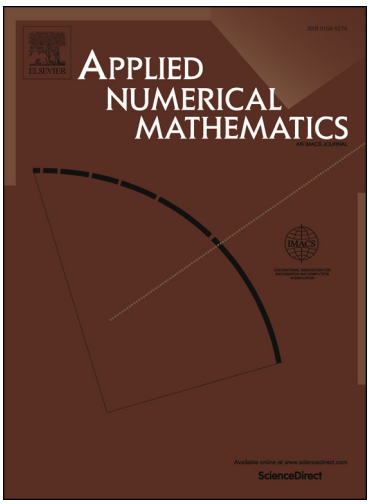
Journal Pre-proof



Sparse Machine Learning in Banach Spaces

Yuesheng Xu

PII: S0168-9274(23)00042-9
DOI: https://doi.org/10.1016/j.apnum.2023.02.011
Reference: APNUM 4534

To appear in: *Applied Numerical Mathematics*

Received date: 7 November 2022
Revised date: 4 February 2023
Accepted date: 11 February 2023

Please cite this article as: Y. Xu, Sparse Machine Learning in Banach Spaces, *Applied Numerical Mathematics*, doi: https://doi.org/10.1016/j.apnum.2023.02.011.

Sparse Machine Learning in Banach Spaces

Yuesheng Xu*

Abstract

The aim of this expository paper is to explain to graduate students and beginning researchers in the field of mathematics, statistics and engineering the fundamental concept of sparse machine learning in Banach spaces. In particular, we use binary classification as an example to explain the essence of learning in a reproducing kernel Hilbert space and sparse learning in a reproducing kernel Banach space (RKBS). We then utilize the Banach space $\ell_1(\mathbb{N})$ to illustrate the basic concepts of the RKBS in an elementary yet rigorous fashion. This paper reviews existing results in the author's perspectives to reflect the state of the art of the field of sparse learning, and includes new theoretical observations on the RKBS. Several open problems critical to the theory of the RKBS are also discussed at the end of this paper.

Key words: sparse machine learning, reproducing kernel Banach space
AMS subject classifications: 46B45, 46N10, 90C30

1 Introduction

Most of machine learning methods are to learn a function from available data. Mathematically, we need a hypothesis space to hold functions to be learned [13]. Choices of the hypothesis space are critical for learning outcomes. Traditionally, a reproducing kernel Hilbert space (RKHS), a Hilbert space where the point-evaluation functionals are continuous, is chosen, due to the existence of an inner product that can be used to measure similarity in data and continuous point-evaluation functionals that often are used to sample data.

While learning in an RKHS offers attractive features, it suffers from a major drawback. To explain it, we define the notion of sparsity. We say a vector or a sequence is sparse, we mean most of its components are zero. When it is not sparse, we call it dense. We stress here that this definition of density is peculiar to the data science community, and it does not mean (topologically) dense in the usual sense. Sparsity also refers to a representation of a function under a basis when the coefficient vector is sparse. A major drawback of learning in a Hilbert space is that a learned solution tends to be dense. This feature is a consequence of the smoothness of Hilbert spaces used as hypothesis spaces. Such a dense learned solution leads to high computational costs when the learned solution is used in prediction or other decision making procedures since once a function is learned it will be used repeatedly many times. The suffering is even more severe in the context of big data analytics. For example, Google would want its search engine to be able to make instant recommendations. Due to increasingly greater amounts of data and related model sizes, demands for more competent data processing models have emerged. As pointed in [21], the future of machine learning is sparsity. It is immensely desirable to learn sparse solutions in the sense that they have substantially fewer nonzero components or fewer terms in their representation.

*Department of Mathematics and Statistics, Old Dominion University, Norfolk, Virginia 23529, USA, y1xu@odu.edu

Most data sets have intrinsic sparsity. As a matter of fact, data that we encounter often have certain embedded sparsity structures in the sense that if they are represented in certain ways, their intrinsic characteristics can concentrate on a few scattering spots. To construct sparse solutions, we require that the norm of the hypothesis space for the learning model have certain sparsity promoting property.

With a sparsity promoting norm for the hypothesis space, when a solution of a learning model is represented in an appropriate basis, the solution can have a sparse representation. It is well-known that the norm of a usual Hilbert space does not promote sparsity due to its smoothness. Norms of certain nonsmooth Banach spaces have the ability to induce sparsity for solutions of learning methods having the spaces as hypothesis spaces, such as the 1-norm. For this reason, some Banach spaces have been used as hypothesis spaces for sparse machine learning methods. With sparse solutions, machine learning models can considerably reduce storage, computing time and communication time.

We are interested in a class of Banach spaces which are spaces of functions because function values are useful in machine learning. In particular, we would like the point-evaluation functionals in such spaces to be continuous. This gives rise to a special class of Banach spaces: the reproducing kernel Banach spaces (RKBSs) which was introduced in [58]. During the last decade, great interest has been paid to understanding of this class of function spaces and their uses in applications. We will discuss these function spaces and learning in the spaces.

The goal of this expository paper is to review recent advances in sparse machine learning in Banach spaces. We will use binary classification, a basic problem in data science, as an example to motivate the notion of sparse learning in Banach spaces. In Section 2, we present a brief review of classification. Section 3 is devoted to a presentation of learning in an RKHS. In Section 4 we illustrate the theory of the RKBS by the example of the $\ell^1(\mathbb{N})$ space. We discuss in Section 5 learning in a Banach space. We present a numerical example in Section 6 to demonstrate a sparse classification method in a Banach space. In Section 7, we elaborate several unsolved mathematical problems related to the theory of RKBS and sparse learning.

2 A Review of Classification

In this section, we review the classical support vector machine for binary classification which was discussed in [52].

A typical binary classification problem maybe described as follows: Given two sets of points in the Euclidean space $\mathbb{R}^d$, we wish to find a surface to separate them. A simplest way to separate two sets of points is to use a hyperplane, since among all types of surfaces a hyperplane is the easiest to construct and to work with. For this purpose, we wish to construct the hyperplane

$$\mathbf{w}^\top \mathbf{x} - b = 0, \quad \mathbf{x} \in \mathbb{R}^d, \tag{2.1}$$

where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are to be determined so that the hyperplane (2.1) has the largest distance to both of the two sets. Specifically, we let the training data $D := \{(\mathbf{x}_k, y_k) : k = 1, 2, \ldots, N\}$ be composed of input data points

$$X := \{\mathbf{x}_k : k = 1, 2, \ldots, N\} \subset \mathbb{R}^d$$

and output data values

$$Y := \{y_k : k = 1, 2, \ldots, N\} \subset \{-1, 1\}.$$

We intend to find a hyperplane determined by the linear function

$$s(\mathbf{x}) := \mathbf{w}_s^\top \mathbf{x} - b_s, \quad \mathbf{x} \in \mathbb{R}^d, \tag{2.2}$$

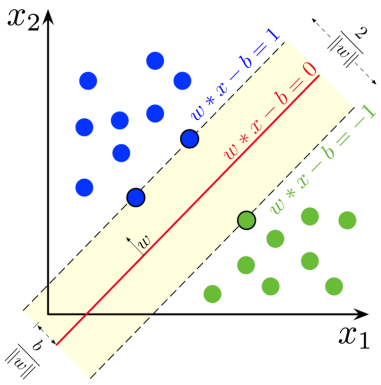


Figure 2.1: Classification (Courtesy: Wikipedia)

which separates the training data D into two groups, one with label $y_k = 1$ which corresponds to $s(\mathbf{x}_k) > 0$ and another with label $y_k = -1$ which corresponds to $s(\mathbf{x}_k) < 0$. The parameters $(\mathbf{w}_s, b_s) \in \mathbb{R}^d \times \mathbb{R}$ in equation (2.2) are chosen such that the hyperplane maximizes its distance to the points in D. This gives us a decision rule to predict the labels for new points, that is,

$$r(\mathbf{x}) := \text{sign}(s(\mathbf{x})), \quad \text{for} \quad \mathbf{x} \in \mathbb{R}^d.$$

This method is called the support vector machine (SVM) [12].

Specifically, as illustrated in Figure 2.1 we select two parallel hyperplanes

$$\mathbf{w}^\top \mathbf{x} - b = 1 \quad \text{and} \quad \mathbf{w}^\top \mathbf{x} - b = -1$$

that separate the two classes of data so that the distance between the two hyperplanes is as large as possible. The region bounded by these two hyperplanes is called the *margin*. We compute the distance $\mathbf{d}$ between these two planes. Let $\mathbf{x}_0 \in \mathbb{R}^d$ be a point on the first plane. That is, $\mathbf{x}_0$ satisfies the equation

$$\mathbf{w}^\top \mathbf{x}_0 - b = 1. \tag{2.3}$$

The distance $\mathbf{d}$ between these two planes is given by

$$\mathbf{d} = \frac{|\mathbf{w}^\top \mathbf{x}_0 - b + 1|}{\|\mathbf{w}\|_2} = \frac{2}{\|\mathbf{w}\|_2}.$$

Maximizing the distance $\mathbf{d}$ is equivalent to minimizing $\frac{1}{\mathbf{d}} = \frac{1}{2}\|\mathbf{w}\|_2$. To prevent data points from falling into the margin, we require

$$\mathbf{w}^\top \mathbf{x}_k - b \geq 1, \quad \text{if} \quad y_k = 1 \quad \text{or} \quad \mathbf{w}^\top \mathbf{x}_k - b \leq -1, \quad \text{if} \quad y_k = -1.$$

These constraints state that each data point must lie on the correct side of the margin, and they may be rewritten as

$$y_k(\mathbf{w}^\top \mathbf{x}_k - b) \geq 1, \quad k = 1, 2, \ldots, N. \tag{2.4}$$

The pair $(\mathbf{w}_s, b_s)$ is solved by a constrained minimization problem

$$\min \left\{ \frac{1}{2}\|\mathbf{w}\|_2 : (\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R} \right\} \tag{2.5}$$

subject to (2.4). Let

$$C := \{(\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R} : y_k(\mathbf{w}^\top \mathbf{x}_k - b) \geq 1, \quad k = 1, 2, \ldots, N\}. \tag{2.6}$$

It can be verified that C is a convex set. The minimization problem (2.5) with constraints (2.4) may be rewritten as

$$\min \left\{ \frac{1}{2}\|\mathbf{w}\|_2 : \mathbf{w} \in C \right\}, \tag{2.7}$$

where the constraint set C is defined by (2.6). Model (2.7) is called the *hard margin support vector machine* (SVM), which is a constrained minimization problem.

It is computationally advantageous to reformulate the constrained minimization problem (2.7) as an unconstrained one. This leads to the soft margin SVM model. Specifically, we rewrite the constraints (2.4) as

$$1 - y_k(\mathbf{w}^\top \mathbf{x}_k - b) \leq 0, \quad k = 1, 2, \ldots, N, \tag{2.8}$$

and make use of the ReLU (Rectified Linear Unit) function

$$\sigma(s) := \max\{0, s\}, \quad \text{for} \quad s \in \mathbb{R}$$

to define the hinge loss function

$$L(y, t) := \sigma(1 - yt), \quad \text{for } y \in \{-1, 1\} \text{ and } t \in \mathbb{R}.$$

When (2.8) is satisfied, we have that $L(y_k, \mathbf{w}^\top \mathbf{x}_k - b) = 0$ and when (2.8) is not satisfied, we have that

$$L(y_k, \mathbf{w}^\top \mathbf{x}_k - b) = 1 - y_k(\mathbf{w}^\top \mathbf{x}_k - b) = |1 - y_k(\mathbf{w}^\top \mathbf{x}_k - b)| > 0,$$

which is the case that we would like to prevent from occurring. For this reason, we would like to minimize the *nonnegative* fidelity term

$$\frac{1}{N} \sum_{k=1}^{N} L(y_k, \mathbf{w}^\top \mathbf{x}_k - b).$$

Combining this with the minimization (2.5) leads to the regularization problem

$$\min \left\{ \frac{1}{N} \sum_{k=1}^{N} L(y_k, \mathbf{w}^\top \mathbf{x}_k - b) + \lambda \|\mathbf{w}\|_2^2 : (\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R} \right\}, \tag{2.9}$$

where $\lambda > 0$ is a regularization parameter. Regularization problem (2.9) is called the *soft margin* SVM. Upon solving minimization problem (2.9), we obtain the pair $(\mathbf{w}_s, b_s)$ which defines the function s for classification.

3 Learning in Reproducing Kernel Hilbert Spaces

We discuss in this section the notion of feature maps in machine learning. In particular, we show the necessity of introducing feature maps in machine learning, which leads to learning in an RKHS.

We first motivate the introduction of feature maps in machine learning by returning to the classification problem reviewed in the last section. When given data sets can be separated by a hyperplane, the hard/soft margin SVM discussed in the last section will do a reasonably good job for classification. However, in most cases of practical applications, it is not possible to separate

two sets of points entirely by a hyperplane in the same space. Certain *misclassification* may occur. In general, we might allow a low degree of misclassification but do not tolerate a high degree of misclassification. We are required to separate two sets with a low degree of misclassification.

One way to alleviate misclassification is to map data sets to a higher (or even an infinite) dimensional space and perform classification in the new space. Note that sparse data sets are *easier* to be separated by a hyperplane, with a low degree of misclassification. Non-sparse data sets in a lower dimensional space may be made sparse if they are projected to a higher or even an infinite dimensional space by an appropriate map. A nice idea is to find a *feature map* Φ to map lower dimensional data sets into a higher dimensional space to gain *sparsity* for the resulting data sets. The sparsity of the mapped data sets in the higher dimensional space can significantly reduce the degree of misclassifications. In other words, the feature map Φ transfers the classification problem in a lower dimensional space to one in a higher dimensional space, where it is possible to separate point sets mapped from $\mathbb{R}^d$ by a "hyperplane", with a lower degree of misclassification. This is illustrated in Figure 3.2. A crucial issue is the choice of the feature map Φ. From a pure theoretical standpoint, "any" function Φ is a feature map. The choice of the feature map Φ pretty much depends on specific applications. Feature maps lead to the notion of *kernel based learning* or learning in RKHSs.

