Reproducing Kernel Banach Spaces for Machine Learning

Haizhang Zhang, Yuesheng Xu and Jun Zhang

Abstract— **Reproducing kernel Hilbert space (RKHS) methods have become powerful tools in machine learning. However, their kernels, which measure similarity of inputs, are required to be symmetric, constraining certain applications in practice. Furthermore, the celebrated representer theorem only applies to regularizers induced by the norm of an RKHS. To remove these limitations, we introduce the notion of reproducing kernel Banach spaces (RKBS) for pairs of reflexive Banach spaces of functions by making use of semi-inner-products and the duality mapping. As applications, we develop the framework of RKBS standard learning schemes including minimal norm interpolation, regularization network, and support vector machines. In particular, existence, uniqueness and representer theorems are established.**

I. Introduction

This note serves as an extended abstract for our recent work [1], which aims at building a theoretical basis for developing kernel methods for learning in Banach spaces.

Learning a function from its finite samples is a fundamental science problem. The essence in achieving this is to choose an appropriate measurement of similarities between elements in the domain of the function. A recent trend in machine learning is to use a positive definite kernel [2] to measure the similarity between elements in an input space X, [3]-[7]. A function $K : X \times X \to \mathbb{C}$ is called a *positive definite kernel* if for all finite subsets $\mathbf{x} := \{x_j : j \in \mathbb{N}_n\} \subseteq X$ the matrix

$$K[\mathbf{x}] := [K(x_j, x_k) : j, k \in \mathbb{N}_n] \tag{1}$$

is hermitian (especially symmetric, when K is real-valued) and positive semi-definite. The reason of using positive definite kernels to measure similarity lies in the celebrated theoretical fact due to Mercer [8] that there is a bijective correspondence between them and *reproducing kernel Hilbert spaces* (RKHS). An RKHS $\mathcal{H}$ on X is a Hilbert space of functions on X for which point evaluations are always continuous linear functionals. One direction of the bijective correspondence says that if K is a positive definite kernel on X then there exists a unique RKHS $\mathcal{H}$ on X such that $K(x, \cdot) \in \mathcal{H}$ for each $x \in X$ and for all $f \in \mathcal{H}$ and $y \in X$

$$f(y) = (f(\cdot), K(y, \cdot))_{\mathcal{H}}, \tag{2}$$

Haizhang Zhang and Yuesheng Xu are with the Department of Mathematics, Syracuse University, Syracuse, NY 13244 (email: {hzhang12, yxu06}@syr.edu). Jun Zhang is with the Department of Psychology, University of Michigan, Ann Arbor, MI 48109 (email: junz@umich.edu). Address all correspondence to Jun Zhang.

This work was supported by the National Science Foundation under grant 0631541 (to J.Z.) and grants CCR-0407476 and DMS-0712827 (to Y.X.), and by the Natural Science Foundation of China under grants 10371122 and 10631080 (to Y.X.), by the Education Ministry of the People's Republic of China under the Changjiang Scholar Chair Professorship Program through Sun Yat-sen University (to Y.X.).

where $(\cdot, \cdot)_{\mathcal{H}}$ denotes the inner product on $\mathcal{H}$. Conversely, if $\mathcal{H}$ is an RKHS on X then there is a unique positive definite kernel K on X such that $\{K(x, \cdot) : x \in X\} \subseteq \mathcal{H}$ and (2) holds. In light of this bijective correspondence, positive definite kernels are usually called *reproducing kernels*.

By taking $f := K(x, \cdot)$ for $x \in X$ in equation (2), we get that

$$K(x, y) = (K(x, \cdot), K(y, \cdot))_{\mathcal{H}}, \quad x, y \in X. \tag{3}$$

Thus $K(x, y)$ is represented as an inner product on an RKHS. This explains why $K(x, y)$ is able to measure similarities of x and y. The advantages brought by the use of an RKHS include: (1) the inputs can be handled and explained geometrically; (2) geometric objects such as hyperplanes are provided by the RKHS for learning; (3) the powerful tool of functional analysis applies, [3]. Based on the theory of reproducing kernels, many effective schemes have been developed for the learning from finite samples, [3][4][5][9][10]. In particular, the widely used regularized learning algorithm works by outputting a predictor function from the training data $\{(x_j, y_j) : j \in \mathbb{N}_n\} \subseteq X \times \mathbb{C}$ as the minimizer of

$$\min_{f \in \mathcal{H}_K} \sum_{j \in \mathbb{N}_n} \mathcal{L}(f(x_j), y_j) + \mu \|f\|^2_{\mathcal{H}_K}, \tag{4}$$

where $\mathcal{H}_K$ denotes the RKHS corresponding to the positive definite kernel K, $\mathcal{L}$ is a prescribed loss function, and μ is a positive regularization parameter.

This paper is motivated from machine learning in Banach spaces. There are advantages of learning in Banach spaces over in Hilbert spaces. Firstly, there is essentially only one Hilbert space once the dimension of the space is fixed. This follows from the well-known fact that any two Hilbert spaces over $\mathbb{C}$ of the same dimension are isometrically isomorphic. By contrast, for $p \neq q \in [1, +\infty]$, $L^p[0, 1]$ and $L^q[0, 1]$ are not isomorphic, namely, there does not exist a bijective bounded linear mapping between them (see [11, pp. 180]). Thus, Banach spaces possess much richer geometric structures than Hilbert spaces. Secondly, in some applications, a norm from a Banach space is invoked without being induced from an inner product. For instance, it is known that minimizing about the ℓ^p norm on $\mathbb{R}^d$ leads to sparsity of the minimizer when p is close to 1 (see, for example, [12]). In the extreme case that $\varphi : \mathbb{R}^d \to [0, +\infty)$ is strictly concave and $\mu > 0$, one can show that the minimizer for

$$\min\{\varphi(x) + \mu \|x\|_{\ell^1} : x \in \mathbb{R}^d\}$$

has at most one nonzero component. The reason is that the extreme points on a sphere in the ℓ^1 norm must lie on axes of the Euclidean coordinate system. Thirdly, RKHS methods require symmetric kernels, which might be violated

978-1-4244-3553-1/09/$25.00 ©2009 IEEE

by certain data structures with asymmetric similarity. Finally, one sometimes considers regularizers other than $\|\cdot\|_{\mathcal{H}_K}$ as in (4), and needs to look for corresponding representer theorems. Hence, there is a need to modify the algorithms by adopting norms in Banach spaces.

There has been considerable work on learning in Banach spaces in the literature. References [13]-[17] considered the problem of minimizing a regularized functional of the form

$$\sum_{j\in\mathbb{N}_n} \mathcal{L}(\lambda_j(f), y_j) + \phi(\|f\|_{\mathcal{B}}), \quad f \in \mathcal{B},$$

where $\mathcal{B}$ is Banach space, λ_j are in the dual $\mathcal{B}^*$, $y_j \in \mathbb{C}$, $\mathcal{L}$ is a loss function, and ϕ is a strictly increasing nonnegative function. In particular, paper [16] considered learning in a Besov space (a special type of Banach spaces). On-line learning in finite dimensional Banach spaces was studied, for example, in [18]. Learning of an L^p function was considered in [19]. Classifications in Banach spaces, and more generally in metric spaces were discussed in [13][20][21][22][23].

