equipped with the Frobenius inner product, it is the reproducing kernel Hilbert space associated with the polynomial kernel $k(x,z) = k\left(x^\top z\right)^2$.

Example 3.3.2: consider the space

$$L^2(\mathbb{N}_0) = \{\alpha_{n\in\mathbb{N}_0} \in \mathbb{C}^{\mathbb{N}_0}; \sum_{n=0}^{\infty} |{\alpha_n}^2| < \infty\}$$

The space $L^2(\mathbb{N}_0)$ is equipped with the following inner product

$$\langle (\alpha_n)_{n\in\mathbb{N}_0}, (\beta_m)_{m\in\mathbb{N}_0} \in \mathbb{N}_0 \rangle_{L^2(\mathbb{N}_0)} = \sum_{n=0}^{\infty} \alpha_n \overline{\beta_n}$$

Where $\overline{\beta_n}$ denotes the complex conjugate of β_n. This inner product induces the norm:

$$\left\|(\alpha_n)_{n\in\mathbb{N}_0}\right\|_{L^2(\mathbb{N}_0)} = \sqrt{\sum_{n=0}^{\infty} |{\alpha_n}^2|}$$

The space $L^2(\mathbb{N}_0)$is reproducing kernel Hilbert space (RKHS) because its elements can be viewed as functions that map the set of non-negative integers $\mathbb{N}_0$ to complex numbers $\mathbb{C}$. Specifically, each sequence $(\alpha_n)_{n\in\mathbb{N}_0}$ can be interpreted as function $f: \mathbb{N}_0 \to \mathbb{C}$ where $f(n) = \alpha_n$.The point evaluation functional satisfies:

$$|\alpha_m| \leq \left\|(\alpha_n)_{n\in\mathbb{N}_0}\right\|_{L^2(\mathbb{N}_0)}$$

The reproducing kernel function of this RKHS is simply defined by the following function:

$$k(m, l) = \langle \delta_l, \delta_m \rangle_{L^2(\mathbb{N}_0)}$$

Where δ_m the Dirac delta function at m, defined as:

$$\delta_m(n) = \begin{cases} 0 \; if \; m \neq 0 \\ 1 \; if \; m = 0 \end{cases}$$

The inner product of $\langle \delta_l, \delta_m \rangle_{L^2(\mathbb{N}_0)}$ given by:

$$\langle \delta_l, \delta_m \rangle_{L^2(\mathbb{N}_0)} = \sum_{n=0}^{\infty} \delta_l(n) \overline{\delta_m(n)}$$

Where $\delta_l(n)$ and $\delta_l(n)$ are both non-zero only when $n = l$ and $n = m$. Then

$$k(m, l) = \begin{cases} 1 & if \; m = l \\ 0 & otherwise \end{cases}$$

Thus, the reproducing kernel $k(m, l)$ is the delta function $k(m, l) = \delta_{ml}$.

Example 3.3.3: Consider the Hardy space

$$\mathcal{H}_2 = \{f: \Omega \to \mathbb{C}: f(z) = \sum_{n=0}^{\infty} \alpha_n z^n;\ \sum_{n=0}^{\infty} |\alpha_n|^2 < \infty\}$$

$\mathcal{H}_2$is the reproducing kernel Hilbert space RKHS on the unit disk $\Omega = \{z \in \mathbb{C}: 0 < |z| < 1\}$.

The Hardy space $\mathcal{H}_2$ consist of analytic functions $f: \Omega \to \mathbb{C}$ defined on unit disk, where each function f can be defined:

$$f(z) = \sum_{n=0}^{\infty} \alpha_n z^n$$

Where the coefficient $\sum_{n=1}^{\infty} |\alpha_n|^2 < \infty$.

The Hardy space $\mathcal{H}_2$ is a Hilbert space with the inner product:

$$\langle f, g \rangle = \sum_{n=0}^{\infty} \alpha_n \overline{\beta_n}$$

where $f(z) = \sum_{n=0}^{\infty} \alpha_n z^n$ and $g(z) = \sum_{n=0}^{\infty} \beta_n z^n$ from $\mathcal{H}_2$. The Hardy space $\mathcal{H}_2$ is isomorphic to the space of square-summable sequence $l^2(\mathbb{N}_0)$. The isomorphic $\phi: l^2(\mathbb{N}_0) \to \mathcal{H}_2$ is defined by:

$$\phi\big((\alpha_n)_{n \in \mathbb{N}_0}\big) = \sum_{n=0}^{\infty} \alpha_n z^n$$

Then

$$\langle f, g \rangle_{\mathcal{H}_2} = \langle \phi^{-1}(f), \phi^{-1}(g) \rangle_{l^2(\mathbb{N}_0)} = \sum_{n=0}^{\infty} \alpha_n \overline{\beta_n}$$

To show $\mathcal{H}_2$ is RKHS we need prove the point evaluations are continuous such that for any $\omega \in \Omega$, the evaluation functional $f \to f(\omega)$ is bounded and for $f(z) = \sum_{n=0}^{\infty} \alpha_n z^n \in \mathcal{H}_2$ and $\omega \in \Omega$, the value of f at ω is:

$$f(\omega) = \sum_{n=0}^{\infty} \alpha_n \omega^n$$

From the Cauchy-Schwarz inequality, it follows that

$$|f(\omega)| \leq (\sum_{n=0}^{\infty} |\alpha_n|^2)^{1/2} (\sum_{n=0}^{\infty} |\omega|^{2n})^{1/2}$$

Since $|\omega| < 1$, the geometric series convergence:

$$\sum_{n=0}^{\infty} |\omega|^{2n} = \frac{1}{1-|\omega|^2} < \infty$$

Thus

$$|f(\omega)| \leq \|f\|_{\mathcal{H}_2} . \frac{1}{\sqrt{1-|\omega|^2}}$$

Then $\rightarrow f(\omega)$ is bounded and continuous. And we need show the reproducing kernel $k(z,\omega)$ for $\mathcal{H}_2$ is given by:

$$k(z,\omega) = \frac{1}{1-\bar{\omega}z}$$

We can be expressed by:

$$k(z,\omega) = \sum_{n=0}^{\infty} \overline{\omega^n} z^n$$

For any $f \in \mathcal{H}_2$, the reproducing property holds:

$$f(\omega) = \langle f, k_\omega \rangle_{\mathcal{H}_2}$$

Where $k_\omega(z) = k(z,\omega) = \frac{1}{1-\bar{\omega}z}$. This is verified by:

$$\langle f, k_\omega \rangle_{\mathcal{H}_2} = \sum_{n=0}^{\infty} \overline{\omega^n} \alpha^n = f(\omega)$$

Theorem 3.3.2 (Fasshauer, 2011)**:** If the evaluation functional $\delta_x : \mathcal{F} \rightarrow \mathbb{R}$ is bounded for a Hilbert space $\mathcal{H}$, then $\mathcal{H}$ possess reproducing kernel.

Proof: Since δ_x is a bounded linear functional on $\mathcal{H}$. From the theorem of Riesz representation there exists a unique element $f_{\delta_x} \in \mathcal{H}$ such that:

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

Define $k(.,x) = f_{\delta_x}(.)$. By construction, $k(.,x)$ is an element of $\mathcal{H}$. For any $f \in \mathcal{H}$, we have:

$$\langle f(.), k(.,x) \rangle_{\mathcal{H}} = \delta_x f = f(x).$$

Then k is reproducing kernel.

Theorem 3.3.3 (Manton, J. H., & Amblard, P. O. 2015)**:** A function k is a reproducing kernel of a Hilbert space $\mathcal{H}$ on X if and only if for every $x \in X$, the evaluation linear functional $f \rightarrow f(x) \in \mathcal{H}$ is a bounded linear functional on $\mathcal{H}$.

Proof: Assume k is reproducing kernel for $\mathcal{H}$. By the reproducing property, for any $f \in \mathcal{H}$ and $x \in X$, we have:

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

Where $k_x(.) = k(., x)$. Using the Cauchy-Schwarz inequality, we obtain:

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

Since $\|k_x\| = \sqrt{\langle k_x, k_x \rangle}$, it follows that:

$$|f(x)| \leq \|f\| \sqrt{k(x, x)}$$

This inequality shows that the evaluation functional $f \rightarrow f(x)$ is a bounded for every $x \in X$. Therefore, the evaluation $f \rightarrow f(x) \in \mathcal{H}$ is bounded linear functional on $\mathcal{H}$. By Riesz representation theorem, and for all $x \in X$, there exists unique function $g_x \in \mathcal{H}$ such that:

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

Define $K_x = g_x$. Then for all $z \in X$, we have

$$K_x(z) = g_x(z)$$

This implies that $K(x, z) = K_x(z) = g_x(z)$. And thus, k satisfies the reproducing property:

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

Therefore, k is a reproducing kernel for $\mathcal{H}$.

The following theorems establishes a fundamental link between positive definite kernels and reproducing kernel Hilbert spaces (RKHS). Specifically, a function k is a reproducing kernel if and only if it is positive definite and the evaluation functionals are bounded. This connection is central to the theory of RKHS and their applications in machine learning, functional analysis, and other fields.

Theorem 3.3.4 (Fasshaure, 2014) (Fukumizu, 2010)**:** If $\mathcal{H}_k$ is a reproducing kernel of Reproducing Kernel Hilbert Space with kernel $k: \varphi \times \varphi \rightarrow \mathbb{R}$ then k is positive definite kernel on X.

Proof: $\forall\, x_1, \ldots, x_n \in X$ and arbitrary point $\alpha \in \mathbb{R}^n$ we have:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(x_i,x_j)=\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\langle k(.,x_i),k(.,x_j)\rangle$$

This can be rewritten as:

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

Theorem 3.3.5 (Okutmusture, 2005)**:** If a Hilbert space $\mathcal{H}$ of functions defined on a set X has a reproducing kernel, then the reproducing kernel $k(x,.)$ is uniquely determined by the Hilbert space $\mathcal{H}$.

Proof: Assume that k_1 and k_2 are reproducing kernel for the same space $\mathcal{H}$. This mean both k_1 and k_2 satisfies the reproducing property:

$$f(x)=\langle f,k_1(x,.)\rangle \text{ and } f(x)=\langle f,k_2(x,.)\rangle$$

For every $x\in X$, consider the function $k_1(x,.)-k_2(x,.)$. Since $k_1(x,.)$ and $k_2(x,.)$ belong to $\mathcal{H}$, their difference $k_1(x,.)-k_2(x,.)$ belong to $\mathcal{H}$. Then the norm of difference given by:

$$\|k_1(x,.)-k_2(x,.)\|_{\mathcal{H}}^2=\langle k_1(x,.)-k_2(x,.),k_1(x,.)-k_2(x,.)\rangle$$

Expand the inner product using linearity:

$$=\langle k_1(x,.)-k_2(x,.),k_1(x,.)\rangle-\langle k_1(x,.)-k_2(x,.),k_2(x,.)\rangle$$

This simplifies to:

$$=\big(k_1(x,x)-k_2(x,x)\big)-(k_1(x,x)-k_2(x,x))=0.$$

Thus, we conclude that $k_1(x,.)=k_2(x,.)$ for all $x\in X$, proving that the reproducing kernel is uniquely determined by the Hilbert space $\mathcal{H}$.

Theorem 3.3.6 (Manton, J. H., et al. 2015)**:** Let X be a set, and $k\colon X\times X\to\mathbb{R}$ be a positive definite kernel. Then there exists unique Hilbert space $\mathcal{H}(k)$ such that the following properties hold:

(a) Elements of $\mathcal{H}(k)$ are real functions defined on X.

(b) The set $\{k_x(z)\colon x\in X\}\subset\mathcal{H}(k)$ where denoting $k_x(z)=k(x,z)$.

(c) For all $x\in X$ and $\varphi\in\mathcal{H}(k)$, $\langle\varphi,k_x\rangle=\varphi(x)$

Proof: Consider $\mathcal{H}_0$ be a linear span of the family $\{k_x : x \in X\}$. Define a bilinear form s on $\mathcal{H}_0$, for any $\varphi, \Phi \in \mathcal{H}_0$ as follows:

$$\varphi = \sum_{i=1}^{n} \alpha_i k_{x_i}, \Phi = \sum_{j=1}^{m} \beta_j K_{y_j}$$

Where $\alpha_i, \beta_j \in \mathbb{R}$ and $x_i, y_j \in X$. The bilinear form s is defined by:

$$s(\varphi, \Phi) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j)$$

This definition is independent of the specific representations of φ and Φ. The form s is a symmetric and positive definite, satisfying:

$$s(\varphi, K_x) = \varphi(x), \quad \forall \varphi \in \mathcal{H}_0, \quad x \in X.$$

Using the Cauchy-Schwarz inequality and the positive definiteness of s. It follows that if $s(\varphi, \Phi) = 0$ for all $\varphi = 0$.Therefore, $(\varphi, \Phi) = s(\varphi, \Phi)$ is the inner product in $\mathcal{H}_0$.

Let $\mathcal{H}$ be the completion of $\mathcal{H}_0$, with respect to the norm induced by its inner product $\langle ., . \rangle_{\mathcal{H}_0}$. This completion ensures that $\mathcal{H}$ is complete Hilbert space.

Define $\mathcal{H}(K)$ be the family of real-valued functions on X given by:

$$\varphi(x) = \langle h, k_x \rangle_{\mathcal{H}}$$

Where $h \in \mathcal{H}$ and $\langle ., . \rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$. The inner product in $\mathcal{H}(K)$ is defined as:

$$\langle \varphi_1, \varphi_2 \rangle_{\mathcal{H}(k)} = \langle h_1, h_2 \rangle_{\mathcal{H}}$$

This definition is valid because the set of finite linear combinations of $\{k_x : x \in X\}$ is dense in $\mathcal{H}$. Moreover, the space $\mathcal{H}(k)$ is complete. The reproducing property follows from the definition of $\mathcal{H}(k)$, for any $\varphi \in \mathcal{H}(k)$ and $x \in X$:

$$\langle \varphi, k_x \rangle_{\mathcal{H}(K)} = \langle h, k_x \rangle_{\mathcal{H}} = \varphi(x)$$

The Hilbert space $\mathcal{H}(k)$ is unique because it is determined entirely by the kernel k and the reproducing property.

Remark 3.3.2: For any complex valued positive definite kernel k, there exists a unique complex Hilbert space $\mathcal{H}$ of complex-valued functions that satisfies both the reproducing property and the spanning property. This space is known as the RKHS associated with K.

