A function representation for learning in Banach spaces*

Charles A. Micchelli[1] and Massimiliano Pontil[2]

[1] Department of Mathematics and Statistics
State University of New York, The University at Albany
1400 Washington Avenue, Albany, NY, 12222, USA
cam@math.albany.edu
[2] Department of Computer Sciences
University College London
Gower Street, London WC1E 6BT, England, UK
m.pontil@cs.ucl.ac.uk

Abstract. Kernel–based methods are powerful for high dimensional function representation. The theory of such methods rests upon their attractive mathematical properties whose setting is in Hilbert spaces of functions. It is natural to consider what the corresponding circumstances would be in Banach spaces. Led by this question we provide theoretical justifications to enhance kernel–based methods with function composition. We explore regularization in Banach spaces and show how this function representation naturally arises in that problem. Furthermore, we provide circumstances in which these representations are dense relative to the uniform norm and discuss how the parameters in such representations may be used to fit data.

1 Introduction

Kernel–based methods have in recent years been a focus of attention in Machine Learning. They consist in choosing a kernel $K : D \times D \to \mathbb{R}$ which provides functions of the form

$$\sum_{j \in \mathbb{Z}_m} c_j K(x_j, \cdot) \tag{1.1}$$

whose parameters $D_m = \{x_j : j \in \mathbb{Z}_m\} \subseteq D$ and $c = \{c_j : j \in \mathbb{Z}_m\} \subset \mathbb{R}$ are used to learn an unknown function f. Here, we use the notation $\mathbb{Z}_m = \{0, \ldots, m-1\}$. Typically K is chosen to be a reproducing kernel of some Hilbert space. Although this is *not* required, it does provide (1.1) with a Hilbert space justification. The *simplicity* of the functional form (1.1) and its ability to address *efficiently high dimensional* learning tasks make it very attractive. Since it arises from Hilbert space considerations it is natural to inquire what may transpire in other Banach spaces. The goal of this paper is to study this question, especially learning algorithms based on regularization in a Banach space. A consequence

* This work was supported by NSF Grant No. ITR-0312113.

of our remarks here is that *function composition* should be introduced in the representation (1.1). That is, we suggest the use of the *nonlinear* functional form

$$\phi\left(\sum_{j\in\mathbb{Z}_m} c_j g_j\right) \tag{1.2}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ and for $j \in Z_m$, $g_j : D \to \mathbb{R}$ are prescribed functions, for example (but not necessarily so) $g_j = K(x_j, \cdot)$. In section 2 we provide an *abstract* framework where in a particular case the functional form (1.2) naturally arises. What we say here is a compromise between the generality in which we work and our desire to provide useful functional forms for Machine Learning.

We consider the problem of learning a function f in a Banach space from a set of continuous linear functionals $L_j(f) = y_j,\ j \in \mathbb{Z}_m$. Typically in Machine Learning there is available function values for learning, that is, the L_j are point evaluation functionals. However, there are many practical problems where such information is not readily available, for example tomography or EXAFS spectroscopy, [15]. Alternatively, it may be of practical advantage to use "local" averages of f as observed information. This idea is investigated in [23, c. 8] in the context of support vector machines. Perhaps, even more compelling is the question of what may be the "best" m observations that should be made to learn a function. For example, is it better to know function values or Fourier coefficients of a periodic function? These and related questions are addressed in [18] and lead us here to deal with linear functionals other than function values for Machine Learning.

We are especially interested in the case when the samples $y_j,\ j \in \mathbb{Z}_m$ are known to be noisy so that it is appropriate to estimate f as the minimizer in some Banach space of a regularization functional of the form

$$E(f) := \sum_{j\in\mathbb{Z}_m} Q(y_j, L_j(f)) + H(\|f\|) \tag{1.3}$$

where $H : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly increasing function, and $Q : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is some prescribed *loss function*. If the Banach space is a reproducing kernel Hilbert space, the linear functionals $L_j, j \in \mathbb{Z}_m$ are chosen to be point evaluations. In this case a minimizer of (1.3) has the form in equation (1.1), a fact which is known as the representer theorem, see e.g. [22, 25], which we generalize here to any Banach space.

We note that the problem of minimizing a regularization functional of the form (1.3) in a *finite dimensional* Banach has been considered in the case of support vector machines in [1] and in more general cases in [26]. Finite dimensional Banach spaces have been also considered in the context of on–line learning, see e.g. [9]. Learning in infinite dimensional Banach spaces has also been considered. For example, [7] considers learning a univariate function in L_p spaces, [2] addresses learning in non–Hilbert spaces using point evaluation with kernels, and [24, 6] propose large margin algorithms in a metric input space by embedding this space into certain Banach spaces of functions.

Since the functions (1.2) do not form a linear space as we vary $c \in \mathbb{R}^m$, we may also enhance them by *linear superposition* to obtain functions of the form

$$\sum\nolimits_{j\in\mathbb{Z}_n} a_j \phi \left(\sum\nolimits_{k\in\mathbb{Z}_m} c_{jk} g_k \right) \tag{1.4}$$

where $\{a_j : j \in \mathbb{Z}_n\} \subset \mathbb{R}$ and $\{c_{jk} : j \in \mathbb{Z}_n, k \in \mathbb{Z}_m\} \subset \mathbb{R}$ are real–valued parameters. This functional form has flexibility and simplicity. In particular, when the functions $\{g_j : j \in \mathbb{Z}_{m+1}\}$ are chosen to be a basis for *linear functions* on $\mathbb{R}^m$, (1.4) corresponds to feed–forward neural networks with one hidden layer, see for example [12].

In section 3 we address the problem of when functions of the form in equation (1.4) are dense in the space of continuous functions in the uniform norm. Finally, in section 4 we present some preliminary thoughts about the problem of choosing the parameters in (1.4) from prescribed linear constraints.

