Journal of Machine Learning Research 10 (2009) 2741-2775 Submitted 2/09; Revised 7/09; Published 12/09

Reproducing Kernel Banach Spaces for Machine Learning

Haizhang Zhang HZHANG12@SYR.EDU
Yuesheng Xu YXU06@SYR.EDU
Department of Mathematics
Syracuse University
Syracuse, NY 13244, USA

Jun Zhang JUNZ@UMICH.EDU
Department of Psychology
University of Michigan
Ann Arbor, MI 48109, USA

Editor: Ingo Steinwart

Abstract

We introduce the notion of reproducing kernel Banach spaces (RKBS) and study special semi-inner-product RKBS by making use of semi-inner-products and the duality mapping. Properties of an RKBS and its reproducing kernel are investigated. As applications, we develop in the framework of RKBS standard learning schemes including minimal norm interpolation, regularization network, support vector machines, and kernel principal component analysis. In particular, existence, uniqueness and representer theorems are established.

Keywords: reproducing kernel Banach spaces, reproducing kernels, learning theory, semi-inner-products, representer theorems

1. Introduction

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 (Aronszajn, 1950) to measure the similarity between elements in an input space X (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Vapnik, 1998; Xu and Zhang, 2007, 2009). Set $\mathbb{N}_n := \{1,2,\ldots,n\}$ for $n \in \mathbb{N}$. 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 and positive semi-definite. The reason of using positive definite kernels to measure similarity lies in the celebrated theoretical fact due to Mercer (1909) 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 $y \in X$

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

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

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

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 (Schölkopf and Smola, 2002). Based on the theory of reproducing kernels, many effective schemes have been developed for learning from finite samples (Evgeniou et al., 2000; Micchelli et al., 2009; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Vapnik, 1998). In particular, the widely used regularized learning algorithm works by generating 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 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, Fabian et al., 2001, page 180). Thus, compared to Hilbert spaces, Banach spaces possess much richer geometric structures, which are potentially useful for developing learning algorithms. 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, Tropp, 2006). 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

$$\min\{\varphi(x) + \mu \|x\|_{\ell^1} : x \in \mathbb{R}^d\} \tag{5}$$

has at most one nonzero element. The reason is that the extreme points on a sphere in the ℓ^1 norm must lie on axes of the Euclidean coordinate system. A detailed proof of this result is provided in the appendix. Thirdly, since many training data come with intrinsic structures that make them impossible to be embedded into a Hilbert space, learning algorithms based on RKHS may not work well for them. Hence, there is a need to modify the algorithms by adopting norms in Banach spaces. For example, one might have to replace the norm $\|\cdot\|_{\mathcal{H}_K}$ in (4) with that of a Banach space.

There has been considerable work on learning in Banach spaces in the literature. References Bennett and Bredensteiner (2000); Micchelli and Pontil (2004, 2007); Micchelli et al. (2003); Zhang (2002) considered the problem of minimizing a regularized functional of the form

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

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, Micchelli et al. (2003) considered learning in Besov spaces (a special type of Banach spaces). On-line learning in finite dimensional Banach spaces was studied, for example, in Gentile (2001). Learning of an L^p function was considered in Kimber and Long (1995). Classifications in Banach spaces, and more generally in metric spaces were discussed in Bennett and Bredensteiner (2000), Der and Lee (2007), Hein et al. (2005), von Luxburg and Bousquet (2004) and Zhou et al. (2002).

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 Canu et al. (2003), 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 in Section 2, and a general construction in Section 3. It will become clear that the lack of an inner product may cause arbitrariness in the properties of the associated reproducing kernel. To overcome this, we shall establish in Section 4 s.i.p. reproducing kernel Banach spaces by making use of semi-inner-products for normed vector spaces first defined by Lumer (1961) and further developed by Giles (1967). Semi-inner-products were first applied to machine learning by Der and Lee (2007) to develop hard margin hyperplane classification in Banach spaces. Here the availability of a semi-inner-product enables us to study basic properties of reproducing kernel Banach spaces and their reproducing kernels. In Section 5, we shall develop in the framework of reproducing kernel Banach spaces standard learning schemes including minimal norm interpolation, regularization network, support vector machines, and kernel principal component analysis. Existence, uniqueness and representer theorems for the learning schemes will be proved. We draw conclusive remarks in Section 6 and include two technical results in Appendix.

2. Reproducing Kernel Banach Spaces

Without specifically mentioned, all vector spaces in this 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. By this definition, $L^p[0,1]$, $1 \le p \le +\infty$, is not a Banach space of functions as it consists of equivalent classes of functions with respect to the Lebesgue measure.

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, since for each $f \in C[0,1]$,

$$f(x) = \delta_x(f), \quad x \in [0,1],$$

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 there should exist a way of identifying the elements in the dual of an RKBS with functions. Recall that two normed vector spaces V_1 and V_2 are said to be *isometric* if there is a bijective linear norm-preserving mapping between them. We call such

V_1 and V_2 an *identification* of each other. We would like the dual space $\mathcal{B}^*$ of an RKBS $\mathcal{B}$ on X to be isometric to a Banach space of functions on X. In addition to this requirement, 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 $(\mathcal{B}^*)^* = \mathcal{B}$. The above discussion leads to the following formal definition.

Definition 1 *A* **reproducing kernel Banach space** *(RKBS) on X is a reflexive Banach space $\mathcal{B}$ of functions on X for which $\mathcal{B}^*$ is isometric to a Banach space $\mathcal{B}^\#$ of functions on X and the point evaluation is continuous on both $\mathcal{B}$ and $\mathcal{B}^\#$.*

Several remarks are in order about this definition. First, whether $\mathcal{B}$ is an RKBS is independent of the choice of the identification $\mathcal{B}^\#$ of $\mathcal{B}^*$. In other words, if the point evaluation is continuous on some identification $\mathcal{B}^\#$ then it is continuous on all the identifications. The reason is that any two identifications of $\mathcal{B}^*$ are isometric. Second, an RKHS $\mathcal{H}$ on X is an RKBS. To see this, we set

$$\mathcal{H}^\# := \{\overline{f} : f \in \mathcal{H}\} \tag{6}$$

with the norm $\|\overline{f}\|_{\mathcal{H}^\#} := \|f\|_{\mathcal{H}}$, where $\overline{f}$ denotes the conjugate of f defined by $\overline{f}(x) := \overline{f(x)}$, $x \in X$. By the Riesz representation theorem (Conway, 1990), each $u^* \in \mathcal{H}^*$ has the form

$$u^*(f) = (f, f_0)_{\mathcal{H}}, \quad f \in \mathcal{H}$$

for some unique $f_0 \in \mathcal{H}$ and $\|u^*\|_{\mathcal{H}^*} = \|f_0\|_{\mathcal{H}}$. We introduce a mapping $\iota : \mathcal{H}^* \to \mathcal{H}^\#$ by setting

$$\iota(u^*) := \overline{f_0}.$$

Clearly, ι so defined is isometric from $\mathcal{H}^*$ to $\mathcal{H}^\#$. We conclude that an RKHS is a special RKBS. Third, the identification $\mathcal{B}^\#$ of $\mathcal{B}^*$ of an RKBS is usually not unique. However, since they are isometric to each other, we shall assume that one of them has been chosen for an RKBS $\mathcal{B}$ under discussion. In particular, the identification of $\mathcal{H}^*$ of an RKHS $\mathcal{H}$ will always be chosen as (6). Fourth, for notational simplicity, we shall still denote the fixed identification of $\mathcal{B}^*$ by $\mathcal{B}^*$. Let us keep in mind that originally $\mathcal{B}^*$ consists of continuous linear functionals on $\mathcal{B}$. Thus, when we shall be treating elements in $\mathcal{B}^*$ as functions on X, we actually think $\mathcal{B}^*$ as its chosen identification. With this notational convention, we state our last remark that if $\mathcal{B}$ is an RKBS on X then so is $\mathcal{B}^*$.

We shall show that there indeed exists a *reproducing kernel* for an RKBS. To this end, we introduce for a normed vector space V the following *bilinear form* on $V \times V^*$ by setting

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

It is called bilinear for the reason that for all $\alpha, \beta \in \mathbb{C}$, $u, v \in V$, and $u^*, v^* \in V^*$ there holds

$$(\alpha u + \beta v, u^*)_V = \alpha(u, u^*)_V + \beta(v, u^*)_V$$

and

$$(u, \alpha u^* + \beta v^*)_V = \alpha(u, u^*)_V + \beta(u, v^*)_V.$$

Note that if V is a reflexive Banach space then for any continuous linear functional T on V^* there exists a unique $u \in V$ such that

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

Theorem 2 *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

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

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

$$f^*(x) = (K(x,\cdot), f^*)_{\mathcal{B}}, \ \text{ for all } f^* \in \mathcal{B}^*. \tag{7}$$

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

$$\overline{\text{span}}\{K(x,\cdot) : x \in X\} = \mathcal{B}. \tag{8}$$

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

$$\overline{\text{span}}\{K(\cdot,x) : x \in X\} = \mathcal{B}^*. \tag{9}$$

(e) For all $x,y \in X$

$$K(x,y) = (K(x,\cdot), K(\cdot,y))_{\mathcal{B}}. \tag{10}$$

Proof For every $x \in X$, since δ_x is a continuous linear functional on $\mathcal{B}$, there exists $g_x \in \mathcal{B}^*$ such that

$$f(x) = (f, g_x)_{\mathcal{B}}, \ \ f \in \mathcal{B}.$$

We introduce a function $\tilde{K}$ on $X \times X$ by setting

$$\tilde{K}(x,y) := g_x(y), \ \ x,y \in X.$$

It follows that $\tilde{K}(x,\cdot) \in \mathcal{B}^*$ for each $x \in X$, and

$$f(x) = (f, \tilde{K}(x,\cdot))_{\mathcal{B}}, \ \ f \in \mathcal{B}, \ \ x \in X. \tag{11}$$

There is only one function on $X \times X$ with the above properties. Assume to the contrary that there is another $\tilde{G} : X \times X \to \mathbb{C}$ satisfying $\{\tilde{G}(x,\cdot) : x \in X\} \subseteq \mathcal{B}^*$ and

$$f(x) = (f, \tilde{G}(x,\cdot))_{\mathcal{B}}, \ \ f \in \mathcal{B}, \ \ x \in X.$$

The above equation combined with (11) yields that

$$(f, \tilde{K}(x,\cdot) - \tilde{G}(x,\cdot))_{\mathcal{B}} = 0, \ \text{ for all } f \in \mathcal{B}, \ x \in X.$$

Thus, $\tilde{K}(x,\cdot) - \tilde{G}(x,\cdot) = 0$ in $\mathcal{B}^*$ for each $x \in X$. Since $\mathcal{B}^*$ is a Banach space of functions on X, we get for every $y \in X$ that

$$\tilde{K}(x,y) - \tilde{G}(x,y) = 0,$$

that is, $\tilde{K} = \tilde{G}$.