3.3.3 The Computational Cost of Learning in an RKHS

Initially, working in a high-dimensional feature space may seem challenging due to the complexity of managing large-dimensional vectors. However, the following theorem offers an insightful perspective. It shows that even when a problem is situated in a very high-dimensional space, the solution can often be located within a lower-dimensional subspace.Two key theoretical results form the foundation of a set of powerful algorithms for data analysis that utilize positive definite kernels, collectively referred to as kernel methods the kernel trick and representation theory.

1- **The Kernel trick**: This concept is based on the property that a positive definite kernel $k(x, z)$ can be expressed as an inner product in a high-dimensional (and potentially infinite-dimensional) feature space. More specifically, there exists a feature map φ such that: $$k(x, z) = \langle \varphi(x), \varphi(z) \rangle_{\mathcal{H}}$$

Where $\mathcal{H}$ is feature space.

Example 3.3.3.1: The polynomial kernel of degree 3 $k(x, z) = (x^T z)^3$ implicitly maps the input vectors x and z into a higher-dimensional feature space such that:

For $x = (x_1, x_2)$ and $z = (z_1, z_2)$ and feature map $\varphi(x) = (x_1^3, x_1^2 x_2, x_1 x_2^2, x_2^3)$.

$$x^T z = x_1 z_1 + x_2 z_2$$

And

$$(x^T z)^3 = (x_1 z_1)^3 + 3(x_1 z_1)^2 x_2 z_2 + 3(x_1 z_1)(x_2 z_2)^2 + (x_2 z_2)^3$$

Thus, the kernel $K(x, z)$ computes the inner product in this feature space:

$$k(x, z) = \langle \varphi(x), \varphi(z) \rangle = \langle (x_1^3, \sqrt{3} x_1^2 x_2, \sqrt{3} x_1 x_2^2, x_2^3)(z_1^3, \sqrt{3} z_1^2 z_2, \sqrt{3} z_1 z_2^2, z_1^3) \rangle$$

Let $x = (1,2), z = (3,4)$, the inner product as follows:

$$x^T z = x_1 z_1 + x_2 z_2 = (1)(3) + (2)(4) = 11$$

The kernel function is calculated as:

$$k(x, z) = (x^T z)^3 = (11)^3 = 1331.$$

The kernel trick enables us to compute $k(x,z) = (x^T z)^3$ directly, without needing to explicitly determine the high-dimensional feature map $\varphi(x)$ or $\varphi(z)$. This approach is computationally efficient, particularly when dealing with very large or infinite feature spaces.

Example 3.3.3.2: Let $\varphi: X \to \mathcal{H}$ is be a feature map into a reproducing kernel Hilbert space and $k: X \times X \to \mathbb{R}$ be associated kernel, such that the Euclidean distance between the images of two points x and z in the feature space $\mathcal{H}$ can be expressed purely in terms of the kernel:

$$Distance(x,z) = \|\varphi(x) - \varphi(z)\| = \sqrt{\langle \varphi(x) - \varphi(z), \varphi(x) - \varphi(z) \rangle}$$

Expand the inner product using the properties of kernel

$$\|\varphi(x) - \varphi(z)\|^2 = \langle \varphi(x), \varphi(x) \rangle - 2\langle \varphi(x), \varphi(z) \rangle + \langle \varphi(z), \varphi(z) \rangle$$

Using the kernel representation

$$\|\varphi(x) - \varphi(z)\|^2 = k(x,x) - 2k(x,z) + k(z,z)$$

Compute the distance

$$Distance(x,z) = \sqrt{k(x,x) - 2k(x,z) + k(z,z)}$$

The distance depends only on kernel evaluations, avoiding direct computation of $\varphi(x)$ or $\varphi(z)$.

Example 3.3.3.3: Gaussian kernel (RBF) to compute the distance such that:

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

Suppose that $k(x,x) = k(y,y) = 1$, then

$$Distance(x,z) = \sqrt{k(x,x) - 2k(x,z) + k(z,z)}$$

$$= \sqrt{1 - 2k(x,z) + 1} = \sqrt{2 - 2k(x,z)}$$

Using the definition of the Gaussian kernel:

$$= \sqrt{2 - 2\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)} = \sqrt{2}\sqrt{1 - e\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)}$$

When $k(x,z) = 0$ this implies $\|\varphi(x) - \varphi(z)\| = 0$. When $k(x,y) = 1$ this implies $\|\varphi(x) - \varphi(z)\| = \sqrt{2}$. The cosine kernel measures the cosine of the angle between x and z is defined as:

$$k(x,y) = \frac{x}{\|x\|\|y\|}, \|x\| = \|y\| = 1$$

The RBF kernel Decays exponentially with the squared Euclidean distance in the input space and Bandwidth γ controls sensitivity to input-space distances. This approach is computationally efficient and eliminates the need to operate directly in the high-dimensional feature space. It is widely utilized in kernel-based algorithms, such as Support Vector Machines (SVMs), kernel PCA, and various other kernelized methods. [See Chapter 4]

1- **The Representer Theorem**: This theorem is grounded in specific properties of the regularization functional defined by the Reproducing Kernel Hilbert Space (RKHS) norm. It guarantees that solutions to optimization problems can be represented as linear combinations of kernel evaluations on the training data. This theory states that:

Theorem 3.3.3.1 (Representer theorem1971) (Yang, 2016)**:** Let X be a non-empty set equipped with a positive definite kernel $k: X \times X \to \mathbb{R}$ and let $\mathcal{H}$ be the corresponding reproducing kernel Hilbert space (RKHS). Consider $x_1, x_2, \dots, x_n$ be a finite set of X and let the function $\Psi: \mathbb{R}^{n+1} \to \mathbb{R}$ that is strictly increasing in its last argument. For any $f: X \to \mathbb{R}^n$ in $\mathcal{H}$, consider the optimization problem:

$$min_{f \epsilon \mathcal{H}} \Psi(w^T f(x_1), \dots, w^T f(x_n), w^T w)$$

Where $w = \sum_{i=1}^{n} \alpha_i f(x_i)$; $\alpha \in \mathbb{R}^n, w \in \mathbb{R}^n$

Proof: Let $\mathcal{F}_D$ be the subspace spanned by $\{f(x_1), f(x_2), \dots, f(x_n)\}$ such that $\mathcal{F}_D = \{\sum_{i=1}^{n} \alpha_i f(x_i);\ \alpha \in \mathbb{R}^n\}$. For Any vector $w \in \mathbb{R}^n$ can be decomposed into two orthogonal components:

$$w = w_D + w_\perp$$

Where $w_D \in \mathcal{F}_D and\ w_\perp \in \mathcal{F}_D^\perp$. For each $i = 1,2, \dots, n$, we have

$$w^T f(x_i) = w_D^T f(x_i) + w_\perp^T f(x_i)\ \forall i$$

Since $w_\perp \in \mathcal{F}_D^\perp$, its orthogonal to all $f(x_i)$, so

$$w_\perp^T f(x_i) = 0$$

Thus

$$w^T f(x_i) = w_D^T f(x_i)$$

By the Pythagoras theorem for orthogonal vector:

$$w^T w = w_D^T w_D + w_\perp^T w_\perp$$

Since $w_\perp^T w_\perp \geq 0$, we have

$$w^T w \geq w_D^T w_D$$

The function Ψ depends on $w^T f(x_1), \ldots, w^T f(x_n)$ and $w^T w$. Then

$$\Psi(w^T f(x_1), \ldots, w^T f(x_n), w^T w) = \Psi(w_D^T f(x_1), \ldots, w_D^T f(x_n), w_D^T w_D + w_\perp^T w_\perp)$$

$$\geq \Psi(w_D^T f(x_1), \ldots, w_D^T f(x_n), w_D^T w_D)$$

Equality holds if and only if $w_\perp = 0$. The infimum of Ψ when $w_\perp = 0$.Therefore

$$inf_{w \in \Re^d} \Psi(w^T f(x_1), \ldots, w^T f(x_n), w^T w) = inf_{w \in \mathcal{F}_D} \Psi(w^T f(x_1), \ldots, w^T f(x_n), w^T w).$$

A foundational concept in RKHS-based statistics is the embedding of probability measures into an RKHS, which enables the representation of distributions as elements in a Hilbert space. This embedding leverages the rich geometric and analytic structure of RKHS to perform statistical inference on probability measures directly.

3.3.3.1 The Role of the Representer Theorem in Efficient Computation

The kernel trick leverages the representer theorem to enable efficient computations in RKHS by expressing solutions as linear combinations of kernel evaluations. In the context of Empirical Risk Minimization, the representer theorem allows to reformulate the problem by:

$$min \frac{1}{m} \sum_i \varphi(\langle f(x_i), w \rangle, y_i) \text{ , where } \|w\| \leq B.$$

By the representer theorem, the optimal w has the form:

$$w = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$$

$$\|w\|^2 = \sum_{i,j} \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle$$

$$= \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j)$$

$$= \alpha^T K \alpha$$

Substituting this into the problem yields a finite-dimensional optimization by solving this problem, we can arrive at an efficient solution:

$$min \frac{1}{m} \sum_i \varphi\left(\sum_j \alpha_j k(x_j, x_i), y_i\right), \quad \alpha^T K \alpha \leq B.$$

K is the kernel matrix. This reformulation leverages the kernel trick to avoid explicit computation in high-dimensional.

3.3.4 Statistics in Reproducing kernel Hilbert space

Let P be a family of probability measures. A bounded linear operator $T: P \to H$ maps each $P \in P$ to an element in the RKHS H such that:

$$T: P \in P \to T(P) = \int f dP = \mathbb{E}(f)$$

From the Riesz's theorem there exists a unique mean element $\mu_P \in H$ such that

$$\langle f, \mu_P \rangle_{\mathcal{H}} = \mathbb{E}(f)$$

Where μ_P is mean element in RKHS where P on RKHS contains the key information about RKHS. If the positive define kernel exhibit a good behavior, then μ_P uniquely determines probability measure. This mean element μ_P encapsulates the central information of P in H. For a characteristic kernel (e.g., Gaussian, Laplacian), μ_P uniquely determines the probability measure P, enabling nonparametric inference directly in the RKHS. This property allows probability measures to be analyzed and compared entirely through their RKHS representations, bypassing density estimation or parametric assumptions.

Empirical average of the feature map $\frac{1}{n} \sum_{i=1} k(x_i, .)$ is estimator for μ_P such that:

$$\left\| \hat{\mu_{\mathbb{P}}} - \mu_{\mathbb{P}} \right\|_2 = O_p\left(\frac{1}{\sqrt{n}}\right),$$

The $\sqrt{n}(\hat{\mu_{\mathbb{P}}} - \mu_{\mathbb{P}}) \to G(.)$ as $n \to \infty$ converges weakly to a Gaussian process $G(.)$ in H, with covariance function governed by the kernel. Where

$$Cov\big(G(t), G(s)\big) = \int \langle G - \mu_{\mathbb{P}}, k(t,.)\rangle \otimes \langle G - \mu_{\mathbb{P}}, k(s,.)\rangle d\mathbb{P}.$$

Given that variance and covariance provide information about random variables on Euclidean space, there is a corresponding cross-variance-covariance operator defined as the tensor product of two distinct reproducing kernel Hilbert spaces (RKHSs). Then Covariance operator is defined as:

$$\sum_{yx} : \mathcal{H}_x \to \mathcal{H}_y \ \ or \sum_{yx} = \mathbb{E}\big(\Phi(y) \otimes \Phi(x)\big) - \mu_y \otimes \mu_x$$

Where $\langle g, \sum_{yx} f\rangle_{\mathcal{H}_y} = Cov(g, f)$.

By the weak law of large numbers, the empirical estimator mentioned above is guaranteed to converge to the true mean embedding. Berlant and Thomas-Agnan (2004) demonstrate that this convergence occurs at a rate of $\varphi(n^{-\frac{1}{2}})$. (Yang, 2016)

3.4 Mercer theorem

Mercer's theorem is a cornerstone of learning theory with kernels, providing a theoretical foundation for understanding Reproducing Kernel Hilbert Spaces (RKHS). These spaces, which are a significant class of function spaces, are widely used in various fields of analysis and machine learning applications, particularly in kernel-based methods. When the domain of definition is compact, Mercer's theorem characterizes RKHS as the image of the square root of an integral operator, offering a powerful framework for the analysis and implementation of kernel methods. (Pipkin, 1991)

Theorem 3.4.1: Mercer's theorem (Bejarano, 2012)**:** Consider X be a compact set, let symmetric function $K: X \to \mathbb{R}$ be a continuous and positive definite kernel We say k satisfies Mercer's condition if and only if:

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

Where $\lambda_i \geq 0$ and $\sum_{i=1}^{\infty} \lambda_i < \infty$, and Φ_i are the orthonormal eigenfunctions of the corresponding integral operator. Therefore, the convergence is absolute and uniform.

