Available online at www.sciencedirect.com

ScienceDirect

Journal of Computational and Applied Mathematics 220 (2008) 456–463

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.com/locate/cam

Derivative reproducing properties for kernel methods in learning theory

Ding-Xuan Zhou*

Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China

Received 27 June 2007

Abstract

The regularity of functions from reproducing kernel Hilbert spaces (RKHSs) is studied in the setting of learning theory. We provide a reproducing property for partial derivatives up to order s when the Mercer kernel is C^{2s}. For such a kernel on a general domain we show that the RKHS can be embedded into the function space C^s. These observations yield a representer theorem for regularized learning algorithms involving data for function values and gradients. Examples of Hermite learning and semi-supervised learning penalized by gradients on data are considered.

MSC: 68T05; 62J02

Keywords: Learning theory; Reproducing kernel Hilbert spaces; Derivative reproducing; Representer theorem; Hermite learning and semi-supervised learning

1. Introduction

Reproducing kernel Hilbert spaces (RKHSs) form an important class of function spaces in *learning theory*. Their reproducing property together with the Hilbert space structure ensures the effectiveness of many practical learning algorithms implemented in these function spaces.

Let X be a separable metric space and $K : X \times X \to \mathbb{R}$ be a continuous and symmetric function such that for any finite set of points $\{x_1, \dots, x_\ell\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^{\ell}$ is positive semidefinite. Such a function is called a *Mercer kernel*.

The RKHS $\mathscr{H}_K$ associated with the kernel K is defined (see [2]) to be the completion of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ given by $\langle K_x, K_y \rangle_K = K(x, y)$. That is, $\langle \sum_i \alpha_i K_{x_i}, \sum_j \beta_j K_{y_j} \rangle_K = \sum_{i,j} \alpha_i \beta_j K(x_i, y_j)$. The *reproducing property* takes the form

$$\langle K_x, f \rangle_K = f(x) \quad \forall x \in X, \quad f \in \mathscr{H}_K. \tag{1.1}$$

* Tel.: +852 2788 9708; fax: +852 2788 8561.
E-mail address: mazhou@cityu.edu.hk.

0377-0427/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2007.08.023

1.1. Learning algorithms by regularization in RKHS

A large family of learning algorithms are generated by *regularization schemes* in RKHS. Such a scheme can be expressed [6] in terms of a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in (X \times \mathbb{R})^m$ for learning and a loss function $V : \mathbb{R}^2 \to \mathbb{R}_+$ as

$$f_{\mathbf{z},\lambda} = \arg \min_{f \in \mathscr{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m V(y_i, f(x_i)) + \lambda \|f\|_K^2 \right\}, \quad \lambda > 0. \tag{1.2}$$

The reproducing property (1.1) makes solution (1.2) to a minimization problem over $\mathscr{H}_K$, a possibly infinite dimensional space, achieved in the finite dimensional subspace spanned by $\{K_{x_i}\}_{i=1}^m$. This is called a *representer theorem* [15]. When V is convex with respect to the second variable, (1.2) can be solved by a convex optimization problem for the coefficients of $f_{\mathbf{z},\lambda} = \sum_{i=1}^m c_i K_{x_i}$ over $\mathbb{R}^m$. For some special loss functions such as those in support vector machines, this convex optimization problem is actually a convex quadratic programming one, hence many efficient computing tools are available. This makes the *kernel method* in learning theory very powerful in various applications.

For regression problems, one often takes the loss function to be $V(y, f(x)) = \psi(y - f(x))$ with $\psi : \mathbb{R} \to \mathbb{R}_+$ a convex function. In particular, for the least square regression, we choose $\psi(t) = t^2$ and for the *support vector machine* regression, we choose $\psi(t) = \max\{0, |t| - \varepsilon\}$, an ε-insensitive loss function with some threshold $\varepsilon > 0$.

For binary classification problems, $y \in \{1, -1\}$, so one usually sets $V(y, f(x)) = \phi(yf(x))$ with $\phi : \mathbb{R} \to \mathbb{R}_+$ a convex function. In particular, for the least square classification, $\psi(t) = (1 - t)^2$; for the support vector machine classification, we choose ϕ to be the hinge loss $\phi(t) = \max\{0, 1 - t\}$ or the support vector machine q-norm loss $\phi_q(t) = (\max\{0, 1 - t\})^q$ with $1 < q < \infty$.

1.2. Learning with gradients

For some applications, one may have gradient data or unlabelled data available for improving learning ability [3,7]. Such situations yield learning algorithms involving data for function values or their gradients. Here $X \subseteq \mathbb{R}^n$ and with n variables $\{x^1, \ldots, x^n\}$ of $\mathbb{R}^n$, the *gradient* of a differentiable function $f : X \to \mathbb{R}$ is a vector formed by its partial derivatives as $\nabla f = (\partial f/\partial x^1, \ldots, \partial f/\partial x^n)^{\mathrm{T}}$.