2 Regularization and minimal norm interpolation

Let $\mathcal{X}$ be a Banach space and $\mathcal{X}^*$ its dual, that is, the space of bounded linear functionals $L : \mathcal{X} \to \mathbb{R}$ with the norm $\|L\| := \sup\{L(x) : \|x\| \leq 1\}$. Given a set of examples $\{(L_j, y_j) : j \in \mathbb{Z}_m\} \subset \mathcal{X} \times \mathbb{R}$ and a prescribed function $V : \mathbb{R}^m \times \mathbb{R}_+ \to \mathbb{R}$ which is *strictly increasing* in its last argument (for every choice of its first argument) we consider the problem of minimizing the functional $E : \mathcal{X} \to \mathbb{R}$ defined for $x \in \mathcal{X}$ as

$$E(x) := V\left((L_j(x) : j \in \mathbb{Z}_m), \|x\|\right) \tag{2.5}$$

over all elements x in $\mathcal{X}$ (here V contains the information about the y_j). A special case of this problem is covered by a functional of the form (1.3). Suppose that x_0 is the solution to the above problem, x is any element of $\mathcal{X}$ such that $L_j(x) = y_j, j \in \mathbb{Z}_m$ where we set $y_j := L_j(x_0), j \in \mathbb{Z}_m$. By the definition of x_0 we have that

$$V(y, \|x_0\|) \leq V(y, \|x\|)$$

and so

$$\|x_0\| = \min\{\|x\| : L_j(x) = y_j, j \in \mathbb{Z}_m,\ x \in \mathcal{X}\}. \tag{2.6}$$

This observation is the motivation for our study of problem (2.6) which is usually called minimal norm interpolation. Note that this conclusion even holds when $\|x\|$ is replaced by *any* functional of x.

We make no claim for originality in our ensuing remarks about this problem which have been chosen to show the usefulness of the representation (1.2). Indeed, we are roaming over well–trodden ground.

Thus, given data $\{y_j : j \in \mathbb{Z}_m\} \subset \mathbb{R}\backslash\{0\}$, we consider the minimum norm interpolation (MNI) problem

$$\mu := \inf\{\|x\| : L_j(x) = y_j,\ j \in \mathbb{Z}_m,\ x \in \mathcal{X}\}. \tag{2.7}$$

We always require in (2.7) that corresponding to the prescribed data $y := (y_j : j \in \mathbb{Z}_m)$ there is at least one $x \in \mathcal{X}$ for which the linear constraints in (2.7) are satisfied. In addition, we may assume that the linear functionals $\{L_j : j \in \mathbb{Z}_m\}$ are *linearly independent*. This means that whenever $a := (a_j : j \in \mathbb{Z}_m)$ is such that $\sum_{j\in\mathbb{Z}_m} a_j L_j = 0$ then $a = 0$. Otherwise, we can "thin" the set of linear functionals to a linearly independent set.

We say that the linear functional $L \in \mathcal{X}^*\backslash\{0\}$ *peaks* at $x \in \mathcal{X}\backslash\{0\}$, if $L(x) = \|L\|\|x\|$. Let us also say that x peaks at L, if L peaks at x. A consequence of the Hahn–Banach Theorem, see for example [21, p. 223], is that for *every* $x \in \mathcal{X}$ there always exists an $L \in \mathcal{X}^*$ which peaks at x and so, $\|x\| = \max\{L(x) : \|L\| \le 1, L \in \mathcal{X}^*\}$, see [21, p. 226, Prop. 6]. On the other hand, the supremum in the definition of $\|L\|$ is not always achieved, unless L peaks at some $x \in \mathcal{X}\backslash\{0\}$. We also recall that $\mathcal{X}$ is *weakly compact* if, for every *norm bounded* sequence $\{x_n : n \in \mathbb{Z}_+\} \subset \mathcal{X}$ there exists a *weakly* convergent subsequence $\{x'_n : n \in \mathbb{N}\}$, that is, there is an $x \in \mathcal{X}$ such that for every $L \in \mathcal{X}^*$ $\lim_{n\to\infty} L(x'_n) = L(x)$. When $\mathcal{X}$ is weakly compact then for *every* $L \in \mathcal{X}^*$ there is always an $x \in \mathcal{X}$ which peaks at L. Recall that a Banach space $\mathcal{X}$ is reflexive, that is, $(\mathcal{X}^*)^* = \mathcal{X}$ if and only if $\mathcal{X}$ is weakly compact, see [16, p. 127, Thm. 3.6] and it is known that any weakly compact normed linear spaces always admit a minimal norm interpolant.

If $\mathcal{M}$ is a closed subspace of $\mathcal{X}$, we define the distance of x to $\mathcal{M}$ as

$$d(x, \mathcal{M}) := \min\{\|x - t\| : t \in \mathcal{M}\}.$$

In particular, if we choose $\mathcal{M}_0 := \{x : x \in \mathcal{X}, L_j(x) = 0, j \in \mathbb{Z}_m\}$ and *any* $w \in \mathcal{X}$ such that $L_j(w) = y_j, j \in \mathbb{Z}_m$ then we have that

$$d(w, \mathcal{M}_0) = \mu. \tag{2.8}$$

Theorem 1. *x_0 is a solution of (2.7) if and only if $L_j(x_0) = y_j, j \in \mathbb{Z}_m$ and there exists $(c_j : j \in \mathbb{Z}_m) \in \mathbb{R}^m$ such that the linear functional $\sum_{j\in\mathbb{Z}_m} c_j L_j$ peaks at x_0.*

Proof. We choose in (2.8) $w = x_0$ so that $L_j(x_0) = y_j, j \in \mathbb{Z}_m$ and $\|x_0\| = d(x_0, \mathcal{M}_0)$. Using the basic duality principle for the distance (2.8), see for example [8], we conclude that

$$\|x_0\| = \max\{L(x_0) : L(x) = 0,\ x \in \mathcal{M}_0, \|L\| \le 1\}. \tag{2.9}$$

However, L vanishes on $\mathcal{M}_0$ if and only if there exists $(c_j : j \in \mathbb{Z}_m) \in \mathbb{R}^m$ such that $L = \sum_{j\in\mathbb{Z}_m} c_j L_j$ and by (2.9) there is such an L which peaks at x_0.

On the other hand, if for some $(c_j : j \in \mathbb{Z}_m) \in \mathbb{R}^m$ the linear functional $\sum_{j\in\mathbb{Z}_m} c_j L_j$ peaks at x_0 with $L_j(x_0) = y_j, j \in \mathbb{Z}_m$ we have, for every $t \in \mathcal{M}_0$, that

$$\left\|\sum_{j\in\mathbb{Z}_m} c_j L_j\right\| \|x_0\| = \sum_{j\in\mathbb{Z}_m} c_j L_j(x_0) = \sum_{j\in\mathbb{Z}_m} c_j L_j(x_0+t) \le \|x_0+t\| \left\|\sum_{j\in\mathbb{Z}_m} c_j L_j\right\|$$

and so, x_0 is a minimal norm interpolant. □

This theorem tells us if x_0 solves the MNI problem then there exists $\{c_j, j \in N_m\} \subset \mathbb{R}$ such that $\|x_0\| = \sum_{j\in\mathbb{Z}_m} c_j L_j(x_0) = \sum_{j\in\mathbb{Z}_m} c_j y_j$. How do we find the the parameters $\{c_j : j \in \mathbb{Z}_m\}$? This is described next.