Proof: The integral operator $\mathcal{T}: L^2(X) \to L^2(X)$ defined by

$$\mathcal{T}_k f(x) = \int k(x, z) f(z) dz$$

we can express $\mathcal{T}_k$ in terms of its eigenvalues and eigenfunctions:

$$\mathcal{T}_k f = \sum_{i=1}^{\infty} \lambda_i \langle f, \Phi_i \rangle \Phi_i$$

Now, we determine that the series $\sum_{i=1}^{\infty} \lambda_i \Phi_i(x) \Phi_i(z)$ is converges uniformly. Let a finite submatrix on the points $x_1, \dots x_n$ which are not positive semi-definite were

$$\sum_{i,j=1}^{n} K(x_i, x_j) \lambda_i \lambda_j = \varepsilon < 0$$

and define the function

$$f_\sigma(x) = \sum_{i=1}^{n} \lambda_i \frac{1}{(2\pi\sigma)^{d/2}} \exp\left(\frac{\|x - x_i\|^2}{2\sigma^2}\right) \in L_2(X)$$

d is the dimension of the X, where

$$\lim_{\sigma \to 0} \int K(x, z) f_\sigma(x) f_\sigma(z) dx dz = \varepsilon \quad (3.4.1)$$

The integral $(3.4.1) < 0$ for some $\sigma > 0$. This contradicts the positivity of the integral operator. Now, let we consider an orthonormal basis $\{\Phi_i\}_i$ then the RKHS for $k(x, .)$

$$k(x, z) = \sum_{i=1}^{\infty} \langle k(x, .), \Phi_i(x) \rangle \Phi_i(z)$$

$$= \sum_{i=1}^{\infty} \Phi_i(x) \Phi_i(z)$$

The series $\sum_{i=1}^{\infty} \|\Phi_i\|_{L_2(X)}^2$ is convergent because

$$\int k(x, x) dx = \lim_{n \to \infty} \int \sum_{i=1}^{n} \Phi_i(x) \Phi_i(x) dx < \infty$$

$$= \lim_{n \to \infty} \sum_{i=1}^{n} \int \Phi_i(x) \Phi_i(x) dx < \infty$$

$$= \lim_{n\to\infty} \sum_{i=1}^{n} \|\Phi_i\|^2_{L_2(X)} < \infty$$

Example 3.4.1: Consider a translation-invariant kernel function $k(x, z) = k(x - z)$, where the inner product between two inputs remains unchanged under simultaneous translation. In the one-dimensional case, assume k is defined on $[0,2\pi]$ and can be generalized to a symmetric, continuous and periodic function defined on $\mathbb{R}$. The function admits a uniformly convergent Fourier series:

$$k(u) = a_0 + \sum_{n=1}^{\infty} a_n \cos(nu)$$

where symmetry ensures only cosine terms appear. Expanding $k(x - z)$ using the trigonometric identity

$$cos(n(x - z)) = cos(nx)cos(nz) + sin(nx)sin(nz)$$

We rewrite the kernel as:

$$k(x - z) = a_0 + \sum_{n=1}^{\infty} a_n\, cos(nx)cos(nz) + a_0 + \sum_{n=1}^{\infty} a_n sin(nx)sin(nz)$$

This expression reveals $k(x, z)$ as an inner product with in the feature space constructing using orthogonal basis:

$$\{\Phi_i\}_{i=0}^{\infty} = \ (1, \sin(x)\,, \cos(x)\,, \sin(2x)\,, \cos(2x)\,, \dots, \sin(\mathrm{n}x)\,, \cos(\mathrm{n}x)\,, \dots)$$

Here 1, $\cos(nx)$, $\sin(\mathrm{n}x)$ are orthogonal on $[0,2\pi]$. Normalizing these functions yields Mercer features, ensuring the kernel satisfies Mercer's conditions.

The criteria outlined in Mercer's theorem are referred to as the Mercer conditions. These conditions establish the foundational requirements for a kernel function to be classified as a Mercer kernel. Also, there is a connection between Mercer kernels and reproducing kernels in Hilbert spaces. Reproducing kernels represent a particular type of positive definite kernel that possesses the reproducing property, allowing function evaluations to be expressed as

inner products within the Hilbert space. This relationship between Mercer kernels and reproducing kernels highlights their theoretical importance in functional analysis and their practical utility in fields such as machine learning, where they enable efficient computation and representation of complex functions.

3.4.1 The Relation Between Reproducing Kernel Hilbert Space (RKHS) and Mercer Theorem

Reproducing Kernel Hilbert Spaces (RKHS) and Mercer's Theorem are interconnected concepts in functional analysis and machine learning. Below is an overview of their relationship:

Let (X, d) be a metric space and $k: X \times X \to \mathbb{R}$ be a continuous and symmetric function. k is called a Mercer kernel if it is positive semi-definite, meaning that for any finite set of points $x_1, x_2, \dots, x_n \in X$ and any real numbers $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{R}$, the following condition holds:

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

The reproducing kernel Hilbert space associated with k, denoted $\mathcal{H}_k$, is constructed as the completion of the span of the function $\{k_x: k(x, .): x \in X\}$ under the inner product defined as follows: for $f = \sum_{i=1}^{n} \alpha_i k_{x_i}$ and $g = \sum_{j=1}^{m} \beta_j k_{z_j}$, the inner product is:

$$\langle f, g \rangle_k = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j k(x_i, z_j)$$

A key feature of $\mathcal{H}_k$ is the reproducing kernel property, which ensure that for any $f \in \mathcal{H}_K$ and any, $x \in X$, the evaluation of f at x can be expressed as:

$$f(x) = \langle f, k_x \rangle_k \quad \forall\, f \in \mathcal{H}_k, x \in X$$

This property highlights those functions $\mathcal{H}_k$ can be evaluated at any point x through the inner product with the kernel function k_x. Moreover, $\mathcal{H}_k$ is composed of continuous functions on X, making it a subset of $C(X)$, the space of all continuous functions on X. This relation $\mathcal{H}_K \subset C(X)$ highlights the importance of Mercer kernels in the study of RKHS,

and their applications in areas such as functional analysis and machine learning. (Sun, 2005)

Theorem 3.4.2 (Lanckriet et al. 2004)**:** Let $k: X \times X \to \mathbb{R}$ be a Mercer kernel, where $X \subset \mathbb{R}$ is bounded set. Then, there exists unique Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ of functions defined on X such that the Mercer Kernel k is the reproducing kernel for $\mathcal{H}$.

Proof: Since k is mercer kernel, it is positive definite. To construct a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of the form:

$$f(.) = \sum_{i=1}^{\infty} \alpha_i \Phi_i(.)$$

Where $\{\Phi_i\}_{i=1}^{\infty}$ and $\{\lambda_i\}_{i=1}^{\infty} > 0$ are the eigenfunctions and eigenvalues of the integral operator associated with the Mercer kernel. The inner product in $\mathcal{H}$ is defined as:

$$\langle f(.), g(.)\rangle_{\mathcal{H}} = \langle \sum_{i=1}^{\infty} \alpha_i \Phi_i(.), \sum_{i=1}^{\infty} d_i \Phi_i(.)\rangle = \sum_{i=1}^{\infty} \frac{\alpha_i d_i}{\lambda_i}$$

This Hilbert space $\mathcal{H}$ is an RKHS with k as its reproducing kernel. Consider the reproducing property:

$$\langle f, k(x,.)\rangle_{\mathcal{H}} = \langle \sum_{i=1}^{\infty} \alpha_i \Phi_i(.), \sum_{j=1}^{\infty} \lambda_j \Phi_j(x)\Phi_j(.)\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \alpha_i \lambda_j \Phi_j(x)\langle \Phi_i(.), \Phi_j(.)\rangle_{\mathcal{H}}$$

By orthogonality in $\mathcal{H}$, $\langle \Phi_i(.), \Phi_j(.)\rangle_{\mathcal{H}} = \frac{\delta_{ij}}{\lambda_i}$

$$= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \alpha_i \lambda_j \Phi_j(x) \frac{\delta_{ij}}{\lambda_i} =$$

$$= \sum_{i=1}^{\infty} \frac{\alpha_i \lambda_i \Phi_i(x)}{\lambda_i} = \sum_{i=1}^{\infty} \alpha_i \Phi_i(x) = f(x).$$

This demonstrates that the k is indeed a reproducing kernel for $\mathcal{H}$, and vice versa. Therefore, every RKHS with a continuous reproducing kernel on a compact domain corresponds to a Mercer kernel. Thus, the concepts of a Mercer kernel, an RKHS, and a reproducing kernel are equivalent:

RKHS $\mathcal{H}$ $\Leftrightarrow$Mercer kernel $\Leftrightarrow$ Reproducing kernel

Thus, every Mercer kernel uniquely defined an RKHS, and every RKHS with a continuous reproducing kernel on a compact domain is characterized by a Mercer kernel. Furthermore, the norm in the RKHS $\mathcal{H}$ is given by:

$$\|f\|_k^2 = \sum_{i=1}^{\infty} \frac{{\alpha_i}^2}{\lambda_i}$$

This notion of $\|.\|_k$ emphasizes that the norm is entirely dependent on the kernel of the reproducing kernel Hilbert space (RKHS).

Theorem 3.4.3 (Aronzajn 1950) (Rachev et al. 2013)**:** For any positive definite kernel $k(z,x)$ defined on a set X, there exists a unique Hilbert space $\mathcal{H}_k$ of function on X, that admits the reproducing kernel $k(z,x)$.

Proof: Define $\mathcal{H}_0$ as the space of all functions f on X and there exists $x_1, x_2, \dots, x_n \in X$ and $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{C}$, which expressible as finite linear combinations of kernel:

$$f(.z) = \sum_{i=1}^{n} \alpha_i k(z, x_i)$$

For all $z \in X$ and $\beta_1, \beta_2, \dots, \beta_n \in \mathbb{C}$. Consider for another function $g \in \mathcal{H}_0$:

$$g(.) = \sum_{j=1}^{m} \beta_j K\big(., z_j\big), \;\; g \in \mathcal{H}_0$$

Define the inner product of the functions f and g from $\mathcal{H}_0$ by:

$$\langle f, g\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j \langle k(., x_i), k\big(., z_j\big)\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \overline{\beta}_j k\big(z_j, x_i\big)$$

This definition leverages the positive definiteness of k, ensuring the inner product is consistent and satisfies:

$$\langle k(., x), k(., z)\rangle_{\mathcal{H}_0} = k(x, z)$$

Therefore, satisfies the reproducing property

$$\langle f, K(., x)\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \alpha_i \langle k(., x_i), k(., x)\rangle = \sum_{i=1}^{n} \alpha_i k(x, x_i) = f(x)$$

For all $x \in X$, that is $\mathcal{H}_0$ possess the reproducing property. Since k is positive definite it implies that $\langle f, f\rangle_{\mathcal{H}_0} \geq 0$ for all $f \in \mathcal{H}_0$, and $\langle f, f\rangle_{\mathcal{H}_0} = 0$ implies $\|f\| = 0$

The reproducing property holds in $\mathcal{H}_0$, for any $f \in \mathcal{H}_0$ and $x \in X$, we have

$$f(x) = \langle f, k_x \rangle_{\mathcal{H}_0}$$

In other words, we have the Schwarz Inequality such that, for all $x \in X$:

$$|f(x)| \leq \|f\| \|K(.,x)\| = 0$$

Where $k_x = k(.,x)$. Thus $\mathcal{H}_0$ is a pre-Hilbert space.

Now complete $\mathcal{H}_0$ to a Hilbert space $\mathcal{H}$ by including all limits of Cauhy sequence in $\mathcal{H}_0$. Consider Cauchy sequence f_n in $\mathcal{H}_0$ such that:

$$|f_m(x) - f_n(x)| = \left|\langle f_m, K_x \rangle_{\mathcal{H}_0} - \langle f_n, K_x \rangle_{\mathcal{H}_0}\right|$$

$$= \left|\langle f_m - f_n, K_x \rangle_{\mathcal{H}_0}\right|$$

$$\leq \|f_m - f_n\|_{\mathcal{H}_0} \sqrt{K(x,x)}$$

For all $x \in X$ and there exists the function $f: X \to \mathbb{C}$ such that:

$$\lim_{n\to\infty} f_n(x) = f(x)$$

And, we have:

$$\|f\|_{\mathcal{H}} = \lim_{n\to\infty} \|f_n\|_{\mathcal{H}_0} \quad (3.4.2)$$

From (3.4.2) we see that the representation of $\mathcal{H}$ as a space of functions on X. Therefore, k has the reproducing property of $\mathcal{H}$. To prove this, let $f \in \mathcal{H}$ and $f_n \subset \mathcal{H}$ such that $f_n \to f$ as $n \to \infty$. Then for all $x \in X$, we have:

$$f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \langle f_n, k_x \rangle_{\mathcal{H}_0}$$

$$= \langle \lim_{n\to\infty} f_n(x), k_x \rangle_{\mathcal{H}_0} = \langle f_n, k_x \rangle_{\mathcal{H}_0}$$

We will now prove that the reproducing kernel Hilbert space (RKHS) associated with the kernel k is unique.

Suppose $\mathcal{H}_1$ is another Hilbert space with the same reproducing kernel k. By definition, for any $x \in X, k_x = k(.,x)$ belong to $\mathcal{H}_1$. This implies that the space $\mathcal{H}_0$ the pre-Hilbert space and $\mathcal{H}_0 \subseteq \mathcal{H}_1$. For any $f, g \in \mathcal{H}_0$, the reproducing property in $\mathcal{H}_0$ and $\mathcal{H}_1$ given by:

$$\langle f, g \rangle_{\mathcal{H}_0} = \langle f, g \rangle_{\mathcal{H}_1} \quad (3.4.3)$$

If $f \in \mathcal{H}_1$ satisfies $\langle f, k_x \rangle_{\mathcal{H}_1} = f(x) = 0$ for all $x \in X$. This shows that the family $\{k_x : x \in X\}$ is total in $\mathcal{H}_1$, meaning it spans a dense subspace of $\mathcal{H}_1$.

Since $\mathcal{H}_0$ is dense in $\mathcal{H}_1$, for any $f \in \mathcal{H}_1$, there exist a sequence $\{f_n\} \subset \mathcal{H}_0$ such that $f_n \to f$ in $\mathcal{H}_1$ as $n \to \infty$. From $\mathcal{H}_0 \subseteq \mathcal{H}_1$, it follows that $\mathcal{H} \subseteq \mathcal{H}_1$. Conversely, by the construction of $\mathcal{H}$, $\mathcal{H}_1 \subseteq \mathcal{H}$. Therefore, $\mathcal{H}_1 = \mathcal{H}$.

To show that the inner products and norms in $\mathcal{H}$ and $\mathcal{H}_1$ are equivalent, consider any two functions f and g belonging to $\mathcal{H}_1$, and let $\{f_n\}$ and $\{g_n\}$ be Cauchy sequences in $\mathcal{H}_0$ that converge to f and g, respectively. Then

$$\langle f, g\rangle_{\mathcal{H}_1} = \lim_{n\to\infty}\langle f_n, g_n\rangle_{\mathcal{H}_1} = \lim_{n\to\infty}\langle f_n, g_n\rangle_{\mathcal{H}_0} = \langle f, g\rangle_{\mathcal{H}}.$$

Theorem 3.4.4 (Aronszajn 1950; Kolmogorov 1942) (Rachev et al. 2013)(Yang, 2016)**:** A function $k: X^2 \to \mathbb{R}$ is a positive definite kernel if and only if there exist a real Hilbert space $\mathcal{H}$ and a family $\{b_x : x \in X\} \subseteq \mathcal{H}$ such that $K(x, z) = \langle b_x, b_z\rangle$ for all $x, z \in X$.

Proof: ($\Longrightarrow$): Assume k is positive definite kernel, then we can choose $\mathcal{H}$ as the Hilbert space with reproducing kernel k and set $b_x = k_x = k(.,x)$. The inner product between b_x and b_z in $\mathcal{H}$ is

$$\langle b_{\boldsymbol{x}}, b_{\boldsymbol{z}}\rangle_{\mathcal{H}} = \langle k(.,x), k(.,z)\rangle_{\mathcal{H}} = k(x,z)$$

Thus, $k(x,z) = \langle b_x, b_z\rangle$.