Example 1 (*Semi-supervised learning with gradients of functions in $\mathscr{H}_K$*). If $\mathbf{z}=\{(x_i, y_i)\}_{i=1}^m \in (X \times \mathbb{R})^m$ are labelled data and $\mathbf{u}=\{x_i\}_{i=m+1}^{m+\ell} \in X^\ell$ are unlabelled data, we introduce a *semi-supervised learning algorithm* involving gradients as

$$f_{\mathbf{z},\mathbf{u},\lambda,\mu} = \arg \min_{f \in \mathscr{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m V(y_i, f(x_i)) + \frac{\mu}{m + \ell} \sum_{i=1}^{m+\ell} |\nabla f(x_i)|^2 + \lambda \|f\|_K^2 \right\}. \tag{1.3}$$

Here $\lambda, \mu > 0$ are two regularization parameters.

Example 2 (*Hermite learning with gradient data*). Assume in addition to the data $\mathbf{z}$ approximating values of a desired function $\tilde{f}$ (i.e. $y_i \approx \tilde{f}(x_i)$), we get sampling values $\mathbf{y}' = \{y_i'\}_{i=1}^m$, $y_i' \in \mathbb{R}^n$, for the gradients of $\tilde{f}$ (i.e. $y_i' \approx \nabla \tilde{f}(x_i)$), then we introduce an *Hermite learning algorithm* by learning the function values and gradients simultaneously as

$$f_{\mathbf{z},\mathbf{y}',\lambda} = \arg \min_{f \in \mathscr{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m [(y_i - f(x_i))^2 + |y_i' - \nabla f(x_i)|^2] + \lambda \|f\|_K^2 \right\}. \tag{1.4}$$

To solve optimization problems like (1.3) and (1.4) with effective computing tools (such as those for convex quadratic programming), we study representer theorems and need *reproducing properties for gradients* similar to (1.1). This is the first purpose of this paper.

1.3. Capacity of RKHS

Learning ability of algorithms in $\mathscr{H}_K$ depends on the kernel K, a loss function measuring errors and probability distributions from which the samples are drawn [11,1]. Its quantitative estimates in terms of $\mathscr{H}_K$ rely on two features of the RKHS: the approximation power and the capacity [12,5,16,17]. The latter can be measured by covering numbers of the unit ball of the RKHS as a subset of $C(X)$. These covering numbers have been extensively studied in learning theory, see e.g. [1,19,18]. In particular, when $X=[0,1]^n$ and K is C^s with s not being an even integer, an explicit bound for the covering numbers was presented in [19]. This was done by showing that $\mathscr{H}_K$ can be embedded in $C^{s/2}(X)$, a Hölder space on X. The embedding result yields error estimates in the $C^{s/2}$ metric by means of bounds in the $\mathscr{H}_K$ metric for learning algorithms. For example, for the least square regularized regression algorithm (1.2) with $V(y,f(x))=(f(x)-y)^2$, when $K\in C^{2+\varepsilon}(X\times X)$ with some $\varepsilon>0$, rates of the error $\|f_{\mathbf{z},\lambda}-f_\rho\|_{C^1(X)}$ for learning the regression function f_ρ was provided in [14] from bounds for $\|f_{\mathbf{z},\lambda}-f_\rho\|_K$. A natural question is whether the extra $\varepsilon>0$ can be omitted. The general problem for $K\in C^s$ is whether $\mathscr{H}_K$ can be embedded in $C^{s/2}(X)$ when X is a general domain in $\mathbb{R}^n$ or when s is an even integer. Solving this general question is the second purpose of this paper. The main difficulty we overcome here is the lack of regularity of the general domain.

2. Reproducing partial derivatives in RKHS

To allow a general situation, we would not assume any regularity for the boundary of X. To consider partial derivatives, we assume that the interior of X is nonempty.

For $s\in\mathbb{Z}_+$, we denote an index set $I_s:=\{\alpha\in\mathbb{Z}_+^n:|\alpha|\leqslant s\}$ where $|\alpha|=\sum_{j=1}^n\alpha^j$ for $\alpha=(\alpha^1,\ldots,\alpha^n)\in\mathbb{Z}_+^n$. For a function f of n variables and $x=(x^1,\ldots,x^n)\in\mathbb{R}^n$, we denote its partial derivative $D^\alpha f$ at x (if it exists) as

$$D^\alpha f(x)=D_1^{\alpha^1}\ldots D_n^{\alpha^n}f(x)=\frac{\partial^{|\alpha|}}{\partial(x^1)^{\alpha^1}\ldots\partial(x^n)^{\alpha^n}}f(x).$$

Definition 1 (*Ziemer [21]*). Let X be a compact subset of $\mathbb{R}^n$ which is the closure of its nonempty interior X^o. Define $C^s(X^o)$ to be the space of functions f on X^o such that $D^\alpha f$ is well-defined and continuous on X^o for each $\alpha\in I_s$. Define $C^s(X)$ to be the space of continuous functions f on X such that $f|_{X^o}\in C^s(X^o)$ and for each $\alpha\in I_s$, $D^\alpha(f|_{X^o})$ has a continuous extension to X denoted as $D^\alpha f$.

