J Glob Optim (2012) 54:235–250
DOI 10.1007/s10898-010-9575-z

Regularized learning in Banach spaces as an optimization problem: representer theorems

Haizhang Zhang · Jun Zhang

Received: 9 September 2009 / Accepted: 22 June 2010 / Published online: 11 July 2010

Abstract We view regularized learning of a function in a Banach space from its finite samples as an optimization problem. Within the framework of reproducing kernel Banach spaces, we prove the representer theorem for the minimizer of regularized learning schemes with a general loss function and a nondecreasing regularizer. When the loss function and the regularizer are differentiable, a characterization equation for the minimizer is also established.

Keywords Reproducing kernel Banach spaces · Semi-inner products · Representer theorems · Regularization networks · Support vector machine classification

1 Introduction

Many scientific questions boil down to the learning of an input-output mapping when only finite samples are known [10,13,27,29,30,33]. Tikhonov regularization [31] is an important methodology for solving such ill-posed inverse problems. The term regularization refers to imposing additional constraints on the function space from where the target function is to be chosen.

Assume that the target function is from domain X to range $Y \subseteq \mathbb{C}$. We call X the input space and thus, Y the output space. Suppose that a finite set $\mathbf{z} := \{(x_j, y_j) : j \in \mathbb{N}_n\} \subseteq X \times Y$ of samples of the target function is available. Here, for the simplicity of enumerating with

This work was supported in part by the National Science Foundation under grant 0631541 (PI: Jun Zhang).

H. Zhang · J. Zhang (✉)
University of Michigan, Ann Arbor, MI 48109, USA
e-mail: junz@umich.edu

Present Address:
H. Zhang
School of Mathematics and Computational Science, Sun Yat-sen University, Guangzhou 510275, China
e-mail: haizhang@umich.edu

finite sets, we set $\mathbb{N}_n := \{1, 2, \ldots, n\}$ for $n \in \mathbb{N}$. We shall also use the notations $\mathbf{x} := (x_j : j \in \mathbb{N}_n) \in X^n$, $\mathbf{y} := (y_j : j \in \mathbb{N}_n) \in Y^n$, and $\mathbb{R}_+ := [0, +\infty)$ throughout the paper. Following the framework of Tikhonov regularization in machine learning [5,10,13,29,30,33,35–37], we let $\mathcal{H}_K$ be a reproducing kernel Hilbert space (RKHS) on X with the reproducing kernel K. In other words, point evaluations are continuous linear functionals on $\mathcal{H}_K$ and K is a complex-valued function on $X \times X$ satisfying $K(x, \cdot) \in \mathcal{H}_K$ for all $x \in X$ and

$$f(x) = (f, K(x, \cdot))_{\mathcal{H}_K}, \quad x \in X, \ f \in \mathcal{H}_K,$$

where $(\cdot, \cdot)_{\mathcal{H}_K}$ is the inner product on $\mathcal{H}_K$, [2,22,25,26]. We also let $\mathcal{L}_{\mathbf{y}} : \mathbb{C}^n \to \mathbb{R}_+$ be a *loss function* [29] that measures how well a candidate function in $\mathcal{H}_K$ fits the sample data $\mathbf{z}$, $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ a *regularizer* that controls the set of functions in $\mathcal{H}_K$ from which we may select the candidate functions, and λ a positive regularization parameter. With these notations, an approximation of the target function is taken as a minimizer of the optimization problem:

$$\inf\{\mathcal{L}_{\mathbf{y}}(f(\mathbf{x})) + \lambda\phi(\|f\|_{\mathcal{H}_K}) : f \in \mathcal{H}_K\}, \tag{1}$$

where $f(\mathbf{x}) := (f(x_j) : j \in \mathbb{N}_n)$ and $\|\cdot\|_{\mathcal{H}_K}$ denotes the norm on $\mathcal{H}_K$. Various choices of the loss function $\mathcal{L}_{\mathbf{y}}$, the regularizer ϕ, the output space Y, and the corresponding learning schemes can be found in [13,29,30,33]. We present three popular ones below:

- regularization networks

$$Y = \mathbb{R}, \ \mathcal{L}_{\mathbf{y}}(f(\mathbf{x})) = \sum_{j \in \mathbb{N}_n} |f(x_j) - y_j|^2, \ \phi(t) = t^2, \ t \in \mathbb{R}_+, \tag{2}$$

- support vector machine regression

$$Y = \mathbb{R}, \ \mathcal{L}_{\mathbf{y}}(f(\mathbf{x})) = \sum_{j \in \mathbb{N}_n} |f(x_j) - y_j|_\varepsilon, \ \phi(t) = t^2, \ t \in \mathbb{R}_+,$$

  where ε is a positive constant and $|t|_\varepsilon := \max(|t| - \varepsilon, 0)$, $t \in \mathbb{R}$, is called Vapnik's epsilon-insensitive norm.
- support vector machine classification

$$Y = \{-1, 1\}, \ \mathcal{L}_{\mathbf{y}}(f(\mathbf{x})) = \sum_{j \in \mathbb{N}_n} \max(1 - y_j f(x_j), 0), \ \phi(t) = t^2, \ t \in \mathbb{R}_+. \tag{3}$$

Conditions such as that $\mathcal{L}_{\mathbf{y}}$ and ϕ are continuous, ϕ is nondecreasing, and $\phi(t)$ tends to infinity as t does ensure that (1) has a minimizer. If, in addition, $\mathcal{L}_{\mathbf{y}}$ and ϕ are convex and ϕ is strictly increasing then the minimizer is unique. In regularization networks (2), the unique minimizer f_0 of (1) has the form

$$f_0 = \sum_{j \in \mathbb{N}_n} c_j K(x_j, \cdot) \tag{4}$$

for some complex constants $c_j \in \mathbb{C}$, $j \in \mathbb{N}_n$. This remarkable result, due to Kimeldorf and Wahba [19], is known as the representer theorem. Two reasons account for its fundamental importance in machine learning. First of all, a function $K : X \times X \to \mathbb{C}$ is a reproducing kernel if and only if there is a mapping Φ from X to a Hilbert space $\mathcal{W}$ such that

$$K(x, y) = (\Phi(x), \Phi(y))_{\mathcal{W}}, \quad x, y \in X. \tag{5}$$

By the above equation, for all $x, y \in X$, $K(x, y)$ is the inner product of vectors $\Phi(x)$ and $\Phi(y)$ in the Hilbert space $\mathcal{W}$. Therefore, $K(x, y)$ is a measurement of similarity between elements x, y in the input space. From this point of view, the predicted value $f_0(x)$ of each input $x \in X$ given by the minimizer (4) can be interpreted as a weighted sum of similarities $K(x, x_j)$ between x and the input sample points x_j, with the weights being c_j. Using input similarities to generate a desired output justifies the regularized learning schemes. Secondly, although the RKHS $\mathcal{H}_K$ in the minimization problem (1) tends to be infinite dimensional, the representer theorem enables us to consider the learning in the linear subspace spanned by $\{K(x_j, \cdot) : j \in \mathbb{N}_n\}$, which is of finite dimension.

The representer theorem for (1) was generalized in [9] to the case where the loss function is non-quadratic, and in [28] for a general nondecreasing regularizer ϕ. Relationships between (1) and the problem of minimal norm interpolation in RKHS were investigated in [1]. As a result, it was proved there that there exists a representer theorem for (1) if and only if the regularizer ϕ is nondecreasing.