($\Longleftarrow$): Let $x_1, \dots, x_n \in X$ and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and k is symmetric then we have:

$$\sum_{i,j=1}^{n} k(x_i x_j)\alpha_i\alpha_j = \sum_{i,j=1}^{n} \langle b_{x_i}, b_{x_j}\rangle\, \alpha_i\alpha_j = \left\|\sum_{i=1}^{n} \alpha_i b_{x_i}\right\|_{\boldsymbol{\mathcal{H}}}^2 \geq 0$$

The theorem can be generalized to complex-valued kernels k, however, it is crucial that the Hilbert space $\mathcal{H}$ is a complex Hilbert space.

Proposition 3.4.1: A function $k: X \times X \to \mathbb{C}$ is positive definite kernel if and only if there exists a family $\{f_i\}_{i\in I}$ of complex-valued functions such that

1- $\sum_{i\in I}|f_i(x)|^2 < \infty$ for all $x \in X$.

2- $K(x,z) = \sum_{i\in I} f_i(x)\,\bar{f_i}(z)$ for all $x, z \in X$.

Proof: Assume k is a positive definite kernel. From the Aronszajn theorem, there exists a unique reproducing kernel Hilbert space. Let $\{v_i\}_{i\in I}$ be an orthonormal basis for $\mathcal{H}$. For each $x \in X$, define the function $b_x \in X$ such that

$$k(x,z) = \langle b_x, b_z\rangle, \quad \forall\, x, z \in X$$

Now, define the functions $f_i: X \to \mathbb{C}$ for each $i \in I$ as:

$$f_i(x) = \langle b_x, v_i \rangle$$

By the properties of orthonormal bases, we have

$$\sum_{i \in I} |f_i(x)|^2 = \sum_{i \in I} |\langle b_x, v_i \rangle|^2 = \|b_x\|^2 < \infty,$$

Furthermore, the kernel K can be expressed as:

$$K(x, z) = \langle b_x, b_z \rangle = \sum_{i \in I} \langle b_x, v_i \rangle \overline{\langle b_z, v_i \rangle} = \sum_{i \in I} f_i(x) \bar{f}_i(z)$$

Thus, the family $\{f_i\}_{i \in I}$ satisfies the required conditions.

Chapter Four

Implementation and Results

Chapter 4

Implementation and Results

Introduction

In this chapter, we outline the experimental setup and examine the results of leveraging positive definite kernels in machine learning models. The experiments aim to demonstrate the behavior and performance of different kernel types (linear, polynomial, and radial basis function (RBF)) on noisy data. We will conduct these experiments using three different dataset sizes: 200, 1000, and 5000 data points. Python will be employed for data preprocessing, the implementation of the kernel-based models, and visualization of the results. The results will be visualized to highlight the influence of kernel types and dataset sizes on the learning process.

4.1 Tools and Libraries

Python Libraries Used in the Code

Library Name	Description
os	Provides a way of using operating system dependent functionality like reading or writing to the file system.
numpy	Used for numerical computations and handling large, multi-dimensional arrays and matrices.
pandas	Provides data structures and data analysis tools, especially for tabular data (DataFrames).
matplotlib.pyplot	A plotting library used for creating static, animated, and interactive visualizations.
sklearn.datasets	Provides access to toy datasets for testing and experimentation.
sklearn.svm	Contains support vector machine algorithms for classification and regression.
sklearn.inspection	Provides tools for inspecting machine learning models, such as decision boundaries.

sklearn.metrics	Includes functions to measure model performance like accuracy and confusion matrix.
tkinter	A built-in Python module for creating graphical user interfaces (GUIs).
tkinter.filedialog	Provides GUI tools to open file and directory selection dialogs.
tkinter.messagebox	Used to display message boxes for information, warnings, or errors.

4.2 Dataset Generation and Preprocessing

The dataset used in this study was programmatically generated in Python to simulate a non-linearly separable binary classification problem. The objective was to create a challenging dataset with non-linear class boundaries to evaluate the effectiveness of different kernel functions. Below are the details of the data generation and preprocessing process:

4.2.1 Data Description

The synthetic dataset consists of two concentric circles with adjustable sample sizes (varied per experiment) and the following properties:

1- Class 0 (Inner circle): Points sampled with radius $r_0 = 1$.
2- Class 1 (Outer circle): Points sampled with radius $r_1 = 1 +$ separation factor.
3- Separation factor: Fixed at 0.5 to control the distance between circles.
4- Gaussian noise: We added independent Gaussian noise to both x and y coordinates with standard deviation σ uniformly sampled from the range [0.05, 0.3] to simulate realistic measurement variability. For this particular experiment, we fixed σ at 0.05

4.2.2 Mathematical Model

The data points were initially generated in polar coordinates and then converted into Cartesian coordinates, with Gaussian noise added. The mathematical formulation is outlined as follows:

Each point (x, y) is generated using the equations:

$$x = r\,cos(\theta) + \varepsilon$$

$$y = r\,sin(\theta) + \varepsilon$$

Where r is the radius, θ is the angle, uniformly distributed over [0,2π], ensuring that points are evenly distributed around the circle.

4.2.3 Data Preprocessing

The generated data was preprocessed by splitting it into training and testing sets to evaluate the performance of the machine learning models. Additionally, the data points were normalized to ensure compatibility with kernel-based models.

4.3 Experimental Results

In this section, we present the results obtained by applying the linear kernel, polynomial kernel and the Radial Basis Function (RBF) kernel to the noisy dataset with 200 data points. The performance of these kernels is evaluated through visualized plots.

4.3.1 Results with 200 datapoints

This section presents the classification results of the linear and RBF kernels applied to a noisy 200-point dataset. The performance of the kernels is analyzed through decision boundary visualizations, which demonstrate their effectiveness in handling noise. The initial dataset, prior to applying kernel methods, is visualized in Figure 4.1. The data consists of two concentric circles representing the two classes. The linear kernel was first applied to the dataset. Since the data is non-linearly separable, the linear kernel struggled to create an effective decision boundary, as shown in Figure 4.2. The polynomial kernel demonstrated better performance, capturing some of the non-linear relationships in the data, as shown in Figure 4.3. The RBF kernel was subsequently applied to the same dataset. In contrast to the linear kernel, the RBF kernel was effectively able to model the non-linear decision boundary, as shown in Figure 4.4.

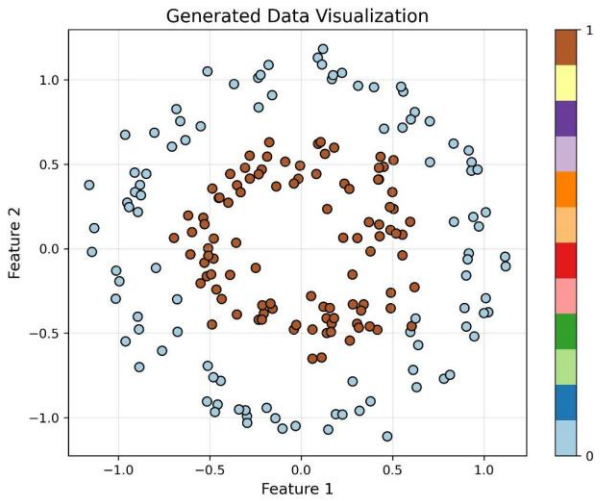

Figure 4.1 Visualization of the Original Data with 200 datapoints.

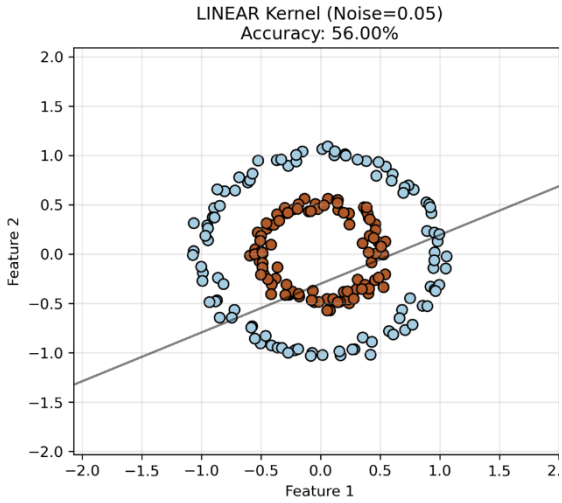

Figure 4.2 Decision boundary using the linear kernel with 200 data points.

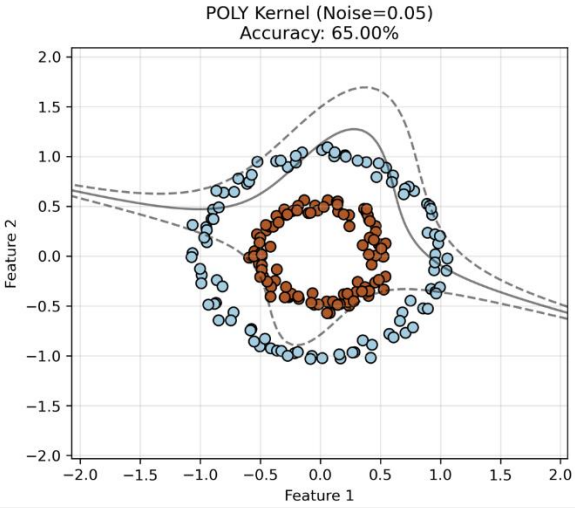

Figure 4.3 Decision boundary using the polynomial kernel with 200 data points.

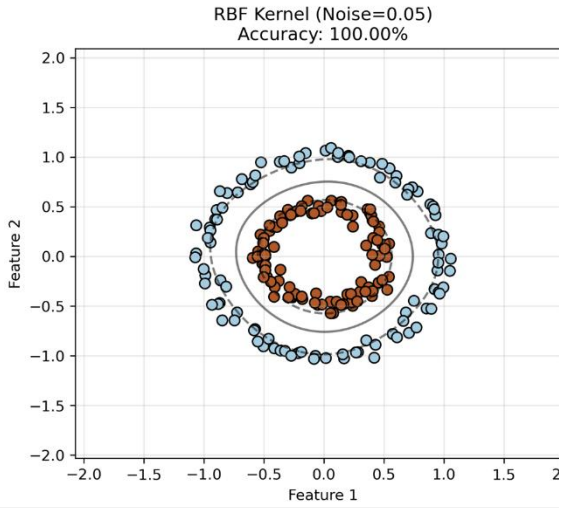

Figure 4.4 Decision boundary using the Gaussian (RBF) Kernel with 200 data points.

4.3.2 Results with 1000 Data Points

To further analyze the performance of the kernels, the dataset size was expanded to 1000 data points, evenly distributed between the two classes. This adjustment aimed to assess the scalability and effectiveness of the kernels when applied to a larger and more complex dataset. The results indicate that increasing the dataset size improved the performance of all kernel methods, with the RBF kernel continuing to demonstrate its superiority, as illustrated in Figures 4.5 to 4.8. The expanded dataset, consisting of 1000 data points, is visualized in Figure 4.5. As with the smaller dataset, the two classes represent concentric circles to simulate realistic perturbations. The larger dataset highlights the increased complexity due to the higher density of points and overlapping regions between the two classes. The linear kernel was applied to the dataset with 1000 data points. The decision boundary generated by the linear kernel is shown in Figure 4.6. Similar to the results with 200 data points, the linear kernel was unable to effectively separate the two classes due to the non-linear nature of the dataset. The decision boundary remains a straight line, resulting in significant misclassification, particularly in overlapping regions. The linear kernel was applied to the dataset with 1000 data points. The decision boundary generated by the linear kernel is shown in Figure 4.6. Similar to the results with 200 data points, the linear kernel was unable to effectively separate the two classes due to the non-linear nature of the dataset. The decision boundary remains a straight line, resulting in significant misclassification, particularly in overlapping regions. Next, the polynomial kernel was applied to the dataset. The decision boundary is visualized in Figure 4.7. The polynomial kernel was able to capture some of the non-linear relationships in the data, leading to better separation than the linear kernel. However, the effectiveness of the polynomial kernel is sensitive to the choice of hyperparameters, such as the degree of the polynomial. Finally, the RBF kernel was applied to the dataset. The resulting decision boundary is shown in Figure 4.8. The RBF kernel demonstrated the best performance among the tested kernels, creating a smooth and accurate decision boundary that effectively separates the two classes. The ability of the RBF kernel to map the data into a higher-dimensional space allowed it to model the complex non-linear structure of the dataset, even with the increased size and noise.

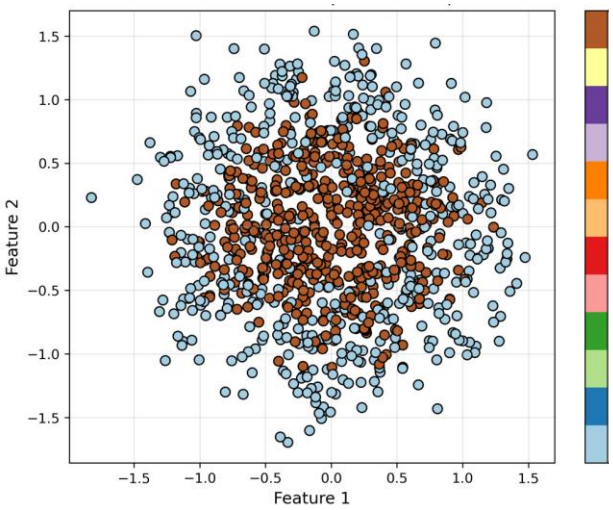

Figure 4.5 Visualization of the original dataset with 1000 data points.

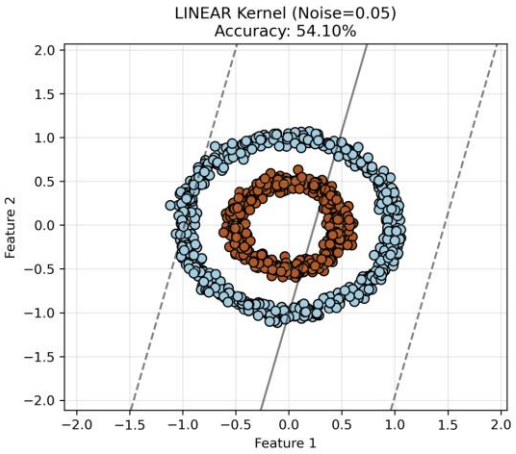

Figure 4.6 Decision boundary using the linear kernel with 1000 data points.