Theorem 2. *If $\mathcal{X}$ be a Banach space then*

$$\min\left\{\left\|\sum_{j\in\mathbb{Z}_m} c_j L_j\right\| : \sum_{j\in\mathbb{Z}_m} c_j y_j = 1\right\} = 1/\mu. \tag{2.10}$$

In addition, if $\mathcal{X}$ is weakly compact and $\hat{c}$ is the solution to (2.10) then there exists $\hat{x} \in \mathcal{X}$ such that $\|\hat{x}\| = 1$, $L_j(\hat{x}) = y_j/\mu, j \in \mathbb{Z}_m$ and $\sum_{j\in\mathbb{Z}_m} \hat{c}_j L_j(\hat{x}) = \|\sum_{j\in\mathbb{Z}_m} \hat{c}_j L_j\|$.

Proof. Since the function $H : \mathbb{R}^m \to \mathbb{R}_+$ defined for each $c = (c_j : j \in \mathbb{Z}_m)$ by $H(c) := \|\sum_{j\in\mathbb{Z}_m} c_j L_j\|$ is continuous, homogeneous and nonzero for $c \neq 0$, it tends to infinity as $c \to \infty$, so the minimum in (2.10) exists. The proof of (2.10) is transparent from our remarks in Theorem 1. Indeed, for every $w \in \mathcal{X}$ such that $L_j(w) = y_j,\ j \in \mathbb{Z}_m$ we have that $\mu = d(w, \mathcal{M}_0)$ and

$$\mu = \max\{L(w) : L(x) = 0, x \in \mathcal{M}_0, \|L\| \leq 1\}.$$

Moreover, since L vanishes on $\mathcal{M}_0$ if and only if $L = \sum_{j\in\mathbb{Z}_m} c_j L_j$ for some $c = (c_j : j \in \mathbb{Z}_m)$, the right hand side of this equation becomes

$$\max\left\{\sum_{j\in\mathbb{Z}_m} c_j y_j : \left\|\sum_{j\in\mathbb{Z}_m} c_j L_j\right\| \leq 1\right\} = \left(\min\left\{\left\|\sum_{j\in\mathbb{Z}_m} c_j L_j\right\| : \sum_{j\in\mathbb{Z}_m} c_j y_j = 1\right\}\right)^{-1}$$

from which equation (2.10) follows.

For vectors $c = (c_j : j \in \mathbb{Z}_m)$, $d = (d_j : j \in \mathbb{Z}_m)$ in $\mathbb{R}^m$, we let $c \cdot d = \sum_{j\in\mathbb{Z}_m} c_j d_j$, the standard inner product on $\mathbb{R}^m$. Let $\hat{c} := (\hat{c}_j : j \in \mathbb{Z}_m)$ be a solution to the minimization problem (2.10) and consider the linear functional

$$\hat{L} := \sum_{j\in\mathbb{Z}_m} \hat{c}_j L_j.$$

This solution is characterized by the fact that the right directional derivative of the function H at $\hat{c}$ along *any* vector $a = (a_j : j \in \mathbb{Z}_m)$ perpendicular to y is nonnegative. That is, we have that

$$H'(\hat{c}; a) := \lim_{\lambda\to 0^+} \frac{H(\hat{c} + \lambda a) - H(\hat{c})}{\lambda} \geq 0 \tag{2.11}$$

when $a \cdot y = 0$. This derivative can be computed to be

$$H'(\hat{c}; a) = \max\left\{\sum_{j=1}^{m} a_j L_j(x) : \|x\| \leq 1\right\} \tag{2.12}$$

see [13]. We introduce the convex and the compact set $\mathcal{C} := \{(L_j(x) : j \in \mathbb{Z}_m) : \|x\| \leq 1\} \subset \mathbb{R}^m$. If a is perpendicular to y then, by the inequality (2.11) and the formula (2.12), we have that

$$\max\{a \cdot v : v \in \mathcal{C}\} \geq 0. \tag{2.13}$$

We shall now prove that the line $\mathcal{L} := \{\lambda y : \lambda \in \mathbb{R}\}$ intersects $\mathcal{C}$. Suppose to the contrary that it does not. So, there exists an hyperplane $\{z : u \cdot z + \tau = 0\}$ where $u \in \mathbb{R}^m$ and $\tau \in \mathbb{R}$, which separates these sets, that is

$$(i)\ \ u \cdot z + \tau > 0,\ z \in \mathcal{L}, \qquad (ii)\ \ u \cdot z + \tau \leq 0,\ z \in \mathcal{C}$$

see [21]. From condition (i) we conclude that u is perpendicular to y and $\tau > 0$ while (ii) implies that $\max\{u \cdot v : v \in \mathcal{C}\} < 0$. This is in contradiction to (2.13). Hence, there is an $\hat{x}$ such that $L_j(\hat{x}) = y_j/\mu, j \in \mathbb{Z}_m$, $\hat{L}(\hat{x}) = \|\hat{L}\|$ and $\|\hat{x}\| = 1$. Therefore, it must be that $x_0 := \mu\hat{x}$ is a MNS. □

This theorem leads us to a method to identify the MNS in a reflexive smooth Banach space $\mathcal{X}$. Recall that a reflexive Banach space $\mathcal{X}$ is *smooth* provided that for every $L \in \mathcal{X}^* \backslash \{0\}$ there is *unique* $x_L \in \mathcal{X}$ which peaks at L.

Corollary 1. *If $\mathcal{X}$ is a smooth reflexive Banach space, $\hat{L} := \sum_{j \in \mathbb{Z}_m} \hat{c}_j L_j$ is the solution to (2.10) and $\hat{L}$ peaks at x_L with $\|x_L\| = 1$ then $x_0 := \mu x_L$ is the unique solution to (2.7) and $\mu = 1/\|\hat{L}\|$.*

We wish to note some important examples of the above results. The first to consider is naturally a Hilbert space $\mathcal{X}$. In this case $\mathcal{X}$ is reflexive and $\mathcal{X}^*$ can be identified with $\mathcal{X}$, that is, for each $L_j \in \mathcal{X}^*$, there is a *unique* $x^j \in \mathcal{X}$ such that $L_j(x) = (x^j, x), x \in \mathcal{X}$. Thus, $\hat{x} = \sum_{j \in \mathbb{Z}_m} c_j x^j$ solves the dual problem when $(x^j, x) = \lambda y_j, j \in \mathbb{Z}_m$, $\lambda = \|\hat{x}\|^2$ and $x_0 = \hat{x}/\|\hat{x}\|$ is the minimal norm solution.