The primary purpose of this note is to establish appropriate representer theorems for the following regularized learning scheme in a Banach space $\mathcal{B}$ of functions on X:

$$\inf\{\mathcal{L}_{\mathbf{y}}(f(\mathbf{x})) + \lambda\phi(\|f\|_{\mathcal{B}}) : f \in \mathcal{B}\}. \tag{6}$$

There are some needs that motivate the study of learning in Banach spaces. Firstly, it is well-known that any two Hilbert spaces over $\mathbb{C}$ of the same dimension are isometrically isomorphic, namely, there exists a bijective linear norm-preserving mapping between them. By contrast, when $p \neq q \in [1, +\infty]$, we can not find a bijective bounded linear mapping from $L^p[0, 1]$ to $L^q[0, 1]$ (see [14], page 180). Thus compared to Hilbert spaces, Banach spaces possess richer geometric structures, which might be useful in the development of learning algorithms. Secondly, a norm from a Banach space is more desirable than one induced from an inner product in some applications. For instance, it is known that regularizing a minimization problem by the ℓ^1 norm leads to sparsity of the minimizer (see, for example, [32]). Thirdly, training data in practice might come with intrinsic structures that make them impossible to be embedded into a Hilbert space.

Learning in Banach spaces has received considerable attention in the literature, [4,7,12,15,17,18,23,24,34,41,42]. Especially, regularized learning schemes were considered in [4,23,24,41]. However, no representer theorems existed due to the lack in Banach spaces of an inner product. In a recent work [39], we established the notion of reproducing kernel Banach spaces (RKBS) and studied in the framework of RKBS standard learning schemes including minimal norm interpolation, regularization networks, support vector machines, and kernel principal component analysis. In particular, we considered minimization problems of the form (6) with $\mathcal{B}$ being an RKBS. We proved a representer theorem for regularization networks, that is, the loss function and regularizer are given in (2). A characterization equation for the minimizer was also established. Moreover, we obtained a representer theorem for the support vector machine classification (3).

The main purpose of this note is to establish a representer theorem and a characterization equation for the minimizer of (6) under the setting that $\mathcal{L}_{\mathbf{y}}$ and ϕ are respectively a general loss function and nondecreasing regularizer. Our study accommodates a large class of loss functions and regularizers, and provides a unified treatment of them. We shall introduce in Sect. 2 the main elements of RKBS. The existence, uniqueness and representer theorem for the minimizer of (6) will be proved in Sect. 3. We shall establish in Sect. 4 a characterization equation for the case when $\mathcal{L}_{\mathbf{y}}$ and ϕ are both differentiable. Finally, we conclude the paper with a discussion of future directions.

2 Reproducing kernel Banach spaces

Let X be a prescribed input space where samples are generated. We call $\mathcal{B}$ a *Banach space of functions* on X if it is a Banach space consisting of functions defined on X and a function $f \in \mathcal{B}$ satisfies $\|f\|_{\mathcal{B}} = 0$ if and only if it vanishes everywhere on X. Note that the space $L^p[0, 1]$ is not a Banach space of functions.

A close examination of the establishment of the representer theorems for RKHS (see, for example, [28]) indicates the vital role of the existence of orthogonal projections. Orthogonal projections are a unique characteristic of an inner product. To have a substitute for inner products in the Banach space setting, we shall focus on Banach spaces of functions that are uniformly Fréchet differentiable and uniformly convex. A Banach space $\mathcal{B}$ is said to be *uniformly Fréchet differentiable* if for all $f, g \in \mathcal{B}$

$$\lim_{t\in\mathbb{R},\, t\to 0} \frac{\|f + tg\|_{\mathcal{B}} - \|f\|_{\mathcal{B}}}{t}$$

exists and the limit is approached uniformly for f, g in the unit sphere of $\mathcal{B}$. *Uniform convexity* of $\mathcal{B}$ requires that for all $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\|f + g\|_{\mathcal{B}} \le 2 - \delta \text{ for all } f, g \in \mathcal{B} \text{ with } \|f\|_{\mathcal{B}} = \|g\|_{\mathcal{B}} = 1 \text{ and } \|f - g\|_{\mathcal{B}} \ge \varepsilon.$$

For more information about these two useful properties of Banach spaces, see [11,21]. For the sake of simplicity, we shall call a Banach space $\mathcal{B}$ *uniform* if it is both uniformly Fréchet differentiable and uniformly convex.

For a uniform Banach space $\mathcal{B}$, there exists a unique function $[\cdot, \cdot]_{\mathcal{B}} : \mathcal{B} \times \mathcal{B} \to \mathbb{C}$ such that for all $f, g, h \in \mathcal{B}$ and $\alpha \in \mathbb{C}$

1. $[f + g, h]_{\mathcal{B}} = [f, h]_{\mathcal{B}} + [g, h]_{\mathcal{B}}$, $[\alpha f, g]_{\mathcal{B}} = \alpha[f, g]_{\mathcal{B}}$,
2. $[f, f]_{\mathcal{B}} = \|f\|^2_{\mathcal{B}}$,
3. (Cauchy-Schwartz) $|[f, g]_{\mathcal{B}}|^2 \le [f, f]_{\mathcal{B}}[g, g]_{\mathcal{B}}$.

The above function $[\cdot, \cdot]_{\mathcal{B}}$ is called a *semi-inner product* on $\mathcal{B}$. It is unique by the fact [16] that

$$\lim_{t\in\mathbb{R},\, t\to 0} \frac{\|f + tg\|_{\mathcal{B}} - \|f\|_{\mathcal{B}}}{t} = \frac{\operatorname{Re}([g, f]_{\mathcal{B}})}{\|f\|_{\mathcal{B}}}, \quad f, g \in \mathcal{B}, \ f \ne 0, \tag{7}$$

where $\operatorname{Re}(\alpha)$ denotes the real part of a complex number α. Semi-inner products were introduced by Lumer [20] for the purpose of extending Hilbert space type arguments to Banach spaces. Fundamental properties of semi-inner products were explored by Giles [16]. Recently, a generalized semi-inner product was introduced [40] that was shown to reflect the generalized duality mapping. Semi-inner products were first applied to machine learning by Der and Lee [12] to develop hard margin hyperplane classification in Banach spaces. The use of semi-inner products for reproducing kernels in Banach spaces was developed by Zhang et al [39].

Let $\mathcal{B}$ be a uniform Banach space with the semi-inner product $[\cdot, \cdot]_{\mathcal{B}}$. Then by the Cauchy-Schwartz inequality, for each $f \in \mathcal{B}$ the function sending $g \in \mathcal{B}$ to $[g, f]_{\mathcal{B}}$ is a bounded linear functional on $\mathcal{B}$, which will be denoted by f^* and called the *dual element* of f. By definition, we have

$$f^*(g) = [g, f]_{\mathcal{B}}, \quad f, g \in \mathcal{B}. \tag{8}$$

The mapping $f \to f^*$ is said to be the *duality mapping* from $\mathcal{B}$ to its dual space $\mathcal{B}^*$. The importance of semi-inner products in our paper lies in the following fact established in [16].