Likewise, there exists a unique $K : X \times X \to \mathbb{C}$ such that $K(y,\cdot) \in \mathcal{B}$, $y \in X$ and

$$f^*(y) = (K(y,\cdot), f^*)_{\mathcal{B}}, \ \ f^* \in \mathcal{B}^*, \ \ y \in X. \tag{12}$$

Letting $f := K(y,\cdot)$ in (11) yields that

$$K(y,x) = (K(y,\cdot),\tilde{K}(x,\cdot))_{\mathcal{B}}, \ \ x,y \in X, \tag{13}$$

and setting $f^* := \tilde{K}(x,\cdot)$ in (12) ensures that

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

Combining the above equation with (13), we get that

$$\tilde{K}(x,y) = K(y,x), \ \ x,y \in X.$$

Therefore, K satisfies (a) and (b) as stated in the theorem. Equation (10) in (e) is proved by letting $f^* = K(\cdot,y)$ in (7). To complete the proof, we shall show (c) only, since (d) can be handled in a similar way. Suppose that (8) does not hold. Then by the Hahn-Banach theorem, there exists a nontrivial functional $f^* \in \mathcal{B}^*$ such that

$$(K(x,\cdot), f^*)_{\mathcal{B}} = 0, \ \ \text{for all } x \in X.$$

We get immediately from (12) that $f^*(x) = 0$ for all $x \in X$. Since $\mathcal{B}^*$ is a Banach space of functions on X, $f^* = 0$ in $\mathcal{B}^*$, a contradiction. ■

We call the function K in Theorem 2 the **reproducing kernel** for the RKBS $\mathcal{B}$. By Theorem 2, an RKBS has exactly one reproducing kernel. However, different RKBS may have the same reproducing kernel. Examples will be given in the next section. This results from a fundamental difference between Banach spaces and Hilbert spaces. To explain this, we let $\mathcal{W}$ be a Banach space and V a subset of $\mathcal{W}$ such that $\text{span}V$ is dense in $\mathcal{W}$. Suppose that a norm on elements of V is prescribed. If $\mathcal{W}$ is a Hilbert space and an inner product is defined among elements in V, then the norm extends in a unique way to $\text{span}V$, and hence to the whole space $\mathcal{W}$. Assume now that $\mathcal{W}$ is only known to be a Banach space and $V^* \subseteq \mathcal{W}^*$ satisfying $\overline{\text{span}}V^* = \mathcal{W}^*$ is given. Then even if a bilinear form is defined between elements in V and those in V^*, the norm may not have a unique extension to the whole space $\mathcal{W}$. Consequently, although we have at hand a reproducing kernel K for an RKBS $\mathcal{B}$, the relationship (13), and denseness conditions (8), (9), we still can not determine the norm on $\mathcal{B}$.

3. Construction of Reproducing Kernels via Feature Maps

In this section, 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. For the corresponding results in the RKHS case, see, for example, Saitoh (1997), Schölkopf and Smola (2002), Shawe-Taylor and Cristianini (2004) and Vapnik (1998).

Theorem 3 *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*

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

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

$$\|(u, \Phi^*(\cdot))_{\mathcal{W}}\|_{\mathcal{B}} := \|u\|_{\mathcal{W}} \tag{15}$$

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

$$((u, \Phi^*(\cdot))_{\mathcal{W}}, (\Phi(\cdot), u^*)_{\mathcal{W}})_{\mathcal{B}} := (u, u^*)_{\mathcal{W}}, \ \ u \in \mathcal{W}, \ u^* \in \mathcal{W}^*. \tag{16}$$

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

$$K(x, y) := (\Phi(x), \Phi^*(y))_{\mathcal{W}}, \ \ x, y \in X. \tag{17}$$

Proof We first show that $\mathcal{B}$ defined above is a Banach space of functions on X. To this end, we set $u \in \mathcal{W}$ and assume that

$$(u, \Phi^*(x))_{\mathcal{W}} = 0, \ \ \text{for all } x \in X. \tag{18}$$

Then by the denseness condition (14), $(u, u^*)_{\mathcal{W}} = 0$ for all $u^* \in \mathcal{W}^*$, implying that $u = 0$. Conversely, if $u = 0$ in $\mathcal{W}$ then it is clear that (18) holds true. These arguments also show that the representer $u \in \mathcal{W}$ for a function $(u, \Phi^*(\cdot))_{\mathcal{W}}$ in $\mathcal{B}$ is unique. It is obvious that (15) defines a norm on $\mathcal{B}$ and $\mathcal{B}$ is complete under this norm. Therefore, $\mathcal{B}$ is a Banach space of functions on X. Similarly, so is $\tilde{\mathcal{B}} := \{(\Phi(\cdot), u^*)_{\mathcal{W}} : u^* \in \mathcal{W}^*\}$ equipped with the norm

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

Define the bilinear form T on $\mathcal{B} \times \tilde{\mathcal{B}}$ by setting

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

Clearly, we have for all $u \in \mathcal{W}$, $u^* \in \mathcal{W}^*$ that

$$|T((u, \Phi^*(\cdot))_{\mathcal{W}}, (\Phi(\cdot), u^*)_{\mathcal{W}})| \le \|u\|_{\mathcal{W}} \|u^*\|_{\mathcal{W}^*} = \|(u, \Phi^*(\cdot))_{\mathcal{W}}\|_{\mathcal{B}} \, \|(\Phi(\cdot), u^*)_{\mathcal{W}}\|_{\tilde{\mathcal{B}}}.$$

Therefore, each function in $\tilde{\mathcal{B}}$ is a continuous linear functional on $\mathcal{B}$. Note that the linear mapping $u \to (u, \Phi^*(\cdot))_{\mathcal{W}}$ is isometric from $\mathcal{W}$ to $\mathcal{B}$. As a consequence, functions in $\tilde{\mathcal{B}}$ exhaust all the continuous linear functionals on $\mathcal{B}$. We conclude that $\mathcal{B}^* = \tilde{\mathcal{B}}$ with the bilinear form (16). Likewise, one can show that $\mathcal{B}$ is the dual of $\mathcal{B}^*$ by the reflexivity of $\mathcal{W}$. We have hence proved that $\mathcal{B}$ is reflexive with dual $\mathcal{B}^*$.

It remains to show that point evaluations are continuous on $\mathcal{B}$ and $\mathcal{B}^*$. To this end, we get for each $x \in X$ and $f := (u, \Phi^*(\cdot))_{\mathcal{W}}$, $u \in \mathcal{W}$ that

$$|f(x)| = |(u, \Phi^*(x))_{\mathcal{W}}| \le \|u\|_{\mathcal{W}} \|\Phi^*(x)\|_{\mathcal{W}^*} = \|f\|_{\mathcal{B}} \|\Phi^*(x)\|_{\mathcal{W}^*},$$

which implies that δ_x is continuous on $\mathcal{B}$. By similar arguments, it is continuous on $\mathcal{B}^*$. Combining all the discussion above, we reach the conclusion that $\mathcal{B}$ is an RKBS on X.

For the function K on $X \times X$ defined by (17), we get that $K(x,\cdot) \in \mathcal{B}$ and $K(\cdot,x) \in \mathcal{B}^*$ for all $x \in X$. It is also verified that for $f := (u, \Phi^*(\cdot))_{\mathcal{W}}$, $u \in \mathcal{W}$

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

Similarly, for $f^* := (\Phi(\cdot), u^*)_{\mathcal{W}}$, $u^* \in \mathcal{W}^*$

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

These facts show that K is the reproducing kernel for $\mathcal{B}$ and complete the proof. ■

We call the mappings Φ, Φ^* in Theorem 3 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 3, we obtain the following characterization of reproducing kernels for RKBS.

Theorem 4 *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 (17), where $\mathcal{W}$ is a reflexive Banach space, and mappings $\Phi : X \to \mathcal{W}$, $\Phi^* : X \to \mathcal{W}^*$ satisfy (14).*

Proof The sufficiency has been shown by the last theorem. For the necessity, we assume that K is the reproducing kernel of an RKBS $\mathcal{B}$ on X, and set

$$\mathcal{W} := \mathcal{B}, \;\; \mathcal{W}^* := \mathcal{B}^*, \;\; \Phi(x) := K(x,\cdot), \;\; \Phi^*(x) := K(\cdot,x), \;\; x \in X.$$

By Theorem 2, $\mathcal{W}, \mathcal{W}^*, \Phi, \Phi^*$ satisfy all the conditions. ■

To demonstrate how we get RKBS and their reproducing kernels by Theorem 3, 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

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

For $f \in L^1(\mathbb{R})$, its Fourier transform $\hat{f}$ is defined as

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

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

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

The Fourier transform and the inverse Fourier transform can be defined on tempered distributions. Since the Fourier transform is injective on $L^1(\mathbb{R})$ (see, Rudin, 1987, page 185), the denseness requirement (14) is satisfied.

By the construction described in Theorem 3, we obtain

$$\mathcal{B} := \{f \in C(\mathbb{R}) : \operatorname{supp} \hat{f} \subseteq \mathbb{I},\ \hat{f} \in L^p(\mathbb{I})\} \tag{19}$$

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

$$K(x,y) = (\Phi(x), \Phi^*(y))_{\mathcal{W}} = \int_{\mathbb{I}} e^{-i2\pi xt} e^{i2\pi yt} dt = \frac{\sin \pi(x-y)}{\pi(x-y)} = \operatorname{sinc}(x-y).$$

We check that for each $f \in \mathcal{B}$

$$(f, K(\cdot,x))_{\mathcal{B}} = \int_{\mathbb{I}} \hat{f}(t)(K(\cdot,x))\check{}(t)dt = \int_{\mathbb{I}} \hat{f}(t)e^{i2\pi xt}dt = f(x),\ \ x \in \mathbb{R}$$

and for each $g \in \mathcal{B}^*$

$$(K(x,\cdot), g)_{\mathcal{B}} = \int_{\mathbb{I}} (K(x,\cdot))\hat{}(t)\check{g}(t)dt = \int_{\mathbb{I}} \check{g}(t)e^{-i2\pi xt}dt = g(x),\ \ x \in \mathbb{R}.$$

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 (19) 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. We make a simple observation below to illustrate this.

Proposition 5 *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.*

Proof Let K be an arbitrary nontrivial function on $X \times X$. Assume that $X = \mathbb{N}_m$ for some $m \in \mathbb{N}$. Let $d \in \mathbb{N}$ be the rank of the matrix $K[X]$ as defined by (1). By elementary linear algebra, there exist nonsingular matrices $P, Q \in \mathbb{C}^{m\times m}$ such that the transpose $(K[X])^T$ of $K[X]$ has the form

$$(K[X])^T = P \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix} Q = P \begin{bmatrix} I_d \\ 0 \end{bmatrix} \begin{bmatrix} I_d & 0 \end{bmatrix} Q, \tag{20}$$

where I_d is the $d \times d$ identity matrix. For $j \in \mathbb{N}_m$, let P_j be the transpose of the jth row of $P \begin{bmatrix} I_d \\ 0 \end{bmatrix}$ and Q_j the jth column of $\begin{bmatrix} I_d & 0 \end{bmatrix} Q$. Choose an arbitrary $p \in (1, +\infty)$. Equation (20) is rewritten as

$$K(j,k) = (Q_j, P_k)_{l^p(\mathbb{N}_d)},\ \ j,k \in \mathbb{N}_m. \tag{21}$$

We set $\mathcal{W} := l^p(\mathbb{N}_d)$, $\mathcal{W}^* := l^q(\mathbb{N}_d)$ and $\Phi(j) := Q_j$, $\Phi^*(j) := P_j$, $j \in \mathbb{N}_m$. Since P, Q are nonsingular, (14) holds true. Also, we have by (21) that

$$K(j,k) = (\Phi(j), \Phi^*(k))_{\mathcal{W}}, \quad j,k \in \mathbb{N}_m.$$

By Theorem 4, K is a reproducing kernel for some RKBS on X. ■

Proposition 5 reveals that due to the lack of an inner product, the reproducing kernel for a general RKBS can be an arbitrary function on $X \times X$. Particularly, it might be nonsymmetric or non-positive definite. 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 (1961). 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.

4. 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 (Giles, 1967; Lumer, 1961).

4.1 Semi-Inner-Products

A **semi-inner-product** on a vector space V is a function, denoted by $[\cdot,\cdot]_V$, from $V \times V$ to $\mathbb{C}$ 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. $[\lambda x, y]_V = \lambda [x,y]_V$, $[x, \lambda y]_V = \overline{\lambda}[x,y]_V$,
3. $[x,x]_V > 0$ for $x \neq 0$,
4. (Cauchy-Schwartz) $|[x,y]_V|^2 \leq [x,x]_V [y,y]_V$.

The property that $[x, \lambda y]_V = \overline{\lambda}[x,y]_V$ was not required in the original definition by Lumer (1961). We include it here for the observation by Giles (1967) that this property can always be imposed.

It is necessary to point out the difference between a semi-inner-product and an inner product. 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 only if it degenerates to an inner product. We show this fact below.