The Hilbert space case does not show the value of function composition appearing in (1.2). A better place to reveal this is in the context of Orlicz spaces. The theory of such spaces is discussed in several books, see e.g [17, 20], and minimal norm interpolation is studied in [3]. We review these ideas in the context of Corollary 1. Let $\omega : [0, \infty) \to [0, \infty)$ be a convex and continuously differentiable function on $[0, \infty)$ such that $\lim_{s \to \infty} \omega'(s) = \infty$ and $\omega(0) = \omega'_+(0) = 0$ where ω'_+ is the right derivative of ω. Such a function is sometimes known as a Young function. We will also assume that the function $s \mapsto s\omega'(s/\omega(s)),\ s \in [0, \infty)$ is bounded on $[k, \infty)$ for some $k \in [0, \infty)$. Let $(D, \mathcal{B}, \mu)$ be a finite measure space, see [21, p. 286], $L^0(\mu)$ the space of measurable functions $f : D \to \mathbb{R}$, and denote by $\mathcal{L}_\omega$ the convex hull of the set

$$\left\{ f \in L^0(\mu) : \int_D \omega(|f(t)|) d\mu(t) < \infty \right\}.$$

The space $\mathcal{L}_\omega$ can be made into a normed space by introducing, for every $f \in \mathcal{L}_\omega$ the norm

$$\|f\|_\omega := \inf \left\{ \lambda \geq 0 : \int_D \omega\left(\frac{|f(t)|}{\lambda}\right) d\mu(t) \leq 1 \right\}.$$

The dual of $\mathcal{L}_\omega$ is the space $\mathcal{L}_{\omega^*}$ where ω^* is the complementary function of ω which is given by the formula

$$\omega^*(s) = \int_0^s (\omega')^{-1}(\xi)d\xi, \ s \in [0,\infty).$$

For every $f \in \mathcal{L}_\omega$ and $g \in \mathcal{L}_{\omega^*}$ there also holds the Orlicz inequality

$$|(f,g)| \leq \|f\|_\omega \|g\|_{\omega^*}$$

where we have defined $(f,g) := \int_D f(t)g(t)d\mu(t)$. The Orlicz inequality becomes an equality if and only if

$$f = \lambda(\omega^*)'(|g|)\text{sign}(g), \tag{2.14}$$

for some $\lambda \in \mathbb{R}$. This means that the linear functional represented by $g \in \mathcal{L}_{\omega^*}$ peaks at f if and only if f satisfies equation (2.14). Moreover, under the above conditions on ω, $\mathcal{L}_\omega$ is reflexive and smooth. Thus the hypothesis of Corollary 1 is satisfied and we conclude that the unique solution to (2.7) is given by $f = \lambda\phi_\omega(\sum_{j\in\mathbb{Z}_m} c_j g_j)$ where ϕ_ω is defined for $t \in \mathbb{R}$ as

$$\phi_\omega(t) = (\omega^*)'(|t|)\text{sign}(t) \tag{2.15}$$

and the coefficients $\lambda, \ c_j, j \in \mathbb{Z}_m$ solve the system of *nonlinear* equations $(f, g_j) = y_j, \ j \in \mathbb{Z}_m$.

As a special case consider the choice $\omega(s) = s^p/p, p > 1, \ s \in [0,\infty)$. In this case $\mathcal{L}_\omega = \mathcal{L}^p$, the space of functions whose p power is integrable, and the dual space is $\mathcal{L}^q$ where $1/p + 1/q = 1$, [21]. Since $\omega^*(s) = s^q/q, \ s \in [0,\infty)$, the solution to equations (2.5) and (2.7) has the form $f = \lambda\phi_q(\sum_{j\in\mathbb{Z}_m} \hat{c}_j g_j)$ where for all $t \in \mathbb{R}$ ϕ_q is defined by the equation

$$\phi_q(t) := |t|^{q-1}\text{sign}(t). \tag{2.16}$$

3 Learning all continuous functions: density

An important feature of any learning algorithm is its ability to enhance accuracy by increasing the *number of parameters* in the model. Below we present a sufficient condition on the functions ϕ and $\{g_j : j \in \mathbb{Z}_m\}$ so that the functions in (1.4) can approximate any continuous real–valued function within any given tolerance on a compact set $D \subseteq \mathbb{R}^d$. For related material see [19]. Let us formulate our observation.

We use $\mathcal{C}(D)$ for the space of all continuous functions on the set D and for any $f \in \mathcal{C}(D)$ we set $\|f\|_D := \max\{|f(x)| : x \in D)\}$. For any subset $\mathcal{T}$ of $\mathcal{C}(D)$ we use $span(\mathcal{T})$ to denote the smallest *closed* linear subspace of $\mathcal{C}(D)$ containing $\mathcal{T}$. We enumerate vectors in $\mathbb{R}^m$ by superscripts and use $g := (g_j : j \in \mathbb{Z}_m)$ for the vector–valued map $g : D \to \mathbb{R}^m$ whose coordinates are built from the functions in $\mathcal{G} := \{g_j : j \in \mathbb{Z}_m\}$. This allows us to write the functions in (1.4) as

$$\sum_{j\in\mathbb{Z}_n} a_j \phi(c^j \cdot g). \tag{3.17}$$

For any two subsets $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{C}(D)$ we use $\mathcal{A}\cdot\mathcal{B}$ for the set defined by $\mathcal{A}\cdot\mathcal{B} := \{fg : f \in \mathcal{A}, g \in \mathcal{B}\}$ and, for every $k \in \mathbb{N}$, $\mathcal{A}^k$ denotes the set $\{f^k : f \in \mathcal{A}\}$. Given any $\phi \in \mathcal{C}(D)$ we let $\mathcal{M}(\phi)$ be the smallest closed linear subspace containing all the functions (3.17). Note that m is fixed while $\mathcal{M}(\phi)$ contains all the functions (3.17) for *any* n. We use $\mathcal{A}_\mathcal{G}$ for the smallest subalgebra in $\mathcal{C}(D)$ which contains $\mathcal{G}$, that is, the direct sum $\oplus_{k\in\mathbb{N}}\mathcal{G}^k$. We seek conditions on ϕ and g so that $\mathcal{M}(\phi) = \mathcal{C}(D)$ and we prepare for our observation with two lemmas.