In particular, the extension of $f|_{X^o}$ is f itself. The linear space $C^s(X)$ is actually a Banach space with the norm

$$\|f\|_{C^s(X)}=\sum_{\alpha\in I_s}\|D^\alpha f\|_\infty=\sum_{\alpha\in I_s}\sup_{x\in X^o}|D^\alpha(f|_{X^o})(x)|.$$

Observe that for $f\in C^1(X)$, the gradient ∇f equals $(D^{e^1}f,\ldots,D^{e^n}f)$ where e^j is the jth standard unit vector in $\mathbb{R}^n$.

The property of reproducing partial derivatives of functions in $\mathscr{H}_K$ is given by partial derivatives of the Mercer kernel K.

If $K\in C^{2s}(X\times X)$, for $\alpha\in I_s$ we extend α to $\mathbb{Z}_+^{2n}$ by adding zeros to the last n components and denote the partial derivative of K as $D^\alpha K$. That is,

$$D^\alpha K(x,y)=\frac{\partial^{|\alpha|}}{\partial(x^1)^{\alpha^1}\ldots\partial(x^n)^{\alpha^n}}K(x^1,\ldots,x^n,y^1,\ldots,y^n),\quad x,y\in X^o$$

and $D^\alpha K$ is a continuous extension of $D^\alpha(K|_{X^o\times X^o})$ to $X\times X$. For $x\in X$, denote $(D^\alpha K)_x$ as the function on X given by $(D^\alpha K)_x(y)=D^\alpha K(x,y)$. By the symmetry of K, we have

$$D^\alpha(K_y)(x)=(D^\alpha K)_x(y)=D^\alpha K(x,y)\quad\forall x,y\in X. \tag{2.1}$$

Now we can give the result on reproducing partial derivatives and embedding of $\mathscr{H}_K$ into $C^s(X)$. Here (1.1) and the weak compactness of a closed ball of a Hilbert space play an important role.

Theorem 1. *Let $s\in\mathbb{N}$ and $K:X\times X\to\mathbb{R}$ be a Mercer kernel such that $K\in C^{2s}(X\times X)$. Then the following statements hold*:

(a) *For any $x\in X$ and $\alpha\in I_s$, $(D^\alpha K)_x\in\mathscr{H}_K$.*

(b) *A partial derivative reproducing property holds true for $\alpha\in I_s$:*

$$D^\alpha f(x)=\langle(D^\alpha K)_x,f\rangle_K\quad\forall x\in X,\quad f\in\mathscr{H}_K. \tag{2.2}$$

(c) *The inclusion* $J : \mathscr{H}_K \hookrightarrow C^s(X)$ *is well-defined and bounded*:

$$\|f\|_{C^s(X)} \leqslant \sqrt{n^s \|K\|_{C^{2s}(X\times X)}}\|f\|_K \quad \forall f \in \mathscr{H}_K. \tag{2.3}$$

(d) *J is compact. More strongly, for any closed bounded subset B of $\mathscr{H}_K$, $J(B)$ is a compact subset of $C^s(X)$.*

Proof. We first prove (a) and (b) together by induction on $|\alpha| = 0, 1, \ldots, s$.

The case $|\alpha| = 0$ is trivial since $\alpha = 0$ and $(D^0 K)_x = K_x$ satisfies (1.1).

Let $0 \leqslant \ell \leqslant s - 1$. Suppose that $(D^\alpha K)_x \in \mathscr{H}_K$ and (2.2) holds for any $x \in X$ and $\alpha \in \mathbb{Z}_+^n$ with $|\alpha| = \ell$. Then (2.2) implies that for any $y \in X$,

$$\langle (D^\alpha K)_y, (D^\alpha K)_x\rangle_K = D^\alpha((D^\alpha K)_x)(y) = D^\alpha(D^\alpha K(x, \cdot))(y) = D^{(\alpha,\alpha)}K(x, y). \tag{2.4}$$

Here $(\alpha, \alpha) \in \mathbb{Z}_+^{2n}$ is formed by α in the first and second sets of n components.

Now we turn to the case $\ell + 1$. Consider the index $\alpha + e^j$ with $|\alpha + e^j| = \ell + 1$. We prove (a) and (b) for this index in four steps.