The above discussion indicates that there is a need of introducing the notion of reproducing kernel Banach spaces for the systematic study of learning in Banach spaces. Such a definition is expected to result in consequences similar to those in an RKHS. A generalization of RKHS to non-Hilbert spaces using point evaluation with kernels was proposed in [24], although the spaces considered there might be too general to have favorable properties of an RKHS. We shall introduce the notion of reproducing kernel Banach spaces and a general construction in Section 2. It will become clear that the lack of an inner product may cause arbitrariness of the associated reproducing kernel. To overcome this, we shall establish in Section 3 s.i.p. reproducing kernel Banach spaces by making use of semi-inner-products for normed vector spaces first defined by G. Lumer [25] and further developed by J. Giles [26] and B. Nath [27]. Semi-inner-products were first used in the context of machine learning by Der and Lee [20] to develop hard margin hyperplane classification in Banach spaces. The availability of a semi-inner-product makes possible the study of basic properties of reproducing kernel Banach spaces and their reproducing kernels. In the last section, we shall develop in the framework of reproducing kernel Banach spaces standard learning schemes including minimal norm interpolation, regularization network, and support vector machines. Existence, uniqueness and representer theorems will be proved. Due to space limitations, we omit all proofs. Interested readers are referred to [1].

II. Reproducing Kernel Banach Spaces

Without specifically mentioned, all vector spaces in the paper are assumed to be complex. Let X be a prescribed input space. A normed vector space $\mathcal{B}$ is called a **Banach space of functions** on X if it is a Banach space whose elements are functions on X, and for each $f \in \mathcal{B}$, its norm $\|f\|_{\mathcal{B}}$ in $\mathcal{B}$ vanishes if and only if f, as a function, vanishes everywhere on X.

Influenced by the definition of RKHS, our first intuition is to define a reproducing kernel Banach space (RKBS) as a Banach space of functions on X on which point evaluations are continuous linear functionals. If such a definition was adopted then the first example that comes to our mind would be $C[0,1]$, the Banach space of continuous functions on $[0,1]$ equipped with the maximum norm. It satisfies the definition. However, the reproducing kernel for $C[0,1]$ would have to be the delta distribution, which is not a function that can be evaluated. This example suggests that the dual of an RKBS should still consist of functions. Note that the dual space V^* of a normed vector space V is a notion by construction. It depends not only on the topology of V but also on how the following bilinear form on $V \times V^*$

$$(u, v^*)_V := v^*(u), \quad u \in V,\ v^* \in V^*$$

is defined. We make the convention throughout this paper that whenever we write V^* we mean that it along with the above bilinear form has been chosen. In particular, if V is a Hilbert space then naturally elements in V^* are identified with those in V by the Riesz representation theorem [28]. In addition to requiring that the dual of an RKBS be a space of functions, later on we will find it very convenient to jump freely between a Banach space and its dual. For this reason, we would like an RKBS $\mathcal{B}$ to be *reflexive* in the sense that $(\mathcal{B}^*)^* = \mathcal{B}$. These considerations lead to the following definition.

We call a reflexive Banach space $\mathcal{B}$ of functions on X a **reproducing kernel Banach space** (RKBS) if $\mathcal{B}^*$ is also a Banach space of functions and point evaluations on both $\mathcal{B}$ and $\mathcal{B}^*$ are continuous.

It follows immediately from the definition that if $\mathcal{B}$ is an RKBS on X then so is $\mathcal{B}^*$. Moreover, an RKHS is a special RKBS. We shall show that there indeed exists a *reproducing kernel* for an RKBS.

Theorem 2.1: *Suppose that $\mathcal{B}$ is an RKBS on X. Then there exists a unique function $K : X \times X \to \mathbb{C}$ such that the following statements hold.*
(a) For every $x \in X$, $K(\cdot, x) \in \mathcal{B}^*$ and

$$f(x) = (f, K(\cdot, x))_{\mathcal{B}}, \quad \textit{for all } f \in \mathcal{B}.$$

(b) For every $x \in X$, $K(x, \cdot) \in \mathcal{B}$ and

$$f^*(x) = (K(x, \cdot), f^*)_{\mathcal{B}}, \quad \textit{for all } f^* \in \mathcal{B}^*.$$

(c) The linear span of $\{K(x, \cdot) : x \in X\}$ is dense in $\mathcal{B}$, namely,

$$\overline{\text{span}}\{K(x, \cdot) : x \in X\} = \mathcal{B}.$$

(d) The linear span of $\{K(\cdot, x) : x \in X\}$ is dense in $\mathcal{B}^*$, namely,

$$\overline{\text{span}}\{K(\cdot, x) : x \in X\} = \mathcal{B}^*. \tag{5}$$

(e) For all $x, y \in X$, $K(x, y) = (K(x, \cdot), K(\cdot, y))_{\mathcal{B}}$.

We call the function K in Theorem 2.1 the **reproducing kernel** for the RKBS $\mathcal{B}$. By Theorem 2.1, an RKBS has exactly one reproducing kernel. However, different RKBS may have the same reproducing kernel. Examples will be given later. This results from a fundamental difference between Banach spaces and Hilbert spaces that a norm defined

on a subset of a Banach space $\mathcal{B}$ whose linear span is dense in $\mathcal{B}$ may not be extended to the whole space in a unique way.

In the following, we shall characterize reproducing kernels for RKBS. The characterization will at the same time provide a convenient way of constructing reproducing kernels and their corresponding RKBS.

Theorem 2.2: *Let $\mathcal{W}$ be a reflexive Banach space with dual space $\mathcal{W}^*$. Suppose that there exists $\Phi : X \to \mathcal{W}$, and $\Phi^* : X \to \mathcal{W}^*$ such that*

$$\overline{\operatorname{span}}\,\Phi(X) = \mathcal{W}, \quad \overline{\operatorname{span}}\,\Phi^*(X) = \mathcal{W}^*. \tag{6}$$

Then $\mathcal{B} := \{(u, \Phi^*(\cdot))_{\mathcal{W}} : u \in \mathcal{W}\}$ with norm

$$\|(u, \Phi^*(\cdot))_{\mathcal{W}}\|_{\mathcal{B}} := \|u\|_{\mathcal{W}}$$

is an RKBS on X with the dual space $\mathcal{B}^* := \{(\Phi(\cdot), u^*)_{\mathcal{W}} : u^* \in \mathcal{W}^*\}$ endowed with the norm

$$\|(\Phi(\cdot), u^*)_{\mathcal{W}}\|_{\mathcal{B}^*} := \|u^*\|_{\mathcal{W}^*}$$

and the bilinear form

$$((u, \Phi^*(\cdot))_{\mathcal{W}}, (\Phi(\cdot), u^*)_{\mathcal{W}})_{\mathcal{B}} = (u, u^*)_{\mathcal{W}}, \ u \in \mathcal{W}, \ u^* \in \mathcal{W}^*.$$

Moreover, the reproducing kernel K for $\mathcal{B}$ is

$$K(x, y) := (\Phi(x), \Phi^*(y))_{\mathcal{W}}, \quad x, y \in X. \tag{7}$$

We call the mappings Φ, Φ^* in Theorem 2.2 a pair of **feature maps** for the reproducing kernel K. The spaces $\mathcal{W}$, $\mathcal{W}^*$ are called the pair of **feature spaces** associated with the feature maps for K. As a corollary to Theorem 2.2, we obtain the following characterization of reproducing kernels for RKBS.