Proposition 6 *A semi-inner-product $[\cdot,\cdot]_V$ on a complex vector space V is an inner product if and only if*

$$[x, y+z]_V = [x,y]_V + [x,z]_V, \text{ for all } x,y,z \in V. \tag{22}$$

Proof Suppose that V has a semi-inner-product $[\cdot,\cdot]_V$ that satisfies (22). It suffices to show that for all $x,y \in V$,

$$[x,y]_V = \overline{[y,x]_V}. \tag{23}$$

Set $\lambda \in \mathbb{C}$. By the linearity on the first and the additivity on the second variable, we get that

$$[x+\lambda y, x+\lambda y]_V = [x,x]_V + [\lambda y, \lambda y]_V + \lambda [y,x]_V + \bar{\lambda}[x,y]_V.$$

Since $[z,z]_V \geq 0$ for all $z \in V$, we must have

$$\lambda [y,x]_V + \bar{\lambda}[x,y]_V \in \mathbb{R}.$$

Choosing $\lambda = 1$ yields that $\operatorname{Im}[y,x]_V = -\operatorname{Im}[x,y]_V$. And the choice $\lambda = i$ results that $\operatorname{Re}[y,x]_V = \operatorname{Re}[x,y]_V$. Therefore, (23) holds, which implies that $[\cdot,\cdot]_V$ is an inner product on V. ■

It was shown in Lumer (1961) that a vector space V with a semi-inner-product is a normed space equipped with

$$\|x\|_V := [x,x]_V^{1/2}, \ \ x \in V. \tag{24}$$

Therefore, if a vector space V has a semi-inner-product, we always assume that its norm is induced by (24) and call V an **s.i.p. space**. Conversely, every normed vector space V has a semi-inner-product that induces its norm by (24) (Giles, 1967; Lumer, 1961). 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

$$[x,y]_V = y^*(x) = (x,y^*)_V, \ \ x,y \in V. \tag{25}$$

In general, a semi-inner-product for a normed vector space may not be unique. However, a differentiation property of the norm will ensure the uniqueness. We call a normed vector space V **Gâteaux differentiable** if for all $x,y \in V \setminus \{0\}$

$$\lim_{t \in \mathbb{R}, t \to 0} \frac{\|x+ty\|_V - \|x\|_V}{t}$$

exists. It is called **uniformly Fréchet differentiable** if the limit is approached uniformly on $\mathcal{S}(V) \times \mathcal{S}(V)$. Here, $S(V) := \{u \in V : \|u\|_V = 1\}$ is the unit sphere of V. The following result is due to Giles (1967).

Lemma 7 *If an s.i.p. space V is Gâteaux differentiable then for all $x,y \in V$ with $x \neq 0$*

$$\lim_{t \in \mathbb{R}, t \to 0} \frac{\|x+ty\|_V - \|x\|_V}{t} = \frac{\operatorname{Re}[y,x]}{\|x\|_V}. \tag{26}$$

The above lemma indicates that a Gâteaux differentiable normed vector space has a unique semi-inner-product. In fact, we have by (26) that

$$[x,y]_V = \|y\|_V \left(\lim_{t \in \mathbb{R}, t \to 0} \frac{\|y+tx\|_V - \|y\|_V}{t} + i \lim_{t \in \mathbb{R}, t \to 0} \frac{\|iy+tx\|_V - \|y\|_V}{t} \right), \ \ x,y \in V \setminus \{0\}. \tag{27}$$

For this reason, if V is a Gâteaux differentiable normed vector space we always assume that it is an s.i.p. space with the semi-inner-product defined as above. Interested readers are referred to the appendix for a proof that (27) indeed defines an s.i.p.

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

$$\|x+y\|_V \leq 2-\delta \text{ for all } x,y \in \mathcal{S}(V) \text{ with } \|x-y\|_V \geq \varepsilon.$$

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 Conway (1990), page 134, a uniformly convex Banach space is reflexive. There is a well-known relationship between uniform Fréchet differentiability and uniform convexity (Cudia, 1963). It states that a normed vector space is uniformly Fréchet differentiable if and only if its dual is uniformly convex. Therefore, if $\mathcal{B}$ is a uniformly convex and uniformly Fréchet differentiable Banach space then so is $\mathcal{B}^*$ since $\mathcal{B}$ is reflexive. The important role of uniform convexity is displayed in the next lemma (Giles, 1967).

Lemma 8 *(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,*

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

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

The above Riesz representation theorem is desirable for RKBS. By Lemma 8 and the discussion right before it, we shall investigate in the next subsection RKBS which are both uniformly convex and uniformly Fréchet differentiable.

Let $\mathcal{B}$ be a uniformly convex and uniformly Fréchet differentiable Banach space. By Lemma 8, $x \to x^*$ 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

$$[x^*,y^*]_{\mathcal{B}^*} = [y,x]_{\mathcal{B}}, \;\; x,y \in \mathcal{B}. \tag{28}$$

We close this subsection with a concrete example of uniformly convex and uniformly Fréchet differentiable Banach spaces. 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

$$f^* = \frac{\overline{f}|f|^{p-2}}{\|f\|^{p-2}_{L^p(\Omega,\mu)}}. \tag{29}$$

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

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

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

4.2 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 8 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.

Theorem 9 *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*

$$f(x) = [f, G(x,\cdot)]_{\mathcal{B}}, \quad \text{for all } f \in \mathcal{B},\ x \in X. \tag{30}$$

Moreover, there holds the relationship

$$K(\cdot,x) = (G(x,\cdot))^*,\ \ x \in X \tag{31}$$

and

$$f^*(x) = [K(x,\cdot), f]_{\mathcal{B}}, \quad \text{for all } f \in \mathcal{B},\ x \in X. \tag{32}$$

Proof By Lemma 8, for each $x \in X$ there exists a function $G_x \in \mathcal{B}$ such that $f(x) = [f, G_x]_{\mathcal{B}}$ for all $f \in \mathcal{B}$. We define $G : X \times X \to \mathbb{C}$ by $G(x,y) := G_x(y)$, $x,y \in X$. We see that $G(x,\cdot) = G_x \in \mathcal{B}$, $x \in X$, and there holds (30). By the uniqueness in the Riesz representation theorem, such a function G is unique. To prove the remaining claims, we recall from Theorem 2 that the reproducing kernel K satisfies for each $f \in \mathcal{B}$ that

$$f(x) = (f, K(\cdot,x))_{\mathcal{B}},\ \ x \in X. \tag{33}$$

and

$$f^*(x) = (K(x,\cdot), f^*)_{\mathcal{B}},\ \ x \in X. \tag{34}$$

By (25), (30) and (33), we have for each $x \in X$ that

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

The above equation implies (31). Equation (25) also implies that

$$(K(x,\cdot), f^*)_{\mathcal{B}} = [K(x,\cdot), f]_{\mathcal{B}}.$$

This together with equation (34) proves (32) and completes the proof. ■

We call the unique function G in Theorem 9 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 9, we call G an **s.i.p. reproducing kernel**. By (30), an s.i.p. reproducing kernel G satisfies the following generalization of (3)

$$G(x,y) = [G(x,\cdot), G(y,\cdot)]_{\mathcal{B}},\ \ x,y \in X. \tag{35}$$

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

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

Theorem 10 *Let $\mathcal{W}$ be a uniformly convex and uniformly Fréchet differentiable Banach space and Φ a mapping from X to $\mathcal{W}$ such that*

$$\overline{\text{span}}\Phi(X) = \mathcal{W}, \ \ \overline{\text{span}}\Phi^*(X) = \mathcal{W}^*. \tag{36}$$

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

$$\Big[[u,\Phi(\cdot)]_{\mathcal{W}}, [v,\Phi(\cdot)]_{\mathcal{W}}\Big]_{\mathcal{B}} := [u,v]_{\mathcal{W}} \tag{37}$$

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

$$\Big([u,\Phi(\cdot)]_{\mathcal{W}}, [\Phi(\cdot),v]_{\mathcal{W}}\Big)_{\mathcal{B}} := [u,v]_{\mathcal{W}}, \ \ u,v \in \mathcal{W}. \tag{38}$$

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

$$G(x,y) = [\Phi(x),\Phi(y)]_{\mathcal{W}}, \ \ x,y \in X, \tag{39}$$

which coincides with its reproducing kernel K.

Proof We shall show (39) only. The other results can be proved using arguments similar to those in Theorem 3 and those in the proof of Theorem 7 in Giles (1967). Let $f \in \mathcal{B}$. Then there exists a unique $u \in \mathcal{W}$ such that $f = [u,\Phi(\cdot)]_{\mathcal{W}}$. By (38), for $y \in X$,

$$f(y) = [u,\Phi(y)]_{\mathcal{W}} = ([u,\Phi(\cdot)]_{\mathcal{W}}, [\Phi(\cdot),\Phi(y)]_{\mathcal{W}})_{\mathcal{B}} = (f, [\Phi(\cdot),\Phi(y)]_{\mathcal{W}})_{\mathcal{B}}.$$

Comparing the above equation with (33), we obtain that

$$K(\cdot,y) = [\Phi(\cdot),\Phi(y)]_{\mathcal{W}}. \tag{40}$$

On the other hand, by (37), for $x \in X$

$$f(x) = [u,\Phi(x)]_{\mathcal{W}} = [[u,\Phi(\cdot)]_{\mathcal{W}}, [\Phi(x),\Phi(\cdot)]_{\mathcal{W}}]_{\mathcal{B}},$$

which implies that the s.i.p. kernel of $\mathcal{B}$ is

$$G(x,\cdot) = [\Phi(x),\Phi(\cdot)]_{\mathcal{W}}. \tag{41}$$

By (40) and (41),

$$K(x,y) = G(x,y) = [\Phi(x),\Phi(y)]_{\mathcal{W}},$$

which completes the proof. ■

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

Theorem 11 *A function G on $X \times X$ is an s.i.p. reproducing kernel if and only if it is of the form (39), where Φ is a mapping from X to a uniformly convex and uniformly Fréchet differentiable Banach space $\mathcal{W}$ satisfying (36).*

Proof The sufficiency is implied by Theorem 10. For the necessity, suppose that G is an s.i.p. reproducing kernel for some s.i.p. RKBS $\mathcal{B}$ on X. We choose $\mathcal{W} = \mathcal{B}$ and $\Phi(x) := G(x, \cdot)$. Then G has the form (39) by equation (35). Moreover, by (8), $\operatorname{span}\Phi(X)$ is dense in $\mathcal{W}$. Assume that $\operatorname{span}\Phi^*(X)$ is not dense in $\mathcal{W}^*$. Then by the Hahn-Banach theorem and Lemma 8, there exists a nontrivial $f \in \mathcal{B}$ such that $[\Phi^*(x), f^*]_{\mathcal{B}^*} = 0$, $x \in X$. Thus, by (28) we get that

$$f(x) = [f, G(x, \cdot)]_{\mathcal{B}} = [f, \Phi(x)]_{\mathcal{W}} = [\Phi^*(x), f^*]_{\mathcal{B}^*} = 0, \ \ x \in X.$$

We end up with a zero function f, a contradiction. The proof is complete. ■

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 (31) and the denseness condition (9), the s.i.p kernel G of an s.i.p. RKBS $\mathcal{B}$ on X satisfies

$$\overline{\operatorname{span}}\{(G(x, \cdot))^* : x \in X\} = \mathcal{B}^*. \tag{42}$$

It is also of the form (35). By Theorem 11, G is identical with the reproducing kernel K for $\mathcal{B}$ if and only if

$$\overline{\operatorname{span}}\{G(x, \cdot) : x \in X\} = \mathcal{B}. \tag{43}$$

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 denseness of a linear span. As a result, (43) would not follow automatically from (42). Here we remark that for most finite dimensional s.i.p. RKBS, (42) implies (43). 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, we shall give an explicit example to illustrate that the two kernels might be different.