Step 1: Proving $(D^{\alpha+e^j} K)_x \in \mathscr{H}_K$ for $x \in X^o$. Since $x \in X^o$, there exists some $r > 0$ such that $x + \{y \in \mathbb{R}^n : |y| \leqslant r\} \subseteq X^o$. Then by (2.4), the set $\{(1/t)((D^\alpha K)_{x+te^j} - (D^\alpha K)_x) : |t| \leqslant r\}$ of functions in $\mathscr{H}_K$ satisfies

$$\begin{aligned}\left\|\frac{1}{t}((D^\alpha K)_{x+te^j} - (D^\alpha K)_x)\right\|_K^2 &= \frac{1}{t^2}\{D^{(\alpha,\alpha)}K(x + te^j, x + te^j) - D^{(\alpha,\alpha)}K(x + te^j, x)\\ &\quad - D^{(\alpha,\alpha)}K(x, x + te^j) + D^{(\alpha,\alpha)}K(x, x)\} \leqslant \|D^{(\alpha+e^j,\alpha+e^j)}K\|_\infty \quad \forall |t| \leqslant r.\end{aligned}$$

Here we have used the assumption $K \in C^{2s}(X \times X)$ and $|(\alpha + e^j, \alpha + e^j)| = 2|\alpha| + 2 = 2\ell + 2 \leqslant 2s$. That means $\{(1/t)((D^\alpha K)_{x+te^j} - (D^\alpha K)_x) : |t| \leqslant r\}$ lies in the closed ball of the Hilbert space $\mathscr{H}_K$ with a finite radius $\|D^{(\alpha+e^j,\alpha+e^j)}K\|_\infty$. Since this ball is weakly compact, there is a sequence $\{t_i\}_{i=1}^\infty$ with $|t_i| \leqslant r$ and $\lim_{i\to\infty} t_i = 0$ such that $\{(1/t_i)((D^\alpha K)_{x+t_ie^j} - (D^\alpha K)_x)\}$ converges weakly to an element g_x of $\mathscr{H}_K$ as $i \to \infty$. The weak convergence tells us that

$$\lim_{i\to\infty}\left\langle \frac{1}{t_i}((D^\alpha K)_{x+t_ie^j} - (D^\alpha K)_x), f\right\rangle_K = \langle g_x, f\rangle_K \quad \forall f \in \mathscr{H}_K. \tag{2.5}$$

In particular, by taking $f = K_y$ with $y \in X$, there holds

$$g_x(y) = \lim_{i\to\infty}\left\langle \frac{1}{t_i}((D^\alpha K)_{x+t_ie^j} - (D^\alpha K)_x), K_y\right\rangle_K.$$

By (2.2) for α and (2.1) we have

$$\begin{aligned}g_x(y) &= \lim_{i\to\infty}\frac{1}{t_i}(D^\alpha(K_y)(x + t_ie^j) - D^\alpha(K_y)(x))\\ &= \lim_{i\to\infty}\frac{1}{t_i}(D^\alpha K(x + t_ie^j, y) - D^\alpha K(x, y)) = D^{\alpha+e^j}K(x, y) = (D^{\alpha+e^j}K)_x(y).\end{aligned}$$

This is true for an arbitrary point $y \in X$. Hence $(D^{\alpha+e^j}K)_x = g_x$ as functions on X. Since $g_x \in \mathscr{H}_K$, we know $(D^{\alpha+e^j}K)_x \in \mathscr{H}_K$.

Step 2: Proving for $x \in X^o$ the convergence

$$\frac{1}{t}((D^\alpha K)_{x+te^j} - (D^\alpha K)_x) \to (D^{\alpha+e^j}K)_x \quad \text{in } \mathscr{H}_K \quad (t \to 0). \tag{2.6}$$

Applying (2.5) and (2.2) for α to the function $(D^{\alpha+e^j}K)_x \in \mathscr{H}_K$ yields

$$\begin{aligned}
&\langle (D^{\alpha+e^j}K)_x, (D^{\alpha+e^j}K)_x\rangle_K \\
&\quad = \lim_{i\to\infty} \frac{1}{t_i}\{D^\alpha((D^{\alpha+e^j}K)_x)(x+t_ie^j) - D^\alpha((D^{\alpha+e^j}K)_x)(x)\} \\
&\quad = \lim_{i\to\infty} \frac{1}{t_i}\{D^\alpha(D^{\alpha+e^j}K(x,\cdot))(x+t_ie^j) - D^\alpha(D^{\alpha+e^j}K(x,\cdot))(x)\} \\
&\quad = D^{(\alpha+e^j,\alpha+e^j)}K(x,x).
\end{aligned}$$