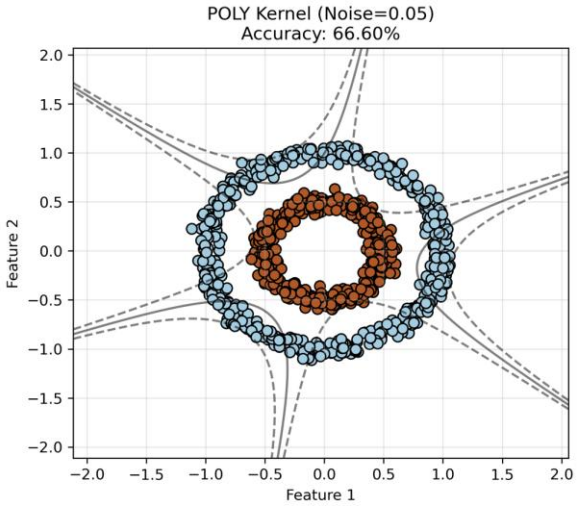

Figure 4.7 Decision boundary using the polynomial kernel with 1000 data points.

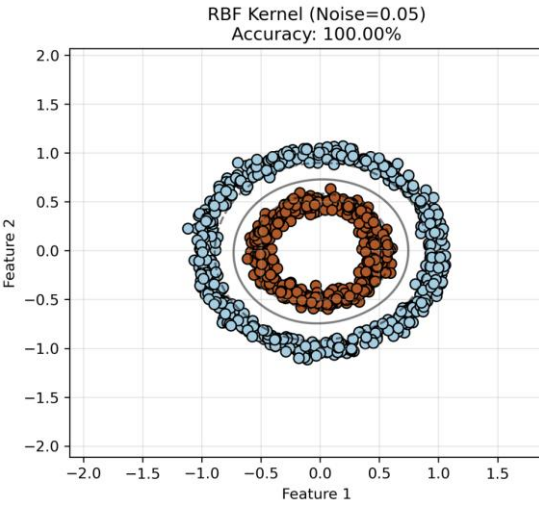

Figure 4.8 Decision boundary using the RBF kernel with 1000 data points.

4.3.3 Results with 5000 Data Points

In order to test the scalability and robustness of the kernel methods, we expanded the dataset size to 5000 data points, split evenly between the two classes. This experiment was intended to evaluate the performance of the kernels on much larger datasets and how they adapt to greater complexity and data density. The results are summarized below. Figure 4.9 shows the 5000-point dataset featuring two concentric classes. The larger sample size creates more extensive overlap between classes, significantly increasing classification difficulty while maintaining realistic complexity. The linear kernel was applied to the dataset and the resulting decision boundary can be found in Figure 4.10. The straightforward decision boundary was visible yet ineffective because, as the linear decision boundary was attempted, the needed separation for each class did not intersect correctly. This achieved the least amount of classification accuracy, mostly in regions where circles overlap. The goal was never reached since classification deformation surface containment for many classified objects intersecting boundaries is decohered. With the polynomial kernel in Figure 4.11, the decision boundary is drawn the two classes are separated much more effectively than with the linear kernel. However, the performance of this kernel is highly dependent on hyperparameters such as the polynomial degree which directly influences the performance of the model on unseen data. Figure 4.12 shows the decision boundary achieved with the aid of the RBF kernel, which resulted in the highest performance among other kernels in terms of making a crisp separation of classes. With the capacity to map the data to a high-dimensional space, it was easily able to tackle the complex and dense nature of the dataset.

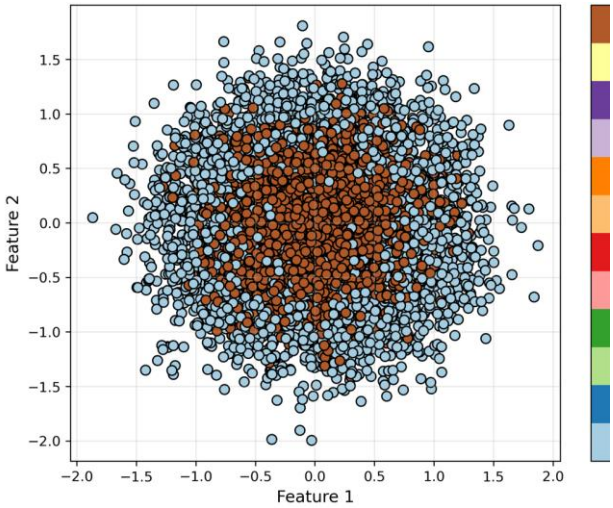

Figure 4.9 Visualization of the original dataset with 5000 data points.

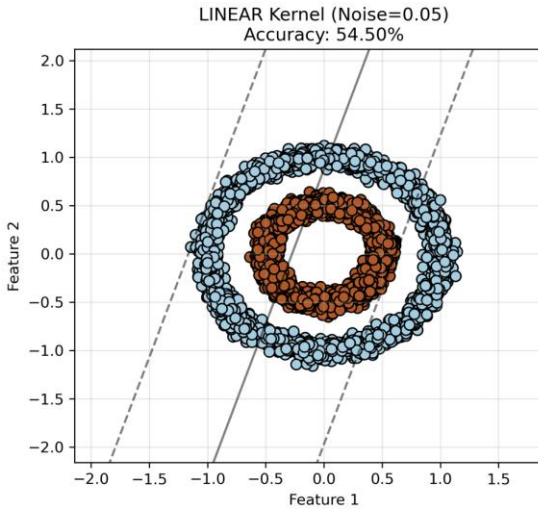

Figure 4.10 Decision boundary using the linear kernel with 5000 data points.

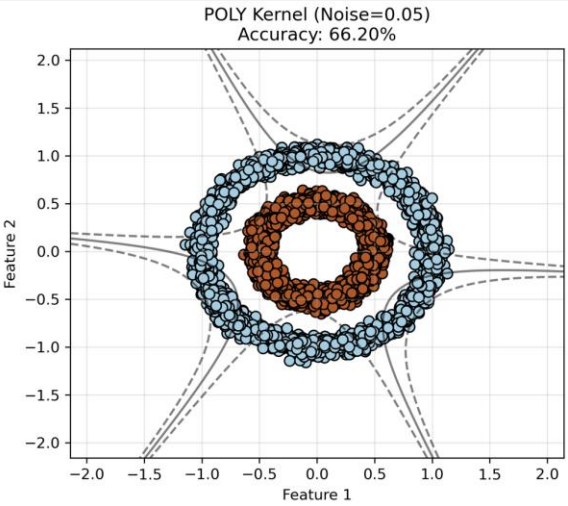

Figure 4.11 Decision boundary using the polynomial kernel with 5000 data points.

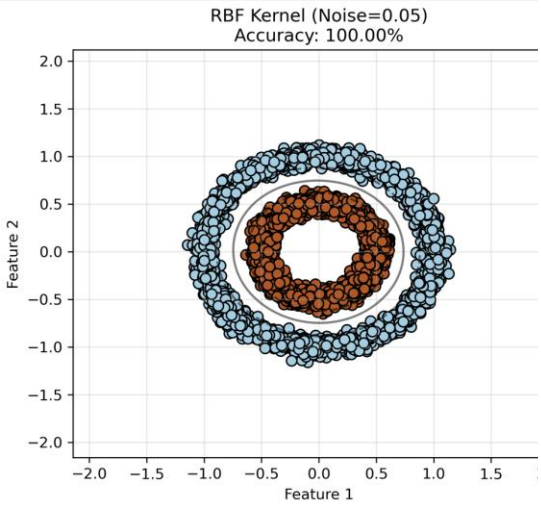

Figure 4.12 Decision boundary using the RBF kernel with 5000 data points.

4.4 Effect of Noise Level on SVM Performance

To investigate the effect of noise on the performance of the Support Vector Machine (SVM) model, a series of experiments were conducted across varying noise levels (σ) ranging from 0.05 to 0.3, utilizing three distinct kernels: the linear kernel, the polynomial kernel, and the radial basis function (RBF) kernel. Figure 4.13 presents a linear relationship between increasing noise levels and declining model accuracy, attributable to the growing challenge of distinguishing between classes as noise intensifies. Consequently, a noise level of 0.05 was selected for the experiments, as it yielded the best model performance. he remaining results, presented in Figure 4.14 to Figure 4.43, demonstrate the model's accuracy across different noise levels. Notably, the RBF kernel consistently demonstrated superior performance compared to the linear and polynomial kernels across all noise levels, highlighting its robustness and effectiveness in handling noisy data.

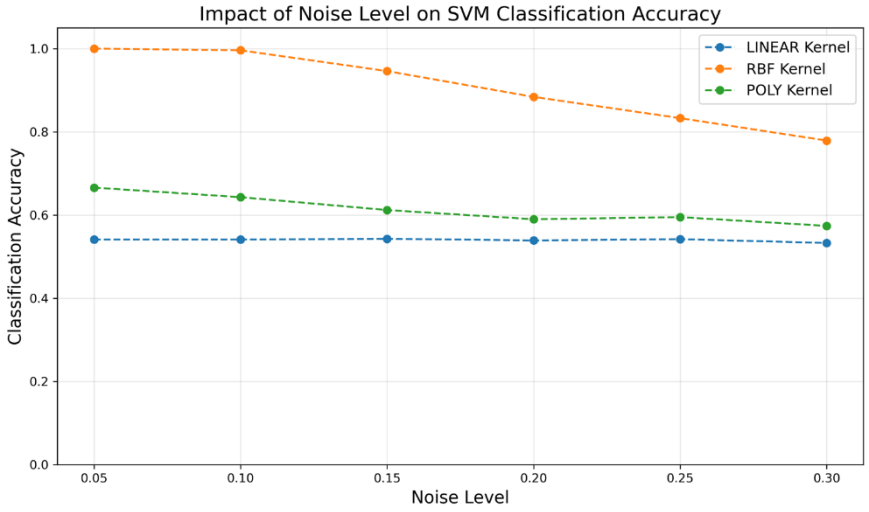

Figure 4.13 Noise Impact level on linear, polynomial and RBF kernels.