For each $p \in (1, +\infty)$ and $n \in \mathbb{N}$, we denote by $\ell^p(\mathbb{N}_n)$ the Banach space of vectors in $\mathbb{C}^n$ with norm

$$\|a\|_{\ell^p(\mathbb{N}_n)} := \left(\sum_{j \in \mathbb{N}_n} |a_j|^p \right)^{1/p}, \ \ a = (a_j : j \in \mathbb{N}_n) \in \mathbb{C}^n.$$

As pointed out at the end of Section 4.1, $\ell^p(\mathbb{N}_n)$ is uniformly convex and uniformly Fréchet differentiable. Its dual space is $\ell^q(\mathbb{N}_n)$. To construct the example, we introduce three vectors in $\ell^3(\mathbb{N}_3)$ by setting

$$e_1 := (2, 9, 1), \ e_2 := (1, 8, 0), \ e_3 := (5, 5, 3).$$

By (29), their dual elements in $\ell^{\frac{3}{2}}(\mathbb{N}_3)$ are

$$e_1^* = \frac{1}{(738)^{1/3}}(4, 81, 1), \ e_2^* = \frac{1}{(513)^{1/3}}(1, 64, 0), \ e_3^* = \frac{1}{(277)^{1/3}}(25, 25, 9).$$

It can be directly verified that $\{e_1, e_2, e_3\}$ is linearly independent but $\{e_1^*, e_2^*, e_3^*\}$ is not. Therefore,

$$\operatorname{span}\{e_1, e_2, e_3\} = \ell^3(\mathbb{N}_3) \tag{44}$$

while

$$\operatorname{span}\{e_1^*, e_2^*, e_3^*\} \subsetneqq \ell^{\frac{3}{2}}(\mathbb{N}_3). \tag{45}$$

With the above preparations, we let $\mathbb{N}_3$ be the input space, Φ the function from $\mathbb{N}_3$ to $\ell^3(\mathbb{N}_3)$ defined by $\Phi(j) = e_j$, $j \in \mathbb{N}_3$, and $\mathcal{B}$ the space of all the functions $\Phi_u := [\Phi(\cdot), u]_{\ell^3(\mathbb{N}_3)}$, $u \in \ell^3(\mathbb{N}_3)$, on $\mathbb{N}_3$. By equation (44),

$$\|\Phi_u\| := \|u\|_{\ell^3(\mathbb{N}_3)}, \ \ u \in \ell^3(\mathbb{N}_3)$$

defines a norm on $\mathcal{B}$. It is clear that point evaluations are continuous on $\mathcal{B}$ under this norm. Furthermore, since the linear mapping $\Phi_u \to u^*$ is isometrically isomorphic from $\mathcal{B}$ to $\ell^{\frac{3}{2}}(\mathbb{N}_3)$, $\mathcal{B}$ is a uniformly convex and uniformly Fréchet differentiable Banach space. By this fact, we obtain that $\mathcal{B}$ is an s.i.p. RKBS with semi-inner-product

$$[\Phi_u, \Phi_v]_{\mathcal{B}} = [v, u]_{\ell^3(\mathbb{N}_3)}, \ \ u, v \in \ell^3(\mathbb{N}_3). \tag{46}$$

The above equation implies that the s.i.p. kernel G for $\mathcal{B}$ is

$$G(j,k) = [e_k, e_j]_{\ell^3(\mathbb{N}_3)}, \ \ j,k \in \mathbb{N}_3. \tag{47}$$

Recall that the reproducing kernel K for $\mathcal{B}$ satisfies the denseness condition (8). Consequently, to show that $G \neq K$, it suffices to show that

$$\operatorname{span}\{G(j,\cdot) : j \in \mathbb{N}_3\} \subsetneqq \mathcal{B}. \tag{48}$$

To this end, we notice by (45) that there exists a nonzero element $v \in \ell^3(\mathbb{N}_3)$ such that

$$[v, e_j]_{\ell^3(\mathbb{N}_3)} = (v, e_j^*)_{\ell^3(\mathbb{N}_3)} = 0, \ \ j \in \mathbb{N}_3.$$

As a result, the nonzero function Φ_v satisfies by (46) and (47) that

$$[G(j,\cdot), \Phi_v]_{\mathcal{B}} = [\Phi_{e_j}, \Phi_v]_{\mathcal{B}} = [v, e_j]_{\ell^3(\mathbb{N}_3)} = 0, \ \ j \in \mathbb{N}_3,$$

which proves (48), and implies that the s.i.p. kernel and reproducing kernel for $\mathcal{B}$ are different. By (45), this is essentially due to the reason that the second condition of (36) is not satisfied.

4.3 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 present below three of these properties.

4.3.1 NON-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, (39) and (36) hold. By properties 3 and 4 in the definition of a semi-inner-product, we have that

$$G(x,x) \geq 0, \ \ x \in X \tag{49}$$

and

$$|G(x,y)|^2 \leq G(x,x)G(y,y), \ \ x,y \in X. \tag{50}$$

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 (49) and (50) 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 stop being positive semi-definite if $\mathbf{x}$ contains more than two points. We shall give an explicit example to explain this phenomenon.

Set $p \in (1,+\infty)$ and $\mathcal{W} := \ell^p(\mathbb{N}_2)$. We let $X := \mathbb{R}_+ := [0,+\infty)$ and $\Phi(x) = (1,x)$, $x \in X$. Thus,

$$\Phi^*(x) = \frac{(1,x^{p-1})}{(1+x^p)^{\frac{p-2}{p}}}, \ \ x \in X.$$

Clearly, Φ satisfies the denseness condition (36). The corresponding s.i.p. reproducing kernel G is constructed as

$$G(x,y) = [\Phi(x),\Phi(y)]_{\mathcal{W}} = \frac{1+xy^{p-1}}{(1+y^p)^{\frac{p-2}{p}}}, \ \ x,y \in X. \tag{51}$$

Proposition 12 *For the s.i.p. reproducing kernel G defined by (51), matrix $G[\mathbf{x}]$ is positive semi-definite for all $\mathbf{x} = \{x,y,z\} \subseteq X$ if and only if $p = 2$.*

Proof If $p = 2$ then $\mathcal{W}$ is a Hilbert space. As a result, G is a positive definite kernel. Hence, for all finite subsets $\mathbf{x} \subseteq X$, $G[\mathbf{x}]$ is positive semi-definite.

Assume that $G[\mathbf{x}]$ is positive semi-definite for all $\mathbf{x} = \{x,y,z\} \subseteq X$. Choose $\mathbf{x} := \{0,1,t\}$ where t is a positive number to be specified later. Then we have by (51) that

$$G[\mathbf{x}] = \begin{bmatrix} 1 & 2^{2/p-1} & \dfrac{1}{(1+t^p)^{1-2/p}} \\ 1 & 2^{2/p} & \dfrac{1+t^{p-1}}{(1+t^p)^{1-2/p}} \\ 1 & \dfrac{1+t}{2^{1-2/p}} & \dfrac{1+t^p}{(1+t^p)^{1-2/p}} \end{bmatrix}.$$

Let M be the symmetrization of $G[\mathbf{x}]$ given as

$$M = \begin{bmatrix} 1 & \dfrac{1}{2}+2^{2/p-2} & \dfrac{1}{2}+\dfrac{1}{2(1+t^p)^{1-2/p}} \\ \dfrac{1}{2}+2^{2/p-2} & 2^{2/p} & \dfrac{1+t}{2^{2-2/p}}+\dfrac{1+t^{p-1}}{2(1+t^p)^{1-2/p}} \\ \dfrac{1}{2}+\dfrac{1}{2(1+t^p)^{1-2/p}} & \dfrac{1+t}{2^{2-2/p}}+\dfrac{1+t^{p-1}}{2(1+t^p)^{1-2/p}} & \dfrac{1+t^p}{(1+t^p)^{1-2/p}} \end{bmatrix}.$$

Matrix M preserves the positive semi-definiteness of $G[\mathbf{x}]$. Therefore, its determinant $|M|$ must be nonnegative. Through an analysis of the asymptotic behavior of the component of M as t goes to infinity, we obtain that

$$|M| = -\frac{t^2}{8}\left(2^{\frac{2}{p}} - 2\right)^2 + \varphi(t), \;\; t > 0,$$

where φ is a function satisfying that

$$\lim_{t\to\infty} \frac{\varphi(t)}{t^2} = 0.$$

Therefore, $|M|$ being always nonnegative forces $2^{\frac{2}{p}} - 2 = 0$, which occurs only if $p = 2$. ■

By Proposition 12, non-positive semi-definiteness is a characteristic of s.i.p. reproducing kernels for RKBS that distincts them from reproducing kernels for RKHS.

4.3.2 POINTWISE CONVERGENCE

If f_n converges to f in an s.i.p. RKBS with its 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. This follows from (30) and the Cauchy-Schwartz inequality by

$$|f_n(x) - f(x)| = |[f_n - f, G(x,\cdot)]_{\mathcal{B}}| \le \|f_n - f\|_{\mathcal{B}} \sqrt{[G(x,\cdot), G(x,\cdot)]_{\mathcal{B}}} = \sqrt{G(x,x)}\|f_n - f\|_{\mathcal{B}}.$$

4.3.3 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 $\mathcal{Z} \subseteq X$, $\operatorname{span}\{G(x,\cdot) : x \in \mathcal{Z}\}$ is dense in $C(\mathcal{Z})$ (Micchelli et al., 2006; Steinwart, 2001). 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 $\operatorname{span}\{G(x,\cdot) : x \in X\}$ is dense in $C(X)$. We shall present a characterization of weak universality. The results in the cases of positive definite kernels and vector-valued positive definite kernels have been established respectively in Micchelli et al. (2006) and Caponnetto et al. (2008).

Proposition 13 *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 (39) is continuous on $X \times X$, and there holds in $C(X)$ the equality of subspaces*

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

Consequently, G is weakly universal if and only if

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

Proof First, we notice by

$$G(x,y) = [\Phi(x), \Phi(y)]_{\mathcal{W}} = (\Phi(x), \Phi^*(y))_{\mathcal{W}}, \;\; x,y \in X$$

that G is continuous on $X \times X$. Similarly, for each $u \in \mathcal{W}$, $[u, \Phi(\cdot)]_{\mathcal{W}} = (u, \Phi^*(\cdot))_{\mathcal{W}} \in C(X)$. Now since

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

we have the inclusion

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

On the other hand, let $u \in \mathcal{W}$. By denseness condition (36), there exists a sequence $v_n \in \text{span}\{\Phi(x) : x \in X\}$ that converges to u. Since G is continuous on the compact space $X \times X$, it is bounded. Thus, by the property of pointwise convergence discussed before, $[v_n, \Phi(\cdot)]_{\mathcal{W}}$ converges in $C(X)$ to $[u, \Phi(\cdot)]_{\mathcal{W}}$. Noting that

$$[v_n, \Phi(\cdot)]_{\mathcal{W}} \in \text{span}\{G(x,\cdot) : x \in X\}, \;\; n \in \mathbb{N},$$

we have the reverse inclusion

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

which proves the result. ■

We remark that in the case that $\mathcal{W}$ is a Hilbert space, the idea in the above proof can be applied to show with less effort the main result in Caponnetto et al. (2008) and Micchelli et al. (2006), that is, for each compact subset $\mathcal{Z} \subseteq X$

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

where the two closures are taken in $C(\mathcal{Z})$. A key element missing in the Banach space is the orthogonal decomposition in a Hilbert space $\mathcal{W}$:

$$\mathcal{W} = (\overline{\text{span}}\Phi(\mathcal{Z})) \oplus \Phi(\mathcal{Z})^{\perp}, \;\; \mathcal{Z} \subseteq X.$$

For a normed vector space V, we denote for each $A \subseteq V$ by $A^{\perp} := \{v^* \in V^* : (a, v^*)_V = 0, \; a \in A\}$, and for each $B \subseteq V^*$, ${}^{\perp}B := \{a \in V : (a, v^*)_V = 0, \; v^* \in B\}$. A Banach space $\mathcal{W}$ is in general not the direct sum of $\overline{\text{span}}\Phi(\mathcal{Z})$ and ${}^{\perp}\Phi^*(\mathcal{Z})$. In fact, closed subspaces in $\mathcal{W}$ may not always have an algebraic complement unless $\mathcal{W}$ is isomorphic to a Hilbert space (see, Conway, 1990, page 94).