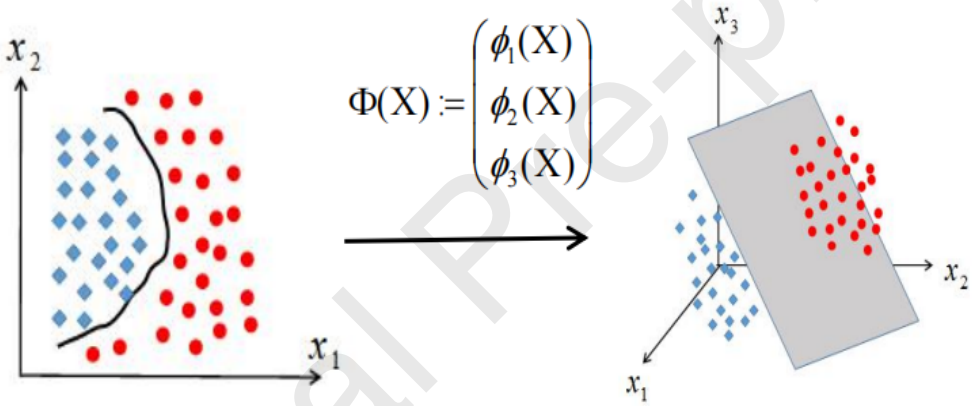


Figure 3.2: An illustration of data mapping

We next demonstrate how a feature map defines a kernel. A feature map has a well-known remarkable property which we show below. We recall that $\mathbf{x}_j \in \mathbb{R}^d$, $j = 1, 2, \ldots, n$, are the n data points.

Proposition 3.1 *Let $\mathcal{H}$ be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. If $\Phi : \mathbb{R}^d \to \mathcal{H}$, then for all $n \in \mathbb{N}$, $X_n := \{\mathbf{x}_j \in \mathbb{R}^d : j = 1, 2, \ldots, n\}$, the matrices*

$$\Phi(X_n) := [\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}} : i, j = 1, 2, \ldots, n]$$

are positive semi-definite.

Proof: For all $n \in \mathbb{N}$, suppose that $c_j \in \mathbb{R}$, $j = 1, 2, \ldots, n$, are arbitrarily given. Let $\mathbf{c} :=$

$[c_1, c_2, \ldots, c_n]^\top$. For all $\mathbf{x}_j \in \mathbb{R}^d$, $j = 1, 2, \ldots, n$, there holds

$$\begin{aligned}\mathbf{c}^\top \Phi(X_n)\mathbf{c} &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)\rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^{n} c_i \Phi(\mathbf{x}_i), \sum_{j=1}^{n} c_j \Phi(\mathbf{x}_j) \right\rangle_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^{n} c_i \Phi(\mathbf{x}_i) \right\|^2 \geq 0.\end{aligned}$$

This confirms that the matrices $\Phi(X_n)$ are positive semi-definite. □

The feature map $\Phi : \mathbb{R}^d \to \mathcal{H}$ naturally leads to a function

$$K(\mathbf{x}, \mathbf{y}) := \langle \Phi(\mathbf{x}), \Phi(\mathbf{y})\rangle_{\mathcal{H}}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \tag{3.1}$$

which is a *kernel* according to Proposition 3.1 and the following definition.

Definition 3.2 *Let $\mathbf{X} \subset \mathbb{R}^d$. A function $K : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ is called a kernel if it is symmetric and positive semi-definite, that is,*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0, \quad \textit{for all} \;\; n \in \mathbb{N}, \; \mathbf{x}_j \in \mathbf{X}, \; c_j \in \mathbb{R}, j = 1, 2, \ldots, n.$$

Kernels were studied over 100 years ago in the context of integral equations [29]. See, also [4]. Kernels can define distances of data sets to measure their similarity.

Function values are often used in machine learning. Hence, it would be desirable to require that the Hilbert space $\mathcal{H}$ of functions defined on $\mathbf{X}$ that we work with has the property that the point-evaluation functionals are continuous in the space. Precisely, we have the following definition.

Definition 3.3 *A Hilbert space $\mathcal{H}$ of functions defined on $\mathbf{X}$ is called an RKHS if for every point-evaluation functional $\delta_{\mathbf{x}}$, $\mathbf{x} \in \mathbf{X}$, there exists a constant $M_{\mathbf{x}}$ such that for all $f \in \mathcal{H}$*

$$|\delta_{\mathbf{x}} f| = |f(\mathbf{x})| \leq M_{\mathbf{x}} \|f\|_{\mathcal{H}}.$$

We now illustrate Definition 3.3 by a simplest example. To this end, we consider the space $\ell_2(\mathbb{N})$ of real sequences $\mathbf{f} := [f_1, f_2, \ldots]$ such that $\|\mathbf{f}\|_2 := \sum_{k\in\mathbb{N}} |f_k|^2 < \infty$. Clearly, elements of $\ell_2(\mathbb{N})$ are functions defined on the set $\mathbf{X} := \mathbb{N}$ and $\ell_2(\mathbb{N})$ is a Hilbert space, and thus, it is a Hilbert space of functions on $\mathbb{N}$. Moreover, $\ell_2(\mathbb{N})$ is isometrically isomorphic to $(\ell_2(\mathbb{N}))^*$. The point-evaluation functional has the form $\delta_{\mathbf{x}}(\mathbf{f}) = \delta_j(\mathbf{f}) := f_j$, for $\mathbf{x} := j \in \mathbb{N}$. It follows that

$$|\delta_j(\mathbf{f})| = |f_j| \leq \|\mathbf{f}\|_2, \quad \text{for all} \;\; \mathbf{f} \in \ell_2(\mathbb{N}),$$

with $M_{\mathbf{x}}$ in Definition 3.3 being $M_j := 1$, for all $j \in \mathbb{N}$. Therefore, according to Definition 3.3, $\ell_2(\mathbb{N})$ is an RKHS on $\mathbb{N}$ with the kernel $K := [\delta_{i,j} : i, j \in \mathbb{N}]$, where $\delta_{i,j} := 1$ if $i = j$ and 0 otherwise, for $i, j \in \mathbb{N}$.

Immediately from Definition 3.3, we observe that the closeness of two functions in an RKHS in the norm of the space implies their pointwise closeness.

An RKHS is associated with a kernel. The following well-known result can be found in [1].

Theorem 3.4 *If a Hilbert space $\mathcal{H}$ of functions on $\mathbf{X}$ is an RKHS, then there exists a unique kernel $K : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ such that $K(\mathbf{x}, \cdot) \in \mathcal{H}$, $\mathbf{x} \in \mathbf{X}$ and for all $f \in \mathcal{H}$,*

$$f(\mathbf{x}) = \langle f(\cdot), K(\mathbf{x}, \cdot)\rangle_{\mathcal{H}}, \quad \mathbf{x} \in \mathbf{X}. \tag{3.2}$$

If $K : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ is a kernel, then there exists a unique RKHS $\mathcal{H}$ on $\mathbf{X}$ such that $K(\mathbf{x}, \cdot) \in \mathcal{H}$, $\mathbf{x} \in \mathbf{X}$ and for all $f \in \mathcal{H}$, the reproducing property (3.2) *holds true.*

Proof: Suppose that $\mathcal{H}$ is an RKHS of functions on $\mathbf{X}$. Since the point-evaluation functional $\delta_{\mathbf{x}} f := f(\mathbf{x})$, for each $\mathbf{x} \in \mathbf{X}$, is continuous, by the Riesz representation theorem, there exists $k_{\mathbf{x}} \in \mathcal{H}$ such that $f(\mathbf{x}) = \delta_{\mathbf{x}} f = \langle f, k_{\mathbf{x}}\rangle_{\mathcal{H}}$, for all $f \in \mathcal{H}$. For each $\mathbf{x}, \mathbf{y} \in \mathbf{X}$, we define a function $K : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ by $K(\mathbf{x}, \mathbf{y}) := k_{\mathbf{x}}(\mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in \mathbf{X}$. Clearly, we have that for all $\mathbf{x} \in \mathbf{X}$, $K(\mathbf{x}, \cdot) \in \mathcal{H}$ and (3.2) holds true. The uniqueness of such a function K can be verified. It remains to show that K is a kernel. The symmetry of K can be seen from the computation

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

It remains to prove that it is positive semi-definite. For any $n \in \mathbb{N}$, $\mathbf{x}_j \in \mathbf{X}$, $c_j \in \mathbb{R}$, $j = 1, 2, \ldots, n$, we let

$$\mathbf{c} := [c_1, c_2, \ldots, c_n]^\top \quad \text{and} \quad K_n := [K(\mathbf{x}_i, \mathbf{x}_j) : i, j = 1, 2, \ldots, n].$$

By the reproducing property and symmetry of the kernel, we have that

$$\begin{aligned}
\mathbf{c}^\top K_n \mathbf{c} &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \langle K(\mathbf{x}_j, \cdot), K(\mathbf{x}_i, \cdot)\rangle_{\mathcal{H}} \\
&= \left\langle \sum_{j=1}^{n} c_j K(\mathbf{x}_j, \cdot), \sum_{i=1}^{n} c_i K(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}} \\
&= \left\| \sum_{i=1}^{n} c_i K(\mathbf{x}_i, \cdot) \right\|^2 \geq 0.
\end{aligned}$$

Thus, K is a kernel. We omit the proof for the converse. □

Theorem 3.4 clearly states that every function in the RKHS can be reproduced by the underlying kernel. Equation (3.2) is called the *reproducing property.* For this reason, the kernel is also called the reproducing kernel. Moreover, the kernel (3.1) defined by a feature map Φ determines an RKHS $\mathcal{H}$. The mapped RKHS $\mathcal{H}$ is more complex than the original space $\mathbb{R}^d$ and thus by using it as a hypothesis space of a learning method we can better represent our learned function. We expect that learning outcomes in this space will be more accurate than those in the original Euclidean space. In particular, data sets mapped from $\mathbb{R}^d$ to $\mathcal{H}$ via the feature map are likely more sparse than the original data sets and thus classification of the mapped data sets may result in substantially less misclassification.

Two most popular kernels in machine learning are the polynomial kernel of degree r

$$K(\mathbf{x}, \mathbf{y}) := (\mathbf{x}^\top \mathbf{y} + c)^r, \quad \text{for} \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \tag{3.3}$$

where $c \in \mathbb{R}$ is a constant and the Gaussian kernel

$$K(\mathbf{x}, \mathbf{y}) := e^{-\theta^2 \|\mathbf{x}-\mathbf{y}\|_2^2}, \quad \text{for} \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{3.4}$$

Google Scholar shows that the term "polynomial kernel" is used in titles of 53,900 articles and the term "Gaussian kernel" is used in titles of 649,000 articles. Gaussian kernels are used widely in application due to its remarkable flexibility in fitting data. In Figures 3.3, we illustrate that by altering the parameter θ we can control the shape of the graphs of the Gaussian kernel.

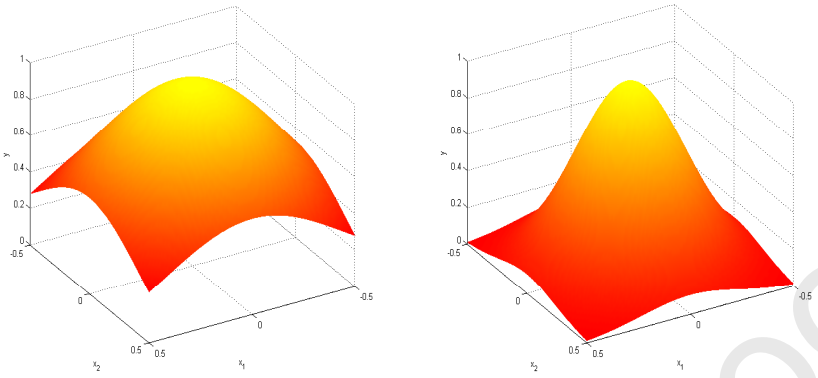


Figure 3.3: $K(\mathbf{x}, 0)$ with $\theta = 1.6$ (left); with $\theta = 3$ (right)

Following the above discussion, we map the data points $\mathbf{x}_j \in \mathbb{R}^d$ to the RKHS $\mathcal{H}$ by the feature map Φ and seek a function f in $\mathcal{H}$ that defines a decision rule for classification. This discussion leads us to extend the regularization problem (2.9) for classification to

$$\min \left\{ \frac{1}{n} \sum_{k=1}^{n} L(y_k, \langle f, \Phi(\mathbf{x}_k) \rangle_{\mathcal{H}}) + \lambda \|f\|_{\mathcal{H}}^2 : f \in \mathcal{H} \right\} \tag{3.5}$$

to determine a decision function f in the RKHS $\mathcal{H}$. It follows from the reproducing property that

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

Thus, the regularization problem (3.5) for classification is rewritten as

$$\min \left\{ \frac{1}{n} \sum_{k=1}^{n} L(y_k, f(\mathbf{x}_k)) + \lambda \|f\|_{\mathcal{H}}^2 : f \in \mathcal{H} \right\}. \tag{3.6}$$

Clearly, the regularization problem (2.9) with $b = 0$ is a special case of (3.6). To see this, we introduce the reproducing kernel

$$K(\mathbf{x}, \mathbf{y}) := \langle \mathbf{x}, \mathbf{y} \rangle, \quad \text{for all} \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

The corresponding RKHS has the form $\mathcal{H} := \{\langle \cdot, \mathbf{x} \rangle : \mathbf{x} \in \mathbb{R}^d\}$, with $\langle \langle \cdot, \mathbf{x} \rangle, \langle \cdot, \mathbf{y} \rangle \rangle_{\mathcal{H}} := \langle \mathbf{x}, \mathbf{y} \rangle$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, as its inner product. According to the reproducing property that for each $f := \langle \cdot, \mathbf{w} \rangle$ with $\mathbf{w} \in \mathbb{R}^d$, we deduce that $f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle = \mathbf{w}^\top \mathbf{x}$, for all $\mathbf{x} \in \mathbb{R}^d$, and $\|f\|_{\mathcal{H}} = \|\mathbf{w}\|_2$. Thus, problem (3.6) reduces to (2.9) with $b = 0$.