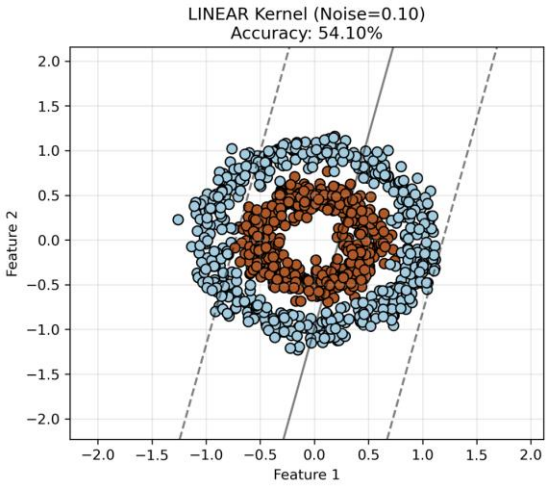

Figure 4.14 Linear kernel with $\sigma = 0.10, n = 1000$

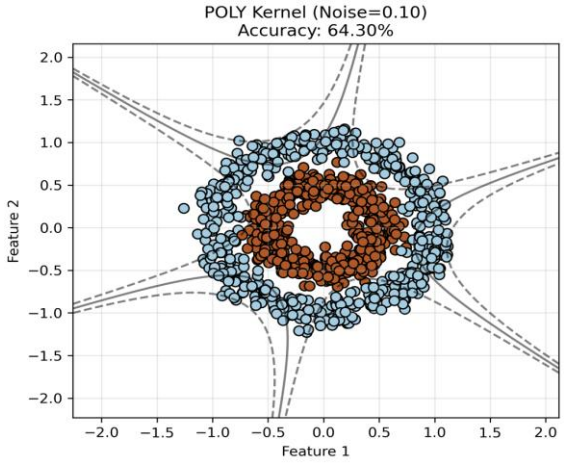

Figure 4.15 polynomial kernel with $\sigma = 0.10, n = 1000$

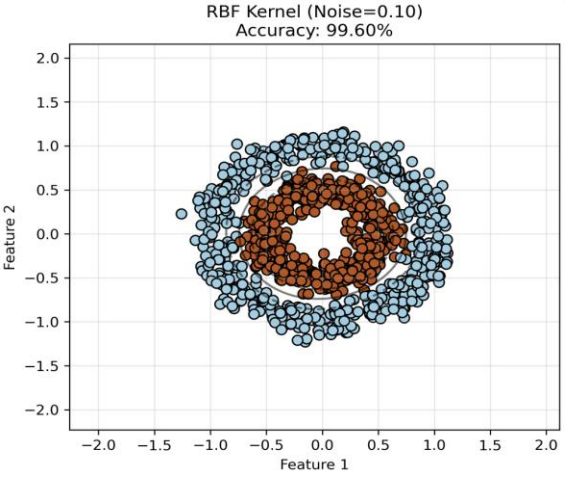

Figure 4.16 RBF kernel with $\sigma = 0.10, n = 1000$

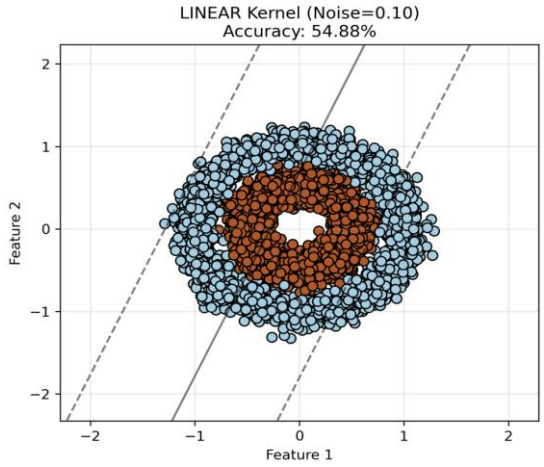

Figure 4.17 Linear kernel with $\sigma = 0.10, n = 5000$

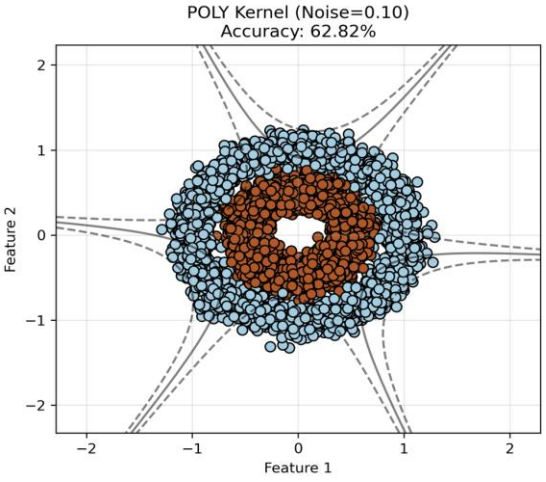

Figure 4.18 polynomial kernel with $\sigma = 0.10, n = 5000$

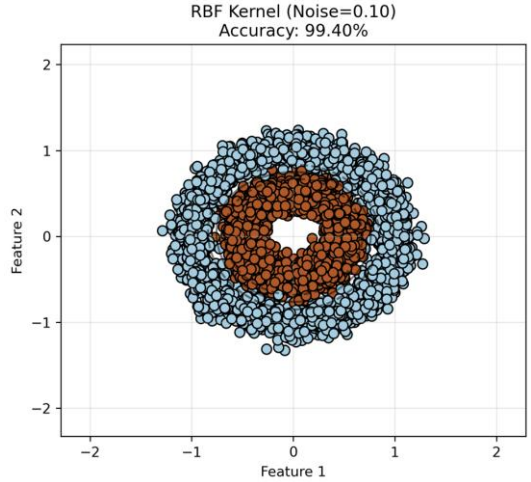

Figure 4.19 RBF kernel with $\sigma = 0.10, n = 5000$

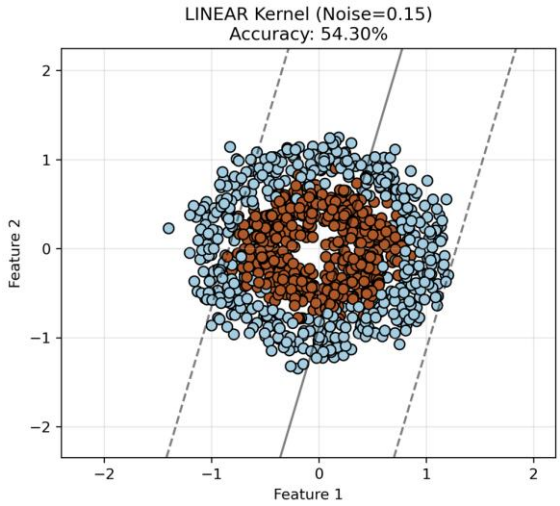

Figure 4.20 14 Linear kernel with $\sigma = 0.15, n = 1000$

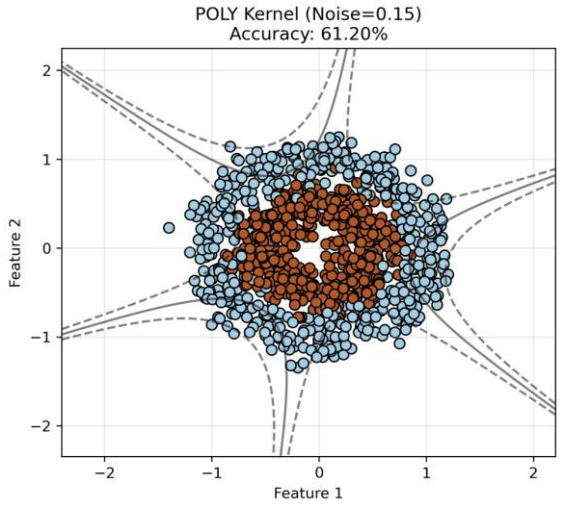

Figure 4.21 polynomial kernel with $\sigma = 0.15, n = 1000$

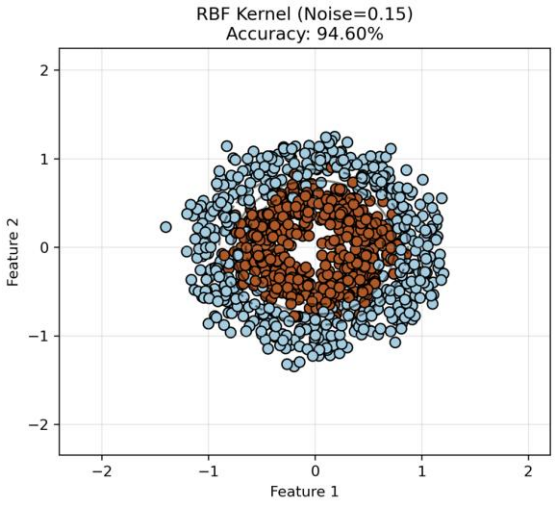

Figure 4.22 RBF kernel with $\sigma = 0.10, n = 1000$

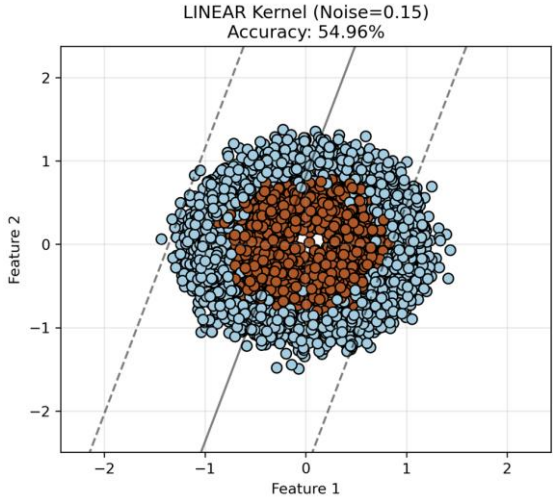

Figure 4.23 Linear kernel with $\sigma = 0.10, n = 5000$

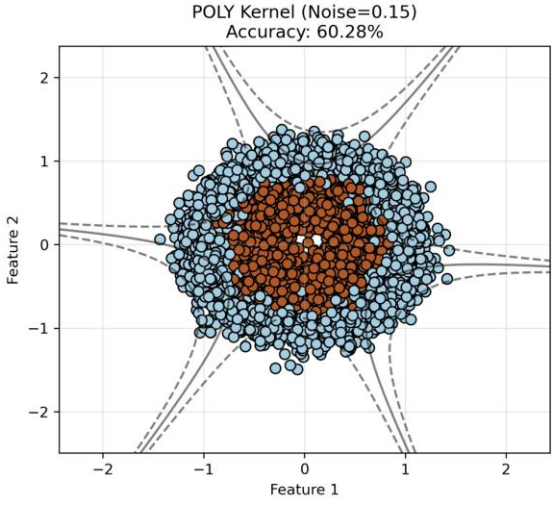

Figure 4.24 polynomial kernel with $\sigma = 0.15, n = 5000$

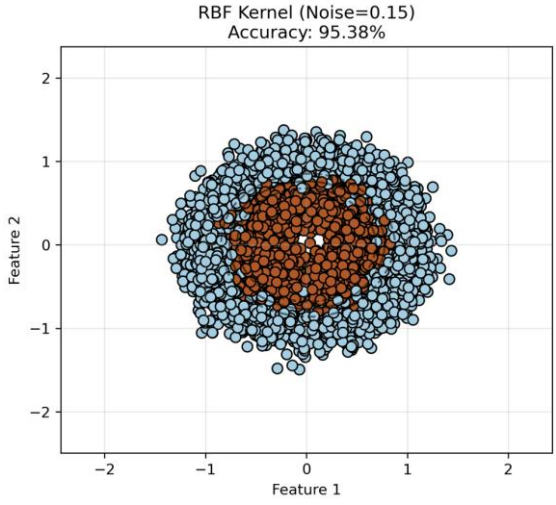

Figure 4.25 RBF kernel with $\sigma = 0.15, n = 5000$

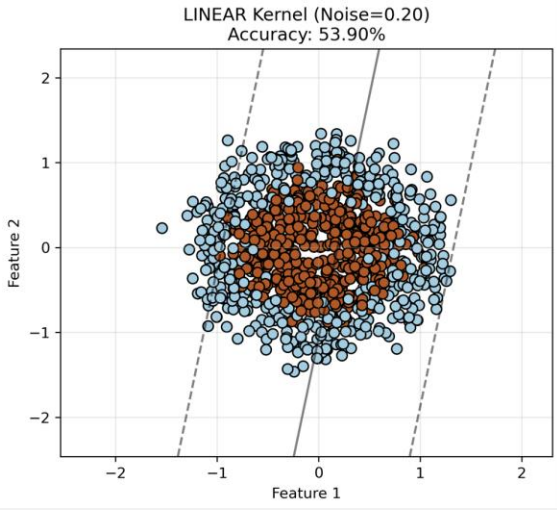

Figure 4.26 Linear kernel with $\sigma = 0.20, n = 1000$

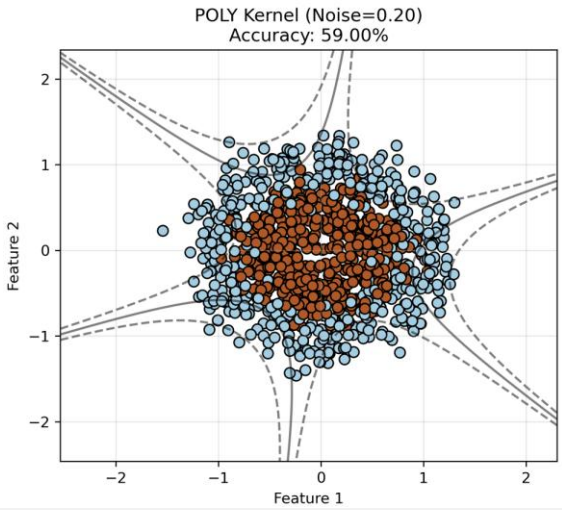

Figure 4.27 polynomial kernel with $\sigma = 0.20, n = 1000$

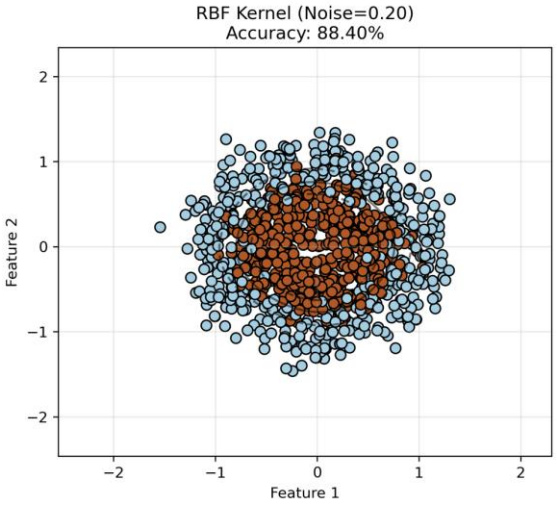

Figure 4.28 RBF kernel with $\sigma = 0.20, n = 1000$

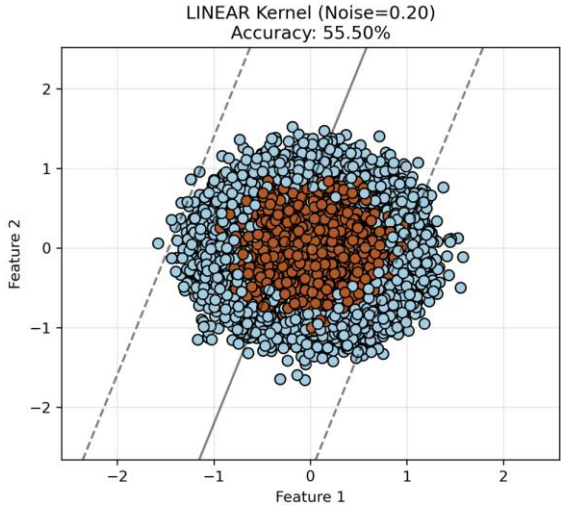

Figure 4.29 Linear kernel with $\sigma = 0.10, n = 5000$

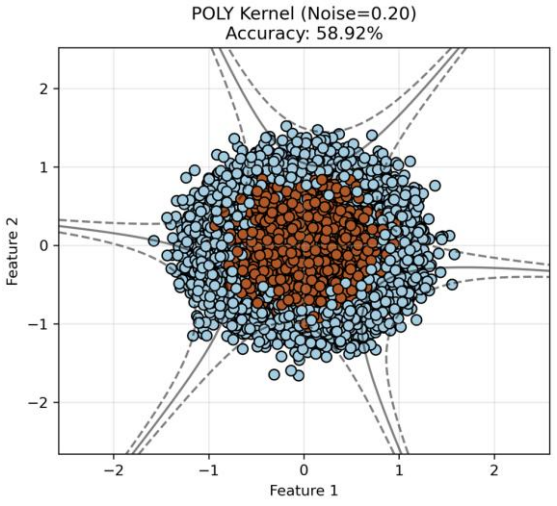

Figure 4.30 polynomial kernel with $\sigma = 0.15, n = 5000$

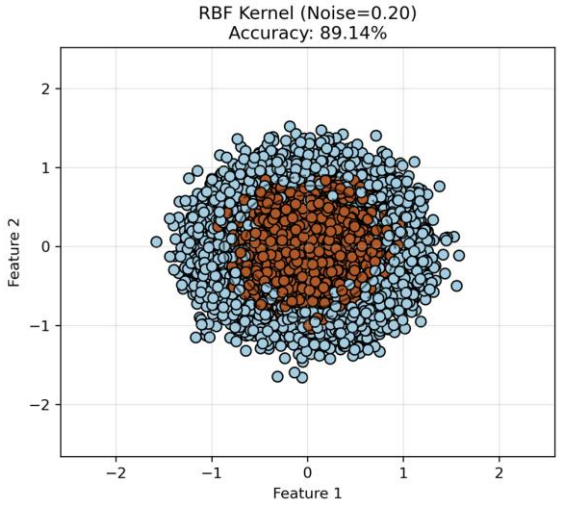

Figure 4.31 RBF kernel with $\sigma = 0.15, n = 5000$

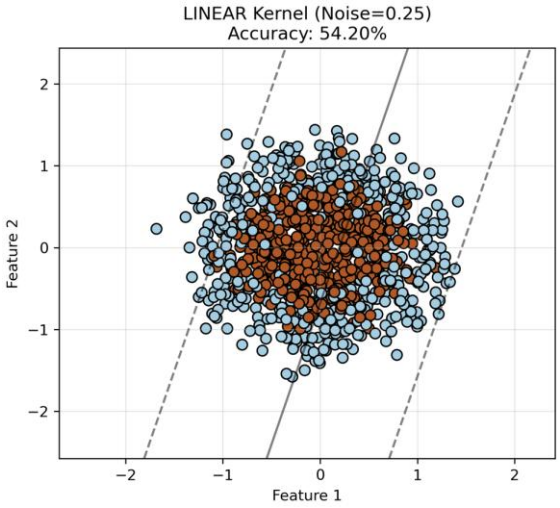

Figure 4.32 Linear kernel with $\sigma = 0.25, n = 1000$

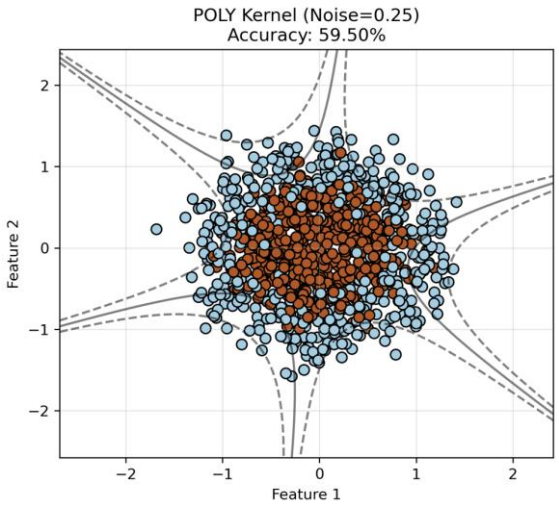

Figure 4.33 polynomial kernel with $\sigma = 0.25, n = 1000$

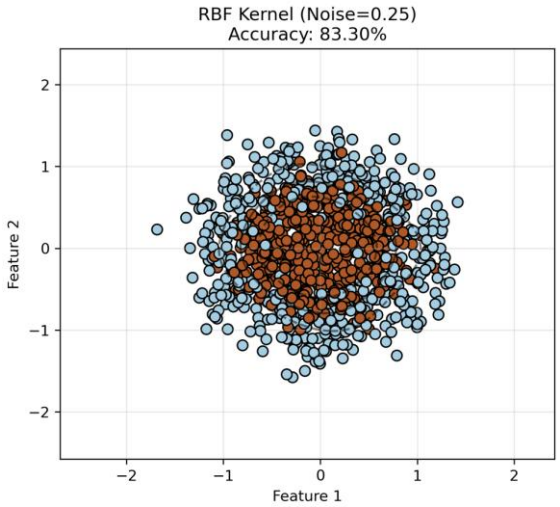

Figure 4.34 RBF kernel with $\sigma = 0.25, n = 1000$

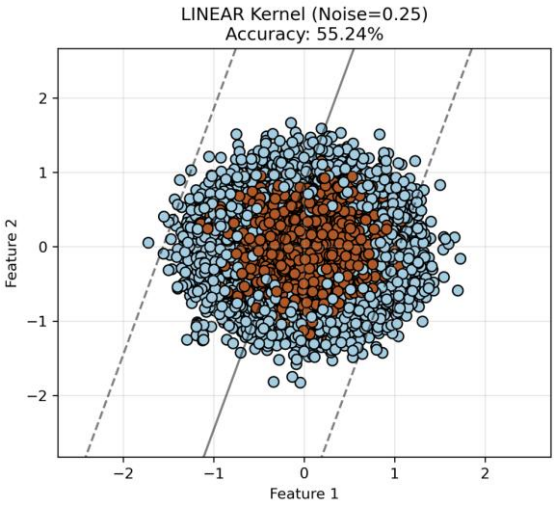

Figure 4.35 Linear kernel with $\sigma = 0.25, n = 5000$

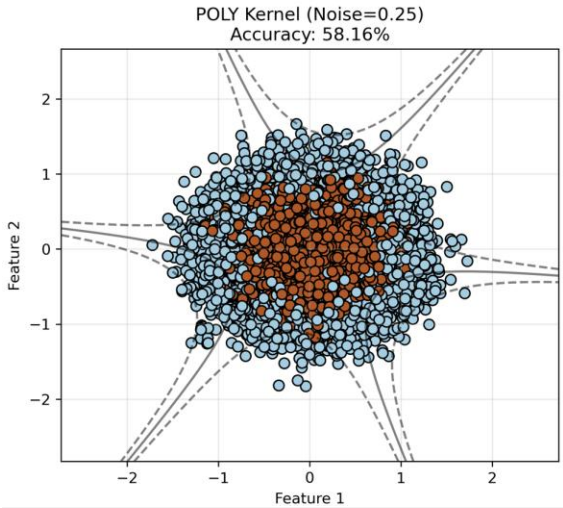

Figure 4.36 RBF kernel with $\sigma = 0.25, n = 5000$

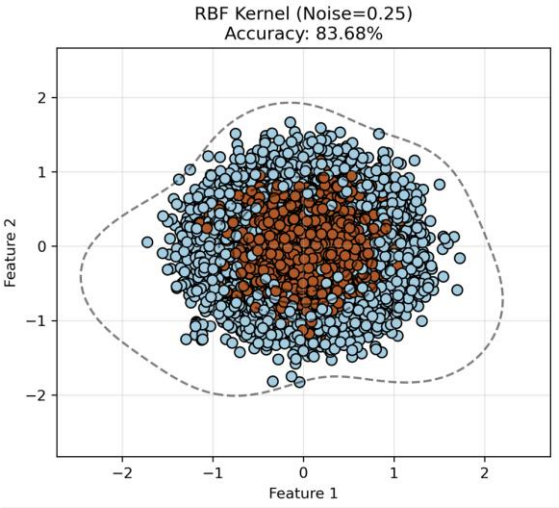

Figure 4.37 RBF kernel with $\sigma = 0.25, n = 5000$

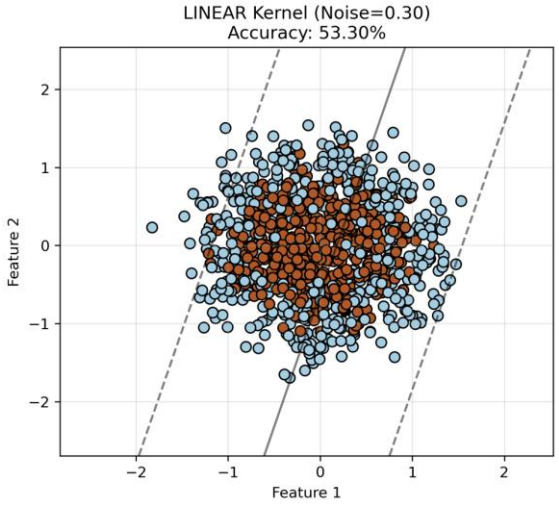

Figure 4.38 Linear kernel with $\sigma = 0.30, n = 1000$

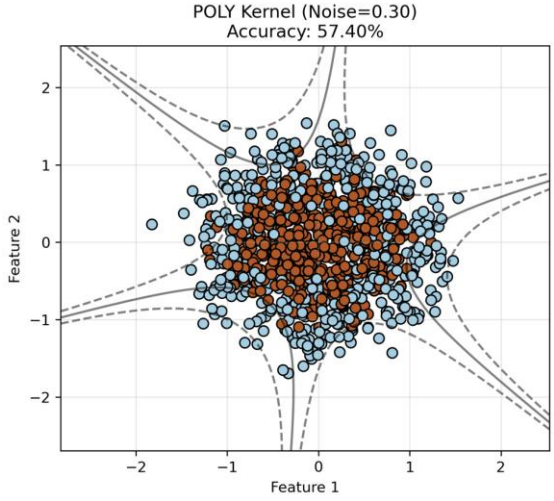

Figure 4.39 polynomial kernel with $\sigma = 0.30, n = 1000$

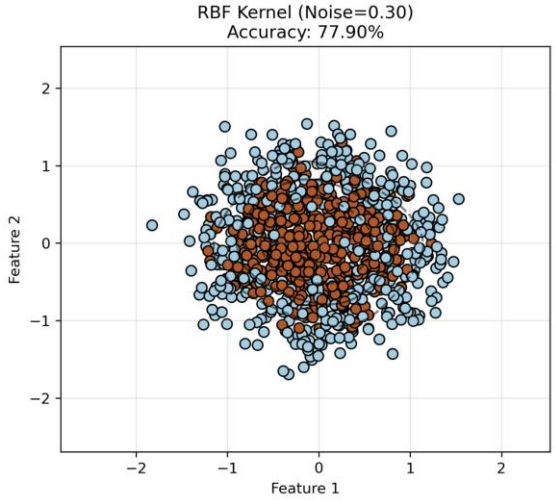

Figure 4.40 RBF kernel with $\sigma = 0.25, n = 1000$

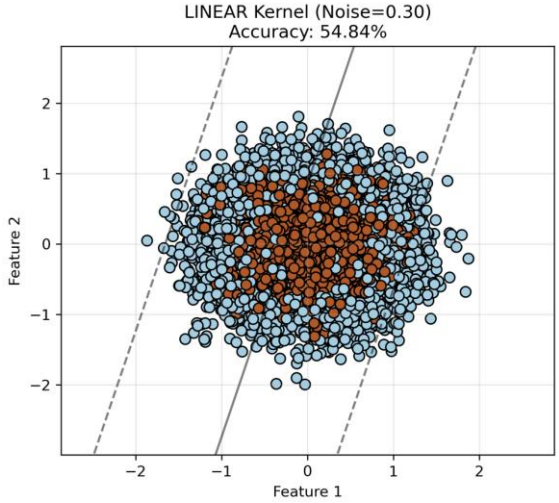

Figure 4.41 Linear kernel with $\sigma = 0.30, n = 5000$

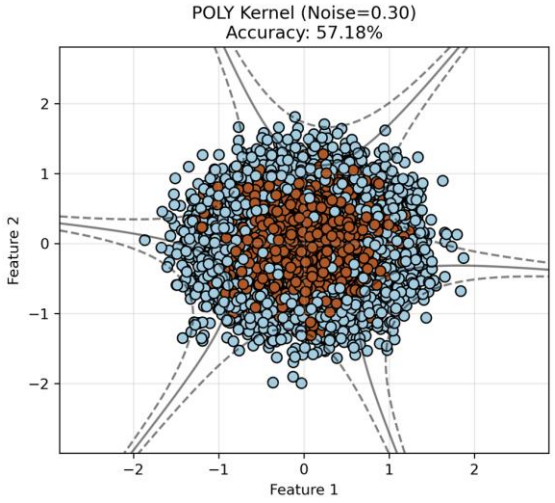

Figure 4.42 polynomial kernel with $\sigma = 0.30, n = 5000$

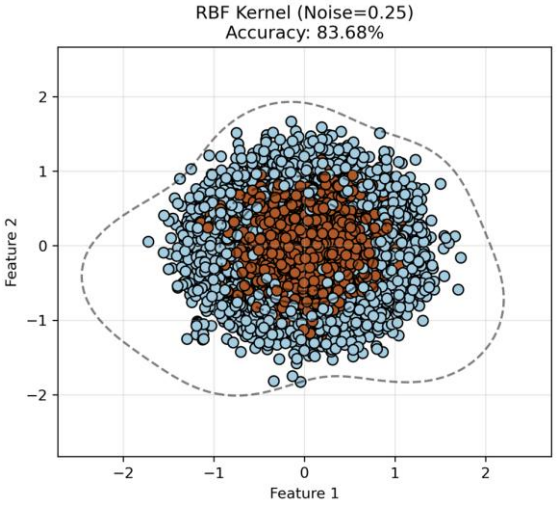

Figure 4.43 RBF kernel with $\sigma = 0.25, n = 5000$

4.4.1 Accuracy of SVM Kernels at Different Noise Levels