Lemma 1. *If $\phi \in \mathcal{C}(D)\backslash\{0\}$ and $1 \in span(\mathcal{G})$ then $1 \in \mathcal{M}(\phi)$.*

Proof. By hypothesis, there is a $t \in \mathbb{R}$ such that $\phi(t) \neq 0$ and a $c \in \mathbb{R}^m$ such that $c\cdot g = t$. Hence we have that $1 = \frac{1}{\phi(t)}\phi(c\cdot g) \in \mathcal{M}(\phi)$. □

Lemma 2. *If $\phi' \in \mathcal{C}(D)$ then $\mathcal{M}(\phi')\cdot\mathcal{G} \subseteq \mathcal{M}(\phi)$.*

Proof. We choose any function f of the form

$$f = \sum_{j\in\mathbb{Z}_n} a_j\phi'(c^j\cdot g)$$

where $a = (a_j : j \in \mathbb{Z}_n) \in \mathbb{R}^n$ and $\{c^j : j \in \mathbb{Z}_n\} \subset \mathbb{R}^m$. For any $d \in \mathbb{R}^m$ we define the function $q = d\cdot g$. Let us show that $f\cdot q \in \mathcal{M}(\phi)$. To this end, we define for $t \in \mathbb{R}$ the function

$$h_t := \sum_{j\in\mathbb{Z}_n} a_j\phi((c^j + td)\cdot g)$$

and observe that $\lim_{t\to 0} t^{-1}(h_t - h_0) = f\cdot q$. Since $\{h_t - h_0 : t \in \mathbb{R}\} \subseteq \mathcal{M}(\phi)$, the result follows. □

We say that $\mathcal{G}$ separates points on $\mathcal{D}$ when the map $g : D \to \mathbb{R}^m$ is injective. Recall that an algebra $\mathcal{A} \subseteq C(D)$ *separates points* provided for each pair of distinct points x and $y \in D$ there is an $f \in \mathcal{A}$ such that $f(x) \neq f(y)$.

Theorem 3. *If $\phi \in \mathcal{C}^\infty(\mathbb{R})$, ϕ is not a polynomial, $1 \in span(\mathcal{G})$ and $\mathcal{G}$ separates points then $\mathcal{M}(\phi) = \mathcal{C}(D)$.*

Proof. Our hypothesis implies that $\mathcal{A}_\mathcal{G}$ separates points and contains constants. Hence, the Stone–Weierstrass Theorem, see for example [21], implies that the algebra $\mathcal{A}_\mathcal{G}$ is dense in $\mathcal{C}(D)$. Thus, the result will follow as soon as we show that $\mathcal{A}_\mathcal{G} \subseteq \mathcal{M}(\phi)$. Since $\phi \in \mathcal{C}^\infty(\mathbb{R})$ Lemma 2 implies for any positive integer k that

$$\mathcal{M}(\phi^{(k)})\cdot\mathcal{G}^k \subseteq \mathcal{M}(\phi).$$

Using Lemma 1 and the fact that ϕ is not a polynomial the above inclusion implies that $\mathcal{G}^k \subseteq \mathcal{M}(\phi)$. Consequently, we conclude that

$$\mathcal{A}_\mathcal{G} = \bigoplus_{k\in\mathbb{N}} \mathcal{G}^k \subseteq \mathcal{M}(\phi).$$

□

We remark that the idea for the proof of Lemma 2 is borrowed from [4] where only the case that $span(\mathcal{G})$ is linear functions on $\mathbb{R}^{m-1}$ and D is a subset of $\mathbb{R}^m$ is treated. We also recommend [12] for a Fourier analysis approach to density and [10] which may allow for the removable of our hypothesis that $\phi \in \mathcal{C}^\infty(D)$.

In Theorem 3 above m is fixed and we enhance approximation of an arbitrary function by functions of the special type (1.4) by adjusting n. Next, we provide another density result where m is allowed to vary, but in this case, g is chosen in a specific fashion from the reproducing kernel of a Hilbert space $\mathcal{H}$ of real–valued functions on D contained in $\mathcal{C}(D)$. Indeed, let K be the reproducing kernel for $\mathcal{H}$ which is jointly continuous on $D \times D$. There are useful cases when $\mathcal{H}$ is endowed with a *semi–norm*, that is, there are nontrivial functions in $\mathcal{H}$ with norm zero, see e.g [25]. To ensure that these cases are covered by our results below we specify a finite number of functions $\{k_j : j \in \mathbb{Z}_r\}$ and consider functions of the form

$$\sum_{j \in \mathbb{Z}_m} c_j K(\cdot, x_j) + \sum_{j \in \mathbb{Z}_r} c_{j+m} k_j. \tag{3.18}$$

We use $\mathcal{K}$ for the smallest closed linear subspace of $\mathcal{C}(D)$ which contains all the functions in (3.18) for any m and $c = (c_j : j \in \mathbb{Z}_{m+r}) \in \mathbb{R}^{m+r}$. Here the samples $D_m := \{x_j : j \in \mathbb{Z}_m\}$ are chosen in D and, in the spirit of our previous discussion we compose the function in (3.18) with a function ϕ to obtain functions of the form

$$\phi\Big(\sum_{j \in \mathbb{Z}_m} c_j K(\cdot, x_j) + \sum_{j \in \mathbb{Z}_r} c_{j+m} k_j\Big).$$

We write this function as $\phi(c \cdot w)$ where $c \in \mathbb{R}^{m+r}$ and the coordinates of the vector map $w : D \to \mathbb{R}^{m+r}$ are defined as $w_j = K(\cdot, x_j), j \in \mathbb{Z}_m$ and $w_{j+m} = k_j, j \in \mathbb{Z}_r$. We let $\mathcal{K}(\phi)$ be the smallest closed linear subspace containing all these functions. Our next result provides a sufficient condition on ϕ and w such that $\mathcal{K}(\phi)$ is dense in $\mathcal{C}(D)$. To this end we write K in the "Mercer form"

$$K(x, y) = \sum_{\ell \in \mathbb{Z}_+} \lambda_\ell \phi_\ell(x) \phi_\ell(y), \ x, y \in D \tag{3.19}$$

where we may as well assume that $\lambda_\ell \neq 0$ for all $\ell \in \mathbb{Z}_+$. Here, we demand that $\{\phi_\ell : \ell \in \mathbb{Z}_+\} \subseteq \mathcal{C}(D)$ and we require the series above converges *uniformly* on $D \times D$. We also require that the set $J = \{\ell : \lambda_\ell < 0\}$ has the property that

$$\{\phi_\ell : \ell \in J\} \subseteq span\{k_j : j \in \mathbb{Z}_r\} \tag{3.20}$$

and that $\mathcal{U} := span\{\phi_\ell : \ell \in \mathbb{Z}_+\} = \mathcal{C}(D)$. When these conditions holds we call K *acceptable.*

Theorem 4. *If K is acceptable, $1 \in \mathcal{K}(\phi')$ and $\phi' \in \mathcal{C}(D)\backslash\{0\}$ then $\mathcal{K}(\phi) = \mathcal{C}(D)$.*