The above discussion of classification motivates us to consider learning methods in RKHSs. For example, given a pair $\{X, Y\}$ of data sets $X := \{\mathbf{x}_j \in \mathbb{R}^d : j = 1, 2, \ldots, m\}$, and $Y := \{y_j \in \mathbb{R} : j = 1, 2, \ldots, m\}$, we wish to learn a function f^* from an RKHS $\mathcal{H}$. This learning method may be described as the minimization problem

$$f^* = \operatorname{argmin} \{F(X, Y, f) + \lambda \|f\|_{\mathcal{H}} : f \in \mathcal{H}\}, \tag{3.7}$$

where F denotes the empirical fidelity term which measures the closeness of a learned solution to the given data and $\lambda > 0$ is a regularization parameter. The learning model (3.7) covers many learning problems including minimal norm interpolation, regularization network, support vector machines, kernel principal component analysis, kernel regression and deep regularized learning. Moreover, the recent work [33] established a framework where distributions are mapped into an RKHS in which the kernel methods can be extended to probability measures.

A great success of learning in Hilbert spaces has been achieved. At the same time theory of the RKHS gains coordinated development [19, 56, 57, 59]. In particular, motivated from needs of developing multiscale bases for an RKHS, refinable kernels were investigated in [56, 57, 59]. A mathematical foundation of learning in Hilbert spaces may be found in [13].

Two fundamental mathematical problems were raised regarding learning in an RKHS. The first one regards the form of a solution of the learning problem (3.7) which seeks a solution in an infinite dimensional space $\mathcal{H}$. The second is about the approximation property of the learning methods. Does the learned solution converge to the "true solution" of the practical problem as data points become dense and their number tends to infinity?

Answer to the first fundamental problem is the celebrated *representer theorem* [40]. A solution of the learning problem (3.7) in an RKHS $\mathcal{H}$ can be represented as a linear combination of a finite number of the *kernel sessions*, kernel K evaluated at the input training data points, namely,

$$K(\mathbf{x}_k, \mathbf{x}), \quad k = 1, 2, \ldots, m. \tag{3.8}$$

That is, a solution of the learning problem (3.7) has the form

$$f^*(\mathbf{x}) = \sum_{k=1}^{m} c_k K(\mathbf{x}_k, \mathbf{x}), \tag{3.9}$$

for some suitable parameters $c_k \in \mathbb{R}$, where K is the kernel associated with the space $\mathcal{H}$. Here the integer m is the number of the data points involved in the fidelity term of (3.7). This result is called the representer theorem for the learning method (3.7). The most remarkable feature of the representer theorem lies in a *finite* number of kernel sessions that represent a learned solution in the *infinite* dimensional space $\mathcal{H}$. In other words, even though we seek a learned solution in the infinite dimensional space $\mathcal{H}$ hoping it will provide much more richness in representing a learned solution, the learned solution is indeed located in the finite dimensional subspace of $\mathcal{H}$, spanned by the kernel sessions (3.8). Although the modern form of the representer theorem was found in [40], its origin dates back to [5] and [23] that appeared fifty years ago.

An answer to the second fundamental problem for the learning method (3.7) is the universality of a kernel. According to the representer theorem (3.9), a solution f^* of the learning method (3.7) is expressed as linear combination of a finite number of the kernel sessions. This motivates us to ask the following question: Can a continuous function on a compact set be approximated arbitrarily well by the kernel sessions as the data points getting dense in the set? It is well-known from the Weiersstrass Theorem in Real Analysis [38] that a continuous function on a compact domain can be arbitrarily approximated by polynomials. However, this is not always true when polynomials are replaced by kernel sessions of an arbitrary kernel. To explain this, we assume that the input space $\mathbf{X}$ is a Hausdorff topological space and that all kernels to be considered are continuous on $\mathbf{X} \times \mathbf{X}$. Moreover, we let $\mathbf{Z}$ be an arbitrarily fixed compact subset of $\mathbf{X}$ and let $C(\mathbf{Z})$ denote the space of all continuous real-valued functions from $\mathbf{Z}$ to $\mathbb{R}$ or $\mathbb{C}$ equipped with maximum norm $\|\cdot\|_{\mathbf{Z}}$. For simplicity, we restrict our discussion to real-valued function.

We first examine the polynomial kernel (3.3). We have the following negative result.

Proposition 3.5 *Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a polynomial kernel. There exists a continuous function defined on a compact set $\mathbf{Z} \subset \mathbb{R}^d$ which cannot be approximated in the uniform norm arbitrarily close by any linear combination of the kernel sessions $K(\mathbf{x}_j, \cdot)$, $j = 1, 2, \ldots, m$, for any m and any $\{\mathbf{x}_j : j = 1, 2, \ldots, m\} \subset \mathbf{Z}$.*

Proof: Since a polynomial kernel has the form (3.3), we will construct a function $f \in C(\mathbf{Z})$ so that f cannot be approximated in the uniform norm arbitrarily close by

$$f^*(\mathbf{x}) := \sum_{j=1}^{m} c_j(\mathbf{x}_j^\top \mathbf{x} + c)^r, \quad \mathbf{x} \in \mathbf{Z}, \tag{3.10}$$

for any positive integer m and any $\{\mathbf{x}_j : j = 1, 2, \ldots, m\} \subset \mathbf{Z}$.

In fact, for simplicity, we choose f as a polynomial of degree $r + 1$ with leading coefficient 1. No matter how large the integer m is and no matter what $\mathbf{x}_j$, $j = 1, 2, \ldots, m$, are chosen, f^* defined by equation (3.10) is a polynomial of degree r. The assertion is proved by the fact that a nondegenerate polynomial of degree $r + 1$ cannot be approximated in the uniform norm arbitrarily close by a polynomial of degree r. For example, we can verify this statement for the special case when $d = 1$, $r = 1$, $\mathbf{Z} = [-1, 1]$, and $f(x) := x^2$. In this case, we have that $f^*(x) = ax + b$, $a, b \in \mathbb{R}$, $x \in [-1, 1]$. Let $e(x) := f(x) - f^*(x)$, for $x \in [-1, 1]$. Clearly, $e(x) = x^2 - ax - b$. By the characterization of the best uniform approximation by the linear polynomial [36], we see that

$$\inf\{\|e\|_{[-1,1]} : a, b \in \mathbb{R}\} = \max\left\{\left|x^2 - \frac{1}{2}\right| : x \in [-1, 1]\right\} = \frac{1}{2} > 0.$$

Hence, no matter how large the positive integer m is chosen, no matter what points x_j, $j = 1, 2, \ldots, m$, are chosen, the corresponding error $\|e\|_{[-1,1]} \geq \frac{1}{2}$. □

This example motivates us to introduce in [32] the notion of universality of a kernel. We now describe the definition of the universal kernel. Given a kernel $K : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$, we introduce the function $K_{\mathbf{y}} : \mathbf{X} \to \mathbb{R}$ defined at every $\mathbf{x} \in \mathbf{X}$ by the equation $K_{\mathbf{y}}(\mathbf{x}) := K(\mathbf{x}, \mathbf{y})$ and form the space of kernel sections $K(\mathbf{Z}) := \overline{\text{span}}\{K_{\mathbf{y}} : \mathbf{y} \in \mathbf{Z}\}$. The set $K(\mathbf{Z})$ consists of all functions in $C(\mathbf{Z})$ which are uniform limits of functions of the form (3.9) where $\{\mathbf{x}_j : j = 1, 2, \ldots, m\} \subset \mathbf{Z}$. We say that a kernel $K : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ is universal, if for any given prescribed compact subset $\mathbf{Z}$ of $\mathbf{X}$, any positive number ϵ and any function $f \in C(\mathbf{Z})$ there is a function $g \in K(\mathbf{Z})$ such that

$$\|f - g\|_{\mathbf{Z}} \leq \epsilon.$$

A characterization of a universal kernel was originally established in [32]. Moreover, convenient sufficient conditions of universal kernels were given there and several classes of commonly used kernels including the Gaussian kernel defined by equation (3.4) were shown to be universal. However, it follows from the definition of the universal kernel and Proposition 3.5 that the polynomial kernels are not universal.

Although a solution of a learning method in an RKHS has nice features, it suffers from the denseness in its representation. To illustrate this point, we consider a simple minimum norm interpolation problem in $\ell_2(\mathbb{N})$. We seek $\mathbf{x}^* \in \ell_2(\mathbb{N})$ for such that

$$\|\mathbf{x}^*\|_2 = \inf\big\{\|\mathbf{x}\|_2 : \mathbf{x} \in \ell_2(\mathbb{N}),\ \langle \mathbf{a}_i, \mathbf{x}\rangle_2 = y_i,\ i = 1, 2\big\}, \tag{3.11}$$

where

$$y_1 := 3,\ y_2 := 4,\ \mathbf{a}_1 := \left(\frac{1}{n} : n \in \mathbb{N}\right),\ \mathbf{a}_2 := \left(\frac{1}{(-2)^{n-1}} : n \in \mathbb{N}\right). \tag{3.12}$$

According to [10], the solution of problem (3.11)-(3.12) is given by

$$\mathbf{x}^* \approx \left(\frac{0.4924584}{n} + \frac{2.7004714}{(-2)^{n-1}} : n \in \mathbb{N} \right). \tag{3.13}$$

Clearly, every component of the solution $\mathbf{x}^*$ is nonzero and thus, the solution is dense.

In the context of big data analytics, the dimension of data is large. It is desirable to obtain a *sparse* solution, a solution with substantial number of zero components. We next illustrate the intrinsic characteristic of a space that may or may not lead to a sparse solution. To this end, we consider a learning problem in a Hilbert space

$$\min\{\|\mathbf{x}\|_2 : \mathbf{x} \in \mathbb{R}^d\}, \quad \text{subject to} \quad A\mathbf{x} = b$$

and a Banach space

$$\min\{\|\mathbf{x}\|_1 : \mathbf{x} \in \mathbb{R}^d\}, \quad \text{subject to} \quad A\mathbf{x} = b.$$

Figures 3.4 and 3.5 illustrate that the minimum norm interpolations in the ℓ_1 space are sparse but in the ℓ_2 they are dense. This is because the geometry of the unit balls of these two different norms are different: The unit balls of the ℓ_2 are smooth, which do not promote sparsity, while those of the ℓ_1 have corners at coordinate axes, which promote sparsity.

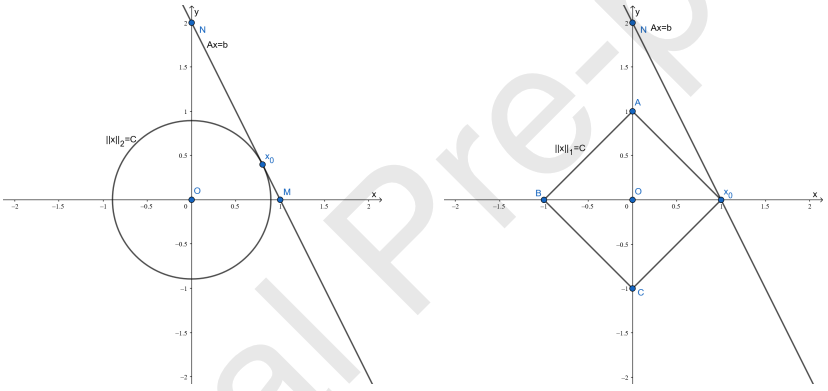


Figure 3.4: The ℓ_2-Norm vs the ℓ_1-Norm: A two dimension illustration.

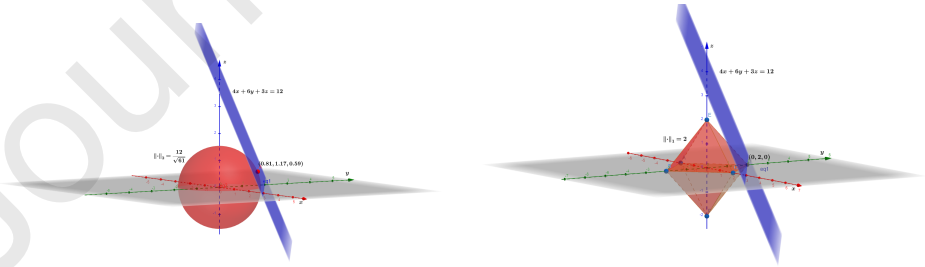


Figure 3.5: The ℓ_2-Norm vs the ℓ_1-Norm: A three dimension illustration.

We now return to the classification problem discussed in Section 2. To obtain a sparse solution, we replace the ℓ_2 norm in (2.9) by the ℓ_1 norm. This gives rise to the ℓ_1 soft margin SVM model

$$\min\left\{ \frac{1}{N} \sum_{k=1}^{N} L\left(y_k, \mathbf{w}^\top \mathbf{x}_k - b\right) + \lambda \|\mathbf{w}\|_1 : (\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R} \right\}. \tag{3.14}$$

To close this section, we summarize the pros and cons of using RKHSs for machine learning. The pros include:

- The canonical *inner product* of an RKHS provides a convenient approach to defining a *measure* for comparison which is an inevitable operation in machine learning.
- A representer theorem of a machine learning method in an RKHS reduces an infinite dimensional problem to a finite dimensional problem, with a canonical finite dimensional space determined by the training data.
- Universal kernels guarantee the approximation property.