This in connection with (2.2) implies

$$\begin{aligned}
&\left\| \frac{1}{t}((D^\alpha K)_{x+te^j} - (D^\alpha K)_x) - (D^{\alpha+e^j}K)_x \right\|_K^2 \\
&\quad = \frac{1}{t^2}\{D^{(\alpha,\alpha)}K(x+te^j, x+te^j) - 2D^{(\alpha,\alpha)}K(x+te^j,x) + D^{(\alpha,\alpha)}K(x,x)\} \\
&\qquad - \frac{2}{t}\{D^\alpha((D^{\alpha+e^j}K)_x)(x+te^j) - D^\alpha((D^{\alpha+e^j}K)_x)(x)\} + D^{(\alpha+e^j,\alpha+e^j)}K(x,x) \\
&\quad = \frac{1}{t^2}\int_0^t\int_0^t D^{(\alpha+e^j,\alpha+e^j)}K(x+ue^j, x+ve^j)\,\mathrm{d}u\,\mathrm{d}v \\
&\qquad - \frac{2}{t}\int_0^t D^{(\alpha+e^j,\alpha+e^j)}K(x, x+ve^j)\,\mathrm{d}v + D^{(\alpha+e^j,\alpha+e^j)}K(x,x) \\
&\quad = \frac{1}{t^2}\int_0^t\int_0^t \{D^{(\alpha+e^j,\alpha+e^j)}K(x+ue^j, x+ve^j) \\
&\qquad - 2D^{(\alpha+e^j,\alpha+e^j)}K(x, x+ve^j) + D^{(\alpha+e^j,\alpha+e^j)}K(x,x)\}\,\mathrm{d}u\,\mathrm{d}v.
\end{aligned}$$

If we define the modulus of continuity for a function $g \in C(X\times X)$ to be a function of $\delta \in (0,\infty)$ as

$$\omega(g,\delta) := \sup\{|g(x_1,y_1) - g(x_2,y_2)| : x_i, y_i \in X \text{ with } |x_1-x_2| \leqslant \delta, |y_1-y_2| \leqslant \delta\}, \tag{2.7}$$

we know from the uniform continuity of g that $\lim_{\delta\to 0_+}\omega(g,\delta) = 0$. Moreover, the function $\omega(g,\delta)$ is continuous on $(0,\infty)$. Using the modulus of continuity for the function $D^{(\alpha+e^j,\alpha+e^j)}K \in C(X\times X)$ we see that

$$\left\| \frac{1}{t}((D^\alpha K)_{x+te^j} - (D^\alpha K)_x) - (D^{\alpha+e^j}K)_x \right\|_K^2 \leqslant 2\omega(D^{(\alpha+e^j,\alpha+e^j)}K, |t|). \tag{2.8}$$

This converges to zero as $t \to 0$. Therefore (2.6) holds true.

Step 3: Proving (2.2) for $x \in X^o$ and $\alpha + e^j$. Let $f \in \mathscr{H}_K$. By (2.6) we have

$$\langle (D^{\alpha+e^j}K)_x, f\rangle_K = \lim_{t\to 0}\left\langle \frac{1}{t}((D^\alpha K)_{x+te^j} - (D^\alpha K)_x), f\right\rangle_K.$$

By (2.2) for α, we see that this equals

$$\langle (D^{\alpha+e^j}K)_x, f\rangle_K = \lim_{t\to 0}\frac{1}{t}\{D^\alpha f(x+te^j) - D^\alpha f(x)\}.$$

That is, $D^{\alpha+e^j}f(x)$ exists and equals $\langle (D^{\alpha+e^j}K)_x, f\rangle_K$. This verifies (2.2) for $\alpha + e^j$.

Step 4: Proving (a) and (b) for $x \in \partial X := X\backslash X^o$. Notice from the first three steps that for $x', x'' \in X^o$, there holds

$$\begin{aligned}
\|(D^{\alpha+e^j}K)_{x'} - (D^{\alpha+e^j}K)_{x''}\|_K^2 &= \{D^{(\alpha+e^j,\alpha+e^j)}K(x',x') - D^{(\alpha+e^j,\alpha+e^j)}K(x',x'') \\
&\quad - D^{(\alpha+e^j,\alpha+e^j)}K(x'',x') + D^{(\alpha+e^j,\alpha+e^j)}K(x'',x'')\} \\
&\leqslant 2\omega(D^{(\alpha+e^j,\alpha+e^j)}K, |x'-x''|).
\end{aligned}$$

It follows that for any sequence $\{x^{(i)} \in X^o\}_{i=1}^{\infty}$ converging to x, the sequence of functions $\{(D^{\alpha+e^j}K)_{x^{(i)}}\}$ is a Cauchy sequence in the Hilbert space $\mathscr{H}_K$. So it converges to a limit function $h \in \mathscr{H}_K$. Applying what we have proved for $x^{(i)} \in X^o$ we get

$$h(y) = \langle h, K_y\rangle_K = \lim_{i\to\infty} \langle (D^{\alpha+e^j}K)_{x^{(i)}}, K_y\rangle_K = (D^{\alpha+e^j}K)_x(y) \quad \forall y \in X.$$

This verifies $(D^{\alpha+e^j}K)_x = h \in \mathscr{H}_K$.