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

5. Representer Theorems for Standard Learning Schemes

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 map $\Phi : X \to \mathcal{W}$ as in (39). We shall develop in this framework several standard learning schemes including minimal norm interpolation, regularization network, support vector machines, and kernel principal component analysis. For introduction and discussions of these widely used algorithms in RKHS, see, for example, Cucker and Smale (2002), Evgeniou et al. (2000), Micchelli and Pontil (2005), Schölkopf and Smola (2002), Shawe-Taylor and Cristianini (2004) and Vapnik (1998).

5.1 Minimal Norm Interpolation

The minimal norm interpolation is to find, among all functions in $\mathcal{B}$ that interpolate a prescribed set of points, a function with the 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 $\mathbf{y} := (y_j : j \in \mathbb{N}_n) \in \mathbb{C}^n$

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

Our purpose is to find $f_0 \in I_{\mathbf{y}}$ such that

$$\|f_0\|_{\mathcal{B}} = \inf\{\|f\|_{\mathcal{B}} : f \in I_{\mathbf{y}}\} \tag{52}$$

provided that I_y is nonempty. Our first concern is of course the condition ensuring that $I_{\mathbf{y}}$ is nonempty. To address this issue, let us recall the useful property of the s.i.p. reproducing kernel G:

$$f(x) = [f, G(x,\cdot)]_{\mathcal{B}} = [G(\cdot,x), f^*]_{\mathcal{B}^*},\ \ x \in X,\ f \in \mathcal{B}. \tag{53}$$

Lemma 14 *The set $I_{\mathbf{y}}$ 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 $\mathcal{B}^*$.*

Proof Observe that $I_{\mathbf{y}}$ is nonempty for any $\mathbf{y} \in \mathbb{C}^n$ if and only if $\operatorname{span}\{(f(x_j) : j \in \mathbb{N}_n) : f \in \mathcal{B}\}$ is dense in $\mathbb{C}^n$. Using the reproducing property (53), we have for each $(c_j : j \in \mathbb{N}_n) \in \mathbb{C}^n$ that

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

Thus,

$$\sum_{j\in\mathbb{N}_n} c_j f(x_j) = 0, \quad \text{for all } f \in \mathcal{B}$$

if and only if

$$\sum_{j\in\mathbb{N}_n} c_j G(\cdot,x_j) = 0.$$

This implies that $G_{\mathbf{x}}$ is linearly independent in $\mathcal{B}^*$ if and only if $\operatorname{span}\{(f(x_j) : j \in \mathbb{N}_n) : f \in \mathcal{B}\}$ is dense in $\mathbb{C}^n$. Therefore, $I_{\mathbf{y}}$ is nonempty for any $\mathbf{y} \in \mathbb{C}^n$ if and only if $G_{\mathbf{x}}$ is linearly independent. ■

We next show that the minimal norm interpolation problem in $\mathcal{B}$ always has a unique solution under the hypothesis that $G_{\mathbf{x}}$ is linearly independent. The following useful property of a uniformly convex Banach space is crucial (see, for example, Istrăţescu, 1984, page 53).

Lemma 15 *If V is a uniformly convex Banach space, then for any nonempty closed convex subset $A \subseteq V$ and any $x \in V$ there exists a unique $x_0 \in A$ such that*

$$\|x - x_0\|_V = \inf\{\|x - a\|_V : a \in A\}.$$

Proposition 16 *If $G_{\mathbf{x}}$ is linearly independent in $\mathcal{B}^*$, then for any $\mathbf{y} \in \mathbb{C}^n$ there exists a unique $f_0 \in I_{\mathbf{y}}$ satisfying (52).*

Proof By Lemma 14, $I_{\mathbf{y}}$ is nonempty. Note also that it is closed and convex. Since $\mathcal{B}$ is uniformly convex, by Lemma 15, there exists a unique $f_0 \in I_{\mathbf{y}}$ such that

$$\|f_0\|_{\mathcal{B}} = \|0 - f_0\|_{\mathcal{B}} = \inf\{\|0 - f\|_{\mathcal{B}} = \|f\|_{\mathcal{B}} : f \in I_{\mathbf{y}}\}.$$

The above equation proves the result. ■

We shall establish a representation of the minimal norm interpolator f_0. For this purpose, a simple observation is made based on the following fact connecting orthogonality with best approximation in s.i.p. spaces (Giles, 1967).

Lemma 17 *If V is a uniformly Fréchet differentiable normed vector space, then $\|x + \lambda y\|_V \geq \|x\|_V$ for all $\lambda \in \mathbb{C}$ if and only if $[y, x]_V = 0$.*

Lemma 18 *If $I_{\mathbf{y}}$ is nonempty then $f_0 \in I_{\mathbf{y}}$ is the minimizer of (52) if and only if*

$$[g, f_0]_{\mathcal{B}} = 0, \quad \text{for all } g \in I_0. \tag{54}$$

Proof Let $f_0 \in I_{\mathbf{y}}$. It is obvious that f_0 is the minimizer of (52) if and only if

$$\|f_0 + g\|_{\mathcal{B}} \geq \|f_0\|_{\mathcal{B}}, \ \ g \in I_0.$$

Since I_0 is a linear subspace of $\mathcal{B}$, the result follows immediately from Lemma 17. ■

The following result is of the representer theorem type. For the representer theorem in learning with positive definite kernels, see, for example, Argyriou et al. (2008), Kimeldorf and Wahba (1971) and Schölkopf et al. (2001).

Theorem 19 *(Representer Theorem) Suppose that $G_{\mathbf{x}}$ is linearly independent in $\mathcal{B}^*$ and f_0 is the solution of the minimal norm interpolation (52). Then there exists $\mathbf{c} = (c_j : j \in \mathbb{N}_n) \in \mathbb{C}^n$ such that*

$$f_0^* = \sum_{j \in \mathbb{N}_n} c_j G(\cdot, x_j). \tag{55}$$

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

$$\left[G(\cdot, x_k), \sum_{j \in \mathbb{N}_n} c_j G(\cdot, x_j) \right]_{\mathcal{B}^*} = y_k, \ \ k \in \mathbb{N}_n. \tag{56}$$

Proof Note that (54) is equivalent to $f_0^* \in I_0{}^\perp$ and that $I_0 = {}^\perp G_{\mathbf{x}}$. Therefore, f_0 satisfies (54) if and only if

$$f_0^* \in ({}^\perp G_{\mathbf{x}})^\perp.$$

Recall a consequence of the Hahn-Banach theorem that in a reflexive Banach space $\mathcal{B}$, for each $B \subseteq \mathcal{B}^*$,

$$({}^\perp B)^\perp = \overline{\operatorname{span} B}. \tag{57}$$

By this fact, (55) holds true for some $\mathbf{c} \in \mathbb{C}^n$.

Suppose that $f \in \mathcal{B}$ is of the form $f^* = \sum_{j\in\mathbb{N}_n} c_j G(\cdot, x_j)$ where $\mathbf{c}$ satisfies (56). Then $f^* \in \operatorname{span} G_{\mathbf{x}}$. By (57), f^* satisfies (54). Furthermore, (56) implies that

$$f(x_k) = [f, G(x_k,\cdot)]_{\mathcal{B}} = [G(\cdot,x_k), f^*]_{\mathcal{B}^*} = \left[G(\cdot,x_k), \sum_{j\in\mathbb{N}_n} c_j G(\cdot,x_j)\right]_{\mathcal{B}^*} = y_k, \ \ k \in \mathbb{N}_n. \tag{58}$$

That is, $f \in I_{\mathbf{y}}$. By Lemma 18, $f = f_0$. On the other hand, f_0 has the form (55). As shown in (58), (56) is simply the interpolation condition that $f_0(x_k) = y_k$, $k \in \mathbb{N}_n$. Thus, it must be true. The proof is complete. ■

Applying the inverse of the duality mapping to both sides of (55) yields a representation of f_0 in the space $\mathcal{B}$. However, since the duality mapping is nonadditive unless $\mathcal{B}$ is an RKHS, this procedure in general does not result in a linear representation.

We conclude that under the condition that $G_{\mathbf{x}}$ is linearly independent, the minimal norm interpolation problem (52) has a unique solution, and finding the solution reduces to solving the system (56) of equations about $\mathbf{c} \in \mathbb{C}^n$. The solution $\mathbf{c}$ of (56) is unique by Theorem 19. Again, the difference from the result for RKHS is that (56) is often nonlinear about $\mathbf{c}$ since by Proposition 6 a semi-inner-product is generally nonadditive about the second variable.

To see an explicit form of (56), 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 10. Then (56) has the equivalent form

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

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 is rewritten as

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

5.2 Regularization Network

We consider learning a predictor function from a finite sample data $\mathbf{z} := \{(x_j, y_j) : j \in \mathbb{N}_n\} \subseteq X \times \mathbb{C}$ in this subsection. The predictor function will yield from a regularized learning algorithm. Let $\mathcal{L} : \mathbb{C} \times \mathbb{C} \to \mathbb{R}_+$ be a *loss function* that is continuous and convex with respect to 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 that we learn from the sample data $\mathbf{z}$ will be the function f_0 satisfying

$$\mathcal{E}_{\mathbf{z},\mu}(f_0) = \inf\{\mathcal{E}_{\mathbf{z},\mu}(f) : f \in \mathcal{B}\}. \tag{59}$$

One can always make this choice as we shall prove that the minimizer of (59) exists and is unique.

Theorem 20 *There exists a unique $f_0 \in \mathcal{B}$ satisfying (59).*

Proof We first show the existence. If $f \in \mathcal{B}$ satisfies $\|f\|_{\mathcal{B}}^2 > \frac{1}{\mu}\mathcal{E}_{\mathbf{z},\mu}(0)$ then

$$\mathcal{E}_{\mathbf{z},\mu}(f) \geq \mu\|f\|_{\mathcal{B}}^2 > \mathcal{E}_{\mathbf{z},\mu}(0).$$

Thus

$$\inf\{\mathcal{E}_{\mathbf{z},\mu}(f) : f \in \mathcal{B}\} = \inf\left\{\mathcal{E}_{\mathbf{z},\mu}(f) : f \in \mathcal{B},\ \|f\|_{\mathcal{B}}^2 \leq \frac{1}{\mu}\mathcal{E}_{\mathbf{z},\mu}(0)\right\}.$$

Let

$$e := \inf\left\{\mathcal{E}_{\mathbf{z},\mu}(f) : f \in \mathcal{B},\ \|f\|_{\mathcal{B}}^2 \leq \frac{1}{\mu}\mathcal{E}_{\mathbf{z},\mu}(0)\right\}$$

and

$$A := \left\{f \in \mathcal{B} : \|f\|_{\mathcal{B}}^2 \leq \frac{1}{\mu}\mathcal{E}_{\mathbf{z},\mu}(0)\right\}.$$

Then, there exists a sequence $f_k \in A$, $k \in \mathbb{N}$, such that

$$e \leq \mathcal{E}_{\mathbf{z},\mu}(f_k) \leq e + \frac{1}{k}. \tag{60}$$

Since $\mathcal{B}$ is reflexive, A is weakly compact, that is, we may assume that there exists $f_0 \in A$ such that for all $g \in \mathcal{B}$

$$\lim_{k\to\infty}[f_k, g]_{\mathcal{B}} = [f_0, g]_{\mathcal{B}}. \tag{61}$$

In particular, the choice $g := G(x_j, \cdot)$, $j \in \mathbb{N}_n$ yields that $f_k(x_j)$ converges to $f_0(x_j)$ as $k \to \infty$ for all $j \in \mathbb{N}_n$. Since the loss function $\mathcal{L}$ is continuous about the first variable, we have that

$$\lim_{k\to\infty}\mathcal{E}_{\mathbf{z}}(f_k) = \mathcal{E}_{\mathbf{z}}(f_0).$$

Also, choosing $g = f_0$ in (61) yields that

$$\lim_{k\to\infty}[f_k, f_0]_{\mathcal{B}} = \|f_0\|_{\mathcal{B}}^2.$$

By the above two equations, for each $\varepsilon > 0$ there exists some N such that for $k \geq N$

$$\mathcal{E}_{\mathbf{z}}(f_0) \leq \mathcal{E}_{\mathbf{z}}(f_k) + \varepsilon$$

and

$$\|f_0\|_{\mathcal{B}}^2 \leq \varepsilon\|f_0\|_{\mathcal{B}}^2 + |[f_k, f_0]_{\mathcal{B}}| \leq \varepsilon\|f_0\|_{\mathcal{B}}^2 + \|f_k\|_{\mathcal{B}}\|f_0\|_{\mathcal{B}}, \text{ if } f_0 \neq 0.$$

If $f_0 = 0$, then trivially we have

$$\|f_0\|_{\mathcal{B}} \leq \varepsilon\|f_0\|_{\mathcal{B}} + \|f_k\|_{\mathcal{B}}.$$

Combining the above three equations, we get for $k \geq N$ that

$$\mathcal{E}_{\mathbf{z},\mu}(f_0) \leq \frac{1}{(1-\varepsilon)^2}\mathcal{E}_{\mathbf{z},\mu}(f_k) + \varepsilon.$$

By (60) and the arbitrariness of ε, we conclude that f_0 is a minimizer of (59).