The cons are mainly on the following aspect:

- Learning methods in an RKHS result in *dense* solutions. It is computationally expensive to use dense solutions in prediction and other practical applications.

4 Reproducing Kernel Banach Spaces for Machine Learning

Discussion in the previous section regarding the denseness of a learned solution in a Hilbert space leads us to explore Banach spaces as hypothesis spaces for machine learning hoping to gain sparsity for a learning solution in the spaces. This is because the class of Banach spaces which includes Hilbert spaces as special cases offers more choices for a hypothesis space for a machine learning method. In particular, certain Banach spaces with special geometric features may lead to sparse learning solutions. Aiming at developing sparse learning methods, the notion of the RKBS was first introduced in [58] in 2009. In this section, we review briefly the development of the RKBS during the last decade.

As pointed out earlier, the aim of most of machine learning methods is to construct functions whose values can be used for prediction or other decision making purposes. For this reason, it would be desirable to consider a Banach space of functions defined over a prescribed set as our hypothesis space for learning. A Banach space $\mathcal{B}$ is called a space of functions on a prescribed set $\mathbf{X}$ if $\mathcal{B}$ is composed of functions defined on $\mathbf{X}$ and for each $f \in \mathcal{B}$, $\|f\|_{\mathcal{B}} = 0$ implies that $f(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathbf{X}$. This implies that in a Banach space $\mathcal{B}$ of functions, the pointwise function evaluation $f(\mathbf{x})$ must be well-defined for all $\mathbf{x} \in \mathbf{X}$ and all $f \in \mathcal{B}$. The standard $L_p(\mathbf{X})$ spaces are Banach function spaces, but not Banach spaces of functions, since their elements are equivalent classes rather than functions, and the pointwise function evaluation is not defined in these spaces. Because function values are crucial in decision making, we would expect that they are stable with respect to functions chosen in the space. For this purpose, we would prefer the point-evaluation functionals $\delta_{\mathbf{x}} : \mathcal{B} \to \mathbb{R}$ defined by

$$\delta_{\mathbf{x}}(f) := f(\mathbf{x}), \quad \text{for all} \quad \mathbf{x} \in \mathbf{X}$$

being continuous with respect to functions in the space $\mathcal{B}$. This is exactly the way in which the RKHS was defined [1]. This consideration gives rise to the following definition of the RKBS .

Definition 4.1 *A Banach space $\mathcal{B}$ of functions defined on a prescribed set $\mathbf{X}$ is called an RKBS if point-evaluation functionals $\delta_{\mathbf{x}}$, for all $\mathbf{x} \in \mathbf{X}$, are continuous on $\mathcal{B}$, that is, for each $\mathbf{x} \in \mathbf{X}$ there exists a constant $c_{\mathbf{x}} > 0$ such that*

$$|\delta_{\mathbf{x}}(f)| \leq c_{\mathbf{x}} \|f\|_{\mathcal{B}}, \quad \textit{for all} \quad f \in \mathcal{B}.$$

The original definition of an RKBS was given in Definition 1 of [59] in a somewhat restricted version, where $\mathcal{B}$ was assumed to reflexive, see [27] for further generalization. In Definition 4.1, we do not have the restriction. It follows from Definition 4.1 that if $\mathcal{B}$ is an RKBS on $\mathbf{X}$, and $f, f_n \in \mathcal{B}$, for $n \in \mathbb{N}$, then $\|f_n - f\|_{\mathcal{B}} \to 0$ implies $f_n(\mathbf{x}) \to f(\mathbf{x})$, as $n \to \infty$, for each $\mathbf{x} \in \mathbf{X}$.

In the RKHS setting, due to the well-known Riesz representation theorem, one can identify a function in the same space to represent each point-evaluation functional. That is, a Hilbert space is isometrically isomorphic to its dual space, by which we refer to the space of all continuous linear functionals on the Hilbert space. This naturally leads to the unique reproducing kernel associated with the RKHS. The kernel in conjunction with the inner product of the Hilbert space yields the reproducing property. However, in general, a Banach space is not isometrically isomorphic to its dual space. Moreover, continuity of the point-evaluation functionals does not guarantee the existence of a kernel. For this reason, we need to pay special attention to the notion of the reproducing kernel for an RKBS.

For a Banach space $\mathcal{B}$, we denote by $\mathcal{B}^*$ its dual space, the space of all continuous linear functionals on $\mathcal{B}$. It is well-known that the dual space of a Banach space is again a Banach space. Definition 4.1 of the RKBS ensures that when $\mathcal{B}$ is an RKBS on $\mathbf{X}$, the point-evaluation functionals

$$\delta_{\mathbf{x}} \in \mathcal{B}^*, \quad \text{for all} \quad \mathbf{x} \in \mathbf{X}. \tag{4.1}$$

When $\mathcal{B}$ is an RKBS on $\mathbf{X}$, we let $\mathcal{B}'$ denote the completion of the linear span $\tilde{\mathcal{B}}$ of all the point-evaluation functionals $\delta_{\mathbf{x}}$ on $\mathcal{B}$ for $\mathbf{x} \in \mathbf{X}$, under the norm of $\mathcal{B}^*$. It follows from (4.1) that $\tilde{\mathcal{B}} \subseteq \mathcal{B}^*$. Thus, $\mathcal{B}' \subseteq \mathcal{B}^*$ since $\mathcal{B}^*$ is complete. Moreover, we have that $\mathcal{B}'$ is the smallest Banach space that contains all point-evaluation functionals on $\mathcal{B}$. We will call $\mathcal{B}'$ the δ-dual space of $\mathcal{B}$. Because in machine learning, we are interested in Banach spaces of functions, we further suppose that the δ-dual space $\mathcal{B}'$ is isometrically isomorphic to a Banach space $\mathcal{B}^{\#}$ of functions on a set $\mathbf{X}'$. This hypothesis is satisfied for most examples that we encounter in application. For instance, the space that we will discuss later in this section has this property. In the rest of this paper, we will not distinguish $\mathcal{B}'$ and $\mathcal{B}^{\#}$.

For a Banach space $\mathcal{B}$ of functions defined on the set $\mathbf{X}$, we let $\langle \cdot, \cdot \rangle_{\mathcal{B} \times \mathcal{B}'}$ denote the dual bilinear form on $\mathcal{B} \times \mathcal{B}'$ induced by restriction of the dual bilinear form on $\mathcal{B} \times \mathcal{B}^*$ to $\mathcal{B} \times \mathcal{B}'$. When there is no ambiguity, we will write it as $\langle \cdot, \cdot \rangle_{\mathcal{B}}$. We now define a reproducing kernel, which provides a closed-form function representation for the point-evaluation functionals, for an RKBS.

Definition 4.2 *Let $\mathcal{B}$ be an RKBS on a set $\mathbf{X}$ with the δ-dual space $\mathcal{B}'$. Suppose that $\mathcal{B}'$ is isometrically isomorphic to a Banach space of functions on a set $\mathbf{X}'$. A function $K : \mathbf{X} \times \mathbf{X}' \to \mathbb{R}$ is called a reproducing kernel for $\mathcal{B}$ if $K(\mathbf{x}, \cdot) \in \mathcal{B}'$ for all $\mathbf{x} \in \mathbf{X}$, and*

$$f(\mathbf{x}) = \langle f(\cdot), K(\mathbf{x}, \cdot) \rangle_{\mathcal{B}}, \quad \textit{for all} \quad f \in \mathcal{B}. \tag{4.2}$$

If in addition $\mathcal{B}'$ is an RKBS on $\mathbf{X}'$, $K(\cdot, \mathbf{y}) \in \mathcal{B}$ for all $\mathbf{y} \in \mathbf{X}'$, and

$$g(\mathbf{y}) = \langle K(\cdot, \mathbf{y}), g(\cdot) \rangle_{\mathcal{B}}, \quad \text{for all} \quad g \in \mathcal{B}', \tag{4.3}$$

we call $\mathcal{B}'$ an adjoint RKBS of $\mathcal{B}$. In this case, the function

$$K'(\mathbf{x}, \mathbf{y}) := K(\mathbf{y}, \mathbf{x}), \quad \text{for} \quad \mathbf{x} \in \mathbf{X}', \mathbf{y} \in \mathbf{X},$$

is a reproducing kernel for $\mathcal{B}'$, and we call $\mathcal{B}$, $\mathcal{B}'$ a pair of RKBSs.

The original definition of a reproducing kernel appeared in [58] with a choice $\mathcal{B}' = \mathcal{B}^*$, which was extended in [55]. Also, see [46, 49] for further development.

Equation (4.2) furnishes a representation of the point-evaluation functional in terms of the kernel and the dual bilinear form. The present form in Definition 4.2 of a reproducing kernel for an RKBS is slightly different from the various forms known in the literature. We feel that the form described in Definition 4.2 better captures the essence of reproducing kernels. Unlike a reproducing kernel for an RKHS, a reproducing kernel for an RKBS is not necessarily symmetric or positive semi-definite. In a special case when an RKBS satisfies the following hypothesis, we can establish the positive semi-definiteness of its kernel.

Hypothesis (H1): Spaces $\mathcal{B}$, $\mathcal{B}'$ are a pair of RKBSs on a set $\mathbf{X}$, the δ-dual space $\mathcal{B}'$ is isometrically isomorphic to a Banach space of functions on $\mathbf{X}$, and a reproducing kernel $K : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ for $\mathcal{B}$ is symmetric.

Suppose that Hypothesis (H1) is satisfied. By symmetry of the kernel K, we have that $K(\cdot, \mathbf{x}) \in \mathcal{B}'$ for all $\mathbf{x} \in \mathbf{X}$ and $K(\mathbf{y}, \cdot) \in \mathcal{B}$ for all $\mathbf{y} \in \mathbf{X}$. In this case, $\langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{x})\rangle_{\mathcal{B}\times\mathcal{B}'}$ is well-defined for $\mathbf{x} \in \mathbf{X}$. We further require that the following hypothesis be satisfied.

Hypothesis (H2): There exists a positive constant A such that

$$\frac{\langle f_n, f_n\rangle_{\mathcal{B}\times\mathcal{B}'}}{\|f_n\|^2_{\mathcal{B}}} \geq A, \text{ for all } f_n := \sum_{i=1}^{n} c_i K(\cdot, \mathbf{x}_i) \neq 0, n \in \mathbb{N},\ \mathbf{x}_i \in \mathbf{X},\ c_i \in \mathbb{R},\ i = 1, 2, \ldots, n. \quad (4.4)$$

In a loose sense, inequality (4.4) has a geometric interpretation: "cosine" of the "angle" formed by $\frac{\langle f_n, f_n\rangle_{\mathcal{B}\times\mathcal{B}'}}{\|f_n\|^2_{\mathcal{B}}}$ is uniformly above zero, for all $f_n \neq 0$, $n \in \mathbb{N}$, $\mathbf{x}_i \in \mathbf{X}$, $c_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$. In other words, the spaces $\mathcal{B}$ and $\mathcal{B}'$ cannot be "perpendicular" to each other, with respect to the dual bilinear form $\langle\cdot,\cdot\rangle_{\mathcal{B}\times\mathcal{B}'}$. When $\mathcal{B}$ is a Hilbert space, inequality (4.4) reduces to an equality with $A = 1$, since in such a case $\mathcal{B} = \mathcal{B}'$. Under Hypotheses (H1) and (H2), we can establish the positive semi-definiteness of a reproducing kernel.

Theorem 4.3 *Suppose that a pair $\mathcal{B}$, $\mathcal{B}'$ of RKBSs satisfy Hypotheses (H1) with a reproducing kernel $K : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ and (H2). For all $n \in \mathbb{N}$, all $\mathbf{x}_j \in \mathbf{X}$, $j = 1, 2, \ldots, n$, let*

$$K_n := [K(\mathbf{x}_i, \mathbf{x}_j) : i, j = 1, 2, \ldots, n]. \quad (4.5)$$

Then, for all $n \in \mathbb{N}$ and all $\mathbf{x}_j \in \mathbf{X}$, $j = 1, 2, \ldots, n$, the matrices K_n are positive semi-definite.

Proof: It suffices to show for all $\mathbf{c} := [c_1, c_2, \ldots, c_n]^\top \in \mathbb{R}^n$ that $\mathbf{c}^\top K_n \mathbf{c} \geq 0$. For $\mathbf{y} \in \mathbf{X}$, we introduce the notation $k_{\mathbf{y}} := K(\cdot, \mathbf{y})$. By the reproducing property (4.2) and symmetry of the kernel K, we observe that

$$\begin{aligned} K(\mathbf{x}_i, \mathbf{x}_j) &= k_{\mathbf{x}_j}(\mathbf{x}_i) \\ &= \left\langle k_{\mathbf{x}_j}(\cdot), K(\mathbf{x}_i, \cdot)\right\rangle_{\mathcal{B}} \\ &= \langle K(\cdot, \mathbf{x}_j), K(\mathbf{x}_i, \cdot)\rangle_{\mathcal{B}} \\ &= \langle K(\cdot, \mathbf{x}_j), K(\cdot, \mathbf{x}_i)\rangle_{\mathcal{B}}\,. \end{aligned}$$

Therefore, according to Hypotheses (H1) and (H2), for all $\mathbf{c} := [c_1, c_2, \ldots, c_n]^\top \in \mathbb{R}^n$, we obtain

that

$$\begin{aligned}\mathbf{c}^\top K_n \mathbf{c} &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \left\langle K(\cdot, \mathbf{x}_j), K(\cdot, \mathbf{x}_i)\right\rangle_{\mathcal{B}} \\ &= \left\langle \sum_{i=1}^{n} c_i K(\cdot, \mathbf{x}_i), \sum_{i=1}^{n} c_i K(\cdot, \mathbf{x}_i) \right\rangle_{\mathcal{B}} \\ &\geq A \left\| \sum_{i=1}^{n} c_i K(\cdot, \mathbf{x}_i) \right\|_{\mathcal{B}}^2 \geq 0,\end{aligned}$$

proving the positive semi-definiteness of K_n. □

In general, the kernel matrix (4.5) is not well-defined unless $\mathbf{X} = \mathbf{X}'$.