Let $f \in \mathscr{H}_K$. We define a function $f^{[j]}$ on X as

$$f^{[j]}(x) = \langle (D^{\alpha+e^j}K)_x, f\rangle_K, \quad x \in X.$$

By the conclusion in Step 3, we know that $f^{[j]}(x) = D^{\alpha+e^j}f(x)$ for $x \in X^o$, hence $f^{[j]}$ is continuous on X^o.

Let us now prove the continuity of $f^{[j]}$ at each $x \in \partial X$. If $\{x^{(i)} \in X^o\}_{i=1}^{\infty}$ is a sequence satisfying $\lim_{i\to\infty} x^{(i)} = x$, then the above proof tells us that $(D^{\alpha+e^j}K)_{x^{(i)}}$ converges $(D^{\alpha+e^j}K)_x$ in the $\mathscr{H}_K$ metric meaning that $\lim_{i\to\infty} \|(D^{\alpha+e^j}K)_{x^{(i)}} - (D^{\alpha+e^j}K)_x\|_K = 0$. So by the definition of $f^{[j]}$ and the Schwarz inequality, we have

$$\begin{aligned}|f^{[j]}(x^{(i)}) - f^{[j]}(x)| &= |\langle (D^{\alpha+e^j}K)_{x^{(i)}} - (D^{\alpha+e^j}K)_x, f\rangle_K| \\ &\leqslant \|(D^{\alpha+e^j}K)_{x^{(i)}} - (D^{\alpha+e^j}K)_x\|_K \|f\|_K \to 0 \quad \text{as } i \to \infty.\end{aligned}$$

Thus the function $f^{[j]}$ is continuous on X, and it is a continuous extension of $D^{\alpha+e^j}f$ from X^o onto X. So (b) holds true for $x \in X$. This completes the induction procedure for proving the statements in (a) and (b).

(c) We use (2.2) and (2.4). For $f \in \mathscr{H}_K$, $x, \tilde{x} \in X$ and $\alpha \in I_s$, the Schwarz inequality implies

$$\begin{aligned}|D^\alpha f(x) - D^\alpha f(\tilde{x})| &= |\langle (D^\alpha K)_x - (D^\alpha K)_{\tilde{x}}, f\rangle_K| \leqslant \|(D^\alpha K)_x - (D^\alpha K)_{\tilde{x}}\|_K \|f\|_K \\ &\leqslant \sqrt{D^{(\alpha,\alpha)}K(x,x) - 2D^{(\alpha,\alpha)}K(x,\tilde{x}) + D^{(\alpha,\alpha)}K(\tilde{x},\tilde{x})}\,\|f\|_K.\end{aligned}$$

Hence

$$|D^\alpha f(x) - D^\alpha f(\tilde{x})| \leqslant \sqrt{2\omega(D^{(\alpha,\alpha)}K, |x-\tilde{x}|)}\,\|f\|_K \quad \forall x, \tilde{x} \in X. \tag{2.9}$$

As $\lim_{\delta\to 0_+} \omega(D^{(\alpha,\alpha)}K, \delta) = 0$, we know that $D^\alpha f \in C(X)$. It means $f \in C^s(X)$ and the inclusion J is well-defined. To see the boundedness, we apply the Schwarz inequality again and have

$$|D^\alpha f(x)| = |\langle (D^\alpha K)_x, f\rangle_K| \leqslant \sqrt{D^{(\alpha,\alpha)}K(x,x)}\,\|f\|_K \leqslant \sqrt{\|D^{(\alpha,\alpha)}K\|_\infty}\,\|f\|_K.$$

It follows that

$$\|f\|_{C^s(X)} = \sum_{\alpha\in I_s} \|D^\alpha f\|_\infty \leqslant \sum_{\alpha\in I_s} \sqrt{\|D^{(\alpha,\alpha)}K\|_\infty}\,\|f\|_K \leqslant \sqrt{n^s \sum_{\alpha\in I_s} \|D^{(\alpha,\alpha)}K\|_\infty}\,\|f\|_K.$$

Then (2.3) is verified.