Since $\mathcal{L}$ is convex with respect to its first variable and $\|\cdot\|_{\mathcal{B}}^2$ is strictly convex, $\mathcal{E}_{\mathbf{z},\mu}$ is strictly convex on $\mathcal{B}$. This implies the uniqueness of the minimizer. ■

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

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

It is continuous and convex with respect to its first variable. Therefore, by Theorem 20, there is a unique minimizer for the regularization network:

$$\min_{f\in\mathcal{B}} \sum_{j\in\mathbb{N}_n} |f(x_j)-y_j|^2 + \mu\|f\|_{\mathcal{B}}^2. \tag{62}$$

We next consider solving the minimizer. To this end, we need the notion of strict convexity of a normed vector space. We call a normed vector space V **strictly convex** if whenever $\|x+y\|_V = \|x\|_V + \|y\|_V$ for some $x,y \neq 0$, there must hold $y = \lambda x$ for some $\lambda > 0$. Note that a uniformly convex normed vector space is automatically strictly convex. The following result was observed in Giles (1967).

Lemma 21 *An s.i.p. space V is strictly convex if and only if whenever $[x,y]_V = \|x\|_V\|y\|_V$ for some $x,y \neq 0$ there holds $y = \lambda x$ for some $\lambda > 0$.*

A technical result about s.i.p. spaces is required for solving the minimizer.

Lemma 22 *Let V be a strictly convex s.i.p. space. Then for all $u,v \in V$*

$$\|u+v\|_V^2 - \|u\|_V^2 - 2\,\mathrm{Re}\,[v,u]_V \geq 0.$$

Proof Assume that there exist $u,v \in V$ such that

$$\|u+v\|_V^2 - \|u\|_V^2 - 2\,\mathrm{Re}\,[v,u]_V < 0.$$

Then we have $u,v \neq 0$. We proceed by the above inequality and properties of semi-inner-products that

$$\begin{aligned}\|u\|_V^2 &= [u+v-v,u]_V = [u+v,u]_V - [v,u]_V \\ &= \mathrm{Re}\,[u+v,u]_V - \mathrm{Re}\,[v,u]_V \leq |[u+v,u]_V| - \mathrm{Re}\,[v,u]_V \\ &\leq \|u+v\|_V\|u\|_V - \frac{\|u+v\|_V^2 - \|u\|_V^2}{2}.\end{aligned}$$

The last inequality above can be simplified as

$$\|u+v\|_V^2 + \|u\|_V^2 \leq 2\|u+v\|_V\|u\|_V.$$

Thus, we must have

$$\|u+v\|_V = \|u\|_V \text{ and } [u+v,u]_V = \|u+v\|_V\|u\|_V.$$

Applying the strict convexity of V and Lemma 21, we obtain that $u+v=u$, namely, $v=0$. This is impossible. ■

Theorem 23 *(Representer Theorem) Let f_0 be the minimizer of (62). Then there exists some $\mathbf{c} \in \mathbb{C}^n$ such that*

$$f_0^* = \sum_{j\in\mathbb{N}_n} c_j G(\cdot, x_j). \tag{63}$$

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

$$\mu\overline{c_k} + \left[G(\cdot,x_k), \sum_{j\in\mathbb{N}_n} c_j G(\cdot,x_j)\right]_{\mathcal{B}^*} = y_k, \ \ k\in\mathbb{N}_n. \tag{64}$$

Proof Let $g\in\mathcal{B}$. Define the function $\phi:\mathbb{R}\to\mathbb{R}$ by $\phi(t) := \mathcal{E}_{\mathbf{z},\mu}(f_0+tg)$, $t\in\mathbb{R}$. We compute by Lemma 7 that

$$\phi'(t) = 2\,\mathrm{Re}\sum_{j\in\mathbb{N}_n} g(x_j)(\overline{f_0(x_j)-y_j} + t\overline{g(x_j)}) + 2\mu\,\mathrm{Re}\,[g, f_0+tg]_{\mathcal{B}}.$$

Since f_0 is the minimizer, $t=0$ is the minimum point of ϕ. Hence $\phi'(0)=0$, that is,

$$\sum_{j\in\mathbb{N}_n} g(x_j)\overline{f_0(x_j)-y_j} + \mu[g,f_0]_{\mathcal{B}} = 0, \ \text{ for all } g\in\mathcal{B}.$$

The above equation can be rewritten as

$$\sum_{j\in\mathbb{N}_n} [\overline{f_0(x_j)-y_j}G(\cdot,x_j), g^*]_{\mathcal{B}^*} + \mu[f_0^*, g^*]_{\mathcal{B}^*} = 0, \ \text{ for all } g\in\mathcal{B},$$

which is equivalent to

$$\mu f_0^* = \sum_{j\in\mathbb{N}_n} \overline{y_j - f_0(x_j)}G(\cdot,x_j). \tag{65}$$

Therefore, f_0^* has the form (63).

If $G_{\mathbf{x}}$ is linearly independent then $f_0^* = \sum_{j\in\mathbb{N}_n} c_j G(\cdot,x_j)$ satisfies (65) if and only if (64) holds. Thus, it remains to show that condition (65) is enough to ensure that f_0 is the minimizer. To this end, we check that (65) leads to

$$\mathcal{E}_{\mathbf{z},\mu}(f_0+g) - \mathcal{E}_{\mathbf{z},\mu}(f_0) = \mu\|f_0+g\|_{\mathcal{B}}^2 - \mu\|f_0\|_{\mathcal{B}}^2 - 2\mu\,\mathrm{Re}\,[g,f_0]_{\mathcal{B}} + \sum_{j\in\mathbb{N}_n}|g(x_j)|^2,$$

which by Lemma 22 is nonnegative. The proof is complete. ■

By Theorem 23, if $G_{\mathbf{x}}$ is linearly independent then the minimizer of (62) can be obtained by solving (64), which has a unique solution in this case. Using the feature map, the system (64) has the following form

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

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

5.3 Support Vector Machines

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 two learning algorithms in RKBS whose RKHS versions are well-known (Evgeniou et al., 2000; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Vapnik, 1998; Wahba, 1999).

5.3.1 SOFT MARGIN HYPERPLANE CLASSIFICATION

We first focus on the soft margin hyperplane classification by studying

$$\inf\left\{\frac{1}{2}\|w\|_{\mathcal{W}}^2 + C\|\xi\|_{\ell^1(\mathbb{N}_n)} : w \in \mathcal{W},\ \xi := (\xi_j : j \in \mathbb{N}_n) \in \mathbb{R}_+^n,\ b \in \mathbb{R}\right\} \tag{66}$$

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 a minimizer $(w_0, \xi_0, b_0) \in \mathcal{W} \times \mathbb{R}_+^n \times \mathbb{R}$ exists, the classifier is taken as $\operatorname{sgn}([\Phi(\cdot), w_0]_{\mathcal{W}} + b_0)$.

Recall by Theorem 10 that functions in $\mathcal{B}^*$ are of the form

$$f^* = [\Phi(\cdot), w]_{\mathcal{W}}, \quad w \in \mathcal{W} \tag{67}$$

and

$$\|f^*\|_{\mathcal{B}^*} = \|w\|_{\mathcal{W}}.$$

Introduce the loss function

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

and the error functional on $\mathcal{B}^*$,

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

where $\mu := 1/(2C)$. Then we observe that (66) can be equivalently rewritten as

$$\inf\{\mathcal{E}_{b,\mathbf{z},\mu}(f^*) : f^* \in \mathcal{B}^*, \quad b \in \mathbb{R}\}. \tag{68}$$

When $b = 0$, (68) is also called the support vector machine classification (Wahba, 1999).

If a minimizer $(f_0^*, b_0) \in \mathcal{B}^* \times \mathbb{R}$ for (68) exists then by (67), the classifier followed from (66) will be taken as $\operatorname{sgn}(f_0^* + b_0)$. It can be verified that for every $b \in \mathbb{R}$, $\mathcal{L}_b$ is convex and continuous with respect to its first variable. This enables us to prove the existence of minimizers for (68) based on Theorem 20.

Proposition 24 *Suppose that $\{y_j : j \in \mathbb{N}_n\} = \{-1, 1\}$. Then there exists a minimizer $(f_0^*, b_0) \in \mathcal{B}^* \times \mathbb{R}$ for (68). Moreover, the first component f_0^* of the minimizer is unique.*

Proof The uniqueness of f_0^* follows from the fact that $\mathcal{E}_{b,\mathbf{z},\mu}$ is strictly convex with respect to its first variable. It remains to deal with the existence. Let e be the infimum (68). Then there exists a sequence $(f_k^*, b_k^*) \in \mathcal{B}^* \times \mathbb{R}$, $k \in \mathbb{N}$ such that

$$\lim_{k\to\infty} \mathcal{E}_{b_k,\mathbf{z},\mu}(f_k^*) = e. \tag{69}$$

Since $\{y_j : j \in \mathbb{N}_n\} = \{-1,1\}$, $\{b_k : k \in \mathbb{N}\}$ is bounded on $\mathbb{R}$. We may hence assume that b_k converges to some $b_0 \in \mathbb{R}$. By Theorem 20, $\min\{\mathcal{E}_{b_0,\mathbf{z},\mu}(f^*) : f^* \in \mathcal{B}^*\}$ has a unique minimizer f_0^*. The last fact we shall need is the simple observation that

$$\max\{1-ay-by,0\} - \max\{1-ay-b'y,0\} \le |b-b'| \text{ for all } a,b,b' \in \mathbb{R} \text{ and } y \in \{-1,1\}.$$

Thus, we get for all $k \in \mathbb{N}$ that

$$\mathcal{E}_{b_0,\mathbf{z},\mu}(f_0^*) \le \mathcal{E}_{b_0,\mathbf{z},\mu}(f_k^*) \le \mathcal{E}_{b_k,\mathbf{z},\mu}(f_k^*) + n|b_0 - b_k|,$$

which together with (69) and $\lim_{k\to\infty} b_k = b_0$ implies that (f_0^*, b_0) is a minimizer for (68). ■

Since the soft margin hyperplane classification (66) is equivalent to (68), we obtain by Proposition 24 that it has a minimizer $(w_0, \xi_0, b_0) \in \mathcal{W} \times \mathbb{R}_+^n \times \mathbb{R}$, where the first component w_0 is unique.

We shall prove a representer theorem for f_0^* using the following celebrated geometric consequence of the Hahn-Banach theorem (see, Conway, 1990, page 111).

Lemma 25 *Let A be a closed convex subset of a normed vector space V and b a point in V that is not contained in A. Then there exist $T \in V^*$ and $\alpha \in \mathbb{R}$ such that*

$$T(b) < \alpha < T(a), \text{ for all } a \in A.$$

Theorem 26 *(Representer Theorem) Let f_0^* be the minimizer of (68). Then f_0 lies in the closed convex cone $\operatorname{co} G_{\mathbf{z}}$ spanned by $G_{\mathbf{z}} := \{y_j G(x_j,\cdot) : j \in \mathbb{N}_n\}$, that is, there exist $\lambda_j \ge 0$ such that*

$$f_0 = \sum_{j\in\mathbb{N}_n} \lambda_j y_j G(x_j,\cdot). \tag{70}$$

Consequently, the minimizer w_0 of (66) belongs to the closed convex cone spanned by $y_j\Phi(x_j)$, $j \in \mathbb{N}_n$.