We next illustrate the notion of the RKBS and its kernel with an example. For this purpose, we consider the space $\ell_1(\mathbb{N})$ of real sequences $\mathbf{f} := [f_1, f_2, \dots]$ such that $\|\mathbf{f}\|_1 := \sum_{k\in\mathbb{N}} |f_k| < \infty$. Elements of $\ell_1(\mathbb{N})$ are functions defined on the set $\mathbf{X} := \mathbb{N}$. It is well-known that $\ell_1(\mathbb{N})$ is a Banach space, and thus, it is a Banach space of functions on $\mathbb{N}$. For each $j \in \mathbb{N}$, the point-evaluation functional δ_j has the form

$$\delta_j(\mathbf{f}) = f_j. \tag{4.6}$$

We first establish that $\ell_1(\mathbb{N})$ is an RKBS, which is stated in the next theorem.

Theorem 4.4 *The space $\ell_1(\mathbb{N})$ is an RKBS on $\mathbb{N}$.*

Proof: It suffices to verify that the point-evaluation functionals on the space $\ell_1(\mathbb{N})$ are continuous. All point-evaluation functionals on space $\ell_1(\mathbb{N})$ are δ_j, for all $j \in \mathbb{N}$. Hence, for all $j \in \mathbb{N}$, it follows from (4.6) that

$$|\delta_j(\mathbf{f})| = |f_j| \leq \|\mathbf{f}\|_1, \quad \text{for all} \quad \mathbf{f} \in \ell_1(\mathbb{N}).$$

This ensures that for all $j \in \mathbb{N}$, the point-evaluation functionals δ_j are all continuous on the space $\ell_1(\mathbb{N})$. By Definition 4.1, the space $\ell_1(\mathbb{N})$ is an RKBS over $\mathbb{N}$. □

We next identify a reproducing kernel for the RKBS $\ell_1(\mathbb{N})$. By $\ell_\infty(\mathbb{N})$ we denote the Banach space of real bounded sequences on $\mathbb{N}$ under the supremum norm. Namely, for any $\mathbf{a} := [a_1, a_2, \dots] \in \ell_\infty(\mathbb{N})$, we have that $\|\mathbf{a}\|_\infty := \sup\{|a_k| : k \in \mathbb{N}\} < \infty$. We further denote by $c_0(\mathbb{N})$ the set of real sequences that are convergent to zero in the sense that for all $\mathbf{a} := [a_1, a_2, \dots] \in c_0(\mathbb{N})$, there holds $\lim_{k\to\infty} a_k = 0$. The set $c_0(\mathbb{N})$ is a Banach space of real convergent sequences under the supremum norm defined on $\mathbb{N}$. Thus, for all $\mathbf{a} \in c_0(\mathbb{N})$, there holds that $\|\mathbf{a}\|_\infty < +\infty$. This ensures that $c_0(\mathbb{N}) \subset \ell_\infty(\mathbb{N})$. Moreover, since a nonzero constant sequence is in $\ell_\infty(\mathbb{N})$ but not in $c_0(\mathbb{N})$, $c_0(\mathbb{N})$ is a true subspace of $\ell_\infty(\mathbb{N})$. It is well-known [37] that $c_0^*(\mathbb{N}) = \ell_1(\mathbb{N})$ and $\ell_1^*(\mathbb{N}) = \ell_\infty(\mathbb{N})$. It follows that $(\ell_1^*(\mathbb{N}))^* = (\ell_\infty(\mathbb{N}))^* \subset c_0^*(\mathbb{N}) = \ell_1(\mathbb{N})$. Thus, $(\ell_1^*(\mathbb{N}))^* \neq \ell_1(\mathbb{N})$. That is, the Banach space $\ell_1(\mathbb{N})$ is not reflexive and its predual is $c_0(\mathbb{N})$.

The dual bilinear form on $\ell_1(\mathbb{N}) \times \ell_\infty(\mathbb{N})$ has a concrete form. Specifically, for any $\mathbf{f} \in \ell_1(\mathbb{N})$ and $\mathbf{a} \in \ell_\infty(\mathbb{N})$, we define

$$\langle \mathbf{f}, \mathbf{a}\rangle_{\ell_1(\mathbb{N})} := \sum_{k\in\mathbb{N}} a_k f_k. \tag{4.7}$$

Restricting the definition (4.7) to the subspace $c_0(\mathbb{N})$ of $\ell_\infty(\mathbb{N})$ yields the dual bilinear form on $\ell_1(\mathbb{N}) \times c_0(\mathbb{N})$. It follows from (4.7) for any $\mathbf{f} \in \ell_1(\mathbb{N})$ and $\mathbf{a} \in c_0(\mathbb{N})$ that $\langle \mathbf{f}, \mathbf{a} \rangle_{\ell_1(\mathbb{N})} = \langle \mathbf{a}, \mathbf{f} \rangle_{c_0(\mathbb{N})}$. For this reason, we will drop the subscript and simply write $\langle \mathbf{f}, \mathbf{a} \rangle_{\ell_1(\mathbb{N})} = \langle \mathbf{f}, \mathbf{a} \rangle$ and $\langle \mathbf{a}, \mathbf{f} \rangle_{c_0(\mathbb{N})} = \langle \mathbf{a}, \mathbf{f} \rangle$. Moreover, since $\langle \mathbf{f}, \mathbf{a} \rangle = \langle \mathbf{a}, \mathbf{f} \rangle$ according to (4.7), we will use $\langle \mathbf{f}, \mathbf{a} \rangle$ and $\langle \mathbf{a}, \mathbf{f} \rangle$ interchangeably when no ambiguity may raise.

Proposition 4.5 *If $\ell_1(\mathbb{N})'$ denotes the δ-dual space of $\ell_1(\mathbb{N})$, then $\ell_1(\mathbb{N})' = c_0(\mathbb{N})$.*

Proof: We first show that $c_0(\mathbb{N}) \subseteq \ell_1(\mathbb{N})'$. Let $\mathbf{a} := [a_1, a_2, \dots] \in c_0(\mathbb{N})$ be arbitrary. For all $\mathbf{f} := [f_1, f_2, \dots] \in \ell_1(\mathbb{N})$, we have that

$$\langle \mathbf{a}, \mathbf{f} \rangle = \sum_{j \in \mathbb{N}} a_j f_j = \sum_{j \in \mathbb{N}} a_j \delta_j(\mathbf{f}).$$

This implies that $\mathbf{a} = \sum_{j \in \mathbb{N}} a_j \delta_j$. Namely, $\mathbf{a}$ is a linear combination of the point-evaluation functionals δ_j, for $j \in \mathbb{N}$. Hence, $c_0(\mathbb{N}) \subseteq \ell_1(\mathbb{N})'$.

Conversely, we establish that $\ell_1(\mathbb{N})' \subseteq c_0(\mathbb{N})$. Let $\mathbf{a} \in \ell_1(\mathbb{N})'$ be arbitrary. Since $\ell_1(\mathbb{N})'$ is the completion of the linear span of all the point-evaluation functionals δ_j on $\ell_1(\mathbb{N})$ for $j \in \mathbb{N}$, under the norm of $\ell_\infty(\mathbb{N})$, there exist $k_n^j \in \mathbb{N}$, for $j = 1, 2, \dots, K_n$, $n \in \mathbb{N}$ such that sequences $\mathbf{a}_n := \sum_{j=1}^{K_n} \gamma_{k_n^j} \delta_{k_n^j}$ satisfies

$$\lim_{n \to \infty} \|\mathbf{a} - \mathbf{a}_n\|_\infty = 0. \tag{4.8}$$

Clearly, $\mathbf{a}_n \in c_0(\mathbb{N})$ for all $n \in \mathbb{N}$. Since $c_0(\mathbb{N})$ is a Banach space, equation (4.8) ensures that $\mathbf{a} \in c_0(\mathbb{N})$, which implies that $\ell_1(\mathbb{N})' \subseteq c_0(\mathbb{N})$. □

Proposition 4.5 guarantees that $\ell_1(\mathbb{N})' = c_0(\mathbb{N})$, which allows us to identify a reproducing kernel for the RKBS $\ell_1(\mathbb{N})$. To this end, we define a function $K : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$, which is a semi-infinite matrix. Specifically, for $i, j \in \mathbb{N}$, we let $\delta_{i,j} := 1$ if $i = j$ and 0 otherwise. We then define

$$K := [K_{i,j} : i, j \in \mathbb{N}] \quad \text{where} \quad K_{i,j} := \delta_{i,j}, \quad \text{for} \quad i, j \in \mathbb{N}. \tag{4.9}$$

Theorem 4.6 *The function $K := [K_{i,j} : i, j \in \mathbb{N}]$ defined by (4.9) is a reproducing kernel for RKBS $\ell_1(\mathbb{N})$.*

Proof: According to Proposition 4.5, we have that $\ell_1(\mathbb{N})' = c_0(\mathbb{N})$, which is a Banach space of functions on $\mathbf{X}' := \mathbb{N}$. By the definition (4.9) of the function K, we find that $K(i, \cdot) = K_{i,\cdot} = \delta_{i,\cdot} \in c_0(\mathbb{N})$, for all $i \in \mathbb{N}$. Moreover, for all $\mathbf{f} := [f_1, f_2, \dots] \in \ell_1(\mathbb{N})$, we have for all $i \in \mathbb{N}$ that

$$f_i = \sum_{j \in \mathbb{N}} \delta_{i,j} f_j = \sum_{j \in \mathbb{N}} K(i, j) f_j = \langle K(i, \cdot), \mathbf{f} \rangle .$$

This proves the reproducing property. By Definition 4.2 with $\mathcal{B} := \ell_1(\mathbb{N})$ and $\mathcal{B}' := c_0(\mathbb{N})$, we conclude that the function K is a reproducing kernel for $\ell_1(\mathbb{N})$. □

Note that Theorem 4.6 gives the simplest example of the RKKS, since any principal (and finite) sub-matrix of the (infinite operator) K is the identity matrix. This result is not covered by the setting described in [58]. In the same manner, we consider the predual space $c_0(\mathbb{N})$.

Theorem 4.7 *The space $c_0(\mathbb{N})$ is an RKBS on $\mathbb{N}$.*

Proof: We establish that the point-evaluation functionals on $c_0(\mathbb{N})$ are continuous. All point-evaluation functionals on space $c_0(\mathbb{N})$ are δ_j, for all $j \in \mathbb{N}$. Hence, we observe for all $\mathbf{a} \in c_0(\mathbb{N})$ that

$$|\delta_j(\mathbf{a})| = |a_j| \leq \|\mathbf{a}\|_\infty, \quad \text{for all} \quad j \in \mathbb{N}.$$

This ensures that the point-evaluation functionals on $c_0(\mathbb{N})$ are continuous. Again, by Definition 4.1, $c_0(\mathbb{N})$ is an RKBS over $\mathbb{N}$. □

We next identify a reproducing kernel for the RKBS $c_0(\mathbb{N})$ of functions defined on $\mathbf{X}' := \mathbb{N}$.

Theorem 4.8 *The function $K := [K_{i,j} : i, j \in \mathbb{N}]$ defined by (4.9) is a reproducing kernel for $c_0(\mathbb{N})$. The space $c_0(\mathbb{N})$ is an adjoint RKBS of $\ell_1(\mathbb{N})$, the spaces $\ell_1(\mathbb{N})$, $c_0(\mathbb{N})$ form a pair of RKBSs, and the kernel K is symmetric and positive semi-definite.*

Proof: By Theorem 4.7, $c_0(\mathbb{N})$ is an RKBS on $\mathbf{X}' := \mathbb{N}$. Furthermore, we have that $K(\cdot, j) := K_{\cdot,j} = \delta_{\cdot,j} \in \ell_1(\mathbb{N})$, for all $j \in \mathbb{N}$. Moreover, we have the dual reproducing property that for all $\mathbf{a} \in c_0(\mathbb{N})$,

$$a_j = \sum_{i \in \mathbb{N}} \delta_{i,j} a_i = \sum_{i \in \mathbb{N}} K(i, j) a_i = \langle K(\cdot, j), \mathbf{a} \rangle .$$

It follows from Definition 4.2 that the function K defined by (4.9) is a reproducing kernel for the RKBS $c_0(\mathbb{N})$. Therefore, $c_0(\mathbb{N})$ is an adjoint RKBS of $\ell_1(\mathbb{N})$.