Theorem 2.3: *A function $K : X \times X \to \mathbb{C}$ is the reproducing kernel of an RKBS on X if and only if it is of the form (7), where $\mathcal{W}$ is a reflexive Banach space, and mappings $\Phi : X \to \mathcal{W}$, $\Phi^* : X \to \mathcal{W}^*$ satisfy (6).*

To demonstrate how we get RKBS and their reproducing kernels by Theorem 2.2, we now present a nontrivial example of RKBS. Set $X := \mathbb{R}$, $\mathbb{I} := [-\frac{1}{2}, \frac{1}{2}]$, and $p \in (1, +\infty)$. We make the convention that q is always the conjugate number of p, that is, $p^{-1} + q^{-1} = 1$. Define $\mathcal{W} := L^p(\mathbb{I})$, $\mathcal{W}^* := L^q(\mathbb{I})$ and $\Phi : X \to \mathcal{W}$, $\Phi^* : X \to \mathcal{W}^*$ as

$$\Phi(x)(t) := e^{-i2\pi xt}, \ \Phi^*(x)(t) := e^{i2\pi xt}, \quad x \in \mathbb{R}, \ t \in \mathbb{I}.$$

Clearly, the density requirement (6) is satisfied. For $f \in L^1(\mathbb{R})$, its Fourier transform $\hat{f}$ is defined as

$$\hat{f}(t) := \int_{\mathbb{R}} f(x) e^{-i2\pi xt} dx, \quad t \in \mathbb{R},$$

and its inverse Fourier transform $\check{f}$ is defined by

$$\check{f}(t) := \int_{\mathbb{R}} f(x) e^{i2\pi xt} dx, \quad t \in \mathbb{R}.$$

The Fourier transform and the inverse Fourier transform can be defined on temperate distributions.

By the construction described in Theorem 2.2, we obtain

$$\mathcal{B} := \{f \in C(\mathbb{R}) : \operatorname{supp} \hat{f} \subseteq \mathbb{I}, \ \hat{f} \in L^p(\mathbb{I})\} \tag{8}$$

with norm $\|f\|_{\mathcal{B}} := \|\hat{f}\|_{L^p(\mathbb{I})}$, and the dual

$$\mathcal{B}^* := \{g \in C(\mathbb{R}) : \operatorname{supp} \hat{g} \subseteq \mathbb{I}, \ \hat{g} \in L^q(\mathbb{I})\}$$

with norm $\|g\|_{\mathcal{B}^*} := \|\check{g}\|_{L^q(\mathbb{I})}$. For each $f \in \mathcal{B}$ and $g \in \mathcal{B}^*$, we have

$$(f, g)_{\mathcal{B}} = \int_{\mathbb{I}} \hat{f}(t) \check{g}(t) dt.$$

The kernel K for $\mathcal{B}$ is given as

$$K(x, y) = (\Phi(x), \Phi^*(y))_{\mathcal{W}} = \frac{\sin \pi(x - y)}{\pi(x - y)} = \operatorname{sinc}(x - y).$$

When $p = q = 2$, $\mathcal{B}$ reduces to the classical space of bandlimited functions.

In the above example, $\mathcal{B}$ is isometrically isomorphic to $L^p(\mathbb{I})$. As mentioned in the introduction, $L^p(\mathbb{I})$ with different p are not isomorphic to each other. As a result, for different indices p the spaces $\mathcal{B}$ defined by (8) are essentially different. However, we see that they all have the sinc function as the reproducing kernel. In fact, if no further conditions are imposed on an RKBS, its reproducing kernel can be rather arbitrary. For instance, one can prove by Theorem 2.3 that if the input space X is a finite set, then any nontrivial function K on $X \times X$ is the reproducing kernel of some RKBS on X. This fact reveals that due to the lack of an inner product, the reproducing kernel for a general RKBS can not only be nonsymmetric, non-positive definite, but also be arbitrary. In order for reproducing kernels of RKBS to have desired properties as those of RKHS, we may need to impose certain structures on RKBS, which in some sense are substitutes of the inner product for RKHS. For this purpose, we shall adopt the semi-inner-product introduced by Lumer [25]. A semi-inner-product possesses some but not all properties of an inner product. Hilbert space type arguments and results become available with the presence of a semi-inner-product. We shall introduce the notion of semi-inner-product RKBS.

III. S.I.P. Reproducing Kernel Banach Spaces

The purpose of this section is to establish the notion of semi-inner-product RKBS and study its properties. We start with necessary preliminaries on semi-inner-products.

A. Semi-Inner-Products

A **semi-inner-product** on a vector space V is a function (generally complex-valued), denoted by $[\cdot, \cdot]_V$, on $V \times V$ such that for all $x, y, z \in V$ and $\lambda \in \mathbb{C}$

1) $[x + y, z]_V = [x, z]_V + [y, z]_V$,
2) $[\lambda x, y]_V = \lambda [x, y]_V$,
3) $[x, x]_V > 0$ for $x \neq 0$,
4) (Cauchy-Schwartz) $|[x, y]_V|^2 \leq [x, x]_V [y, y]_V$.

In general, a semi-inner-product $[\cdot, \cdot]_V$ does not satisfy the conjugate symmetry $[x, y]_V = \overline{[y, x]_V}$ for all $x, y \in V$. As a consequence, there always exist $x, y, z \in V$ such that

$$[x, y + z]_V \neq [x, y]_V + [x, z]_V.$$

In fact, a semi-inner-product is always additive about the second variable when it degenerates to an inner product.