Lemma 1 *Let $\mathcal{B}$ be a uniform Banach space. Then the duality mapping is bijective from $\mathcal{B}$ to $\mathcal{B}^*$. Moreover,*

$$[f^*, g^*]_{\mathcal{B}^*} := [g, f]_{\mathcal{B}}, \quad f, g \in \mathcal{B} \tag{9}$$

defines a semi-inner product on $\mathcal{B}^*$.

The theory of RKBS has been established in [39]. In this paper we shall adopt a simplified definition by calling $\mathcal{B}$ a *reproducing kernel Banach space* (RKBS) on X if it is a uniform Banach space of functions on X such that the point evaluations are continuous linear functionals on $\mathcal{B}$. We prove below that there does exist a reproducing kernel for Banach spaces satisfying this definition.

Theorem 1 *Suppose that $\mathcal{B}$ is an RKBS on the input space X. Then there exists a unique function $G : X \times X \to \mathbb{C}$ such that $G(x, \cdot) \in \mathcal{B}$ for all $x \in X$ and*

$$f(x) = [f, G(x, \cdot)]_{\mathcal{B}} \text{ for all } x \in X, \ f \in \mathcal{B}. \tag{10}$$

Proof Since the point evaluations are continuous linear functionals on $\mathcal{B}$, by Lemma 1, for each $x \in X$ there exists a unique function $g_x \in \mathcal{B}$ such that

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

Introduce a bivariate function G on $X \times X$ by setting

$$G(x, y) := g_x(y), \quad x, y \in X.$$

By the above two equations, for each $x \in X, G(x, \cdot) = g_x \in \mathcal{B}$ and (10) holds true. Assume that there is another function $\tilde{G} : X \times X \to \mathbb{C}$ satisfying the two properties as G does. Then by (8) we have for each $x \in X$ that

$$(G(x, \cdot)^*)(f) = [f, G(x, \dot{)}]_{\mathcal{B}} = f(x) = [f, \tilde{G}(x, \cdot)]_{\mathcal{B}} = (\tilde{G}(x, \cdot)^*)(f) \text{ for all } f \in \mathcal{B}.$$

It implies that $G(x, \cdot)^* = \tilde{G}(x, \cdot)^*$ for all $x \in X$. By Lemma 1, the duality mapping from $\mathcal{B}$ to $\mathcal{B}^*$ is injective. Thus, $G(x, \cdot)$ and $\tilde{G}(x, \cdot)$ as vectors in the Banach space $\mathcal{B}$ are identical. Since $\mathcal{B}$ is a Banach space of functions, we have the equality as functions on X:

$$G(x, y) = \tilde{G}(x, y) \text{ for all } x, y \in X,$$

which completes the proof. □

Let $\mathcal{B}$ be an RKBS on X. We call the function G in the above theorem the *s.i.p. reproducing kernel* of $\mathcal{B}$. When $\mathcal{B}$ is an RKHS, the semi-inner product on it becomes an inner product and G is the reproducing kernel of $\mathcal{B}$ in the usual sense. The function G has the property that

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

Furthermore, we have by (9) and (10) that

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

We close this preparation section with two concrete examples of RKBS. Let $X := \mathbb{R}^d$, μ a finite positive Borel measure on $\mathbb{R}^d$, and $L^p_\mu(\mathbb{R}^d)$, $1 < p < +\infty$, the Banach space of Borel measurable functions u on $\mathbb{R}^d$ such that the norm

$$\|u\|_{L^p_\mu(\mathbb{R}^d)} := \left(\int_{\mathbb{R}^d} |u(t)|^p d\mu(t) \right)^{1/p}$$

is finite. It is uniformly Fréchet differentiable and uniformly convex with the dual space $L^q_\mu(\mathbb{R}^d)$, where q is the conjugate number of p satisfying that $1/p + 1/q = 1$. With $(\cdot,\cdot)$ being the standard inner product on $\mathbb{R}^d$, we set $\mathcal{B}$ the Banach space of functions of the form

$$f_u(x) := \frac{1}{\mu(\mathbb{R}^d)^{\frac{p-2}{p}}} \int\limits_{\mathbb{R}^d} u(t) e^{i(x,t)} d\mu(t), \quad x \in \mathbb{R}^d, \ u \in L^p_\mu(\mathbb{R}^d)$$

equipped with the norm

$$\|f_u\|_{\mathcal{B}} := \|u\|_{L^p_\mu(\mathbb{R}^d)}.$$

Then $\mathcal{B}$ is an RKBS with the semi-inner product

$$[f_u, f_v]_{\mathcal{B}} = \frac{\int_{\mathbb{R}^d} u(t)\overline{v(t)}|v(t)|^{p-2} d\mu(t)}{\|v\|^{p-2}_{L^p_\mu(\mathbb{R}^d)}}, \quad u, v \in L^p_\mu(\mathbb{R}^d)$$

and the s.i.p. reproducing kernel

$$G(x, y) = \frac{1}{\mu(\mathbb{R}^d)^{\frac{p-2}{p}}} \int\limits_{\mathbb{R}^d} e^{i(y-x,t)} d\mu(t), \quad x, y \in \mathbb{R}^d.$$

In the special case that $d = 1$ and μ is the Lebesgue measure supported on $[-\pi, \pi]$, we get that

$$G(x, y) = \frac{1}{(2\pi)^{\frac{p-2}{p}}} \frac{2 \sin \pi(x - y)}{x - y}, \quad x, y \in \mathbb{R}.$$

Our second example is the space $\mathbb{E}^p_\tau$, $p \in (1, +\infty)$, $\tau > 0$ consisting of all entire functions f on $\mathbb{C}$ of exponential type at most τ for which

$$\|f\|_{\mathbb{E}^p_\tau} := \left(\int\limits_{\mathbb{R}} |f(t)|^p dt \right)^{1/p} < +\infty.$$

Point evaluations are continuous on $\mathbb{E}^p_\tau$ as there is a constant C depending on p and τ only such that (see, [38], page 99)

$$|f(x + iy)| \le Ce^{\tau|y|} \|f\|_{\mathbb{E}^p_\tau} \quad \text{for all } x, y \in \mathbb{R}, \ f \in \mathbb{E}^p_\tau.$$

By the above two equations, $\mathbb{E}^p_\tau$ is a Banach space isometrically isomorphic to a closed subspace of $L^p(\mathbb{R})$. Consequently, $\mathbb{E}^p_\tau$ is uniform, and is thus an RKBS on $\mathbb{C}$.

3 Representer theorems

Let $\mathcal{B}$ be an RKBS on X with the s.i.p. reproducing kernel G. We consider the minimization problem (6), where $\mathcal{L}_{\mathbf{y}} : \mathbb{C}^n \to \mathbb{R}_+$ is a loss function, $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is nondecreasing, and λ is a positive regularization parameter. Introduce $\mathcal{E}_{\mathbf{z}} : \mathcal{B} \to \mathbb{R}_+$ by setting

$$\mathcal{E}_{\mathbf{z}}(f) := \mathcal{L}_{\mathbf{y}}(f(\mathbf{x})) + \lambda\phi(\|f\|_{\mathcal{B}}), \quad f \in \mathcal{B}.$$

We start with presenting the following representer theorem provided the existence of a minimizer of (6).

Theorem 2 (Representer theorem, ϕ strictly increasing) *If ϕ is strictly increasing then every minimizer f_0 of (6), provided that it exists, must have the form*

$$f_0^* = \sum_{j \in \mathbb{N}_n} c_j G(x_j, \cdot)^*, \tag{12}$$

where c_j, $j \in \mathbb{N}_n$ are complex constants.