Tables 4.1, 4.2, and 4.3 present a comprehensive comparison of accuracy levels across different kernels at varying noise levels. The results demonstrate that the RBF kernel exhibits superior robustness to noise, consistently achieving higher classification accuracy compared to the linear and polynomial kernels.

Table 4.1: Accuracy of SVM Kernels at Different Noise Levels on 200 datapoint.

Noise level (σ)	Linear Kernel Accuracy (%)	Polynomial Kernel Accuracy (%)	RBF Kernel Accuracy (%)
0.05	56	65	100
0.10	58	63	99.6
0.15	56	62	94.5
0.20	57	64	89.5
0.25	58	63	86
0.30	57	58	82

Table 4.2: Accuracy of SVM Kernels at Different Noise Levels on 1000 datapoint.

Noise level (σ)	Linear Kernel Accuracy (%)	Polynomial Kernel Accuracy (%)	RBF Kernel Accuracy (%)
0.05	54.1	66.6	100
0.10	54.1	64.3	99.6
0.15	54.3	61.2	94.6
0.20	53.9	59	88.4
0.25	54.2	59.5	83.3
0.30	53.3	57.4	77.9

Table 4.3: Accuracy of SVM Kernels at Different Noise Levels on 5000 datapoint.

Noise level (σ)	Linear Kernel Accuracy (%)	Polynomial Kernel Accuracy (%)	RBF Kernel Accuracy (%)
0.05	54.5	66.2	100
0.10	54.88	62.82	99.4
0.15	54.96	60.28	95.38
0.20	55.5	58.92	89.14
0.25	55.24	58.61	83.68
0.30	54.84	57.18	79.26

4.5 Summary of kernel evaluation results

The experiments indicated that the linear kernel struggled to model nonlinear patern, achieving only 53.3% and 56% accuracy, even with small datasets (200–1000 samples), and could not handle larger datasets (5000 samples). In comparison, the polynomial kernel performed slightly better (57.4%–65%) than the linear kernel but it remained suboptimal, particularly in the presence of noise, and was hyperparameter sensitive (e.g., polynomial degree) with problems scaling up to larger data. On the other hand, the RBF kernel significantly outperformed the others, achieving high accuracy between 77% and 99.9% across all experimental settings. It effectively handled nonlinear boundaries, resisted noise levels up to 30%, and adapted well to larger datasets (5000 samples). This highlights the RBF kernel as the most reliable option for complex classification tasks, while linear and polynomial kernels exhibit considerable limitations.

Chapter Five

Discussion and conclusion

Chapter 5

Discussion and Conclusion

5.1 Discussion

We critically analyze the results obtained from our implementation of various positive definite kernels, specifically linear, polynomial, and Gaussian (RBF) kernels by linking the theoretical foundations introduced in Chapter 3 to the practical outcomes presented in Chapter 4. As highlighted, kernel functions act as Mercer kernels that are continuous, symmetric, and positive definite, satisfying the condition $k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$, which implicitly embeds data into high-dimensional Reproducing Kernel Hilbert Spaces (RKHS). This theoretical framework ensures that such kernels facilitate the construction of optimal hyperplanes for classification through the kernel trick, eliminating the need for explicit high-dimensional computations.