It was shown in [25] that a vector space V with a semi-inner-product is a normed space equipped with

$$\|x\|_V := [x,x]_V^{1/2}, \quad x \in V. \tag{9}$$

Therefore, if a vector space V has a semi-inner-product, we always assume that its norm is induced by (9) and call V an **s.i.p. space**. Conversely, every normed vector space V has a semi-inner-product that induces its norm by (9) and satisfies the following homogeneous condition [26]

$$[x, \lambda y]_V = \bar{\lambda}\,[x,y]_V, \text{ for all } x, y \in V, \ \lambda \in \mathbb{C}. \tag{10}$$

Thus, we shall always impose the above property on semi-inner-products. It is worthwhile to mention that if the Cauchy-Schwartz inequality is replaced by a Hölder inequality [27]

$$|[x,y]_V| \le [x,x]_V^{1/p} [y,y]_V^{1/q}$$

with $p, q > 1, \frac{1}{p} + \frac{1}{q} = 1$, then we have the notion of semi-inner-product of order p generalizing [**?**]. In this case, (9) becomes

$$\|x\|_V := [x,x]_V^{1/p}, \quad x \in V.$$

while (10) becomes

$$[x, \lambda y]_V = \bar{\lambda}|\lambda|^{p-2}\,[x,y]_V, \text{ for all } x, y \in V, \ \lambda \in \mathbb{C}.$$

By the Cauchy-Schwartz inequality, if V is an s.i.p. space then for each $x \in V$, $y \to [y,x]_V$ is a continuous linear functional on V. We denote this linear functional by x^*. Following this definition, we have that

$$[x,y]_V = y^*(x) = (x, y^*)_V, \quad x, y \in V.$$

In general, a semi-inner-product for a normed vector space may not be unique. However, a uniformly Fréchet differentiable normed vector space always has a unique semi-inner-product [26]. We shall impose one more condition on an s.i.p. space that will lead to a Riesz representation theorem. A normed vector space V is **uniformly convex** if for all $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\|x+y\|_V \le 2-\delta \text{ for all } \|x\|_V = \|y\|_V = 1 \text{ with } \|x-y\|_V \ge \varepsilon.$$

The space $L^p(\Omega, \mu)$, $1 < p < +\infty$, on a measure space $(\Omega, \mathcal{F}, \mu)$ is uniformly convex. In particular, by the parallelogram law, any inner product space is uniformly convex. By a remark in [28, pp. 134], a uniformly convex Banach space is reflexive. If $\mathcal{B}$ is a uniformly convex and uniformly Fréchet differentiable Banach space then so is $\mathcal{B}^*$, [29]. The important role of uniform convexity is displayed in the next lemma [26].

Lemma 3.1: (Riesz Representation Theorem) *Suppose that $\mathcal{B}$ is a uniformly convex, uniformly Fréchet differentiable Banach space. Then for each $f \in \mathcal{B}^*$ there exists a unique $x \in \mathcal{B}$ such that $f = x^*$, that is,*

$$f(y) = [y,x]_{\mathcal{B}}, \quad y \in \mathcal{B}.$$

Moreover, $\|f\|_{\mathcal{B}^*} = \|x\|_{\mathcal{B}}$.

Let $\mathcal{B}$ be a uniformly convex and uniformly Fréchet differentiable Banach space. By Lemma 3.1, $x \to x^*$ defines a bijection from $\mathcal{B}$ to $\mathcal{B}^*$ that preserves the norm. Note that this **duality mapping** is in general nonlinear. We call x^* the **dual element** of x. Since $\mathcal{B}^*$ is uniformly Fréchet differentiable, it has a unique semi-inner-product, which is given by

$$[x^*, y^*]_{\mathcal{B}^*} = [y,x]_{\mathcal{B}}, \quad x, y \in \mathcal{B}.$$

We close this subsection with a concrete example of uniformly convex and uniformly Fréchet differentiable Banach space. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $\mathcal{B} := L^p(\Omega, \mu)$ for some $p \in (1, +\infty)$. It is uniformly convex and uniformly Fréchet differentiable with dual $\mathcal{B}^* = L^q(\Omega, \mu)$. For each $f \in \mathcal{B}$, its dual element in $\mathcal{B}^*$ is

$$f^* = \frac{\bar{f}|f|^{p-2}}{\|f\|_{L^p(\Omega,\mu)}^{p-2}}.$$

Consequently, the semi-inner-product on $\mathcal{B}$ is

$$[f,g]_{\mathcal{B}} = g^*(f) = \frac{\int_\Omega f\bar{g}|g|^{p-2}d\mu}{\|g\|_{L^p(\Omega,\mu)}^{p-2}}.$$

With the above preparation, we shall study a special kind of RKBS which have desired properties.

B. S.i.p. RKBS

Let X be a prescribed input space. We call a uniformly convex and uniformly Fréchet differentiable RKBS on X an **s.i.p. reproducing kernel Banach space** (**s.i.p. RKBS**). Again, we see immediately that an RKHS is an s.i.p. RKBS. Also, the dual of an s.i.p. RKBS remains an s.i.p. RKBS. An s.i.p. RKBS $\mathcal{B}$ is by definition uniformly Fréchet differentiable. Therefore, it has a unique semi-inner-product, which by Lemma 3.1 represents all the interaction between $\mathcal{B}$ and $\mathcal{B}^*$. This leads to a more specific representation of the reproducing kernel. Precisely, we have the following consequences due essentially to Lemma 3.1.

Theorem 3.2: *Let $\mathcal{B}$ be an s.i.p. RKBS on X and K its reproducing kernel. Then there exists a unique function $G : X \times X \to \mathbb{C}$ such that $\{G(x,\cdot) : x \in X\} \subseteq \mathcal{B}$ and*

$$f(x) = [f, G(x,\cdot)]_{\mathcal{B}}, \quad \textit{for all } f \in \mathcal{B}, \ x \in X. \tag{11}$$

Moreover, there holds the relationship

$$K(\cdot, x) = (G(x,\cdot))^*, \quad x \in X \tag{12}$$

and

$$f^*(x) = [K(x,\cdot), f]_{\mathcal{B}}, \quad \textit{for all } f \in \mathcal{B}, \ x \in X.$$

We call the unique function G in Theorem 3.2 the **s.i.p. kernel** of the s.i.p. RKBS $\mathcal{B}$. It coincides with the reproducing kernel K when $\mathcal{B}$ is an RKHS. In general, when $G = K$ in Theorem 3.2, we call G an **s.i.p. reproducing kernel**. By (11), an s.i.p. reproducing kernel G satisfies the following generalization of (3)

$$G(x,y) = [G(x,\cdot), G(y,\cdot)]_{\mathcal{B}}, \quad x, y \in X. \tag{13}$$

We shall give a characterization of an s.i.p. reproducing kernel in terms of its corresponding feature map. To this end, for a mapping Φ from X to a uniformly convex and

uniformly Fréchet differentiable Banach space $\mathcal{W}$, we denote by Φ^* the mapping from X to $\mathcal{W}^*$ defined as

$$\Phi^*(x) := (\Phi(x))^*, \quad x \in X.$$

Theorem 3.3: *Let $\mathcal{W}$ be a uniformly convex and uniformly Fréchet differentiable Banach space and Φ a mapping from X to $\mathcal{W}$ such that*

$$\overline{\text{span}}\, \Phi(X) = \mathcal{W}, \quad \overline{\text{span}}\, \Phi^*(X) = \mathcal{W}^*. \tag{14}$$