Clearly, spaces $\ell_1(\mathbb{N})$, $c_0(\mathbb{N})$ with the kernel K satisfy Hypotheses (H1)-(H2). The positive definiteness of K can be seen immediately since for any $n \in \mathbb{N}$, $\mathbf{x}_i \in \mathbb{N}$, $i = 1, 2, \ldots, n$, the kernel matrix $K_n := [K(\mathbf{x}_i, \mathbf{x}_j) : i, j = 1, 2, \ldots, n]$ is the identity matrix. □

Several nontrivial RKBSs may be found in [46, 53, 55, 58]. In particular, an RKBS isometrically isomorphic to the space $\ell_1(\mathbb{N})$ was constructed in [46]. We now briefly describe a RKBS on a continuous function space. We choose $\mathbf{X}$ as a locally compact Hausdorff space. By $C_0(\mathbf{X})$ we denote the space of all continuous functions $f : \mathbf{X} \to \mathbb{R}$ such that for each $\epsilon > 0$, the set $\{\mathbf{x} \in \mathbf{X} : |f(\mathbf{x})| \geq \epsilon\}$ is compact. We equip the maximum norm on $C_0(\mathbf{X})$, that is,

$$\|f\|_\infty := \sup_{\mathbf{x} \in \mathbf{X}} |f(\mathbf{x})|, \quad \text{for all} \quad f \in C_0(\mathbf{X}).$$

The Riesz-Markov representation theorem states that the dual space of $C_0(\mathbf{X})$ is isometrically isomorphic to the space $\mathfrak{M}(\mathbf{X})$ of the real-valued regular Borel measures on $\mathbf{X}$ endowed with the total variation norm $\|\cdot\|_{\mathrm{TV}}$. We suppose that $K : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ satisfies $K(\mathbf{x}, \cdot) \in C_0(\mathbf{X})$ for all $\mathbf{x} \in \mathbf{X}$ and the density condition $\overline{\mathrm{span}}\{K(\mathbf{x}, \cdot) : \mathbf{x} \in \mathbf{X}\} = C_0(\mathbf{X})$. We then introduce the space of functions on $\mathbf{X}$ by

$$\mathcal{B} := \left\{ f_\mu := \int_{\mathbf{X}} K(\cdot, \mathbf{x}) d\mu(\mathbf{x}) : \mu \in \mathfrak{M}(\mathbf{X}) \right\} \tag{4.10}$$

equipped with $\|f_\mu\|_{\mathcal{B}} := \|\mu\|_{\mathrm{TV}}$. Clearly, $\mathcal{B}$ defined by (4.10) is a Banach space of functions defined on $\mathbf{X}$. It was established in [53] that $\mathcal{B}$ is an RKBS, the δ-dual $\mathcal{B}'$ of $\mathcal{B}$ is isometrically isomorphic to the space $C_0(\mathbf{X})$. Moreover, the function K is a reproducing kernel for the RKBS $\mathcal{B}$ in the sense of Definition 4.2. More information about the space $\mathcal{B}$ may be found in [3, 27, 47, 53].

Last decade has witnessed the rapid development of the theory of the RKBS and its applications since the publication of the original paper [58]. Continued understanding proper definition and construction of the RKBS and related theoretical issues remains research topics of great interest

[18, 20, 22, 45, 46, 50, 51, 55, 60, 61, 62]. In particular, construction of RKBSs was proposed in [55] by using Mercer's kernels. A class of vector-valued reproducing kernel Banach spaces with the ℓ_1 norm was constructed in [26] and it was used in multi-task learning. Separability of RKBSs was studied in [34]. Statistical margin error bounds were given in [8] for L_1-norm support vector machines. Lipschitz implicit function theorems in RKBSs were established in [43]. A converse sampling theorem in RKBSs was given in [7]. The notion of the RKBS has been used in statistics [39, 44], game theory [48], and machine learning [3, 11, 17, 25, 35, 52].

5 Learning in a Banach Space

Most of learning methods may be formulated as a minimum norm interpolation problem or a regularization problem. We consider both of the cases in this section. We will focus on solution representation of these learning problems.

Recently, representer theorems of learning methods in Banach spaces received considerable attention. In the framework of a semi-inner-product RKBS [58], the representer theorem was derived from the dual elements and the semi-inner-product [58, 61]. In [55], for a reflexive and smooth RKBS, a representer theorem of a solution of the regularized learning problem was established using the Gâteaux derivative of the norm function and the reproducing kernel. In addition, the representer theorem was generalized to a non-reflexive and non-smooth Banach space which has a predual space [22, 50, 51, 52]. Due to lack of the Gâteaux derivative, other tools such as the subdifferential of the norm function and the duality mapping need to be used to describe the representer theorem.

In this section, we assume that $\mathcal{B}$ is a general Banach space with the dual space $\mathcal{B}^*$ or predual space $\mathcal{B}_*$ unless it is stated otherwise. We first study the minimum norm interpolation problem. Given $\mathbf{y} := [y_1, y_2, \ldots, y_m] \in \mathbb{R}^m$ and $\nu_j \in \mathcal{B}^*$, $j = 1, 2, \ldots, m$, we seek $f^* \in \mathcal{B}$ such that

$$\|f^*\|_{\mathcal{B}} = \inf\big\{\|f\|_{\mathcal{B}} : f \in \mathcal{B},\ \langle \nu_j, f\rangle_{\mathcal{B}} = y_j,\ j = 1, 2, \ldots, m\big\}. \tag{5.1}$$

While minimum norm interpolation in a Hilbert space is a classical topic [14, 15], its counterpart in a Banach space has drawn much attention in the literature [9, 10, 30, 50, 52] due to its connection with compressed sensing [6, 16]. An existence theorem of a solution of the minimum norm interpolation problem was established in Proposition 1 of [52].

Minimum norm interpolation problem (5.1) with $\mathcal{B} := \ell_1(\mathbb{N})$ was investigated in [10], which may be stated as follows: Given $y_j \in \mathbb{R}$ and $\mathbf{a}_j \in c_0(\mathbb{N})$, $j = 1, 2, \ldots, m$, we seek $\mathbf{x}_1^* \in \ell_1(\mathbb{N})$ such that

$$\|\mathbf{x}_1^*\|_1 = \inf\big\{\|\mathbf{x}\|_1 : \mathbf{x} \in \ell_1(\mathbb{N}),\ \langle \mathbf{a}_j, \mathbf{x}\rangle = y_j,\ j = 1, 2, \ldots, m\big\}. \tag{5.2}$$

Problem (5.2) is an infinite dimensional version of compressed sensing considered in [6, 16]. According to [10], it may be reduced to a finite dimensional linear programming problem by a duality approach, leading to a sparse solution. Sparse learning methods in $\ell_1(\mathbb{N})$ were studied in [2, 10, 50, 52].

The solution of problem (5.2) with interpolation conditions (3.12) has a sparse solution, which is given by $\mathbf{x}_1^* = \big(\frac{7}{2}, -1, 0, 0, \ldots\big)$ (see, [10] for details about this example). This solution has only two nonzero components while the solution of the minimum ℓ_2 norm interpolation problem (3.11) with the same interpolation conditions has infinite numbers of nonzero components. The solution $\mathbf{x}_1^*$ is sparse because the Banach space $\ell_1(\mathbb{N})$ promotes sparsity while the solution $\mathbf{x}^*$, given by (3.13), of problem (3.11) is dense because the Hilbert space $\ell_2(\mathbb{N})$ does not promote sparsity. Several forms of the representer theorem for a solution of problem (5.1) were established in [52].

We now turn to investigating the regularized learning problem in a Banach space. Such a problem may be formulated from the ill-posed problem

$$\mathcal{L}(f) = \mathbf{y}, \tag{5.3}$$

where $\mathbf{y}$ is given data and $\mathcal{L}$ represents either a physical system or a learning system. We define a data fidelity term $\mathcal{Q}_{\mathbf{y}}(\mathcal{L}(f))$ from (5.3) by using a loss function $\mathcal{Q}_{\mathbf{y}} : \mathbb{R}^m \to \mathbb{R}_+$ and choose a regularization term $\lambda\varphi(\|f\|_{\mathcal{B}})$ with a regularizer $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$. We then solve the regularization problem

$$\inf\{\mathcal{Q}_{\mathbf{y}}(\mathcal{L}(f)) + \lambda\varphi(\|f\|_{\mathcal{B}}) : f \in \mathcal{B}\}, \tag{5.4}$$

where λ is a positive regularization parameter. Here, $\mathcal{Q}_{\mathbf{y}}(\mathbf{u})$, for $\mathbf{u} \in \mathbb{R}^m$, measures the "lost" of $\mathbf{u}$ from given data $\mathbf{y}$. For example, one may choose $\mathcal{Q}_{\mathbf{y}}(\mathbf{u}) := \|\mathbf{u} - \mathbf{y}\|_2$. Regularized learning problems in Banach spaces were originated in [30] and since then, desired representer theorems for such learning problems have received considerable attention in the literature. For existence results of a solution of problem (5.4), the readers are referred to [52].

When the operator $\mathcal{L} : \mathcal{B} \to \mathbb{R}^m$ in (5.3) is defined by

$$\mathcal{L}(f) := \langle \nu_j, f\rangle_{\mathcal{B}}, \quad \text{for} \quad j = 1, 2, \ldots, m, \tag{5.5}$$

the regularization problem (5.4) is intimately related to the minimum norm interpolation problem (5.1). Their relation is described in the next proposition.

Proposition 5.1 *Suppose that $\mathcal{B}$ is a Banach space with the dual space $\mathcal{B}^*$, $\nu_j \in \mathcal{B}^*$, $j \in \mathbb{N}_m$ and $\mathcal{L}$ is defined by* (5.5)*. For a given $\mathbf{y}_0 \in \mathbb{R}^m$, let $\mathcal{Q}_{\mathbf{y}_0} : \mathbb{R}^m \to \mathbb{R}_+$ be a loss function, $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ be an increasing regularizer and $\lambda > 0$. Let $\hat{f} \in \mathcal{B}$ be a solution of the regularization problem (5.4) with $\mathbf{y} := \mathbf{y}_0$. Then the following statements hold true:*

(i) A solution $\hat{g} \in \mathcal{B}$ of problem (5.1) *with $\mathbf{y} := \mathcal{L}(\hat{f})$ is a solution of the regularization problem (5.4) with $\mathbf{y} := \mathbf{y}_0$.*

(ii) If φ is strictly increasing, then $\hat{f}$ is a solution of problem (5.1) *with $\mathbf{y} := \mathcal{L}(\hat{f})$.*

Statement (ii) of Proposition 5.1 was claimed in [30] without details of proof. A complete proof for Proposition 5.1 may be found in [52].

We are interested in characterizing a solution of the regularization problem (5.4) with an operator $\mathcal{L}$ defined by (5.5) having an adjoint operator $\mathcal{L}^* : \mathbb{R}^m \to \mathcal{B}_*$. The following two representer theorems are due to [52]. For a given $\mathbf{y}_0 \in \mathbb{R}^m$, let $\mathcal{Q}_{\mathbf{y}_0} : \mathbb{R}^m \to \mathbb{R}_+$ be a loss function. We first consider the case when the Banach space $\mathcal{B}$ has a smooth predual.

Theorem 5.2 *Suppose that $\mathcal{B}$ is a Banach space having the smooth predual space $\mathcal{B}_*$, and $\nu_j \in \mathcal{B}_*$, $j \in \mathbb{N}_m$. Let $\mathcal{G}_*(\nu)$ denote the Gâteaux derivative of the norm $\|\cdot\|_{\mathcal{B}_*}$ at $\nu \in \mathcal{B}_*$. Let $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ be a strictly increasing regularizer and $\lambda > 0$. Then*

$$f_0 := \|\mathcal{L}^*(\hat{\mathbf{c}})\|_{\mathcal{B}_*}\mathcal{G}_*(\mathcal{L}^*(\hat{\mathbf{c}})), \quad \hat{\mathbf{c}} \in \mathbb{R}^m, \tag{5.6}$$

is a solution of the regularization problem (5.4) with $\mathbf{y} := \mathbf{y}_0$ if and only if $\hat{\mathbf{c}} \in \mathbb{R}^m$ is a solution of the finite dimensional minimization problem

$$\inf\{\mathcal{Q}_{\mathbf{y}_0}(\|\mathcal{L}^*(\mathbf{c})\|_{\mathcal{B}_*}\mathcal{L}(\mathcal{G}_*(\mathcal{L}^*(\mathbf{c}))) + \lambda\varphi(\|\mathcal{L}^*(\mathbf{c})\|_{\mathcal{B}_*}) : \mathbf{c} \in \mathbb{R}^m\}. \tag{5.7}$$

We now consider the case when the predual of the Banach space $\mathcal{B}$ is not necessarily smooth.

Theorem 5.3 *Suppose that $\mathcal{B}$ is a Banach space having the predual space $\mathcal{B}_*$, $\nu_j \in \mathcal{B}_*$, $j \in \mathbb{N}_m$. Let $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ be a regularizer and $\lambda > 0$.*

(i) If φ is increasing, then there exists a solution f_0 of the regularization problem (5.4) with $\mathbf{y} := \mathbf{y}_0$ such that

$$f_0 \in \rho \partial \| \cdot \|_{\mathcal{B}_*} \left(\sum_{j \in \mathbb{N}_m} c_j \nu_j \right), \tag{5.8}$$

for some $c_j \in \mathbb{R}$, $j \in \mathbb{N}_m$, with $\rho := \| \sum_{j \in \mathbb{N}_m} c_j \nu_j \|_{\mathcal{B}_*}$, where $\partial \| \cdot \|_{\mathcal{B}_*}$ denotes the subdifferential of the norm.

(ii) If φ is strictly increasing, then every solution f_0 of the regularization problem (5.4) with $\mathbf{y} := \mathbf{y}_0$ satisfies (5.8) *for some $c_j \in \mathbb{R}$, $j \in \mathbb{N}_m$.*

Item (i) of Theorem 5.3 indicates that there exists a solution f_0 of the regularization problem (5.4) with $\mathbf{y} := \mathbf{y}_0$ that satisfies (5.8), which is a generalization of the stationary point condition to minimization problems with non-differentiable objective function. The essence of Theorems 5.2 and 5.3 is that a solution f_0, defined by (5.6) or (5.8), of the *infinite* dimensional regularization problem (5.4) is determined completely by a *finite* number of parameters. Values of these parameters can be obtained by solving either a *finite dimensional* optimization problem (5.7) or a nonlinear system. When the space $\mathcal{B}$ is a Hilbert space, the nonlinear system reduces to a linear one. In particular, when $\mathcal{B}$ is an RKBS with a kernel K, the functionals ν_j in Theorems 5.2 and 5.3 have convenient representations in terms of the kernel K. We will demonstrate this point later in this section. These representer theorems serve as a base for further development of efficient numerical solvers of the regularization problem (5.4). One may find more discussions about representer theorems in Banach spaces in [41].