Proof Assume there exists a minimizer f_0 of (6) for which (12) is not true for any choice of constants $c_j \in \mathbb{C}$, $j \in \mathbb{N}_n$, that is,

$$f_0^* \notin \operatorname{span}\{G(x_j, \cdot)^* : j \in \mathbb{N}_n\}.$$

Note that as a finite dimensional subspace of $\mathcal{B}^*$, $\operatorname{span}\{G(x_j, \cdot)^* : j \in \mathbb{N}_n\}$ is closed and convex. Thus, by a geometric consequence of the Hahn-Banach theorem in functional analysis (see, for example, [8], page 111), there exists a continuous linear functional T on $\mathcal{B}^*$ and real number α such that

$$\operatorname{Re}(T(f_0^*)) < \alpha \le \operatorname{Re}(T(u)), \quad \text{for all } u \in \operatorname{span}\{G(x_j, \cdot)^* : j \in \mathbb{N}_n\}.$$

Firstly, since for all $\beta \in \mathbb{C}$ and $u \in \operatorname{span}\{G(x_j, \cdot)^* : j \in \mathbb{N}_n\}$

$$\alpha \le \operatorname{Re}(T(\beta u)) = \operatorname{Re}(\beta T(u)),$$

we must have $T(u) = 0$ for all $u \in \operatorname{span}\{G(x_j, \cdot)^* : j \in \mathbb{N}_n\}$. Consequently, $\alpha \le 0$. Secondly, since $\mathcal{B}$ is reflexive there exists a $g \in \mathcal{B}$ such that

$$T(v) = v(g), \quad v \in \mathcal{B}^*.$$

By (8) and (10), we get that

$$0 = T(G(x_j, \cdot)^*) = (G(x_j, \cdot)^*)(g) = [g, G(x_j, \cdot)]_{\mathcal{B}} = g(x_j), \quad j \in \mathbb{N}_n \tag{13}$$

and

$$\operatorname{Re}([g, f_0]_{\mathcal{B}}) = \operatorname{Re}(f_0^*(g)) = \operatorname{Re}(T(f_0^*)) < 0. \tag{14}$$

Consider the function $f_0 + tg$ where $t \in \mathbb{R}_+$ is to be specified. By (13),

$$\mathcal{L}_{\mathbf{y}}((f_0 + tg)(\mathbf{x})) = \mathcal{L}_{\mathbf{y}}(f_0(\mathbf{x})). \tag{15}$$

Equation (14) implies that $f_0 \neq 0$. An application of equations (7) and (14) yields that

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

Therefore, by choosing $t \in \mathbb{R}_+$ close enough to 0, we obtain that $\|f_0 + tg\|_{\mathcal{B}} < \|f_0\|_{\mathcal{B}}$, which together with the assumption that ϕ is strictly increasing implies immediately that

$$\phi(\|f_0 + tg\|_{\mathcal{B}}) < \phi(\|f_0\|_{\mathcal{B}}). \tag{16}$$

Combining equations (15) and (16), we obtain another candidate function $f_0 + tg$ such that

$$\mathcal{E}_{\mathbf{z}}(f_0 + tg) < \mathcal{E}_{\mathbf{z}}(f_0),$$

contradicting that f_0 is a minimizer of (6). The contradiction proves the result. □

To deal with the case where ϕ is only known to be nondecreasing, we shall need the tool of minimal norm interpolation in RKBS. Set

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

The following result was proved in [39] for RKBS defined therein. It is easy to verify that the proof still applies to the definition used in this paper.

Lemma 2 *If $\mathbb{I}_{\mathbf{y}}$ is nonempty then the following minimal norm interpolation*

$$\inf\{\|f\|_{\mathcal{B}} : f \in \mathbb{I}_{\mathbf{y}}\}$$

has a unique minimizer f_0. Furthermore, f_0 has the form (12) for some $c_j \in \mathbb{C}$, $j \in \mathbb{N}_n$.

Theorem 3 (Representer theorem, ϕ nondecreasing) *If ϕ is nondecreasing and the minimization problem (6) has at least one minimizer then there exists a minimizer f_0 of (6) that has the form (12).*

Proof Let $f \in \mathcal{B}$ be a minimizer of (6). Then $\mathbb{I}_{f(\mathbf{x})}$ is nonempty as $f \in \mathbb{I}_{f(\mathbf{x})}$. We set

$$f_0 := \arg\min\{\|g\|_{\mathcal{B}} : g \in \mathbb{I}_{f(\mathbf{x})}\}.$$

Then $f_0(\mathbf{x}) = f(\mathbf{x})$. Thus,

$$\mathcal{L}_{\mathbf{y}}(f_0(\mathbf{x})) = \mathcal{L}_{\mathbf{y}}(f(\mathbf{x})).$$

Clearly, $\|f_0\|_{\mathcal{B}} \leq \|f\|_{\mathcal{B}}$. Since ϕ is nondecreasing, it follows that

$$\phi(\|f_0\|_{\mathcal{B}}) \leq \phi(\|f\|_{\mathcal{B}}).$$

We get from the above two equations that

$$\mathcal{E}_{\mathbf{z}}(f_0) \leq \mathcal{E}_{\mathbf{z}}(f).$$

Therefore, f_0 is a minimizer of (6). By Lemma 2, it satisfies (12) for some $c_j \in \mathbb{C}$, $j \in \mathbb{N}_n$. □

We remark by Theorems 2 and 3 that the essence of a representer theorem is representing the dual element of the minimizer as a linear combination of the point evaluation functionals at x_j, $j \in \mathbb{N}_n$. Note that in the case when $\mathcal{B}$ is an RKHS, the dual function of $f \in \mathcal{B}$ is itself. This is why the representer theorem for RKHS has the form (4).

By examining the proofs of Theorems 2 and 3, one obtains generalized representer theorems for regularized learning of a function $g \in \mathcal{B}$ from its generalized sample data

$$(\nu_j(g) : j \in \mathbb{N}_n) \text{ for some } \nu_j \in \mathcal{B}^*, \quad j \in \mathbb{N}_n.$$

Specifically, one may consider the following optimization problem:

$$\inf\{\mathcal{L}_{\mathbf{y}}((\nu_j(f) : j \in \mathbb{N}_n)) + \lambda\phi(\|f\|_{\mathcal{B}}) : f \in \mathcal{B}\}. \tag{17}$$

Theorem 4 (Representer theorem, generalized sample data) *Suppose that (17) has at least one minimizer. If ϕ is strictly increasing then every minimizer f_0 of (17) must satisfy for some complex constants c_j, $j \in \mathbb{N}_n$ that*

$$f_0^* = \sum_{j \in \mathbb{N}_n} c_j \nu_j.$$

If ϕ is nondecreasing then there exists a minimizer f_0 of (17) that has the above form.

We next consider sufficient conditions for the existence and uniqueness of the minimizer. Special cases of the following result have appeared in the literature [23,39].

Proposition 5 (Sufficient conditions for the existence of the minimizer) *If $\mathcal{L}_{\mathbf{y}}$ and ϕ are continuous, and ϕ is nondecreasing with*

$$\lim_{t\to\infty}\phi(t)=+\infty \tag{18}$$

then there exists a minimizer for (6).