Then $\mathcal{B} := \{[u, \Phi(\cdot)]_{\mathcal{W}} : u \in \mathcal{W}\}$ equipped with

$$\Big[[u, \Phi(\cdot)]_{\mathcal{W}}, [v, \Phi(\cdot)]_{\mathcal{W}}\Big]_{\mathcal{B}} := [u, v]_{\mathcal{W}}$$

and $\mathcal{B}^* := \{[\Phi(\cdot), u]_{\mathcal{W}} : u \in \mathcal{W}\}$ with

$$\Big[[\Phi(\cdot), u]_{\mathcal{W}}, [\Phi(\cdot), v]_{\mathcal{W}}\Big]_{\mathcal{B}^*} := [v, u]_{\mathcal{W}}$$

are uniformly convex and uniformly Fréchet differentiable Banach spaces. And $\mathcal{B}^*$ is the dual of $\mathcal{B}$ with the bilinear form

$$\Big([u, \Phi(\cdot)]_{\mathcal{W}}, [\Phi(\cdot), v]_{\mathcal{W}}\Big)_{\mathcal{B}} := [u, v]_{\mathcal{W}}, \quad u, v \in \mathcal{W}.$$

Moreover, the s.i.p. kernel G of $\mathcal{B}$ is given by

$$G(x, y) = [\Phi(x), \Phi(y)]_{\mathcal{W}}, \quad x, y \in X, \tag{15}$$

which coincides with its reproducing kernel K.

As a direct consequence of the above theorem, we have the following characterization of s.i.p. reproducing kernels.

Theorem 3.4: *A function G on $X \times X$ is an s.i.p. reproducing kernel if and only if it is of the form (15), where Φ is a mapping from X to a uniformly convex and uniformly Fréchet differentiable Banach space $\mathcal{W}$ satisfying (14).*

The mapping Φ and space $\mathcal{W}$ in the above theorem will be called a **feature map** and **feature space** of the s.i.p. reproducing kernel G, respectively.

By the duality relation (12) and the density condition (5), the s.i.p kernel G of an s.i.p. RKBS $\mathcal{B}$ on X satisfies

$$\overline{\text{span}}\{(G(x, \cdot))^* : x \in X\} = \mathcal{B}^*. \tag{16}$$

It is also of the form (13). By Theorem 3.4, G is identical with the reproducing kernel K for $\mathcal{B}$ if and only if

$$\overline{\text{span}}\{G(x, \cdot) : x \in X\} = \mathcal{B}. \tag{17}$$

If $\mathcal{B}$ is not a Hilbert space then the duality mapping from $\mathcal{B}$ to $\mathcal{B}^*$ is nonlinear. Thus, it may not preserve the density of a linear span. As a result, (17) would not follow automatically from (16). Here we remark that for most finite dimensional s.i.p. RKBS, (16) implies (17). This is due to the well-known fact that for all $n \in \mathbb{N}$, the set of $n \times n$ singular matrices has Lebesgue measure zero in $\mathbb{C}^{n \times n}$. Therefore, the s.i.p. kernel for most finite dimensional s.i.p. RKBS is the same as the reproducing kernel. Nevertheless, an explicit example was presented in [1] to demonstrate that the two kernels might be different.

C. Properties of S.i.p. Reproducing Kernels

The existence of a semi-inner-product makes it possible to study properties of RKBS and their reproducing kernels. For illustration, we shall present three of them below.

1. Positive Definiteness. An $n \times n$ matrix M over a number field $\mathbb{F}$ ($\mathbb{C}$ or $\mathbb{R}$) is said to be positive semi-definite if for all $(c_j : j \in \mathbb{N}_n) \in \mathbb{F}^n$

$$\sum_{j \in \mathbb{N}_n} \sum_{k \in \mathbb{N}_n} c_j \overline{c_k} M_{jk} \geq 0.$$

We shall consider positive semi-definiteness of matrices $G[\mathbf{x}]$ as defined in (1) for an s.i.p. reproducing kernel G on X.

Let $\Phi : X \to \mathcal{W}$ be a feature map for G, that is, (15) and (14) hold. By properties 3 and 4 in the definition of a semi-inner-product, we have that

$$G(x, x) \geq 0, \quad x \in X \tag{18}$$

and

$$|G(x, y)|^2 \leq G(x, x) G(y, y), \quad x, y \in X. \tag{19}$$

Notice that if a complex matrix is positive semi-definite then it must be hermitian. Since a semi-inner-product is in general not an inner product, we can not expect a complex s.i.p. kernel to be positive definite. In the real case, inequalities (18) and (19) imply that $G[\mathbf{x}]$ is positive semi-definite for all $\mathbf{x} \subseteq X$ with cardinality less than or equal to two. However, $G[\mathbf{x}]$ might cease to be positive semi-definite if $\mathbf{x}$ contains more than two points. For instance, for the s.i.p. reproducing kernel G defined for all $x, y \in \mathbb{R}_+ := [0, +\infty)$ as

$$G(x, y) = [\Phi(x), \Phi(y)]_{\mathcal{W}} = \frac{1 + xy^{p-1}}{(1 + y^p)^{\frac{p-2}{p}}},$$

the matrix $G[\mathbf{x}]$ is positive semi-definite for all $\mathbf{x} = \{x, y, z\} \subseteq X$ if and only if $p = 2$.

By this example, non-positive semi-definiteness is a characteristic of s.i.p. reproducing kernels for RKBS that distinct them from reproducing kernels for RKHS.

2. Pointwise Convergence. If f_n converges to f in an s.i.p. RKBS with the s.i.p. kernel G then $f_n(x)$ converges to $f(x)$ for any $x \in X$ and the limit is uniform on the set where $G(x, x)$ is bounded.

3. Weak Universality. Suppose that X is metric space and G is an s.i.p. reproducing kernel on X. We say that G is **universal** if G is continuous on $X \times X$ and for all compact subsets $\mathcal{Z} \subseteq X$, $\text{span}\{G(x, \cdot) : x \in \mathcal{Z}\}$ is dense in $C(\mathcal{Z})$, [30][31]. Universality of a kernel ensures that it can approximate any continuous target function uniformly on compact subsets of the input space. This is crucial for the consistence of the learning algorithms with the kernel. We shall discuss the case when X is itself a compact metric space. Here we are concerned with the ability of G to approximate any continuous target function on X uniformly. For this purpose, we call a continuous kernel G on a compact metric space X **weakly universal** if $\text{span}\{G(x, \cdot) : x \in X\}$ is dense in $C(X)$. We shall present a characterization of

weak universality. The result in the cases of positive definite kernels and vector-valued positive definite kernels has been proved respectively in [30] and [32].

Proposition 3.5: *Let Φ be a feature map from a compact metric space X to $\mathcal{W}$ such that both $\Phi : X \to \mathcal{W}$ and $\Phi^* : X \to \mathcal{W}^*$ are continuous. Then the s.i.p. reproducing kernel G defined by (15) is continuous on $X \times X$, and there holds in $C(X)$ the equality of subspaces*

$$\overline{\text{span}}\{G(x,\cdot) : x \in X\} = \overline{\text{span}}\{[u, \Phi(\cdot)]_{\mathcal{W}} : u \in \mathcal{W}\}.$$

Consequently, G is weakly universal if and only if

$$\overline{\text{span}}\{[u, \Phi(\cdot)]_{\mathcal{W}} : u \in \mathcal{W}\} = C(X).$$

Universality and other properties of s.i.p. reproducing kernels will be treated specially in a future work. A main purpose of this study to apply the tool of s.i.p. reproducing kernels to learning in Banach spaces. To be specific, we shall develop in the framework of s.i.p. RKBS several standard learning schemes.