Proof. We establish this fact by showing that there is no *nontrivial* linear functional L which has the property that

$$L(g) = 0 \tag{3.21}$$

for every $g \in \mathcal{K}(\phi)$, see for example [21]. Let c and w be as above. We choose $b \in \mathbb{R}$, $y \in D$ and $g = \phi(c \cdot w + bK(\cdot, y))$. Now, differentiate both sides of equation (3.21) with respect to b and evaluate the resulting equation at $b = 0$ to obtain the equation

$$L(\phi'(c \cdot w)K(\cdot, y)) = 0, \ y \in D. \tag{3.22}$$

On the other hand, differentiating (3.21) with respect to c_{j+m}, $j \in \mathbb{Z}_r$ gives the equation

$$L(\phi'(c \cdot w)k_\ell) = 0, \ \ell \in \mathbb{Z}_r. \tag{3.23}$$

We shall use these equations in a moment. First, we observe that by hypothesis there exists a $t \in \mathbb{R}$ such that $\phi'(t) \neq 0$ and for every $\epsilon > 0$ there exists $f \in \mathcal{K}(\phi')$ given, for some $m \in \mathbb{N}$, $\{a_j : j \in \mathbb{Z}_n\} \subset \mathbb{R}$, $\{d_j : j \in \mathbb{Z}_n\} \subset \mathbb{R}^m$, by the formula

$$f = \sum_{j \in \mathbb{Z}_n} a_j \phi'(d^j \cdot w) \tag{3.24}$$

such that $|\phi'(t) - f| \leq \epsilon$ on D. We now evaluate the equations (3.22) and (3.23) at $c = d^j, j \in \mathbb{Z}_n$ and combine the resulting equations to obtain

$$L(fK(\cdot, y)) = 0, \ y \in D, \qquad L(fk_\ell) = 0, \ \ell \in \mathbb{Z}_r.$$

We let M be a constant chosen big enough so that for all x and $y \in D$, $|K(x, y)| \leq M$, and $|k_\ell(x)| \leq M, \ell \in \mathbb{Z}_r$. We rewrite (3.22) in the form

$$0 = L((f - \phi'(t))K(\cdot, y)) + \phi'(t)L(K(\cdot, y))$$

from which we obtain the inequalities

$$|\phi'(t)L(K(\cdot, y))| \leq \epsilon \|L\| M, \ y \in D, \qquad |\phi'(t)L(k_\ell)| \leq \epsilon \|L\| M, \ \ell \in \mathbb{Z}_r.$$

Since ϵ is arbitrary we conclude for all $y \in D$ that $L((K(\cdot, y)) = 0, \ y \in D$ and $L(k_\ell) = 0, \ \ell \in \mathbb{Z}_r$. Thus, using the Mercer representation for K we conclude, for all $y \in D$, that

$$\sum_{j \notin J} \lambda_j \phi_j(y) L(\phi_j) = 0. \tag{3.25}$$

Next, we apply L to both sides of (3.25) and obtain that $\sum_{j \notin J} \lambda_j |L(\phi_j)|^2 = 0$ which implies that $L(\phi_j) = 0$, $j \in \mathbb{Z}_+$. However, since $span\{\phi_j : j \in \mathbb{Z}_+\} = \mathcal{C}(D)$, it follows that $L = 0$, which proves the result. □

We remark that the proof of this theorem yields for any $f \in \mathcal{C}(D)$ the fact that

$$d(f, \mathcal{K}(\phi)) \leq d(f, \mathcal{K}) = d(f, \mathcal{U}).$$

Note that if $\phi(t) = t$ the hypothesis that $1 \in \mathcal{K}(\phi')$ is automatically satisfied. We provide another sufficient condition for this requirement to hold.

Lemma 3. *If* $1 \in \mathcal{K}$ *and* $\phi \in \mathcal{C}(\mathbb{R})\backslash\{0\}$ *then* $1 \in \mathcal{K}(\phi)$.

Proof. We choose some $t \in \mathbb{R}$ such that $\phi(t) \neq 0$ and some $\epsilon > 0$. There is a $\delta > 0$ such that whenever $|t - s| \leq \epsilon,\ s \in \mathbb{R}$ it follows that $|\phi(t) - \phi(s)| \leq \epsilon$. Since $1 \in \mathcal{K}$, there is a $d \in \mathbb{R}^{m+r}$ and $D_m \subset \mathcal{D}$ so that $|t - d \cdot w| \leq \delta$ uniformly on D. Hence it follows that $|\phi(t) - \phi(d \cdot w)| \leq \epsilon$ uniformly on D which proves the result. □

As an example of the theorem above we choose $D = [-\pi, \pi]^d,\ d \in \mathbb{N},\ \phi(t) = t,\ t \in \mathbb{R}$, K a $2\pi-$periodic translation kernel, that is, $K(x, y) = h(x - y),\ x, y \in D$, where $h : [-\pi, \pi]^d \to \mathbb{R}$ is even, continuous, and $2\pi-$periodic, and $r = 0$. To ensure that K is a reproducing kernel we assume h has a uniformly convergent Fourier series,

$$h(x) = \sum_{n \in \mathbb{Z}_+^d} a_n \cos(n \cdot x), \quad x \in \mathbb{R}^d \tag{3.26}$$

where $a_n \geq 0,\ n \in \mathbb{Z}_+^d$. In this case we have the Mercer representation for K

$$K(x, y) = \sum_{n \in \mathbb{Z}_+^d} a_n \sin(n \cdot x) \sin(n \cdot y) + \sum_{n \in \mathbb{Z}_+^d} a_n \cos(n \cdot x) \cos(n \cdot y), \quad x, y \in \mathbb{R}^d$$

In addition, if $a_n > 0$ for all $n \in \mathbb{Z}_+^d$, the functions appearing in this representation are dense in the $2\pi-$periodic functions in $\mathcal{C}(D)$, we conclude that $\mathcal{K}$ is dense in $\mathcal{C}(D)$ as well.

We remark that the method of proof of Theorem 4 can be extended to other function spaces, for instance $\mathcal{L}^p$ spaces. This would require that (3.19) holds relative to the convergence in that space and that the set of functions $\{\phi_n : n \in \mathbb{Z}_+\}$ are dense in it.