Proof Set

$$e:=\inf\{\mathcal{E}_{\mathbf{z}}(f):f\in\mathcal{B}\}.$$

Clearly, $e\le\mathcal{E}_{\mathbf{z}}(0)$. By the assumptions that ϕ is nondecreasing with (18), there exists a positive number t_0 such that for all $f\in\mathcal{B}$ with $\|f\|_{\mathcal{B}}>t_0$

$$\mathcal{E}_{\mathbf{z}}(f)\ge\lambda\phi(\|f\|_{\mathcal{B}})\ge\lambda\phi(t_0)>e.$$

Thus, with $S:=\{f\in\mathcal{B}:\|f\|_{\mathcal{B}}\le t_0\}$,

$$e=\inf\{\mathcal{E}_{\mathbf{z}}(f):f\in S\}.$$

By the above equality, there exists a sequence $f_m\in S$ such that

$$e\le\mathcal{E}_{\mathbf{z}}(f_m)\le e+\frac{1}{m},\quad m\in\mathbb{N}. \tag{19}$$

Since $\mathcal{B}$ is uniformly convex, it is reflexive. As a consequence, S is weakly compact, that is, there exists a function $f_0\in S$ such that

$$\lim_{m\to\infty}[f_m,g]_{\mathcal{B}}=[f_0,g]_{\mathcal{B}},\quad \text{for all } g\in\mathcal{B}. \tag{20}$$

Choosing $g=G(x_j,\cdot)$, $j\in\mathbb{N}_n$ in the above equation and invoking (10) yields that

$$\lim_{m\to\infty}f_m(x_j)=\lim_{m\to\infty}[f_m,G(x_j,\cdot)]_{\mathcal{B}}=[f_0,G(x_j,\cdot)]_{\mathcal{B}}=f_0(x_j),\quad j\in\mathbb{N}_n.$$

By the continuity of $\mathcal{L}_{\mathbf{y}}$,

$$\lim_{m\to\infty}\mathcal{L}_{\mathbf{y}}(f_m)=\mathcal{L}_{\mathbf{y}}(f_0). \tag{21}$$

If $f_0=0$ then it is obvious that

$$\|f_0\|_{\mathcal{B}}\le\|f_m\|_{\mathcal{B}},\quad m\in\mathbb{N}.$$

If $f_0\ne0$ then we substitute $g=f_0$ in (20) to get for each $\delta>0$ some $m_0\in\mathbb{N}$ such that for $m>m_0$

$$[f_0,f_0]_{\mathcal{B}}\le|[f_m,f_0]_{\mathcal{B}}|+\delta\|f_0\|_{\mathcal{B}},$$

which implies by the Cauchy-Schwartz inequality of semi-inner products that

$$\|f_0\|_{\mathcal{B}}^2\le\|f_m\|_{\mathcal{B}}\|f_0\|_{\mathcal{B}}+\delta\|f_0\|_{\mathcal{B}}.$$

Thus, we have in both cases that for sufficiently large m

$$\|f_0\|_{\mathcal{B}}\le\|f_m\|_{\mathcal{B}}+\delta.$$

Since ϕ is continuous, it is uniformly continuous on $[0, t_0]$. Consequently, for every $\varepsilon > 0$, we may choose δ small enough so that for sufficiently large m

$$\lambda\phi(\|f_0\|_{\mathcal{B}}) \leq \lambda\phi(\|f_m\|_{\mathcal{B}}) + \varepsilon \tag{22}$$

Combining (19), (21), and (22) proves that

$$\mathcal{E}_{\mathbf{z}}(f_0) = e,$$

which shows that f_0 is a minimizer for (6) and completes the proof. □

As a corollary of Proposition 5, we address the issue of uniqueness of the minimizer.

Corollary 6 (Sufficient conditions for the uniqueness of the minimizer) *If $\mathcal{L}_{\mathbf{y}}$ and ϕ are continuous and convex, and ϕ is strictly increasing with (18) then there exists a unique minimizer for* (6).

Proof The existence is justified by the above proposition. Note that since $\mathcal{B}$ is uniformly convex, its norm is strictly convex. In other words, we have for all $f \neq g \in \mathcal{B}$ and $t \in (0, 1)$ that

$$\|tf + (1-t)g\|_{\mathcal{B}} < t\|f\|_{\mathcal{B}} + (1-t)\|g\|_{\mathcal{B}}.$$

As a consequence, $\phi(\|\cdot\|_{\mathcal{B}})$ is strictly convex on $\mathcal{B}$. We hence obtain the strict convexity of $\mathcal{E}_{\mathbf{z}}$, which ensures the uniqueness of the minimizer of (6). □

Suppose that $\mathcal{L}_{\mathbf{y}}$ and ϕ satisfy the conditions in the statement of Corollary 6. Then the minimizer of the optimization problem (6) can be obtained by solving the coefficient vector $\mathbf{c} := (c_j : j \in \mathbb{N}_n) \in \mathbb{C}^n$ after substituting the representer theorem (12) into (6). To this end, we observe from (11) that

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

which together with $\|f_0\|_{\mathcal{B}} = \|f_0^*\|_{\mathcal{B}^*}$ implies that the coefficients $c_j \in \mathbb{C}$, $j \in \mathbb{N}_n$ in (12) is the minimizer of

$$\min_{(a_j : j\in\mathbb{N}_n)\in\mathbb{C}^n} \mathcal{L}_{\mathbf{y}}\left(\left([G(x_j, \cdot)^*, \sum_{k\in\mathbb{N}_n} a_k G(x_k, \cdot)^*]_{\mathcal{B}^*} : j \in \mathbb{N}_n\right)\right) + \lambda\phi\left(\left\|\sum_{j\in\mathbb{N}_n} a_j G(x_j, \cdot)^*\right\|_{\mathcal{B}^*}\right). \tag{23}$$

By Corollary 6, (23) has a unique solution provided that $G(x_j, \cdot)^*$, $j \in \mathbb{N}_n$ are linearly independent in $\mathcal{B}^*$. The second summand of the error functional in (23) is strictly convex with respect to $(a_j : j \in \mathbb{N}_n) \in \mathbb{C}^n$ by the assumption that ϕ is convex and strictly increasing, and the fact that the norm in $\mathcal{B}^*$ is strictly convex. However, the first summand may not be convex with respect to $(a_j : j \in \mathbb{N}_n) \in \mathbb{C}^n$ due to the reason that a semi-inner product is generally non-additive about its second variable.

As a special case of (6), we consider in the final part of this section the support vector machine classification where the output space is $\{-1, 1\}$ and a classifier from X to Y is desirable. Define the loss function by

$$\mathcal{L}_{\mathbf{y}}(a) = \sum_{j\in\mathbb{N}_n} \max(1 - a_j y_j, 0), \quad a = (a_j : j \in \mathbb{N}_n) \in \mathbb{R}^n. \tag{24}$$

Clearly, $\mathcal{L}_{\mathbf{y}}$ is continuous and convex. If ϕ is continuous, convex and strictly increasing with (18) then (6) has a unique minimizer f_0, which provides the classifier $\operatorname{sgn} f_0$. Similar arguments as those in the proof of Theorem 2 are able to prove the following special representer theorem for support vector machine classification.