IV. Applications to Machine Learning

In this section, we assume that $\mathcal{B}$ is an s.i.p. RKBS on X with the s.i.p. reproducing kernel G defined by a feature $\Phi : X \to \mathcal{W}$ as in (15). We shall develop in this framework several standard learning schemes including minimal norm interpolation, regularization network, and support vector machines. For introduction and discussions of these widely used algorithms in RKHS, see, for example, [3][4][5][9][33][34].

A. Minimal Norm Interpolation (MNI)

The minimal norm interpolation is to find, among all functions in $\mathcal{B}$ that interpolate a prescribed set of points, a function with minimal norm. Let $\mathbf{x} := \{x_j : j \in \mathbb{N}_n\}$ be a fixed finite set of distinct points in X and set for each $\mathbf{y} := (y_j : j \in \mathbb{N}_n) \in \mathbb{C}^n$

$$\mathcal{I}_{\mathbf{y}} := \{f \in \mathcal{B} : f(x_j) = y_j,\ j \in \mathbb{N}_n\}.$$

Our purpose is to find $f_0 \in \mathcal{I}_{\mathbf{y}}$ such that

$$\|f_0\|_{\mathcal{B}} = \inf\{\|f\|_{\mathcal{B}} : f \in \mathcal{I}_{\mathbf{y}}\} \tag{20}$$

provided that $\mathcal{I}_y$ is nonempty. The set $\mathcal{I}_{\mathbf{y}}$ is nonempty for any $\mathbf{y} \in \mathbb{C}^n$ if and only if $G_{\mathbf{x}} := \{G(\cdot, x_j) : j \in \mathbb{N}_n\}$ is linearly independent in $\mathcal{B}^*$. Existence, uniqueness and a representer theorem for the solution are addressed in the following result. For the representer theorem in learning with positive definite kernels, see, for example, [35][36].

Theorem 4.1: (Representer Theorem for MNI) *Suppose that $G_{\mathbf{x}}$ is linearly independent in $\mathcal{B}^*$. Then for any $\mathbf{y} \in \mathbb{C}^n$ there exists a unique $f_0 \in \mathcal{I}_{\mathbf{y}}$ satisfying (20). If f_0 is the solution of the minimal norm interpolation (20) then there exists $\mathbf{c} = (c_j : j \in \mathbb{N}_n) \in \mathbb{C}^n$ such that*

$$f_0^* = \sum_{j \in \mathbb{N}_n} c_j G(\cdot, x_j).$$

Conversely, a function of the form in the right hand side above is the solution if and only if $\mathbf{c}$ satisfies

$$\left[G(\cdot, x_k), \sum_{j \in \mathbb{N}_n} c_j G(\cdot, x_j)\right]_{\mathcal{B}^*} = y_k, \quad k \in \mathbb{N}_n. \tag{21}$$

We conclude that under the condition that $G_{\mathbf{x}}$ is linearly independent, the minimal norm interpolation problem (20) has a unique solution, and finding the solution reduces to solving the system (21) of equations about $\mathbf{c} \in \mathbb{C}^n$. The solution $\mathbf{c}$ of (21) is unique by Theorem 4.1. Again, the difference from the result for RKHS is that (21) is often nonlinear about $\mathbf{c}$ since a semi-inner-product is generally nonadditive about the second variable.

To see an explicit form of (21), we shall reformulate it in terms of the feature map Φ from X to $\mathcal{W}$. Let $\mathcal{B}$ and $\mathcal{B}^*$ be identified as in Theorem 3.3. Then (21) has the equivalent form

$$\left[\Phi^*(x_k), \sum_{j \in \mathbb{N}_n} c_j \Phi^*(x_j)\right]_{\mathcal{B}^*} = y_k, \quad k \in \mathbb{N}_n.$$

In the particular case that $\mathcal{W} = L^p(\Omega, \mu)$, $p \in (1, +\infty)$ on some measure space $(\Omega, \mathcal{F}, \mu)$, and $\mathcal{W}^* = L^q(\Omega, \mu)$, the above equation rewrites as

$$\int_\Omega \Phi^*(x_k) \overline{\sum_{j \in \mathbb{N}_n} c_j \Phi^*(x_j)} \left|\sum_{j \in \mathbb{N}_n} c_j \Phi^*(x_j)\right|^{q-2} d\mu$$
$$= y_k \left\|\sum_{j \in \mathbb{N}_n} c_j \Phi^*(x_j)\right\|_{L^q(\Omega,\mu)}^{q-2}, \quad k \in \mathbb{N}_n.$$

B. Regularization Network (RN)

In this subsection, we consider learning a predictor function $f_0 : X \to \mathbb{C}$ from a finite sample data $\mathbf{z} := \{(x_j, y_j) : j \in \mathbb{N}_n\} \subseteq X \times \mathbb{C}$ through a regularized learning algorithm. Let $\mathcal{L} : \mathbb{C} \times \mathbb{C} \to \mathbb{R}_+$ be a *loss function* that is continuous and convex about its first variable. For each $f \in \mathcal{B}$, we set

$$\mathcal{E}_{\mathbf{z}}(f) := \sum_{j \in \mathbb{N}_n} \mathcal{L}(f(x_j), y_j) \text{ and } \mathcal{E}_{\mathbf{z},\mu}(f) := \mathcal{E}_{\mathbf{z}}(f) + \mu \|f\|_{\mathcal{B}}^2,$$

where μ is a positive regularization parameter. The predictor function learned from the sample data $\mathbf{z}$ will be taken as the function f_0 such that

$$\mathcal{E}_{\mathbf{z},\mu}(f_0) = \inf\{\mathcal{E}_{\mathbf{z},\mu}(f) : f \in \mathcal{B}\}. \tag{22}$$

Existence and uniqueness of the minimizer of (22) were proved in [1].

In the rest of this subsection, we shall consider the regularization network in $\mathcal{B}$, that is, the loss function $\mathcal{L}$ is specified as

$$\mathcal{L}(a, b) = |a - b|^2, \quad a, b \in \mathbb{C}.$$

It is continuous and convex about the first variable. As mentioned before, there is a unique minimizer for the regularization network:

$$\min_{f \in \mathcal{B}} \sum_{j \in \mathbb{N}_n} |f(x_j) - y_j|^2 + \mu \|f\|_{\mathcal{B}}^2. \tag{23}$$

Theorem 4.2: (Representer Theorem for RN) *Let f_0 be the minimizer of (23). Then there exists some $\mathbf{c} \in \mathbb{C}^n$ such that*

$$f_0^* = \sum_{j\in\mathbb{N}_n} c_j G(\cdot, x_j).$$

If $G_{\mathbf{x}}$ is linearly independent then the right hand side of the above equation is the minimizer if and only if

$$\mu\overline{c_k} + \left[G(\cdot, x_k), \sum_{j\in\mathbb{N}_n} c_j G(\cdot, x_j)\right]_{\mathcal{B}^*} = y_k, \quad k \in \mathbb{N}_n. \quad (24)$$

By Theorem 4.2, if $G_{\mathbf{x}}$ is linearly independent then the minimizer of (23) can be obtained by solving (24), which has a unique solution in this case. Using the feature map, the system (24) has the following form

$$\mu\overline{c_k} + \left[\Phi^*(x_k), \sum_{j\in\mathbb{N}_n} c_j \Phi^*(x_j)\right]_{\mathcal{B}^*} = y_k, \quad k \in \mathbb{N}_n.$$

As remarked before, this is in general nonlinear about $\mathbf{c}$.