(d) If B is a closed bounded subset of $\mathscr{H}_K$, there is some $R > 0$ such that $B \subseteq \{f \in \mathscr{H}_K : \|f\|_K \leqslant R\}$. To show that $J(B)$ is compact, let $\{f_j\}_{j=1}^{\infty}$ be a sequence in B. The estimate (2.9) tells us that for each $\alpha \in I_s$ and $j \in \mathbb{N}$,

$$|D^\alpha f_j(x) - D^\alpha f_j(\tilde{x})| \leqslant \sqrt{2\omega(D^{(\alpha,\alpha)}K, |x-\tilde{x}|)R} \quad \forall x, \tilde{x} \in X.$$

It says that the sequence of functions $\{D^\alpha f_j\}_{j=1}^{\infty}$ is uniformly continuous. This is true for each $\alpha \in I_s$. So by taking subsequences for α (one-after-one), we know that there is a subsequence $\{f_{j_\ell}\}_{\ell=1}^{\infty}$ which converges to a function $f^* \in C^s(X)$ in the metric $\|\cdot\|_{C^s(X)}$. Observe that $\{f_{j_\ell}\}_{\ell=1}^{\infty}$ lies in the ball of $\mathscr{H}_K$ with radius R which is weakly compact, it contains a subsequence $\{f_{j_{\ell_k}}\}_{k=1}^{\infty}$ which is also a subsequence of $\{f_j\}_{j=1}^{\infty}$ and converges weakly to a function $\tilde{f} \in \mathscr{H}_K$ in the metric $\|\cdot\|_K$. According to (2.2), the weak convergence in $\mathscr{H}_K$ tells us that $\{f_{j_{\ell_k}}\}_{k=1}^{\infty}$ converges to $\tilde{f}$ in the metric$\|\cdot\|_{C^s(X)}$. Therefore, $f^* = \tilde{f} \in \mathscr{H}_K$ and $\{f_j\}_{j=1}^{\infty} \subseteq J(B)$ contains a subsequence which converges in $C^s(X)$ to f^*. This proves that $J(B)$ is compact. The proof of Theorem 1 is complete. □

Theorem 1 can be extended to other kernels [10]. Relation (2.3) tells us that the error bounds in the norm $\|\cdot\|_K$ can be used to estimate convergence rates of learning algorithms in the norm $\|\cdot\|_{C^s(X)}$, as done in [14].

3. Representer theorems for learning with derivative data

A general learning algorithm of regularization in $\mathscr{H}_K$ involving partial derivative data takes the form

$$f_{\mathbf{x},\vec{\mathbf{y}},\lambda} = \arg\min_{f\in\mathscr{H}_K}\left\{\sum_{i=1}^{m} V_i(\vec{y}_i, \{D^\alpha f(x_i)\}_{\alpha\in J_i}) + \lambda\|f\|_K^2\right\}, \tag{3.1}$$

where for each $i \in \{1,\ldots,m\}$, $x_i \in X$, $\vec{y}_i$ is a vector, J_i is a subset of I_s and V_i is a loss function with values in $\mathbb{R}_+$ of compatible variables. Denote the number of elements in the set J_i as $\#(J_i)$.

The partial derivative reproducing property (2.2) stated in Theorem 1 enables us to derive a representer theorem for the learning algorithm (3.1), which asserts that the minimization over the possibly infinite dimensional space $\mathscr{H}_K$ can be achieved in a finite dimensional subspace generated by $\{K_{x_i}\}$ and their partial derivatives.

Theorem 2. *Let $s \in \mathbb{N}$ and $K : X \times X \to \mathbb{R}$ be a Mercer kernel such that $K \in C^{2s}(X\times X)$. If $\lambda > 0$, then the solution $f_{\mathbf{x},\vec{\mathbf{y}},\lambda}$ of scheme (3.1) exists and lies in the subspace spanned by $\{(D^\alpha K)_{x_i} : \alpha \in J_i, i = 1,\ldots,m\}$. If we write $f_{\mathbf{x},\vec{\mathbf{y}},\lambda} = \sum_{i=1}^m \sum_{\alpha\in J_i} c^*_{i,\alpha}(D^\alpha K)_{x_i}$ with $c^* = (c^*_{i,\alpha})_{\alpha\in J_i, i=1,\ldots,m} \in \mathbb{R}^N$ where $N = \sum_{i=1}^m \#(J_i)$, then*

$$c^* = \arg\min_{c\in\mathbb{R}^N}\left\{\sum_{i=1}^m V_i\left(\vec{y}_i, \left\{\sum_{j=1}^m\sum_{\beta\in J_j} c_{j,\beta} D^{(\beta,\alpha)}K(x_j,x_i)\right\}_{\alpha\in J_i}\right)\right.$$
$$\left. +\lambda\sum_{i=1}^m\sum_{\alpha\in J_i}\sum_{j=1}^m\sum_{\beta\in J_j} c_{i,\alpha}c_{j,\beta}D^{(\beta,\alpha)}K(x_j,x_i)\right\}.$$