Proposition 7 (Representer theorem for support vector machine classification) *If ϕ is strictly increasing and the minimization problem* (6) *has a minimizer f_0 then f_0^* lies inside the closed convex cone spanned by $\{y_j G(x_j,\cdot)^* : j \in \mathbb{N}_n\}$, that is, there exist $\lambda_j \geq 0$ such that*

$$f_0^* = \sum_{j\in\mathbb{N}_n} \lambda_j y_j G(x_j,\cdot)^*. \tag{25}$$

Similarly, one may substitute (25) and the loss function (24) into (6) to convert it into an optimization problem about the coefficients λ_j, $j \in \mathbb{N}_n$.

4 Characterization equations

We seek for a characterization equation for the minimizer of (6) in this section. One such equation exists for the minimization problem (1) in an RKHS $\mathcal{H}_K$ when the loss function and regularizer are given by (2). In fact, it is well-known [10,29] that $f_0 \in \mathcal{H}_K$ is the minimizer for this problem if and only if

$$\lambda f_0 = \sum_{j\in\mathbb{N}_n} (y_j - f_0(x_j))K(x_j,\cdot). \tag{26}$$

Consequently, when $K(x_j,\cdot)$, $j \in \mathbb{N}_n$ are linearly independent in $\mathcal{H}_K$, the coefficient vector $c = (c_j : j \in \mathbb{N}_n) \in \mathbb{C}^n$ in (4) can be solved from the linear system of equations:

$$c(\lambda I_n + K[\mathbf{x}]) = \mathbf{y}, \tag{27}$$

where I_n denotes the $n \times n$ identity matrix and $K[\mathbf{x}] := [K(x_j, x_k) : j,k \in \mathbb{N}_n]$. This demonstrates the usefulness of the representer theorem combined with a characterization equation.

For the purpose of establishing a characterization equation for the minimizer of (6), we shall only work with real numbers and loss functions of the form

$$\mathcal{L}_{\mathbf{y}}(a) := \sum_{j\in\mathbb{N}_n} L_j(a_j, y_j), \quad a = (a_j : j \in \mathbb{N}_n) \in \mathbb{R}^n, \tag{28}$$

where $L_j : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ are prescribed bivariate loss functions. We start with a technical lemma.

Lemma 3 *If ϕ is nondecreasing, differentiable, and convex on $\mathbb{R}_+$ then for all $f, g \in \mathcal{B}$*

$$\|f\|_{\mathcal{B}}(\phi(\|f+g\|_{\mathcal{B}}) - \phi(\|f\|_{\mathcal{B}})) - \phi'(\|f\|_{\mathcal{B}})\operatorname{Re}([g,f]_{\mathcal{B}}) \geq 0. \tag{29}$$

Proof Inequality (29) clearly holds true if $f = 0$. Let $f, g \in \mathcal{B}$ with $f \neq 0$. Introduce the function ψ on $\mathbb{R}_+$ by setting

$$\psi(t) := \phi(\|f + tg\|_{\mathcal{B}}).$$

We shall show that ψ is convex. To this end, we verify for all $s, t \in \mathbb{R}_+$ and $\alpha \in [0, 1]$ that

$$\begin{aligned}\|f + \alpha tg + (1-\alpha)sg\|_{\mathcal{B}} &= \|\alpha(f + tg) + (1-\alpha)(f + sg)\|_{\mathcal{B}}\\ &\le \alpha\|f + tg\|_{\mathcal{B}} + (1-\alpha)\|f + sg\|_{\mathcal{B}}.\end{aligned}$$

Thus, $\varphi(t) := \|f + tg\|_{\mathcal{B}}$ is convex on $\mathbb{R}_+$. Since ϕ is nondecreasing and convex, $\psi = \phi(\varphi)$ is convex. Therefore,

$$\psi'(0) \le \psi(1) - \psi(0).$$

By (7),

$$\psi'(0) = \frac{\phi'(\|f\|_{\mathcal{B}})}{\|f\|_{\mathcal{B}}} \operatorname{Re}([g, f]_{\mathcal{B}}).$$

Note also that

$$\psi(1) - \psi(0) = \phi(\|f + g\|_{\mathcal{B}}) - \phi(\|f\|_{\mathcal{B}}).$$

Combining the above three equations proves (29). □

Theorem 8 (Characterization equations) *Let L_j be differentiable and convex with respect to its first variable for all $j \in \mathbb{N}_n$, the loss function $\mathcal{L}_{\mathbf{y}}$ given by (28), and ϕ a strictly increasing, differentiable and convex function on $\mathbb{R}_+$ satisfying (18). Then $f_0 \ne 0$ is the minimizer of (6) if and only if*

$$\sum_{j\in\mathbb{N}_n} \frac{\partial L_j}{\partial a}(f_0(x_j), y_j)G(x_j, \cdot)^* + \lambda\frac{\phi'(\|f_0\|_{\mathcal{B}})}{\|f_0\|_{\mathcal{B}}} f_0^* = 0, \tag{30}$$

where $\frac{\partial L_j}{\partial a}$ denotes the first partial derivative of L_j with respect to its first variable, $j \in \mathbb{N}_n$. The zero function $f_0 = 0$ is the minimizer of (6) if and only if

$$\|T\|_{\mathcal{B}^*} \le \lambda\phi'(0), \tag{31}$$

where T is a continuous linear functional on $\mathcal{B}$ defined by

$$T(f) := \sum_{j\in\mathbb{N}_n} \frac{\partial L_j}{\partial a}(0, y_j) f(x_j), \quad f \in \mathcal{B}.$$

Proof By Corollary 6, (6) has a unique minimizer under the hypotheses. Assume that $f_0 \ne 0$ is the minimizer. Then for each $g \in \mathcal{B}$ the function

$$\varphi(t) := \mathcal{E}_{\mathbf{z}}(f_0 + tg), \quad t \in \mathbb{R}$$

achieves its minimum at $t = 0$. Thus, $\varphi'(0) = 0$. We compute by (7) that

$$\varphi'(0) = \sum_{j\in\mathbb{N}_n} \frac{\partial L_j}{\partial a}(f_0(x_j), y_j)g(x_j) + \lambda\frac{\phi'(\|f_0\|_{\mathcal{B}})}{\|f_0\|_{\mathcal{B}}}[g, f_0]_{\mathcal{B}}. \tag{32}$$

By (11), we get that

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

Similarly, $[g, f_0]_{\mathcal{B}} = [f_0^*, g^*]_{\mathcal{B}^*}$. Therefore, it follows from (32) that

$$\sum_{j\in\mathbb{N}_n} \frac{\partial L_j}{\partial a}(f_0(x_j), y_j)[G(x_j, \cdot)^*, g^*]_{\mathcal{B}^*} + \lambda \frac{\phi'(\|f_0\|_{\mathcal{B}})}{\|f_0\|_{\mathcal{B}}}[f_0^*, g^*]_{\mathcal{B}^*} = 0. \tag{33}$$

Since the above equation must hold true for all $g \in \mathcal{B}$, we obtain (30).