Not all Banach spaces will produce sparse learning solutions. Only Banach spaces with certain geometric features can lead to a sparse learning solution. It has been shown in [10, 50] that minimum norm interpolation and regularization problems in $\ell_1(\mathbb{N})$ have sparse solutions. As it is shown in Theorems 4.4 and 4.6, the space $\ell_1(\mathbb{N})$ is an RKBS on $\mathbb{N}$ with the kernel K defined by (4.9). Hence, a learning solution in the RKBS $\ell_1(\mathbb{N})$ can be expressed in terms of the kernel K. Integrating Theorems 4.4 and 4.6 with those known in the literature [50], we have the following sparse representer theorem for a learning solution in $\ell_1(\mathbb{N})$. For any learning solution $\mathbf{f} \in \ell_1(\mathbb{N})$, there exists a vector of positive integers $[n_1, \ldots, n_q] \in \mathbb{N}^q$ with $q \leq m$ such that

$$\mathbf{f} = \sum_{i \in \mathbb{N}_q} c_i K(\cdot, n_i). \tag{5.9}$$

Clearly, the solution $\mathbf{f}$ has only q nonzero components. Here, the vector $[n_1, \ldots, n_q] \in \mathbb{N}^q$ represents the support of the learning solution $\mathbf{f}$ defined by (5.9) and the coefficients c_i are determined by the fidelity data. The paper [10] suggests that we can first convert the minimum norm interpolation problem to its dual problem which allows us to identify the positive integer q and the support $[n_1, \ldots, n_q] \in \mathbb{N}^q$, and then solve a linear system to obtain the coefficients c_i. The idea of [10] has been extended in [9] to solve a class of regularization problems.

We now return to the binary classification problem and describe a model based on an RKBS for it. We choose the RKBS $\mathcal{B}$ defined by (4.10) with $\mathbf{X} = \mathbb{R}^d$ and

$$K(\mathbf{x}, \mathbf{y}) := \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{2\mu^2} \right), \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

as the hypothesis space. The classification problem in the RKBS $\mathcal{B}$ may be described by

$$\min\left\{\frac{1}{n}\sum_{k=1}^{n} L\left(y_k, f(\mathbf{x}_k)\right) + \lambda\|f\|_{\mathcal{B}} : f \in \mathcal{B}\right\}, \tag{5.10}$$

where the fidelity term is the same as in problem (3.6). A representer theorem for a solution of the learning method (5.10) was obtained in [53]. To state it, for a function $f : \mathbb{R}^d \to \mathbb{R}$ with $\|f\|_\infty < \infty$, we define the set $\mathcal{N}(f) := \{\mathbf{x} \in \mathbb{R}^d : |f(\mathbf{x})| = \|f\|_\infty\}$. We denote by h^* a solution of minimization problem (5.10) and let $\mathbf{y} := [h^*(\mathbf{x}_k) : k \in \mathbb{N}_n] \in \mathbb{R}^n$. It was proved in [53] that there exists a solution of (5.10) having the form

$$f^* := \sum_{k=1}^{n} \alpha_k K(\cdot, \mathbf{z}_k), \quad \text{for some } \ \mathbf{z}_k \in \mathcal{N}\left(\sum_{j=1}^{n} c_j^* K(\mathbf{x}_j, \cdot)\right) \ \text{ and } \alpha_k \in \mathbb{R},\ k \in \mathbb{N}_n, \tag{5.11}$$

where $\mathbf{c}^* \in \mathbb{R}^n$ is a solution of optimization problem

$$\max\left\{\mathbf{y}^\top \mathbf{c} : \mathbf{c} = [c_j : j \in \mathbb{N}_n] \in \mathbb{R}^n, \left\|\sum_{j=1}^{n} c_j K(\mathbf{x}_j, \cdot)\right\|_\infty = 1\right\}. \tag{5.12}$$

Although the points $\mathbf{z}_k$, $k \in \mathbb{N}_n$ may not be the same as the points $\mathbf{x}_k$, $k \in \mathbb{N}_n$, motivated from (5.11), we seek a solution of the regularization problem (5.10) in the form $\tilde{f}^* := \sum_{k=1}^{n} \alpha_k K(\cdot, \mathbf{x}_k)$. By plugging $\tilde{f}^*$ into (5.10) and adding a bias term $b \in \mathbb{R}$, we obtain the ℓ_1-SVM classification model has the form

$$\min\left\{L_D(\alpha, b) + \lambda\|\alpha\|_1 : \alpha \in \mathbb{R}^n, b \in \mathbb{R}\right\}, \tag{5.13}$$

where

$$L_D(\alpha, b) := \sum_{j \in \mathbb{N}_n} \max\left\{1 - y_j\left(\sum_{k \in \mathbb{N}_n} \alpha_k K(\mathbf{x}_k, \mathbf{x}_j) + b\right), 0\right\}.$$

The discrete minimization problem (5.13) is the same as (3.14). We comment that the space $\mathbb{R}^n$ with the ℓ_1-norm is a finite dimensional RKBS according to Theorem 4.4.

Next, we elaborate the numerical solution of the regularized learning problem. Due to the use of a Banach space as the hypothesis space in learning, we are inevitably facing solving the regularized learning problem (5.4) with the Banach norm. Major challenges for solving this problem include infinite dimensionality, nonlinearity and nondifferentiability. The regularization problem (5.4) by nature is infinite dimensional, since we seek a solution in the hypothesis space which is an infinite dimensional Banach space, even though the fidelity term is determined by a finite number of data points. The use of a Banach space as a regularization term leads to nonlinearity. Moreover, employing a sparsity promotion norm results in a nondifferentiable optimization problem. Representer theorems presented in this section for a solution of this problem indicate that the regularization problem (5.4) is essentially of finite dimension. However, how the representer theorems can be used in developing practical algorithms requires further investigation. A recent paper [9] is making an attempt toward this direction. Specifically, we reformulated the regularization problem (5.4) in the following way: When the fidelity term of the regularization problem (5.4) has a form in a Banach norm and a regularization term in another Banach norm, we construct a direct sum space based on the two Banach spaces for the data fidelity term and the regularization term, and then recast the objective function as the norm of a suitable quotient space of the direct sum space. In this way, we express the original regularized problem as a best approximation problem in the direct sum

space, which is in turn reformulated as a dual optimization problem in the dual space of the direct sum space. The dual problem is to find the maximum of a linear function on a convex polytope, which is of finite dimension and may be solved by numerical methods such as linear programming. The solution of the dual optimization problem provides related extremal properties of norming functionals, by which the original problem is reduced to a finite dimensional optimization problem, which is then reformulated as a finite dimensional fixed-point equation [24, 31], solved by iterative algorithms.

6 Numerical Results

This section is devoted to a numerical example about classification of handwriting digits by using the ℓ_1 SVM model (5.13) with a comparison to the ℓ_2 SVM model. The handwriting digit classification by using the ℓ_1-SVM was studied in [28].

In our experiment, we use the database MNIST of handwriting digits, which is originally composed of $60,000$ training samples and $10,000$ testing samples of the digits "0" through "9". We consider classifying two handwriting digits "7" and "9" taken from MNIST. Handwriting digits "7" and "9" are often easy to cause obfuscation in comparison to other pairs of digits. We take $8,141$ training samples and $2,037$ testing samples of these two digits from the database. Specifically, we consider training data set $D := \{(\mathbf{x}_j, y_j) : j \in \mathbb{N}_n\} \subset \mathbb{R}^d \times \{\pm 1\}$, where $\mathbf{x}_j$ are images of digits 7 or 9. We wish to find a function that defines a hyperplane separating the data D into two groups with labels $y_j = 1$ (digit 7) and $y_j = -1$ (digit 9).

We employ the model (5.10) and its related discrete form (5.13) described in section 5 for our experiment. For a comparison purpose, we also test the learning model (5.13) with the squared loss function

$$L_D(\alpha, b) := \frac{1}{2} \sum_{j \in \mathbb{N}_n} \left(\sum_{k \in \mathbb{N}_n} \alpha_k K(\mathbf{x}_k, \mathbf{x}_j) + b - y_j \right)^2$$

as a fidelity term in problem (5.13). We compare the classification performance and solution sparsity of problem (5.13) with those of its Hilbert space counterpart (3.6), which leads to problem (5.13) with the ℓ_1-norm being replaced by the ℓ_2-norm.

The objective functions of both the minimization problem (5.13) and its ℓ_2-norm counterpart are not differentiable and thus they can not be solved by a gradient based iteration algorithm. Instead, we employ the Fixed Point Proximity Algorithm (FPPA) developed in [24, 31] to solve these non-smooth minimization problems. Specifically, we first use the Fermat rule to reformulate a solution of the minimization (5.13) as a solution of an inclusion relation defined by the subdifferential of the objective function of (5.13). We then rewrite the inclusion relation as an equivalent fixed-point equation involving the proximity operator of the functions L_D and $\|\cdot\|_1$ which appear in the objective function of (5.13). The fixed-point equation leads to the FPPA, for solving (5.13), which is guaranteed to converge under a mild condition. We refer the interested readers to [24, 31] for more information on the FPPA.

Numerical results are reported in Tables I-IV. In these tables, λ represents values of the regularization parameter, #NZ the sparsity level (number of nonzero entries), and TRA and TEA the learning accuracy (correction rate) for training datasets and testing datasets, respectively. Specifically,

$$\text{TRA} := \frac{NP_{\text{train}}}{N_{\text{train}}} \quad \text{and} \quad \text{TEA} := \frac{NP_{\text{test}}}{N_{\text{test}}},$$

where NP_{train} and NP_{test} denote the number of predicted labels equal to the given labels of the

λ	0.5	0.7	1.4	1.8	4.0	6.0	487.7
#NZ	1762	880	481	340	187	142	0
TRA	99.46%	99.19%	98.61%	98.34%	97.57%	97.16%	50.67%
TEA	98.77%	98.82%	98.48%	98.38%	97.79%	97.40%	50.47%

Table I: Classification using the Gaussian Kernel with $\mu = 4.8$: Results for the ℓ_1-SVM model with the square loss function.

λ	0.005	1.0	2.3	6.0	43.0	96.0	490
#NZ	8,141	8,141	8,141	8,141	8,141	8,141	8,141
TRA	100%	99.50%	99.18%	98.69%	97.56%	97.16%	95.93%
TEA	99.21%	98.87%	98.72%	98.43%	97.69%	97.25%	96.41%

Table II: Classification using the Gaussian Kernel with $\mu = 4.8$: Results for the ℓ_2-SVM model with the square loss function.

training and testing data points, respectively, and N_{train} and N_{test} denote the total number of the training and testing data points, respectively.

Explanations on the choice of the regularization parameter λ are in order. In Table I, we randomly selected seven different values of λ from the interval $(0, \lambda_{\max}]$, where $\lambda_{\max}$ is determined by Remark 4.12 in [28]. In each of the remaining tables, we selected seven different values of λ in an increasing order such that the ℓ_1-SVM classification model (5.13) and the related ℓ_2 model produce the same accuracy. In this way, we can compare the sparsity of the corresponding solutions of these two models. Moreover, the largest λ in Table I is given by the smallest value such that $\#NZ = 0$, since for the ℓ_1-SVM model with the square loss function, we have Remark 4.12 of [28] to determine such a λ value. However, in Table III, which is for the ℓ_1-SVM model with the hinge loss function, we do not have a theoretical result similar to Remark 4.12 of [28] to determine the smallest value that ensures $\#NZ = 0$ and hence, we empirically choose the λ value such that $\#NZ = 0$.

The numerical results presented in Tables I-IV confirm that the two models generate learning results with comparable accuracy, while the solution of the ℓ_1-SVM model can have different levels of sparsity according to values of λ, but that of the ℓ_2-SVM model is always dense disregarding the choice of λ. In passing, we point it out that the hinge loss function outperforms the square loss function as a fidelity term for the classification problem.

7 Remarks on Future Research Problems

In this section, we elaborate several open mathematical problems critical to the theory of RKBSs and machine learning methods on a Banach space.

It has been demonstrated that learning in certain Banach spaces leads to a sparse solution and learning in a Hilbert space yields a dense solution. Sparse learning solutions can save tremendous computing effort and storage when they are used in decision making processes. However, learning in Banach spaces introduces new challenges. First of all, the theory of the RKBS is far away from completion. Construction of reproducing kernels for RKBSs suitable for different applications is needed. For example, related to the definition of the reproducing kernel, we have the following specific problem: Suppose that $\mathcal{B}$ is a Banach space of functions on $\mathbf{X}$ and point-evaluation functionals on $\mathcal{B}$ are all continuous. Let $\mathcal{B}'$ denote the the completion of the linear span of all the point-evaluation functionals on $\mathcal{B}$ under the norm of $\mathcal{B}^*$. In general, a Banach space $\mathcal{B}$ of functions

λ	0.1	0.2	1	2	4	10	435
♯NZ	552	481	167	92	56	34	0
TRA	99.99%	99.99%	99.08%	98.17%	97.53%	96.30%	50.67%
TEA	98.72%	98.77%	98.38%	98.09%	97.45%	96.27%	50.47%

Table III: Classification using the Gaussian Kernel with $\mu = 4$: Results for the ℓ_1-SVM model with the hinge loss function.