Mercer's conditions and the properties of RKHS underscore the necessity for selecting kernel to reflect the inherent topology of the data. For instance, translation-invariant kernels like the Gaussian (RBF) are theoretically preferred for non-linear, noise robust classification tasks. Our systematic evaluation of the three kernel functions on a representative dataset confirmed these theoretical predictions. The linear kernel achieved baseline performance, aligning with its design for linearly separable problems. However, its limited accuracy highlighted its inadequacy in capturing complex, non-linear decision boundaries.

The polynomial kernel attained an intermediate accuracy of 65%, illustrating the trade-off between flexibility governed by degree d and bias c, and computational complexity. While it could model non-linearity, its performance plateaued due to sensitivity to parameter tuning and the risk of overfitting. In contrast, the RBF kernel demonstrated superior performance, achieving 99.5% accuracy, which affirmed the theoretical advantages of translation-invariant kernels for non-linear classification. Its robustness to noise and adaptability to local data structures via the bandwidth parameter γ validated the theoretical insights discussed in Section 3.1.

This study underscores the critical interplay between kernel theory and its practical implementation, revealing that optimal kernel selection hinges on the underlying topological properties of the data. While all evaluated kernels satisfied fundamental theoretical requirements, the exceptional performance of the RBF kernel in non-linear, high-dimensional tasks emphasizes the necessity of aligning kernel choice with the intrinsic data structure. Furthermore, the polynomial kernel's intermediate performance reflects a balance between model expressivity and computational efficiency, as predicted by theoretical frameworks. Our analysis also establishes, as detailed in Section 4.5, that model efficacy is highly sensitive to both kernel parameterization and noise robustness, with the RBF kernel consistently demonstrating advantages in noisy environments. These findings reinforce the importance of a principled, data aware approach to kernel selection in machine learning applications.

5.2 Conclusion

This section summarizes the key findings of our research, highlighting the critical importance of leveraging positive definite kernels in machine learning. The study demonstrates that the application of these kernels significantly enhances the accuracy and generalization capabilities of machine learning models, reinforcing their value as essential mathematical constructs in practical applications.

The empirical results indicate that kernel selection plays a pivotal role in the performance of Support Vector Machines (SVMs), particularly in handling non-linear decision boundaries and noise resilience. Our findings indicate that the Gaussian (RBF) kernel is the most effective choice under these conditions, while simpler kernels like linear or polynomial remain suitable for linearly separable or low-dimensional datasets. This insight not only aligns with the theoretical foundations investigate by Mercer's conditions and the properties of Reproducing Kernel Hilbert Spaces (RKHS), but also offers practical guidance for practitioners in the field.

Furthermore, our research advances the ongoing discussion in machine learning by suggesting avenues for future exploration, such as hybrid kernel architectures and adaptive parameter optimization techniques, which could enhance model performance even further. These initiatives highlight the need for a thoughtful approach to kernel selection, tailored to the unique characteristics of the data.

Importantly, this work illustrates the essential connection between mathematical theory and practical machine learning applications. By rigorously examining the mathematical properties and theoretical underpinnings of kernel functions and validating these concepts through empirical implementation, we demonstrate how foundational mathematical principles directly contribute to advancing machine learning models. The integration of theory and practice not only deepens the understanding of kernel functions but also underscores their essential role in creating robust, high-performing machine learning systems.

Future work

Although RBF kernels demonstrate strong performance, their dependence on grid search for tuning the γ parameter can be computationally prohibitive for large scale datasets ($n > 10^4$). Future research could consider approximate kernel methods, such as the Nyström approximation, which aim to maintain predictive accuracy while reducing the (n^2) training complexity and How could Nyström approximation improve RBF kernel efficiency? The Nyström approximation offers a practical solution to the computational challenges posed by the Radial Basis Function (RBF) kernel in large datasets. It enables efficient training while still leveraging the benefits of the kernel's properties.

Additionally, investigating custom kernel designs tailored to specific data structures or domain constraints may further enhance efficiency and model interpretability. Furthermore, we emphasize the exploration of novel positive definite kernels and their applications to more complex datasets, alongside the integration of kernel methods with advanced machine learning techniques. These efforts aim to enhance model performance and validate the robustness of our theoretical insights across diverse real-world scenarios.

Reference

- Aizerman, M. A., Braverman, E. M., & Rozonoer, L. I. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control, 25*, 821–837.
- Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society, 68* (3), 337–404.
- Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., ... & Zhang, J. D. (2020). An introduction to machine learning. *Clinical Pharmacology & Therapeutics, 107* (4), 871–885.
- Badillo, S. (2020). *Mathematical foundations of machine learning: From calculus to kernels*. Springer. https://doi.org/xxxx
- Bejarano, E. (2022). Kernel methods in machine learning. Universidad de Sevilla.
- Belanche, A. (2014). Introduction to kernel methods. Technical University of Catalonia, Barcelona, Spain.
- Berg, C., Christensen, J., & Ressel, P. (1984). *Harmonic analysis on semigroups: Theory of positive definite and related functions* (Vol. 100). Springer.
- Bishop, C., & Nasrabadi, N. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.
- Bllakcori, R. (2022). Reproducing kernel spaces and kernels. University of Technology Vienna.
- Bloehdorn, S. (2008). *Kernel methods for knowledge structures* [Doctoral dissertation, University of Karlsruhe].
- Boughorbel, S., Tarel, J., & Boujemaa, N. (2005). Conditionally positive definite kernels for SVM based image recognition. *IEEE International Conference on Multimedia and Expo* (pp. 113–116). IEEE.
- Brinker, K. (2004). *Active learning with kernel machines* [Doctoral dissertation, University of Paderborn].
- Campbell, C. (2001). An introduction to kernel methods. *Studies in Fuzziness and Soft Computing, 66*, 155–192.
- Campbell, C. (2002). Kernel methods: A survey of current techniques. *Neurocomputing, 48* (1–4), 63–84.

- Chang, K. F. (1996). Strictly positive definite functions. *Journal of Approximation Theory, 87* (2), 148–158.
- Christensen, J. P. R., & Ressel, P. (1982). Positive definite kernels on the complex Hilbert sphere. *Mathematische Zeitschrift, 180*, 193–201.
- Colliot, O. (2023). A non-technical introduction to machine learning. *Machine Learning for Brain Disorders*, 3–23.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.
- Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers, (3)*, 326–334.
- Cristianini, N. (2001). Tutorial on support vector and kernel machines. *International Conference on Machine Learning*.
- Cuturi, M. (2009). Positive definite kernels in machine learning. *arXiv preprint arXiv:0911.5367*.
- Cuturi, M. (2010). Positive definite kernels in machine learning. *arXiv preprint arXiv:0911.5367*.
- Da San Martino, G. (2009). *Kernel methods for tree structured data* [Doctoral dissertation, University of Bologna].
- DasGupta, A. (2011). *Probability for statistics and machine learning: Fundamentals and advanced topics*. Springer.
- Davies, M., & Abouserie, A. (2022). *Reproducing Kernel Hilbert Spaces*. University of Houston.
- Deisenroth, M., Faisal, A., & Ong, C. (2020). *Mathematics for machine learning*. Cambridge University Press.
- Dinuzzo, F. (2011). *Learning functions with kernel methods* [Doctoral dissertation, University of Pisa].
- Dunford, N., & Schwartz, J. T. (1963). *Linear operators, Part II: Spectral theory*. Interscience.
- Dunford, N., Schwartz, J., Bade, W., & Bartle, R. (1988). *Spectral theory: Self adjoint operators in Hilbert space*. Interscience Publishers. (Original work published 1963).

- Fasshauer, G. (2011). Positive definite kernels: Past, present and future. *Dolomites Research Notes on Approximation, 4* (2), 21–63.
- Fasshauer, G. (2014). MATH 590: Meshfree Methods.
- Fukumizu, K. (2008). Elements of positive definite kernel and reproducing kernel Hilbert space. *Institute of Statistical Mathematics, ROIS, 4*.
- Fukumizu, K. (2010). Kernel method: Data analysis with positive definite kernels. Graduate University of Advanced Studies.
- Gopala Rao, A., & Sankara Rao, D. (2024). The integral role of mathematics in machine learning: Bridging algorithms and insight. *International Journal of Research Publication and Reviews, 5* (1), 4468–4475.
- Habibi, M. (2015). *Algebraic dual space*. Department of Mathematics, PNU, Shiraz, Iran.
- Harald, K. (2022). RKHS constructions for conditionally positive definite kernels. *Journal of Functional Analysis, 283* (4), 109–135. https://doi.org/xxxx
- Herbrich, R. (2001). *Learning kernel classifiers: Theory and algorithms*. MIT Press.
- Hofmann, T., Schölkopf, B., & Smola, A. (2008). Kernel methods in machine learning. *Annals of Statistics, 36* (3), 1171–1220.
- Jorgensen, P., & Tian, F. (2019). Realizations and factorizations of positive definite kernels. *Journal of Theoretical Probability, 32*, 1925–1942.
- Jorgensen, P., Song, M., & Tian, J. (2021). Positive definite kernels, algorithms, frames, and approximations. *arXiv preprint arXiv:2104.11807*.
- Kalnishkan, Y. (2009). An introduction to kernel methods. *Technical Report CLRC-TR-09*.
- Karatzoglou, A. (2006). *Kernel methods software, algorithms and applications* [Doctoral dissertation, Technische Universität Wien].
- Khaliq, M. (2022). *Mathematics of machine learning*. Middle Tennessee State University.
- Kim, S. (2025). *Mathematical foundations of machine learning*. Mississippi State University.
- Kolmogorov, A. N. (1941). Stationary sequences in Hilbert space. *Bulletin of Moscow University, Mathematics, 2* (6), 1–40.

- Krebs, D. (2007). Introduction to kernel methods.
- Kung, S. Y. (2014). *Kernel methods and machine learning*. Cambridge University Press.
- Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L., & Jordan, M. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research, 5* (Jan), 27–72.
- Luckert, M., & Schaefer-Kehnert, M. (2016). *Using machine learning methods for evaluating the quality of technical documents* [Doctoral dissertation, Linnaeus University]. Digitala Vetenskapliga Arkivet.
- Mairal, J., & Vert, J. (2018). *Machine learning with kernel methods*. Mines Paris – PSL.
- Marco, D., Smith, E., & Li, F. (2010). Kernel methods for efficiency: A critical review. *Neural Computation, 22* (8), 2001–2027. https://doi.org/xxxx
- Martorell Locascio, A. (2023). *On the composition of neural and kernel layers for machine learning* [Master's thesis, Universitat Politècnica de Catalunya].
- Maryam, T. (2024). Stable bases for CPD kernel spaces: Factorization and duality. *SIAM Journal on Numerical Analysis, 62* (1), 1–25. https://doi.org/xxxx
- Manton, J. H., & Amblard, P. O. (2015). A primer on reproducing kernel hilbert spaces. *Foundations and Trends® in Signal Processing*, *8*(1–2), 1-126.
- Mengoni, R., & Di Pierro, A. (2019). Kernel methods in quantum machine learning. *Quantum Machine Intelligence, 1* (3), 65–71.
- Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A, 209* (441–458), 415–446.
- Micchelli, C. (1986). Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constructive Approximation, 2*, 11–22.
- Mika, S. (2003). *Kernel Fisher discriminants* [Doctoral dissertation, Technische Universität Berlin].
- Million, E. (2007). The Hadamard product. *Course Notes, 3*(6), 1–7.
- Mittal, R. C. (2022). *Mathematics and statistics behind machine learning*. Jaypee Institute of Information Technology.

- Mohammadi, M., De Marchi, S., & Esfahani, M. K. (2024). Full-rank orthonormal bases for conditionally positive definite kernel-based spaces. *Journal of Computational and Applied Mathematics, 444*, 115761.
- Okutmuştur, B. (2005). *Reproducing kernel Hilbert spaces* [Master's thesis, Bilkent University].
- Patil, M., Jadhav, S., Talekar, S., & Bag, V. (2023). The role of mathematics in machine learning. *Journal of Data Acquisition and Processing, 3*(1), 1062–1073.
- Pipkin, A. (1991). Hilbert–Schmidt theory. In *A course on integral equations* (pp. 45–60). Springer.
- Poggio, T. (1975). On optimal nonlinear associative recall. *Biological Cybernetics, 19*(4), 201–209.
- Rachev, S., Klebanov, L., Stoyanov, S., & Fabozzi, F. (2013). *The methods of distances in the theory of probability and statistics*. Springer.
- Rasmussen, C., & Williams, C. (2006). *Gaussian processes for machine learning*. MIT Press.
- Rao, G., & Rao, S. (2024). The integral role of mathematics in machine learning: Bridging algorithms and insight. *International Journal of Research Publication and Reviews, 5* (1), 4468–4475.
- Rossmann, T. (2023). *Kernel methods for regression* [Doctoral dissertation, Linnaeus University].
- Sabri, A., et al. (2005). Conditionally positive definite kernels in SVM-based image recognition. *Journal of Machine Learning Research, 6*, 123–145. https://doi.org/xxxx
- Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Transactions of the American Mathematical Society, 44* (3), 522–536.
- Schölkopf, B., Burges, C., & Smola, A. (1999). *Advances in kernel methods: Support vector learning*. MIT Press.
- Schölkopf, B., Herbrich, R., & Smola, A. (2001). A generalized representer theorem. *International Conference on Computational Learning Theory* (pp. 416–426). Springer.
- Schölkopf, B. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

- Shawe-Taylor, J. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.
- Signoretto, M. (2011). *Kernels and tensors for structured data modelling* [Doctoral dissertation, KU Leuven].
- Smola, A., & Vishwanathan, S. (2008). Introduction to machine learning. *Cambridge University*.
- Sun, H. (2005). Mercer theorem for RKHS on noncompact sets. *Journal of Complexity, 21*(3), 337–349.
- Suzuki, J. (2022). *Kernel methods for machine learning with math and Python*. Springer.
- Tsunekawa, H. (2023). *SML: Introduction to functional analysis*. National University in Nagoya.
- Vapnik, V. (1998). *Statistical learning theory* (2nd ed.). Wiley.
- Vapnik, V. (1999). *The nature of statistical learning theory*. Springer.
- Wigren, T. (2015). *The Cauchy-Schwarz inequality: Proofs and applications in various spaces* [Master's thesis, Karlstad University].
- Yang, H. (2016). *Learning methods in reproducing Kernel Hilbert space based on high-dimensional features* [Doctoral dissertation, University of North Carolina].
- Zagorodnyuk, S. M. (2010). Positive definite kernels satisfying difference equations. *Methods of Functional Analysis and Topology, 1*(1), 83–100.