Proof Assume that $f_0 \notin \operatorname{co} G_{\mathbf{z}}$. By Lemmas 25 and 8, there exists $g \in \mathcal{B}$ and $\alpha \in \mathbb{R}$ such that for all $\lambda \ge 0$

$$[f_0, g]_{\mathcal{B}} < \alpha < [\lambda y_j G(x_j,\cdot), g]_{\mathcal{B}} = \lambda y_j g^*(x_j), \quad j \in \mathbb{N}_n.$$

Choosing $\lambda = 0$ above yields that $\alpha < 0$. That $\lambda y_j g^*(x_j) > \alpha$ for all $\lambda \ge 0$ implies $y_j g^*(x_j) \ge 0$, $j \in \mathbb{N}_n$. We hence obtain that

$$[f_0, g]_{\mathcal{B}} < 0 \le y_j g^*(x_j), \quad j \in \mathbb{N}_n.$$

We choose $f^* = f_0^* + tg^*$, $t > 0$. First, observe from $y_j g^*(x_j) \ge 0$ that

$$1 - y_j f^*(x_j) \le 1 - y_j f_0^*(x_j), \quad j \in \mathbb{N}_n. \tag{71}$$

By Lemma 7,

$$\lim_{t\to 0^+} \frac{\|f_0^* + tg^*\|_{\mathcal{B}^*}^2 - \|f_0^*\|_{\mathcal{B}^*}^2}{t} = 2[g^*, f_0^*]_{\mathcal{B}^*} = 2[f_0, g]_{\mathcal{B}} < 0.$$

Therefore, there exists $t > 0$ such that $\|f^*\|_{\mathcal{B}^*}^2 < \|f_0^*\|_{\mathcal{B}^*}^2$. This combined with (71) implies that $\mathcal{E}_{b,\mathbf{z},\mu}(f^*) < \mathcal{E}_{b,\mathbf{z},\mu}(f_0^*)$ for every $b \in \mathbb{R}$, contradicting the hypothesis that f_0^* is the minimizer. ■

To solve (68), by Theorem 26 one substitutes equation (70) into (68) to obtain a convex optimization problem about λ_j subject to the constraint that $\lambda_j \geq 0$, $j \in \mathbb{N}_n$.

5.3.2 HARD MARGIN HYPERPLANE CLASSIFICATION

Consider in the feature space $\mathcal{W}$ the following hard margin classification problem

$$\inf\{\|w\|_{\mathcal{W}} : w \in \mathcal{W},\ b \in \mathbb{R}\} \tag{72}$$

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 Der and Lee (2007). Applying the results in our setting tells that if b is fixed then (72) has a unique minimizer w_0 and $w_0 \in \operatorname{co}\{y_j\Phi(x_j) : j \in \mathbb{N}_n\}$. As a corollary of Proposition 24 and Theorem 26, we obtain here that if $\{y_j : j \in \mathbb{N}_n\} = \{-1, 1\}$ then (72) has a minimizer (w_0, b_0), where w_0 is unique and belongs to the set $\operatorname{co}\{y_j\Phi(x_j) : j \in \mathbb{N}_n\}$.

We draw the conclusion that the support vector machine classifications in this subsection all reduce to a convex optimization problem.

5.4 Kernel Principal Component Analysis

Kernel principal component analysis (PCA) plays a foundational role in data preprocessing for other learning algorithms. We shall present an extension of kernel PCA for RKBS. To this end, let us briefly review the classical kernel PCA (see, for example, Schölkopf and Smola, 2002; Schölkopf et al., 1998).

Let $\mathbf{x} := \{x_j : j \in \mathbb{N}_n\} \subseteq X$ be a set of inputs. We denote by $d(w, V)$ the distance from $w \in \mathcal{W}$ to a closed subspace V of $\mathcal{W}$. Fix $m \in \mathbb{N}$. For each subspace $V \subseteq \mathcal{W}$ with dimension $\dim V = m$, we define the distance from V to $\Phi(\mathbf{x})$ as

$$\mathcal{D}(V, \Phi(\mathbf{x})) := \frac{1}{n} \sum_{j\in\mathbb{N}_n} (d(\Phi(x_j), V))^2.$$

Suppose that $\{u_j : j \in \mathbb{N}_m\}$ is a basis for V. Then for each $v \in \mathcal{W}$ the best approximation v_0 in V of v exists. Assume that $v_0 = \sum_{j\in\mathbb{N}_m} \lambda_j u_j$, $\lambda_j \in \mathbb{C}$, $j \in \mathbb{N}_m$. By Lemma 17, the coefficients λ_j's are uniquely determined by

$$\left[u_k, v - \sum_{j\in\mathbb{N}_m} \lambda_j u_j\right]_{\mathcal{W}} = 0, \quad k \in \mathbb{N}_m. \tag{73}$$

In the case when $\mathcal{W}$ is a Hilbert space, the system (73) of equations resulting from best approximation is linear about λ_j's. This enables us to construct a unique m-dimensional subspace $V_0 \subseteq \mathcal{W}$ such that

$$\mathcal{D}(V_0, \Phi(\mathbf{x})) = \min\{\mathcal{D}(V, \Phi(\mathbf{x})) : V \subseteq \mathcal{W} \text{ subspace, } \dim V = m\}. \tag{74}$$

Let T be the compact operator on $\mathcal{W}$ determined by

$$(Tu, v)_{\mathcal{W}} = \frac{1}{n} \sum_{j \in \mathbb{N}_n} (u, \Phi(x_j))_{\mathcal{W}} (\Phi(x_j), v)_{\mathcal{W}}, \quad u, v \in \mathcal{W}. \tag{75}$$

We let v_j, $j \in \mathbb{N}_m$ be the unit eigenvectors of T corresponding to its first m largest eigenvalues. Then v_j's form an orthonormal basis for V_0 and are called the *principal components* of $\Phi(\mathbf{x})$. For each $x \in X$, $((\Phi(x), v_j)_{\mathcal{W}} : j \in \mathbb{N}_m) \in \mathbb{C}^m$ is its new *feature*. Therefore, kernel PCA amounts to selecting the new feature map from X to $\mathbb{C}^m$. The dimension m is usually chosen to be much smaller than the original dimension of $\mathcal{W}$. Moreover, by (74), the new features of $\mathbf{x}$ are expected to become sparser under this mapping.

The analysis of PCA in Hilbert spaces breaks in s.i.p. spaces where (73) is nonlinear. To tackle this problem, we suggest using a class of linear functionals to measure the distance between two elements in $\mathcal{W}$. Specifically, we choose $B \subseteq \mathcal{W}^*$ and set for all $u, v \in \mathcal{W}$

$$d_B(u, v) := \left(\sum_{b \in B} |(u - v, b)_{\mathcal{W}}|^2 \right)^{1/2}.$$

The idea is that if $d_B(u, v)$ is small for a carefully chosen set B of linear functionals then $\|u - v\|_{\mathcal{W}}$ should be small, and vice versa. In particular, if $\mathcal{W}$ is a Hilbert space and B is an orthonormal basis for $\mathcal{W}$ then $d_B(u, v) = \|u - v\|_{\mathcal{W}}$. From the practical consideration, we shall use what we have at hand, that is, $\Phi(\mathbf{x})$. Thus, we define for each $u, v \in \mathcal{W}$

$$d_{\Phi(\mathbf{x})}(u, v) := \left(\sum_{j \in \mathbb{N}_n} \left| [u - v, \Phi(x_j)]_{\mathcal{W}} \right|^2 \right)^{1/2}.$$

This choice of distance is equivalent to mapping X into $\mathbb{C}^n$ by

$$\widetilde{\Phi}(x) := ([\Phi(x), \Phi(x_j)]_{\mathcal{W}} : j \in \mathbb{N}_n), \quad x \in X.$$

Consequently, new features of elements in X will be obtained by applying the classical PCA to $\widetilde{\Phi}(x_j)$, $j \in \mathbb{N}_n$ in the Hilbert space $\mathbb{C}^n$.

In our method the operator T defined by (75) on $\mathbb{C}^n$ is of the form

$$Tu = \frac{1}{n} \sum_{j \in \mathbb{N}_n} (\widetilde{\Phi}(x_j)^* u) \widetilde{\Phi}(x_j), \quad u \in \mathbb{C}^n,$$

where $\widetilde{\Phi}(x_j)^*$ is the conjugate transpose of $\widetilde{\Phi}(x_j)$. One can see that T has the matrix representation $Tu = M_{\mathbf{x}} u$, $u \in \mathbb{C}^n$, where

$$M_{\mathbf{x}} := \frac{1}{n} (G[\mathbf{x}]^* G[\mathbf{x}])^T.$$

Let λ_k, $k \in \mathbb{N}_n$, be the eigenvalues of $M_{\mathbf{x}}$ arranged in nondecreasing order. We find for each $k \in \mathbb{N}_m$ the unit eigenvector $\alpha^k := (\alpha_j^k : j \in \mathbb{N}_n) \in \mathbb{C}^n$ corresponding to λ_k, that is,

$$M_{\mathbf{x}}\alpha^k = \lambda_k \alpha^k.$$

Vectors α^k, $k \in \mathbb{N}_m$, form an orthonormal sequence. The new feature for $x \in X$ is hence

$$((\widetilde{\Phi}(x), \alpha^k)_{\mathbb{C}^n} : k \in \mathbb{N}_m) \in \mathbb{C}^m.$$

We compute explicitly that

$$(\widetilde{\Phi}(x), \alpha^k)_{\mathbb{C}^n} = \sum_{j \in \mathbb{N}_n} \overline{\alpha_j^k} G(x, x_j), \quad k \in \mathbb{N}_m.$$

We remark that unlike the previous three learning algorithms, the kernel PCA presented here only makes use of the kernel G and is independent of the semi-inner-product on $\mathcal{W}$. The *kernel trick* can hence be applied to this algorithm.

6. Conclusion

We have introduced the notion of reproducing kernel Banach spaces and generalized the correspondence between an RKHS and its reproducing kernel to the setting of RKBS. 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 in the RKBS setting the minimal norm interpolation, regularization network, support vector machines, and kernel PCA. Various representer 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 if 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 systems 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. Finally, section 5.4 only provides a possible implementation of kernel PCA for RKBS. We are interested in further careful analysis and efficient algorithms for these problems. We shall return to these issues in future work.

Acknowledgments

Yuesheng Xu was supported in part by the US National Science Foundation under grant DMS-0712827. Jun Zhang was supported in part by the US National Science Foundation under grant

0631541. This research was accomplished when the last author (J.Z.) was on leave from the University of Michigan to AFOSR under an IPA assignment.

Send all correspondence to Yuesheng Xu.

Appendix A.

In this appendix, we provide proofs of two results stated in the previous sections of this paper. The first one is about the minimization problem (5) in the introduction.

Proposition 27 *If $\varphi : \mathbb{R}^d \to [0,+\infty)$ is strictly concave and $\mu > 0$, then every minimizer of*

$$\min\{\varphi(x) + \mu\|x\|_{\ell^1} : x \in \mathbb{R}^d\} \tag{76}$$

has at most one nonzero element.

Proof Assume to the contrary that $x_0 \in \mathbb{R}^d$ is a minimizer of (76) with more than one nonzero elements. Then x_0 is not an extreme points of the sphere $\{x \in \mathbb{R}^d : \|x\|_{\ell^1} = \|x_0\|_{\ell^1}\}$. In other words, there exist two distinct vectors $x_1, x_2 \in \mathbb{R}^d$ and some $\lambda \in (0,1)$ such that

$$x_0 = \lambda x_1 + (1-\lambda)x_2 \text{ and } \|x_1\|_{\ell^1} = \|x_2\|_{\ell^1} = \|x_0\|_{\ell^1}.$$

By the strict concavity of φ, we get that

$$\begin{aligned} \varphi(x_0) + \mu\|x_0\|_{\ell^1} &> \lambda\varphi(x_1) + (1-\lambda)\varphi(x_2) + \mu\|x_0\|_{\ell^1} \\ &= \lambda(\varphi(x_1) + \mu\|x_1\|_{\ell^1}) + (1-\lambda)(\varphi(x_2) + \mu\|x_2\|_{\ell^1}). \end{aligned}$$

Therefore, we must have either

$$\varphi(x_0) + \mu\|x_0\|_{\ell^1} > \varphi(x_1) + \mu\|x_1\|_{\ell^1}$$

or

$$\varphi(x_0) + \mu\|x_0\|_{\ell^1} > \varphi(x_2) + \mu\|x_2\|_{\ell^1}.$$

Either case contradicts the hypothesis that x_0 is a minimizer of (76). ■

The second result confirms that (27) indeed defines a semi-inner-product.