λ	0.001	0.2	2	12	22	47	128
♯NZ	8,141	8,141	8,141	8,141	8,141	8,141	8,141
TRA	99.90%	99.84%	99.18%	98.13%	97.68%	97.07%	96.31%
TEA	99.02%	98.97%	98.33%	97.89%	97.59%	97.05%	96.41%

Table IV: Classification using the Gaussian Kernel with $\mu = 4$: Results for the ℓ_2-SVM model with hinge loss.

is not isometrically isomorphic to its dual $\mathcal{B}^*$ or its δ-dual $\mathcal{B}'$. We take $\mathcal{B} := \ell_1(\mathbb{N})$ as an example. We have known from section 4 that $\mathcal{B}^* = \ell_\infty(\mathbb{N})$ and $\mathcal{B}' = c_0(\mathbb{N})$. Clearly, the space $\ell_1(\mathbb{N})$ is neither isometrically isomorphic to $\ell_\infty(\mathbb{N})$ nor $c_0(\mathbb{N})$. In Definition 4.2 of a kernel, we need the assumption that $\mathcal{B}'$ is isometrically isomorphic to a Banach space of functions on some set $\mathbf{X}'$. This assumption holds true when $\mathcal{B} := \ell_1(\mathbb{N})$ since $\mathcal{B}' := c_0(\mathbb{N})$ is a Banach space of functions (sequences) defined on $\mathbb{N}$. We are interested in knowing to what extent is this true in general?

Theorem 4.3 reveals that under Hypotheses (H1) and (H2), a reproducing kernel as defined in Definition 4.2 is positive semi-definite. Can the two hypotheses be weakened? We are curious to know how a given function of two arguments will lead to a pair of RKBSs.

A problem related to sparse learning in a Banach space may be stated as follows: It has been understood that the Banach space $\ell_1(\mathbb{N})$ promotes sparsity for a learning solution in the space. What is the essential geometric feature for a general Banach space of functions that can lead to a sparse learning solution in the space?

Finally, practical numerical algorithms for learning a function from an RKBS require systematical investigation.

Answers to these challenging issues would contribute to completing the theory of the RKBS and making it a useful hypothesis space for machine learning, hoping leading to practical numerical methods for learning in an RKBS.

Acknowledgement: This paper is based on the author's plenary invited lecture delivered online at the international conference "Functional Analysis, Approximation Theory and Numerical Analysis" that took place at Matera, Italy, July 5-8, 2022. The author is indebted to Professor Rui Wang for helpful discussion on issues related to the notion of the reproducing kernel Banach space, and to Professor Raymond Cheng for careful reading of the manuscript and providing a list of corrections. The author is grateful to two anonymous referees whose constructive comments improve the presentation of this paper. The author is supported in part by the US National Science Foundation under grants DMS-1912958 and DMS-2208386, and by the US National Institutes of Health under grant R21CA263876.

References

[1] N. Aronszajn, Theory of reproducing kernels, *Transactions of the American Mathematical Society*, **68** (1950), 337–404.

[2] S. Aziznejad and M. Unser, Multikernel regression with sparsity constraint, *SIAM Journal on Mathematics of Data Science* **3** (2021) 10.1137/20M1318882.

[3] F. Bartolucci, E. De Vito, L. Rosasco and S. Vigogna, Understanding neural networks with reproducing kernel Banach spaces, *Applied and Computational Harmonic Analysis* **62** (2023), 194–236.

[4] S. Bochner, Hilbert distances and positive definite functions, *Annals of Mathematics* **42** (1941), 647–656.

[5] C. de Boor and R. E. Lynch, On splines and their minimum properties, *Journal of Mathematics and Mechanics*, **15** (1966), 953–969.

[6] E. J. Candés, J. Romberg and T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, *IEEE Transactions on Information Theory*, **52** (2006), 489–509.

[7] H. Centeno and J. M. Medina, A converse sampling theorem in reproducing kernel Banach spaces, *Sampling Theory, Signal Processing, and Data Analysis*, **20** (2022), 8, 19 pages.

[8] L. Chen and H. Zhang, Statistical margin error bounds for L_1-norm support vector machines, *Neurocomputing*, **339** (2019), 210–216.

[9] R. Cheng, R. Wang, and Y. Xu, A duality approach to regularization learning problems in Banach spaces, preprint, 2022.

[10] R. Cheng and Y. Xu, Minimum norm interpolation in the $\ell_1(\mathbb{N})$ space, *Analysis and Applications*, **19** (2021), 21–42.

[11] P. L. Combettes, S. Salzo and S. Villa, Regularized learning schemes in feature Banach spaces, *Analysis and Applications*, **16** (2018), 1–54.

[12] C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning*, **20 (3)** (1995), 273–297.

[13] F. Cucker and S. Smale, On the mathematical foundations of learning, *Bulletin of the American Mathematical Society*, **39** (2002), 1–49.

[14] F. Deutsch, *Best Approximation in Inner Product Spaces*, Springer, New York, 2001.

[15] F. Deutsch, V. A. Ubhaya, J. D. Ward and Y. Xu, Constrained best approximation in Hilbert space III. Applications to n-convex functions, *Constructive Approximation*, **12** (1996), 361–384.

[16] D. L. Donoho, Compressed sensing, *IEEE Transactions on Information Theory*, **52** (2006), 1289–1306.

[17] G. E. Fasshauer, F. J. Hickernell and Q. Ye, Solving support vector machines in reproducing kernel Banach spaces with positive definite functions, *Applied Computational Harmonic Analysis*, **38** (2015), 115–139.

[18] K. Fukumizu, G. Lanckriet and B. K. Sriperumbudur, Learning in Hilbert vs. Banach spaces: A measure embedding viewpoint, Advances in Neural Information Processing Systems 24 (NIPS 2011).

[19] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, A kernel two-sample test, *Journal of Machine Learning Research*, **13** (2012), 723–773.

[20] P. G. Georgiev, L. Sánchez-González and P. M. Pardalos, Construction of pairs of reproducing kernel Banach spaces, in "Constructive Nonsmooth Analysis and Related Topics" the Springer Optimization and Its Applications book series (SOIA,volume 87), pp. 39–57, September 2013.

[21] T. Hoefler, D. A. Alistarh, N. Dryden, and T. Ben-Nun, The future of deep learning will be sparse, SIAM News, May 03, 2021.

[22] L. Huang, C. Liu, L. Tan and Q. Ye, Generalized representer theorems in Banach spaces, *Analysis and Applications*, **19** (2021), 125–146.

[23] G. S. Kimeldorf and G. Wahba, A correspondence between Bayesian estimation on stochastic processes and smoothing by splines, *Annals of Mathematical Statistics*, **41** (1970), 495–502.

[24] Q. Li, L. Shen, Y. Xu and N. Zhang, Multi-step fixed-point proximity algorithms for solving a class of optimization problems arising from image processing, *Advances in Computational Mathematics*, **41** (2015), 387–422.

[25] Z. Li, Y. Xu and Q. Ye, Sparse support vector machines in reproducing kernel Banach spaces, in "Contemporary Computational Mathematics - A Celebration of the 80th Birthday of Ian Sloan" pp. 869–887, Springer, 2018.

[26] R. Lin, G. Song and H. Zhang, Multi-task learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, *Journal of Complexity* **63** (2021), 101514.

[27] R. Lin, H. Zhang and J. Zhang, On reproducing kernel Banach spaces: Generic definitions and unified framework of constructions, *Acta Mathematica Sinica*, **38** (2022) 1459–1483.

[28] Q. Liu, R. Wang, Y. Xu and M. Yan, Parameter choices for sparse regularization with the ℓ_1 norm, *Inverse Problems* **39** (2023) 025004 (34pp)

[29] J. Mercer, Functions of positive and negative type and their connection with the theory of integral equations, *Philosophical Transactions of the Royal Society of London, Series A* **209** (1909), 415–446.

[30] C. A. Micchelli and M. Pontil, A function representation for learning in Banach spaces, in Learning theory, 255–269, Lecture Notes in Computer Science 3120, Springer, Berlin, 2004.

[31] C. A. Micchelli, L. Shen and Y. Xu, Proximity algorithms for image models: denoising, *Inverse Problems*, **27** (2011), 045009.

[32] C. A. Micchelli, Y. Xu and H. Zhang, Universal kernels, *Journal of Machine Learning Research*, **7** (2006), 2651–2667.

[33] K. Muandet, K. Fukumizu, B. Sriperumbudur and B. Schölkopf, (2017), Kernel mean embedding of distributions: A review and beyond, *Foundations and Trends® in Machine Learning* **10 (1-2)** (2017), 1–141.

[34] H. Owhadi and C. Scovel, Separability of reproducing kernel spaces, *Proceedings of American Mathematical Society*, **145** (2017), 2131–2138.

[35] R. Parhi and R. D. Nowak, Banach space representer theorems for neural networks and ridge splines, *Journal of Machine Learning Research*, **22** (2021), 1–40.

[36] M. J. D. Powell, *Approximation Theory and Methods*, 1st Edition, Cambridge University Press, Cambridge, 1981.

[37] M. Reed and B. Simon, *Functional Analysis*, Academic Press, New York, 1980.

[38] H. L. Royden, *Real Analysis*, 3rd Edition, Macmillan Publishing Company, New York, 1988.

[39] S. Salzo and J. A. K. Suykens, Generalized support vector regression: Duality and tensor-kernel representation, *Analysis and Applications*, **18** (2020), 149–183.

[40] B. Schölkopf, R. Herbrich and A. J. Smola, A generalized representer theorem, Computational Learning Theory, *Lecture Notes in Computer Science*, **2111** (2001), 416–426.

[41] K. Schlegel, When is there a representer theorem? *Advances in Computational Mathematics* **47**, 54 (2021).

[42] T. Schuster, B. Kaltenbacher, B. Hofmann and K. S. Kazimierski, *Regularization Methods in Banach Spaces*, Vol. 10 in "Radon Series on Computational and Applied Mathematics", De Gruyter, Berlin, 2012 (https://doi.org/10.1515/9783110255720).

[43] C. Shannon, On Lipschitz implicit function theorems in Banach spaces and applications, *Journal of Mathematical Analysis and Applications*, **494** (2021), 124589.

[44] B. Sheng and L. Zuo, Error analysis of the kernel regularized regression based on refined convex losses and RKBSs, *International Journal of Wavelets, Multiresolution and Information Processing*, **19** (2021), 2150012.

[45] G. Song and H. Zhang, Reproducing kernel Banach spaces with the ℓ_1 norm ii: error analysis for regularized least square regression, *Neural Computation*, **23** (2011), 2713-2729.

[46] G. Song, H. Zhang, F.J. Hickernell, Reproducing kernel Banach spaces with the ℓ_1 norm, *Applied and Computational Harmonic Analysis*, **34** (2013), 96-116.

[47] L. Spek, T. J. Heeringa and C. Brune, Duality for neural networks through reproducing kernel banach spaces, arXiv preprint arXiv:2211.05020, 2022.

[48] K. Sridharan and A. Tewari, Convex games in Banach spaces, In "Proceedings of the 23rd Annual Conference on Learning Theory," pp. 1–13, Omnipress, 2010.

[49] B. K. Sriperumbudur, K. Fukumizu and G. R. G. Lanckriet, Learning in Hilbert vs. Banach spaces: A measure embedding viewpoint, *Advances in Neural Information Processing Systems (Cambridge)*, MIT Press, pp. 1773–1781, 2011.

[50] M. Unser, Representer theorems for sparsity-promoting ℓ_1 regularization, *IEEE Transactions on Information Theory*, **62** (2016), 5167–5180.

[51] M. Unser, A unifying representer theorem for inverse problems and machine learning, *Foundations of Computational Mathematics*, **21** (2021), 941–960.

[52] R. Wang and Y. Xu, Representer theorems in Banach spaces: Minimum norm interpolation, regularized learning and semi-discrete inverse problems, *Journal of Machine Learning Research*, **22** (2021), 1–65.

[53] R. Wang, Y. Xu and M. Yan, Representer theorems for sparse learning in Banach spaces, preprint, 2023.

[54] W. Wang, S. Lu, B. Hofmann and J. Cheng, Tikhonov regularization with ℓ_0-term complementing a convex penalty: ℓ_1-convergence under sparsity constraints, *Journal of Inverse Ill-Posed Problems*, **27** (2019), 575–590.

[55] Y. Xu and Q. Ye, Generalized Mercer kernels and reproducing kernel Banach spaces, *Memoirs of the American Mathematical Society*, **258** (2019), Number 1243, 122 pages.

[56] Y. Xu and H. Zhang, Refinable kernels, *Journal of Machine Learning Research*, **8** (2007), 2083–2120.

[57] Y. Xu and H. Zhang, Refinement of reproducing kernels, *Journal of Machine Learning Research*, **10** (2009) 107–140.

[58] H. Zhang, Y. Xu and J. Zhang, Reproducing kernel Banach spaces for machine learning, *Journal of Machine Learning Research*, **10** (2009), 2741–2775.

[59] H. Zhang, Y. Xu and Q. Zhang, Refinement of operator-valued reproducing kernels, *Journal of Machine Learning Research*, **13** (2012), 91–136.

[60] H. Zhang and J. Zhang, Frames, Riesz bases, and sampling expansions in Banach spaces via semi-inner products, *Applied and Computational Harmonic Analysis*, **31** (2011), 1–25.

[61] H. Zhang and J. Zhang, Regularized learning in Banach spaces as an optimization problem: representer theorems, *Journal of Global Optimization*, **54** (2012), 235–250.

[62] H. Zhang and J. Zhang, Vector-valued reproducing kernel Banach spaces with applications to multi-task learning, *Journal of Complexity*, **29** (2013), 195–215.