C. Support Vector Machines (SVM)

In this subsection, we assume that all the spaces are over the field $\mathbb{R}$ of real numbers, and consider learning a classifier from the data $\mathbf{z} := \{(x_j, y_j) : j \in \mathbb{N}_n\} \subseteq X \times \{-1, 1\}$. We shall establish for this task three learning algorithms in RKBS whose RKHS versions are well-known [3][4][5][9][37].

1) Support Vector Machine Classification: Following [37], we define the loss function

$$\mathcal{L}(a, y) = \max\{1 - ay, 0\}, \quad (a, y) \in \mathbb{R} \times \{-1, 1\}$$

and minimize

$$\min_{f\in\mathcal{B}} \mathcal{E}_{\mathbf{z},\mu}(f) = \sum_{j\in\mathbb{N}_n} \max\{1 - f(x_j)y_j, 0\} + \mu\|f\|_{\mathcal{B}}^2. \quad (25)$$

If the minimizer f_0 exists then the classifier will be taken as $\operatorname{sgn} f_0$. It can be verified that $\mathcal{L}$ is convex and continuous about the first variable. Therefore, there exists a unique minimizer $f_0 \in \mathcal{B}$ for (25).

We can prove the following representer theorem for f_0 using the celebrated geometric consequence of the Hahn-Banach theorem (see [28, pp. 111]) that in a normed vector space, a closed convex subset and a point outside of it can be strictly separated.

Theorem 4.3: (Representer Theorem for SVM) *Let f_0 be the minimizer of (25). Then f_0^* lies inside the closed convex cone* $\operatorname{cone} G_{\mathbf{z}}$ *spanned by $G_{\mathbf{z}} := \{y_j G(\cdot, x_j) : j \in \mathbb{N}_n\}$, that is, there exist $\lambda_j \geq 0$ such that*

$$f_0^* = \sum_{j\in\mathbb{N}_n} \lambda_j y_j G(\cdot, x_j). \quad (26)$$

To solve (25), one substitutes by Theorem 4.3 equation (26) into (25), which becomes a convex optimization about λ_j subject to the constraint that $\lambda_j \geq 0$, $j \in \mathbb{N}_n$. Standard convex optimization algorithms thus apply.

2) Soft Margin Hyperplane Classification (SMHC): We next focus on the soft margin hyperplane classification by studying

$$\inf\left\{\frac{1}{2}\|w\|_{\mathcal{W}}^2 + C\|\xi\|_{\ell^1(\mathbb{N}_n)} : w \in \mathcal{W}, \xi \in \mathbb{R}_+^n, b \in \mathbb{R}\right\} \quad (27)$$

subject to

$$y_j([\Phi(x_j), w]_{\mathcal{W}} + b) \geq 1 - \xi_j, \quad j \in \mathbb{N}_n.$$

Here, C is a fixed positive constant controlling the tradeoff between margin maximization and training error minimization. If the minimizer $(w_0, \xi_0, b_0) \in \mathcal{W} \times \mathbb{R}_+^n \times \mathbb{R}$ exists, the classifier is taken as $\operatorname{sgn}([\Phi(\cdot), w_0]_{\mathcal{W}} + b_0)$.

The same problem as (27) for the dual space $\mathcal{W}^*$ is

$$\inf\left\{\frac{1}{2}\|w^*\|_{\mathcal{W}^*}^2 + C\|\xi\|_{\ell^1(\mathbb{N}_n)} : w^* \in \mathcal{W}^*,\ \xi \in \mathbb{R}_+^n,\ b \in \mathbb{R}\right\}$$

subject to

$$y_j([w^*, \Phi^*(x_j)]_{\mathcal{W}^*} + b) \geq 1 - \xi_j, \quad j \in \mathbb{N}_n.$$

By considering this equivalent problem, we prove the following result.

Theorem 4.4: (Representer Theorem for SMHC) *Suppose that $\{y_j : j \in \mathbb{N}_n\} = \{-1, 1\}$. Then the minimizer w_0 of (27) uniquely exists. Moreover, the minimizer w_0 of (27) belongs to the closed convex cone spanned by $y_j\Phi(x_j)$, $j \in \mathbb{N}_n$.*

3) Hard Margin Hyperplane Classification: Consider in the feature space $\mathcal{W}$ the following hard margin classification problem

$$\inf\{\|w\|_{\mathcal{W}} : w \in \mathcal{W},\ b \in \mathbb{R}\} \quad (28)$$

subject to

$$y_j([\Phi(x_j), w]_{\mathcal{W}} + b) \geq 1, \quad j \in \mathbb{N}_n.$$

Provided that the minimizer $(w_0, b_0) \in \mathcal{W} \times \mathbb{R}$ exists, the classifier is $\operatorname{sgn}([\Phi(\cdot), w_0]_{\mathcal{W}} + b_0)$.

Hard margin classification in s.i.p. spaces was discussed in [20]. Applying the results in our setting tells that if b is fixed then (28) has a unique minimizer w_0 and $w_0 \in \operatorname{cone}\{y_j\Phi(x_j) : j \in \mathbb{N}_n\}$. As a corollary of Theorem 4.4, we obtain here that if $\{y_j : j \in \mathbb{N}_n\} = \{-1, 1\}$ then (28) has a minimizer (w_0, b_0), where w_0 is unique and belongs to $\operatorname{cone}\{y_j\Phi(x_j) : j \in \mathbb{N}_n\}$.

By Theorems 4.3 and 4.4, we come to the conclusion that the support vector machine classifications discussed in this subsection all reduce to a convex optimization problem.

V. Conclusion

We have introduced the notion of reproducing kernel Banach spaces and generalized under this setting the correspondence between an RKHS and its reproducing kernel. S.i.p. RKBS were specially treated by making use of semi-inner-products and the duality mapping. A semi-inner-product shares many useful properties of an inner product. These properties and the general theory of semi-inner-products make it possible to develop many learning algorithms in

RKBS. As illustration, we discussed minimal norm interpolation, regularization network, and support vector machines. Various represented theorems were established.