Proposition 28 *Let V be a normed vector space over $\mathbb{C}$. If for all $x, y \in V \setminus \{0\}$ the limit $\lim_{t\in\mathbb{R}, t\to 0} \dfrac{\|x+ty\|_V - \|x\|_V}{t}$ exists then $[\cdot,\cdot]_V : V \times V \to \mathbb{C}$ defined by*

$$[x,y]_V := \|y\|_V \left(\lim_{t\in\mathbb{R}, t\to 0} \frac{\|y+tx\|_V - \|y\|_V}{t} + i \lim_{t\in\mathbb{R}, t\to 0} \frac{\|iy+tx\|_V - \|y\|_V}{t} \right) \text{ if } x, y \neq 0 \tag{77}$$

and $[x,y]_V := 0$ if $x = 0$ or $y = 0$ is a semi-inner-product on V.

Proof First, we obtain for $x \neq 0$ that

$$\begin{aligned}[x,x]_V &= \|x\|_V \left(\lim_{t\in\mathbb{R},t\to 0} \frac{\|(1+t)x\|_V - \|x\|_V}{t} + i \lim_{t\in\mathbb{R},t\to 0} \frac{\|(i+t)x\|_V - \|x\|_V}{t} \right) \\ &= \|x\|_V^2 \left(\lim_{t\in\mathbb{R},t\to 0} \frac{|1+t|-1}{t} + i \lim_{t\in\mathbb{R},t\to 0} \frac{|i+t|-1}{t} \right) \\ &= \|x\|_V^2 (1+0) = \|x\|_V^2 > 0. \end{aligned} \tag{78}$$

We then deal with the remaining three conditions of a semi-inner-product. Clearly, they are true if one of the arguments involved is the zero vector. Let $x,y,z \in V \setminus \{0\}$. We start with the estimate:

$$\begin{aligned}\operatorname{Re}[x+y,z]_V &= \|z\|_V \lim_{t\in\mathbb{R},t\to 0^+} \frac{\|z+tx+ty\|_V - \|z\|_V}{t} \\ &\leq \|z\|_V \lim_{t\in\mathbb{R},t\to 0^+} \frac{\|\frac{z}{2}+tx\|_V + \|\frac{z}{2}+ty\|_V - \|z\|_V}{t} \\ &= \|z\|_V \left(\lim_{t\in\mathbb{R},t\to 0^+} \frac{\|\frac{z}{2}+tx\|_V - \|\frac{z}{2}\|_V}{t} + \lim_{t\in\mathbb{R},t\to 0^+} \frac{\|\frac{z}{2}+ty\|_V - \|\frac{z}{2}\|_V}{t} \right) \\ &= \|z\|_V \left(\lim_{t\in\mathbb{R},t\to 0^+} \frac{\|z+2tx\|_V - \|z\|_V}{2t} + \lim_{t\in\mathbb{R},t\to 0^+} \frac{\|z+2ty\|_V - \|z\|_V}{2t} \right). \end{aligned}$$

The above equation implies that

$$\operatorname{Re}[x+y,z]_V \leq \operatorname{Re}[x,z]_V + \operatorname{Re}[y,z]_V. \tag{79}$$

It can be easily verified that $[-x,y]_V = -[x,y]_V$. Replacing y with $x+y$, and x with $-x$ in the above equation yields that

$$\operatorname{Re}[y,z]_V \leq \operatorname{Re}[-x,z]_V + \operatorname{Re}[x+y,z]_V = -\operatorname{Re}[x,z]_V + \operatorname{Re}[x+y,z]_V. \tag{80}$$

Combining (79) and (80), we get that $\operatorname{Re}[x+y,z]_V = \operatorname{Re}[x,z]_V + \operatorname{Re}[y,z]_V$. Similar arguments lead to that $\operatorname{Im}[x+y,z]_V = \operatorname{Im}[x,z]_V + \operatorname{Im}[y,z]_V$. Therefore,

$$[x+y,z]_V = [x,z]_V + [y,z]_V. \tag{81}$$

Next we see for all $\lambda \in \mathbb{R} \setminus \{0\}$ that

$$[\lambda x,y]_V = \|y\|_V \lim_{t\in\mathbb{R},t\to 0} \frac{\|y+t\lambda x\|_V - \|y\|_V}{t} = \lambda \|y\|_V \lim_{t\in\mathbb{R},t\to 0} \frac{\|y+t\lambda x\|_V - \|y\|_V}{\lambda t} = \lambda [x,y]_V.$$

It is also clear from the definition (77) that $[ix,y]_V = i[x,y]_V$. We derive from these two facts and (81) for every $\lambda = \alpha + i\beta$, $\alpha,\beta \in \mathbb{R}$ that

$$\begin{aligned}[\lambda x,y]_V &= [\alpha x + i\beta x,y]_V = [\alpha x,y]_V + [i\beta x,y]_V = \alpha[x,y]_V + i[\beta x,y]_V \\ &= \alpha[x,y]_V + i\beta[x,y]_V = (\alpha+i\beta)[x,y]_V = \lambda[x,y]_V. \end{aligned} \tag{82}$$

We then proceed for $\lambda \in \mathbb{C} \setminus \{0\}$ by (82) that

$$\begin{aligned}[x,\lambda y]_V &= \|\lambda y\|_V \lim_{t\in\mathbb{R},t\to 0} \frac{\|\lambda y + tx\|_V - \|\lambda y\|_V}{t} = \|\lambda y\|_V |\lambda| \lim_{t\in\mathbb{R},t\to 0} \frac{\|y + t\frac{x}{\lambda}\|_V - \|y\|_V}{t} \\ &= |\lambda|^2 [\frac{x}{\lambda},y]_V = \frac{|\lambda|^2}{\lambda}[x,y]_V = \overline{\lambda}[x,y]_V. \end{aligned} \tag{83}$$

Finally, we find some $\lambda \in \mathbb{C}$ such that $|\lambda| = 1$ and $\lambda[x,y]_V = |[x,y]_V|$, and then obtain by (82) and (77) that

$$\begin{aligned} |[x,y]_V| &= \lambda[x,y]_V = [\lambda x,y]_V = \|y\|_V \lim_{t\in\mathbb{R},\, t\to 0^+} \frac{\|y+t\lambda x\|_V - \|y\|_V}{t} \\ &\le \|y\|_V \lim_{t\in\mathbb{R},\, t\to 0^+} \frac{\|y\|_V + t\|\lambda x\|_V - \|y\|_V}{t} = \|y\|_V \|\lambda x\|_V = \|x\|_V \|y\|_V. \end{aligned}$$

By (78), the above inequality has the equivalent form

$$|[x,y]_V| \le [x,x]_V^{1/2} [y,y]_V^{1/2}. \tag{84}$$

Combining Equations (78), (81), (82), (83), and (84) proves the proposition. ■

References

A. Argyriou, C. A. Micchelli and M. Pontil. When is there a representer theorem? Vector versus matrix regularizers. arXiv:0809.1590v1.

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68: 337–404, 1950.

K. P. Bennett and E. J. Bredensteiner. Duality and geometry in SVM classifiers. In *Proceeding of the Seventeenth International Conference on Machine Learning*, pages 57–64, Morgan Kaufmann, San Francisco, 2000.

S. Canu, X. Mary and A. Rakotomamonjy. Functional learning through kernel. In *Advances in Learning Theory: Methods, Models and Applications*, pages 89–110, NATO Science Series III: Computer and Systems Sciences, Volume 190, IOS Press, Amsterdam, 2003.

A. Caponnetto, C. A. Micchelli, M. Pontil and Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9: 1615–1646, 2008.

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

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39: 1–49, 2002.

D. F. Cudia. On the localization and directionalization of uniform convexity. *Bull. Amer. Math. Soc.*, 69: 265–267, 1963.

R. Der and D. Lee. Large-margin classification in Banach spaces. *JMLR Workshop and Conference Proceedings*, 2: AISTATS: 91–98, 2007.

T. Evgeniou, M. Pontil and T. Poggio. Regularization networks and support vector machines. *Adv. Comput. Math.*, 13: 1–50, 2000.

M. J. Fabian, P. Habala, P. Hajek and J. Pelant. *Functional Analysis and Infinite-Dimensional Geometry*. Springer, New York, 2001.

C. Gentile. A new approximate maximal margin classification algorithm. *Journal of Machine Learning Research*, 2: 213–242, 2001.

J. R. Giles. Classes of semi-inner-product spaces. *Trans. Amer. Math. Soc.*, 129: 436–446, 1967.

M. Hein, O. Bousquet and B. Schölkopf. Maximal margin classification for metric spaces. *J. Comput. System Sci.*, 71: 333–359, 2005.

V. I. Istrăţescu. *Strict Convexity and Complex Strict Convexity: Theory and Applications*. Lecture Notes in Pure and Applied Mathematics 89, Marcel Dekker, New York, 1984.

D. Kimber and P. M. Long. On-line learning of smooth functions of a single variable. *Theoret. Comput. Sci.*, 148: 141–156, 1995.

G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.*, 33: 82–95, 1971.

G. Lumer. Semi-inner-product spaces. *Trans. Amer. Math. Soc.*, 100: 29–43, 1961.

J. Mercer. Functions of positive and negative type and their connection with the theorey of integral equations. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 209: 415–446, 1909.

C. A. Micchelli and M. Pontil. A function representation for learning in Banach spaces. In *Learning Theory*, pages 255–269, Lecture Notes in Computer Science 3120, Springer, Berlin, 2004.

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Comput.*, 17: 177–204, 2005.

C. A. Micchelli and M. Pontil. Feature space perspectives for learning the kernel. *Machine Learning*, 66: 297–319, 2007.

C. A. Micchelli, Y. Xu and P. Ye. Cucker Smale learning theory in Besov spaces. In *Advances in Learning Theory: Methods, Models and Applications*, pages 47–68, IOS Press, Amsterdam, The Netherlands, 2003.

C. A. Micchelli, Y. Xu and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7: 2651–2667, 2006.

C. A. Micchelli, Y. Xu and H. Zhang. Optimal learning of bandlimited functions from localized sampling. *J. Complexity*, 25: 85–114, 2009.

W. Rudin. *Real and Complex Analysis*. 3rd Edition, McGraw-Hill, New York, 1987.

S. Saitoh. *Integral Transforms, Reproducing Kernels and Their Applications*. Pitman Research Notes in Mathematics Series 369, Longman, Harlow, 1997.

B. Schölkopf, R. Herbrich and A. J. Smola. A generalized representer theorem. In *Proceeding of the 14th Annual Conference on Computational Learning Theory and the 5th European Conference on Computational Learning Theory*, pages 416–426, Springer-Verlag, London, UK, 2001.

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

B. Schölkopf, A. J. Smola and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10: 1299–1319, 1998.

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

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2: 67–93, 2001.

J. A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory*, 52: 1030–1051, 2006.

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

U. von Luxburg and O. Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5: 669–695, 2004.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In *Advances in Kernel Methods–Support Vector Learning*, pages 69–86, MIT Press, Cambridge, Mass, 1999.

Y. Xu and H. Zhang. Refinable kernels. *Journal of Machine Learning Research*, 8: 2083–2120, 2007.

Y. Xu and H. Zhang. Refinement of reproducing kernels. *Journal of Machine Learning Research*, 10: 107–140, 2009.

T. Zhang. On the dual formulation of regularized linear systems with convex risks. *Machine Learning*, 46: 91–129, 2002.

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.