4 Learning any set of finite data: interpolation

In this section we discuss the possibility of adjusting the parameters in our model (1.4) to satisfy some prescribed linear constraints. This is a complex issue as it leads to the problem of solving *nonlinear* equations. Our observations, although incomplete, provide some instances in which this may be accomplished as well as an algorithm which may be useful to accomplish this goal. Let us first describe our setup. We start with the function

$$f := \sum_{j \in \mathbb{Z}_n} a_j \phi(c^j \cdot g)$$

where $\{a_j : j \in \mathbb{Z}_n\} \subset \mathbb{R}$ and $\{c^j : j \in \mathbb{Z}_n\} \subset \mathbb{R}^m$ are to be specified by some linear constraint. The totality of scalar parameters in this representation is $n(m + 1)$. To use these parameters we suppose there is available data vectors $\{y^j : j \in \mathbb{Z}_n\} \subset \mathbb{R}^m$ and linear operators $L^s : \mathcal{C}(D) \to \mathbb{R}^m, s \in \mathbb{Z}_n$ that lead to the nonlinear equations

$$\sum_{j \in \mathbb{Z}_n} a_j L^s(\phi(c^j \cdot g)) = y^s,\ s \in \mathbb{Z}_n. \tag{4.27}$$

There are mn scalar equations here and the remaining degrees of freedom will be used to specify the Euclidean norm of the vectors $c^j, j \in \mathbb{Z}_n$. We shall explain this in a moment. It is convenient to introduce for each $s \in \mathbb{Z}_m$ the operator $B_s : \mathbb{R}^m \to \mathbb{R}^m$ defined for any $c \in \mathbb{R}^m$ by the equation

$$B_s(c) = L^s(\phi(c \cdot g)), \ s \in \mathbb{Z}_n. \tag{4.28}$$

Therefore, the equations (4.27) take the form

$$\sum_{j \in \mathbb{Z}_n} a_j B_s(c^j) = y^s, \ s \in \mathbb{Z}_n. \tag{4.29}$$

Our first result covers the case $n = 1$.

Theorem 5. *If $\phi \in \mathcal{C}(\mathbb{R})$ is an odd function and B_0 only vanishes on $\mathbb{R}^m$ at 0 then for any $y^0 \in \mathbb{R}^m$ and $r > 0$ there is a $c^0 \in \mathbb{R}^m$ with $c^0 \cdot c^0 = r_0^2$ and $a_0 \in \mathbb{R}$ such that $a_0 B_0(c^0) = y^0$.*

Proof. We choose linearly independent vectors $\{w^j : j \in \mathbb{Z}_{m-1}\} \subset \mathbb{R}^m$ perpendicular to y^0 and construct the map $H : \mathbb{R}^m \to \mathbb{R}^{m-1}$ by setting for $c \in \mathbb{R}^m$

$$H(c) := (w^j \cdot B_0(c) : j \in \mathbb{Z}_{m-1}).$$

We restrict H to the sphere $c \cdot c = r_0$. Since H is an odd continuous map by the Borsuk antipodal mapping theorem, see for example [11], there is a $c^0 \in \mathbb{R}^m$ with $c^0 \cdot c^0 = r_0^2$ such that $H(c^0) = 0$. Hence, $B_0(c^0) = uy^0$ for some scalar $u \in \mathbb{R}$. Since B_0 vanishes only at the origin we have that $u \neq 0$ and, so, setting $a_0 = u^{-1}$ proves the result. □

We remark that the above theorem extends our observation in (2.16). Indeed, if we choose $\phi := \phi_q$ and use the linear operator $L^0 : \mathcal{L}^p \to \mathbb{R}^m$ defined for each $f \in \mathcal{L}^p$ as $L^0(f) := ((f, g_j) : j \in \mathbb{Z}_m)$, then the above result reduces to (2.16). However, note that Theorem 5 even in this special case is *not* proven by the analysis of a variational problem.

We use Theorem 5 to propose an iterative method to solve the system of equations (4.29). We begin with an initial guess $a^0 = (a_j^0 : j \in \mathbb{Z}_n)$ and vectors $\{c^{j,0} : j \in \mathbb{Z}_n\}$ with $c^{j,0} \cdot c^{j,0} = r_j, j \in \mathbb{Z}_n$. We now update these parameters by explaining how to construct $a^1 = (a_j^1 : j \in \mathbb{Z}_n)$ and vectors $\{c^{j,1} : j \in \mathbb{Z}_n\}$. First, we define a_0^1 and $c^{0,1}$ by solving the equation

$$a_0^1 B_0(c^{0,1}) + \sum_{j \in \mathbb{Z}_{n-1}} a_{j+1}^0 B_0(c^{j+1,0}) = y^0.$$

whose solution is assured by Theorem 5. Now, suppose we have found $a_0^1, \ldots, a_{r-1}^1$, $c^{0,1}, \ldots, c^{r-1,1}$ for some integer $1 \leq r < n-1$. We then solve the equation

$$\sum_{j \in \mathbb{Z}_{r+1}} a_j^1 B_r(c^{j,1}) + \sum_{j \in \mathbb{Z}_{n-r-1}} a_{j+r+1}^0 B_r(c^{j+r+1,0}) = y^r$$

for a_r^1 and $c^{r,1}$ until we reach $r = n-1$. In this manner, we construct a sequence of vectors $a^k \in \mathbb{R}^n$ and $c^{j,k} \in \mathbb{R}^m, k \in \mathbb{Z}_+, j \in \mathbb{Z}_n$ such that for all $k \in \mathbb{Z}_+$ and $r \in \mathbb{Z}_n$,

$$\sum_{j \in \mathbb{Z}_{r+1}} a_j^{k+1} B_r(c^{j,k+1}) + \sum_{j \in \mathbb{Z}_{n-r-1}} a_{j+r+1}^{k+1} B_r(c^{j+r+1,k}) = y^r. \tag{4.30}$$

We do not know whether or not this iterative method converges in the generality presented. However, below we provide a sufficient condition for which the sequences generated above remain bounded.

Corollary 2. *If there is an $s \in \mathbb{Z}_n$ such that whenever $\{c^j : j \in \mathbb{Z}_n\} \subset \mathbb{R}^m$, $b = (b_j : j \in \mathbb{Z}_n) \in \mathbb{R}^n$ with $c^j \cdot c^j > 0, j \in \mathbb{Z}_n$ and*

$$\sum_{j \in \mathbb{Z}_n} b_j B_s(c^j) = 0 \tag{4.31}$$

it follows that $b = 0$, then the sequence $\{a_j^k : j \in \mathbb{Z}_n\}$ defined in (4.28) is bounded.