This work attempts to provide an appropriate mathematical foundation of kernel methods for learning in Banach spaces. Many theoretical and practical issues are left for future research. An immediate challenge is to construct a class of useful RKBS and the corresponding reproducing kernels. By the classical theory of RKHS, a function K is a reproducing kernel if and only the finite matrix (1) is always hermitian and positive semi-definite. This function property characterization brings great convenience to the construction of positive definite kernels. Thus, we ask what characteristics a function must possess so that it is a reproducing kernel for some RKBS. Properties of RKBS and their reproducing kernels also deserve a systematic study. For the applications, we have seen that minimum norm interpolation and regularization network reduce to some system of nonlinear equations. Dealing with the nonlinearity requires algorithms specially designed for the underlying s.i.p. space. On the other hand, support vector machines can be reformulated into certain convex optimization problems. We are interested in further careful analysis and efficient algorithms for these problems. We shall return to these issues in future work.

REFERENCES

[1] Y. Xu, H. Zhang and J. Zhang, "Reproducing kernel Banach spaces for machine learning," submitted.

[2] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337-404, 1950.

[3] B. Schölkopf and A.J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, Mass, 2002.

[4] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis.* Cambridge University Press, Cambridge, 2004.

[5] V.N. Vapnik, *Statistical Learning Theory.* Wiley, New York, 1998.

[6] Y. Xu and H. Zhang, "Refinable kernels," *Journal of Machine Learning Research*, vol. 8, pp. 2083-2120, 2007.

[7] Y. Xu and H. Zhang, "Refinement of reproducing kernels," *Journal of Machine Learning Research*, vol. 10, pp. 107-140, 2009.

[8] J. Mercer, "Functions of positive and negative type and their connection with the theorey of integral equations," *Philosophical Transactions of the Royal Society of London. Series A. Mathematical, Physical and Engineering Sciences*, vol. 209, pp. 415-446, 1909.

[9] T. Evgeniou, M. Pontil and T. Poggio, "Regularization networks and support vector machines," *Advances in Computational Mathematics*, vol. 13, pp. 1-50, 2000.

[10] C.A. Micchelli, Y. Xu and H. Zhang, "Optimal learning of bandlimited functions from localized sampling," *Journal of Complexity*, to appear.

[11] M. Fabian, et al., *Functional Analysis and Infinite-Dimensional Geometry.* Springer, New York, 2001.

[12] J.A. Tropp, "Just relax: convex programming methods for identifying sparse signals in noise," *IEEE Transactions on Information Theory*, vol. 52, pp. 1030-1051, 2006.

[13] K. Bennett and E. Bredensteiner, "Duality and geometry in SVM classifier," *Proceeding of the Seventeenth International Conference on Machine Learning*, P. Langley, eds., Morgan Kaufmann, San Francisco, pp. 57-64, 2000.

[14] C.A. Micchelli and M. Pontil, "A function representation for learning in Banach spaces," *Learning Theory*, pp. 255-269, Lecture Notes in Computer Science, 3120, Springer, Berlin, 2004.

[15] C.A. Micchelli and M. Pontil, "Feature space perspectives for learning the kernel," *Machine Learning*, vol. 66, pp. 297-319, 2007.

[16] C.A. Micchelli, Y. Xu and P. Ye, "Cucker Smale learning theory in Besov spaces," *Advances in Learning Theory: Methods, Models and Applications*, J. Suykens, G. Horvath, S. Basu, C. A. Micchelli and J. Vandewalle, eds., IOS Press, Amsterdam, The Netherlands, pp. 47-68, 2003.

[17] T. Zhang, "On the dual formulation of regularized linear systems with convex risks," *Machine Learning*, vol. 46, pp. 91-129, 2002.

[18] C. Gentile, "A new approximate maximal margin classification algorithm," *Journal of Machine Learning Research*, vol. 2, pp. 213-242, 2001.

[19] D. Kimber and P.M. Long, "On-line learning of smooth functions of a single variable," *Theoretical Computer Science*, vol. 148, pp. 141-156, 1995.

[20] R. Der and D. Lee, "Large-margin classification in Banach spaces," *JMLR Workshop and Conference Proceedings*, vol. 2: AISTATS, pp. 91-98, 2007.

[21] M. Hein, O. Bousquet and B. Schölkopf, "Maximal margin classification for metric spaces," *Journal of Computer and System Sciences*, vol. 71, pp. 333-359, 2005.

[22] U. von Luxburg and O. Bousquet, "Distance-based classification with Lipschitz functions," *Journal of Machine Learning Research*, vol. 5, pp. 669-695, 2004.

[23] D. Zhou, B. Xiao, H. Zhou and R. Dai, "Global geometry of SVM classifiers," *Technical Report 30-5-02*, Institute of Automation, Chinese Academy of Sciences, 2002.

[24] S. Canu, X. Mary and A. Rakotomamonjy, "Functional learning through kernel," J. Suykens, G. Horvath, S. Basu, C. Micchelli, J. Vandewalle, eds., *Advances in Learning Theory: Methods, Models and Applications*, NATO Science Series III: Computer and Systems Sciences, Volume 190, IOS Press, Amsterdam, pp. 89-110, 2003.

[25] G. Lumer, "Semi-inner-product spaces," *Transactions of the American Mathematical Society*, vol. 100, pp. 29-43, 1961.

[26] J.R. Giles, "Classes of semi-inner-product spaces," *Transactions of the American Mathematical Society*, vol. 129, pp. 436-446, 1967.

[27] B. Nath, "On a generalization of semi-inner product spaces," *Mathematical Journal of Okayama University*, vol. 15, pp. 1-6, 1971/72.

[28] J.B. Conway, *A Course in Functional Analysis.* 2nd Edition, Springer-Verlag, New York, 1990.

[29] D.F. Cudia, "On the localization and directionalization of uniform convexity," *Bulletin of the American Mathematical Society*, vol. 69, pp. 265-267, 1963.

[30] C.A. Micchelli, Y. Xu and H. Zhang, "Universal kernels," *Journal of Machine Learning Research*, vol. 7, pp. 2651-2667, 2006.

[31] I. Steinwart, "On the influence of the kernel on the consistency of support vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 67-93, 2001.

[32] A. Caponnetto, C.A. Micchelli, M. Pontil and Y. Ying, "Universal multi-task kernels," *Journal of Machine Learning Research*, vol. 9, pp. 1615-1646, 2008.

[33] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, pp. 1-49, 2002.

[34] C.A. Micchelli and M. Pontil, "On learning vector-valued functions," *Neural Computation*, vol. 17, pp. 177-204, 2005.

[35] G. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications*, vol. 33, pp. 82-95, 1971.

[36] B. Schölkopf, R. Herbrich and A.J. Smola, "A generalized representer theorem," *Proceeding of the Fourteenth Annual Conference on Computational Learning Theory and the Fifth European Conference on Computational Learning Theory*, Springer-Verlag, London, UK, pp. 416-426, 2001.

[37] G. Wahba, "Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV," *Advances in Kernel Methods–Support Vector Learning*, B. Schölkopf, C. Burges and A. J. Smola, eds., MIT Press, Cambridge, Mass, pp. 69-88, 1999.