Proof. By Theorem 1, we know that for any α in J_i, the function $(D^\alpha K)_{x_i}$ lies in $\mathscr{H}_K$. Denote the subspace of $\mathscr{H}_K$ spanned by $\{(D^\alpha K)_{x_i} : \alpha \in J_i, i = 1,\ldots,m\}$ as $\mathscr{H}_{K,\mathbf{x}}$. Let P be the orthogonal projection onto this subspace. Then for any $f \in \mathscr{H}_K$, the function $f - P(f)$ is orthogonal to $\mathscr{H}_{K,\mathbf{x}}$. In particular, $\langle f - P(f), (D^\alpha K)_{x_i}\rangle_K = 0$ for any $\alpha \in J_i$ and $1 \leqslant i \leqslant m$. This in connection with the partial derivative reproducing property (2.2) tells us that

$$D^\alpha(f - P(f))(x_i) = D^\alpha f(x_i) - D^\alpha(P(f))(x_i) = 0 \quad \forall \alpha \in J_i,\ \ i = 1,\ldots,m.$$

Thus, if we denote $\mathscr{E}_\mathbf{z}(f) = \sum_{i=1}^m V_i(\vec{y}_i, \{D^\alpha f(x_i)\}_{\alpha\in J_i})$, we see that $\mathscr{E}_\mathbf{z}(f) = \mathscr{E}_\mathbf{z}(P(f))$. Notice that $\|P(f)\|_K \leqslant \|f\|_K$ and the strict inequality holds unless $f = P(f)$, i.e., $f \in \mathscr{H}_{K,\mathbf{x}}$. Therefore,

$$\min_{f\in\mathscr{H}_K}\{\mathscr{E}_\mathbf{z}(f) + \lambda\|f\|_K^2\} = \min_{f\in\mathscr{H}_{K,\mathbf{x}}}\{\mathscr{E}_\mathbf{z}(f) + \lambda\|f\|_K^2\}$$

and a minimizer $f_{\mathbf{x},\vec{\mathbf{y}},\lambda}$ exists and lies in $\mathscr{H}_{K,\mathbf{x}}$ since the subspace is finite dimensional.

The second statement is trivial. □

We shall not discuss learning rates of the learning algorithms (1.3) and (1.4). Though rough estimates can be given using methods from [4,8,13], satisfactory error analysis for more general learning algorithms [9,20] will be done later.

Acknowledgements

The work described in this paper was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 103704) and a Joint Research Fund for Hong Kong and Macau Young Scholars from the National Science Fund for Distinguished Young Scholars (Project No. 10529101).

References

[1] M. Anthony, P.L. Bartlett, Neural Network Learning: Theoretical Foundations, Cambridge University Press, Cambridge, 1999.

[2] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.

[3] M. Belkin, P. Niyogi, Semisupervised learning on Riemannian manifolds, Mach. Learn. 56 (2004) 209–239.

[4] E. De Vito, A. Caponnetto, L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, Found. Comput. Math. 5 (2005) 59–85.

[5] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, A. Verri, Some properties of regularized kernel methods, J. Mach. Learn. Res. 5 (2004) 1363–1390.

[6] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.

[7] D. Hardin, I. Tsamardinos, C.F. Aliferis, A theoretical characterization of linear SVM-based feature selection, Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004.

[8] S. Mukherjee, P. Niyogi, T. Poggio, R. Rifkin, Learning theory: stability is sufficient for generalization and necessary and sufficient for empirical risk minimization, Adv. Comput. Math. 25 (2006) 161–193.

[9] S. Mukherjee, Q. Wu, Estimation of gradients and coordinate covariation in classification, J. Mach. Learn. Res. 7 (2006) 2481–2514.

[10] R. Schaback, J. Werner, Linearly constrained reconstruction of functions by kernels with applications to machine learning, Adv. Comput. Math. 25 (2006) 237–258.

[11] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, M. Anthony, Structural risk minimization over data-dependent hierarchies, IEEE Trans. Inform. Theory 44 (1998) 1926–1940.

[12] S. Smale, D.X. Zhou, Estimating the approximation error in learning theory, Anal. Appl. 1 (2003) 17–41.

[13] S. Smale, D.X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Amer. Math. Soc. 41 (2004) 279–305.

[14] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. 26 (2007) 153–172.

[15] G. Wahba, Spline Models for Observational Data, SIAM, Philadelphia, PA, 1990.

[16] Y. Yao, On complexity issue of online learning algorithms, IEEE Trans. Inform. Theory, to appear.

[17] Y. Ying, Convergence analysis of online algorithms, Adv. Comput. Math. 27 (2007) 273–291.

[18] D.X. Zhou, The covering number in learning theory, J. Complexity 18 (2002) 739–767.

[19] D.X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory 49 (2003) 1743–1752.

[20] D.X. Zhou, K. Jetter, Approximation with polynomial kernels and SVM classifiers, Adv. Comput. Math. 25 (2006) 323–344.

[21] W.P. Ziemer, Weakly Differentiable Functions, Springer, New York, 1989.