Conversely, assume that $f_0 \neq 0$ satisfies (30). Then (33) holds for all $g \in \mathcal{B}$. As a result, we have for all $g \in \mathcal{B}$ that

$$\sum_{j\in\mathbb{N}_n} \frac{\partial L_j}{\partial a}(f_0(x_j), y_j)g(x_j) + \lambda \frac{\phi'(\|f_0\|_{\mathcal{B}})}{\|f_0\|_{\mathcal{B}}}[g, f_0]_{\mathcal{B}} = 0. \tag{34}$$

We shall prove that f_0 indeed is the minimizer by showing that $\mathcal{E}_{\mathbf{z}}(f_0 + g) \geq \mathcal{E}_{\mathbf{z}}(f_0)$. We simplify by (34) that

$$\begin{aligned}
&\mathcal{E}_{\mathbf{z}}(f_0+g) - \mathcal{E}_{\mathbf{z}}(f_0)\\
&\quad = \sum_{j\in\mathbb{N}_n} (L_j(f_0(x_j)+g(x_j), y_j) - L_j(f_0(x_j), y_j)) + \lambda\phi(\|f_0+g\|_{\mathcal{B}}) - \lambda\phi(\|f_0\|_{\mathcal{B}})\\
&\quad = \sum_{j\in\mathbb{N}_n} \left(L_j(f_0(x_j)+g(x_j), y_j) - L_j(f_0(x_j), y_j) - \frac{\partial L_j}{\partial a}(f_0(x_j), y_j)g(x_j)\right)\\
&\qquad\qquad + \lambda\left(\phi(\|f_0+g\|_{\mathcal{B}}) - \phi(\|f_0\|_{\mathcal{B}}) - \frac{\phi'(\|f_0\|_{\mathcal{B}})}{\|f_0\|_{\mathcal{B}}}[g, f_0]_{\mathcal{B}}\right).
\end{aligned}$$

By Lemma 3,

$$\phi(\|f_0+g\|_{\mathcal{B}}) - \phi(\|f_0\|_{\mathcal{B}}) - \frac{\phi'(\|f_0\|_{\mathcal{B}})}{\|f_0\|_{\mathcal{B}}}[g, f_0]_{\mathcal{B}} \geq 0.$$

It remains to show that for each $j \in \mathbb{N}_n$

$$L_j(f_0(x_j)+g(x_j), y_j) - L_j(f_0(x_j), y_j) - \frac{\partial L_j}{\partial a}(f_0(x_j), y_j)g(x_j) \geq 0. \tag{35}$$

To this end, we obtain by the mean value theorem that there exists some $t \in [0, 1]$ such that the left hand side above equals

$$\left(\frac{\partial L_j}{\partial a}(f_0(x_j)+tg(x_j), y_j) - \frac{\partial L_j}{\partial a}(f_0(x_j), y_j)\right) g(x_j). \tag{36}$$

Since L_j is convex with respect to the first variable, $\frac{\partial L_j}{\partial a}(\cdot, y_j)$ is nondecreasing. If $g(x_j) \geq 0$ then $f_0(x_j) + tg(x_j) \geq f_0(x_j)$, implying that

$$\frac{\partial L_j}{\partial a}(f_0(x_j)+tg(x_j), y_j) - \frac{\partial L_j}{\partial a}(f_0(x_j), y_j) \geq 0.$$

Thus, (36) is nonnegative. If $g(x_j) < 0$ then $f_0(x_j) + tg(x_j) \leq f_0(x_j)$ implies that

$$\frac{\partial L_j}{\partial a}(f_0(x_j)+tg(x_j), y_j) - \frac{\partial L_j}{\partial a}(f_0(x_j), y_j) \leq 0.$$

In this case, (36) is also nonnegative. We hence obtain (35). Thus, f_0 is the minimizer of (6).

We now turn to the case that $f_0 = 0$. Suppose that $f_0 = 0$ is the minimizer of (6). Then for each $f \in \mathcal{B}$ and $t \in \mathbb{R}$, we have

$$\mathcal{E}_{\mathbf{z}}(tf) \geq \mathcal{E}_{\mathbf{z}}(0).$$

As a result, there holds

$$\lim_{t \to 0^+} \frac{\mathcal{E}_{\mathbf{z}}(tf) - \mathcal{E}_{\mathbf{z}}(0)}{t} \geq 0,$$

which by direct computations has the following equivalent form

$$T(f) + \lambda\phi'(0)\|f\|_{\mathcal{B}} \geq 0.$$

Since the above equation holds true for all $f \in \mathcal{B}$, we obtain (31).

On the other hand, assume that (31) holds true. Let $f \in \mathcal{B}$. Clearly,

$$\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(0) = \sum_{j \in \mathbb{N}_n} (L_j(f(x_j), y_j) - L_j(0, y_j)) + \lambda(\phi(\|f\|_{\mathcal{B}}) - \phi(0)). \tag{37}$$

Similar techniques as those used in proving the nonnegativity of (36) lead to that

$$\sum_{j \in \mathbb{N}_n} (L_j(f(x_j), y_j) - L_j(0, y_j)) \geq \sum_{j \in \mathbb{N}_n} \frac{\partial L_j}{\partial a}(0, y_j) f(x_j) = T(f).$$

Combining the above equation and (37) yields that

$$\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(0) \geq \lambda(\phi(\|f\|_{\mathcal{B}}) - \phi(0)) + T(f) \geq \lambda\phi'(0)\|f\|_{\mathcal{B}} - \|T\|_{\mathcal{B}^*}\|f\|_{\mathcal{B}},$$

which together with the assumption (31) proves that $\mathcal{E}_{\mathbf{z}}(f) \geq \mathcal{E}_{\mathbf{z}}(0)$. Since this is true for an arbitrary $f \in \mathcal{B}$, $f_0 = 0$ is the minimizer of (6). The proof is complete. □

Clearly, when $\mathcal{B}$ is an RKHS and the loss function and regularizer are specified by (2), the characterization equation (30) or (31) reduces to the classical one (26). Suppose that $G(x_j, \cdot)^*$, $j \in \mathbb{N}_n$, are linearly independent in $\mathcal{B}^*$ then we substitute the representer theorem (2) into (30) to get that the coefficients c_j, $j \in \mathbb{N}_n$ satisfy the system of equations:

$$\begin{aligned} &\frac{\partial L_j}{\partial a}\left(\left[G(x_j, \cdot)^*, \sum_{k \in \mathbb{N}_n} c_k G(x_k, \cdot)^*\right]_{\mathcal{B}^*}, y_j\right) \\ &\quad + \lambda \frac{\phi'(\|\sum_{k \in \mathbb{N}_n} c_k G(x_k, \cdot)^*\|_{\mathcal{B}^*})}{\|\sum_{k \in \mathbb{N}_n} c_k G(x_k, \cdot)^*\|_{\mathcal{B}^*}} c_j = 0, \quad j \in \mathbb{N}_n. \end{aligned} \tag{38}$$

The above system has a unique solution $(c_j : j \in \mathbb{N}_n) \in \mathbb{C}^n$ by Corollary 6 and the assumption that $G(x_j, \cdot)^*$, $j \in \mathbb{N}_n$ are linearly independent in $\mathcal{B}^*$. When $\mathcal{B}$ is a real RKHS, $L_j(t, s) = (t - s)^2$ for each $j \in \mathbb{N}_n$, and $\phi(t) = t^2$, the system (38) of equations has the form (27).

5 Conclusion

By making use of semi-inner products, we have proved the representer theorem for the regularized learning in RKBS with a general loss function $\mathcal{L}_{\mathbf{y}}$ and a nondecreasing regularizer ϕ. A characterization equation for the case when ϕ is differentiable and $\mathcal{L}_{\mathbf{y}}$ is a sum of differentiable bivariate loss functions is also obtained. The established results have a similar form as those for RKHS. The striking difference is that the resulting optimization problem or characterization equation about the coefficients in the representer theorem (2) is usually non-convex or nonlinear. This is due to the fact that a semi-inner product is in general non-additive with respect to its second variable. We shall leave the design of practical algorithms for (23) and (38) for future study. Here we briefly mention some possible approaches. The two problems have equivalent formulations in the feature space of the s.i.p. reproducing kernel. Conditions ensuring the convexity of resulting formulation of the minimization problem (23) should be investigated, as one can then try to apply the theory of nonlinear convex optimization. As far as (38) is concerned, one might first consider the case when the feature space is L^p or ℓ^p with p being an even integer. In this situation, (38) becomes a system of polynomial equations in the feature space and our goal is to find the common zero of these polynomials. The Gröbner basis theory in computational commutative algebra might then be engaged (see, for example, [3,6]). Finally, we remark that generalizations of the results in Sects. 3 and 4 can be obtained by relaxing the conditions on the Banach space $\mathcal{B}$ of functions and by making use of generalized semi-inner products developed in [40].

References

1. Argyriou, A., Micchelli, C.A., Pontil, M.: When is there a representer theorem? Vector versus matrix regularizers. Preprint, arXiv:0809.1590v1 (2008)
2. Aronszajn, N.: Theory of reproducing kernels. Trans. Amer. Math. Soc. **68**, 337–404 (1950)
3. Becker, T., Weispfenning, V.: Gröbner Bases: A Computational Approach to Commutative Algebra. Springer-Verlag, New York (1993)
4. Bennett, K., Bredensteiner, E.: Duality and geometry in SVM classifier. In: Langley, P. (ed.) Proceeding of the Seventeenth International Conference on Machine Learning, pp. 57–64. Morgan Kaufmann, San Francisco (2000)
5. Berlinet, A., Thomas-Agnan, C.: Reproducing Kernel Hilbert Spaces in Probability and Statistics. Kluwer Academic Publishers, Norwell, MA (2004)
6. Boege, W., Gebauer, R., Kredel, H.: Some examples for solving systems of algebraic equations by calculating Gröbner bases. J. Symb. Comput. **1**, 83–98 (1986)
7. Canu, S., Mary, X., Rakotomamonjy, A.: Functional learning through kernel. In: Suykens, J., Horvath, G., Basu, S., Micchelli, C.A., Vandewalle, J. (eds.) Advances in Learning Theory: Methods, Models and Applications. NATO Science Series III: Computer and Systems Sciences, vol. 190, pp. 89–110. IOS Press, Amsterdam (2003)
8. Conway, J.B.: A Course in Functional Analysis, 2nd edn. Springer-Verlag, New York (1990)
9. Cox, D., O'Sullivan, F.: Asymptotic analysis of penalized likelihood and related estimators. Ann. Statist. **18**, 1676–1695 (1990)
10. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bull. Amer. Math. Soc. **39**, 1–49 (2002)
11. Cudia, D.F.: On the localization and directionalization of uniform convexity. Bull. Amer. Math. Soc. **69**, 265–267 (1963)
12. Der, R., Lee, D.: Large-margin classification in Banach spaces. JMLR Workshop and Conference Proceedings 2: AISTATS, 91–98 (2007)
13. Evgeniou, T., Pontil, M., Poggio, T.: Regularization networks and support vector machines. Adv. Comput. Math. **13**, 1–50 (2000)
14. Fabian, M. et al.: Functional Analysis and Infinite-Dimensional Geometry. Springer, New York (2001)
15. Gentile, C.: A new approximate maximal margin classification algorithm. J. Mach. Learn. Res. **2**, 213–242 (2001)
16. Giles, J.R.: Classes of semi-inner-product spaces. Trans. Amer. Math. Soc. **129**, 436–446 (1967)
17. Hein, M., Bousquet, O., Schölkopf, B.: Maximal margin classification for metric spaces. J. Comput. System Sci. **71**, 333–359 (2005)
18. Kimber, D., Long, P.M.: On-line learning of smooth functions of a single variable. Theoret. Comput. Sci. **148**, 141–156 (1995)
19. Kimeldorf, G., Wahba, G.: Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. **33**, 82–95 (1971)

20. Lumer, G.: Semi-inner-product spaces. Trans. Amer. Math. Soc. **100**, 29–43 (1961)
21. Megginson, R.E.: An Introduction to Banach Space Theory. Springer, New York (1998)
22. Mercer, J.: 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)
23. Micchelli, C.A., Pontil, M.: A function representation for learning in Banach spaces. In: Learning Theory, pp. 255–269. Lecture Notes in Computer Science, 3120, Springer, Berlin (2004)
24. Micchelli, C.A., Pontil, M.: Feature space perspectives for learning the kernel. Machine Learning **66**, 297–319 (2007)
25. Moore, E.H.: On properly positive Hermitian matrices. Bull. Amer. Math. Soc. **23**, 59 (1916)
26. Moore, E.H.: General Analysis. Memoirs of the American Philosophical Society, Part I (1935), Part II (1939)
27. Pardalos, P.M., Hansen, P. (eds.): Data Mining and Mathematical Programming. Papers from the workshop held in Montreal, QC, October 10–13, 2006. CRM Proceedings & Lecture Notes 45. American Mathematical Society, Providence, RI (2008)
28. Schölkopf, B., Herbrich, R., Smola, A.J.: A generalized representer theorem. Proceeding of the Fourteenth Annual Conference on Computational Learning Theory and the Fifth European Conference on Computational Learning Theory, pp. 416–426. Springer-Verlag, London, UK (2001)
29. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, Mass (2002)
30. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
31. Tikhonov, A.N., Arsenin, V.Y.: Solutions of Ill-posed Problems. V. H. Winston & Sons (distributed by Wiley), New York (1977)
32. Tropp, J.A.: Just relax: convex programming methods for identifying sparse signals in noise. IEEE Trans. Inform. Theory **52**, 1030–1051 (2006)
33. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
34. von Luxburg, U., Bousquet, O.: Distance-based classification with Lipschitz functions. J. Mach. Learn. Res. **5**, 669–695 (2004)
35. Wahba, G.: Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In: Schölkopf, B., Burge, C., Smola, A.J. (eds.) Advances in Kernel Methods–Support Vector Learning, pp. 69–88. MIT Press, Cambridge, Mass (1999)
36. Xu, Y., Zhang, H.: Refinable kernels. J. Mach. Learn. Res. **8**, 2083–2120 (2007)
37. Xu, Y., Zhang, H.: Refinement of reproducing kernels. J. Mach. Learn. Res. **10**, 107–140 (2009)
38. Young, R.M.: An Introduction to Nonharmonic Fourier Series. Academic Press, New York (1980)
39. Zhang, H., Xu, Y., Zhang, J.: Reproducing kernel Banach spaces for machine learning. J. Mach. Learn. Res. **10**, 2741–2775 (2009)
40. Zhang, H., Zhang, J.: Generalized semi-inner products with applications to regularized learning. J. Math. Anal. Appl., accepted
41. Zhang, T.: On the dual formulation of regularized linear systems with convex risks. Machine Learning **46**, 91–129 (2002)
42. Zhou, D., Xiao, B., Zhou, H., Dai, R.: Global geometry of SVM classifiers. Technical Report 30-5-02. Institute of Automation, Chinese Academy of Sciences (2002)