Proof. Without loss of generality we assume, by reordering the equations, that $s = n-1$. The last equation in (4.30), corresponding to $r = n-1$, allows us to observe that the coefficients $\{a_j^{k+1} : j \in \mathbb{Z}_n\}$ remain bounded during the updating procedure. To confirm this, we set $\gamma_k = \sum_{j \in \mathbb{Z}_n} |a_j^{k+1}|$ and divide both sides of (4.30) by γ_k. If the sequence $\{a_j^{k+1} : k \in \mathbb{N}\}$ is not bounded we obtain, in the limit as $k \to \infty$ through a subsequence, that

$$\sum_{j \in \mathbb{Z}_n} \tilde{a}_j B_s(\tilde{c}^j) = 0 \tag{4.32}$$

where the constants $\tilde{a}_j$, $j \in \mathbb{Z}_n$ satisfy $\sum_{j \in \mathbb{Z}_n} |\tilde{a}_j| = 1$, which in contradiction with our hypothesis. □

5 Discussion

We have proposed a framework for learning in a Banach space and establish a representation theorem for the solution of regularization–based learning algorithms. This naturally extends the representation theorem in Hilbert spaces which is central in developing kernel–based methods. The framework builds on a link between regularization and minimal norm interpolation, a key concept in function estimation and interpolation. For concrete Banach spaces such as Orlicz spaces, our result leads to the functional representation (1.2). We have studied the density property of this functional representation and its extension.

There are important directions that should be explored in the context presented in this paper. First, it would be valuable to extend on–line and batch

learning algorithms which have already been studies for finite dimensional Banach spaces (see e.g. [1, 9, 26]) within the general framework discussed here.

For example, in [14] we consider the hinge loss function used in support vector machines and an appropriate H to identify the dual of the minimization problem (1.3) and report of our numerical experience with it.

Second, it would be interesting to study error bounds for learning in Banach spaces. This study will involve both the sample as well the approximation error, and should uncover advantage or disadvantages of learning in Banach spaces in comparison to Hilbert spaces which are not yet understood.

Finally, we believe that the framework presented here remains valid when problems (2.5) and (2.7) are studied subject to additional *convex constraints*. These may be available in form of prior knowledge on the function we seek to learn. Indeed constrained minimal norm interpolation has been studied in Hilbert spaces, see [15] and [5] for a review. It would be interesting to extend these idea to regularization in Banach spaces. As an example, consider the problem of learning a *nonnegative* function f in the Hilbert space $\mathcal{H} := \mathcal{L}^2(D)$ from the data $\{y_j = \int_D f(t)g_j(t)dt : j \in \mathbb{Z}_m\}$. Then, any minimizer of the regularization functional of the form (1.3) in $\mathcal{H}$ (where $L_j(f) := \int_D f(t)g_j(t)dt$) subject to the additional nonnegativity constraint, has the form in equation (1.2) where $\phi(t) = \max(t, 0)$, $t \in \mathbb{R}$, see Theorem 2.3 in [15] for a proof.

Acknowledgements: We are grateful to Benny Hon and Ding-Xuan Zhou of the Mathematics Department at City University of Hong Kong for providing both of us with the opportunity to complete this work in a scientifically stimulating and friendly environment.

References

1. K. Bennett and Bredensteiner. Duality and geometry in support vector machine classifiers. Proc. of the 17–th Int. Conf. on Machine Learning, P. Langley Ed., Morgan Kaufmann, pp. 57–63, 2000.
2. S. Canu, X. Mary, and A. Rakotomamonjy. Functional learning through kernel. In *Advances in Learning Theory: Methods, Models and Applications*, J. Suykens et al. Eds., NATO Science Series III: Computer and Systems Sciences, Vol. 190, pp 89–110, IOS Press, Amsterdam 2003.
3. J.M. Carnicer and J. Bastero. On best interpolation in Orlicz spaces. *Approx. Theory and its Appl.*, 10(4), pp. 72–83, 1994.
4. W. Dahmen and C.A. Micchelli. Some remarks on ridge functions. *Approx. Theory and its Appl.*, 3, pp. 139–143, 1987.
5. F. Deutsch. *Best Approximation in inner Product Spaces* CMS Books in Mathematics, Springer, 2001.
6. M. Hein and O. Bousquet. Maximal Margin Classification for Metric Spaces. In Proc. of the 16–th Annual Conference on Computational Learning Theory (COLT), 2003.
7. D. Kimber and P. M. Long. On-line learning of smooth functions of a single variable. *Theoretical Computer Science*, 148(1), pp. 141–156, 1995.
8. G.G. Lorenz. *Approximation of Functions.* Chelsea, 2nd ed., 1986.

9. C. Gentile. A new approach to maximal margin classification algorithms. *Journal of Machine Learning Research*, 2, pp. 213–242, 2001.
10. M. Leshno, V. Ya. Lin, A. Pinkus, and S. Schocken. Multilayer Feedforward Networks with a Non–Polynomial Activation Function can Approximate any Function. *Neural Networks*, 6, pp. 861–867, 1993.
11. J. Matousek. *Using the Borsuk-Ulam Theorem: Lectures on Topological Methods in Combinatorics and Geometry*. Springer-Verlag, Berlin, 2003.
12. H.N. Mhaskar and C.A. Micchelli. Approximation by superposition of sigmoidal functions. *Advances in Applied Mathematics*, 13, pp. 350–373, 1992.
13. C.A. Micchelli and M. Pontil. A function representation for learning in Banach spaces. Research Note RN/04/05, Dept of Computer Science, UCL, February 2004.
14. C.A. Micchelli and M. Pontil. Regularization algorithms for learning theory. Working paper, Dept of Computer Science, UCL, 2004.
15. C. A. Micchelli and F. I. Utreras. Smoothing and interpolation in a convex subset of a hilbert space. *SIAM J. of Scientific and Statistical Computing*, 9, pp. 728–746, 1988.
16. T.J. Morrison. *Functional Analysis: An Introduction to Banach Space Theory*. John Wiley Inc., New York, 2001.
17. W. Orlicz. *Linear Functional Analysis*. World Scientific, 1990.
18. A. Pinkus. *n–Widths in Approximation Theory*. Ergebnisse, Springer–Verlag, 1985.
19. A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica* 8, pp. 143–196, 1999.
20. M.M. Rao and Z.D. R. Ren. *Theory of Orlicz Spaces*. Marcel Dekker, Inc. 1992.
21. H.L. Royden. *Real Analysis*. Macmillan Publishing Company, New York, 3rd edition, 1988.
22. B. Schölkopf and A.J. Smola. *Learning with Kernels*. The MIT Press, Cambridge, MA, USA, 2002.
23. V. Vapnik. *The Nature of Statistical Learning Theory*. 2–nd edition, Springer, New York, 1999.
24. U. von Luxburg and O. Bousquet. Distance–based classification with Lipschitz functions. In Proc. of the 16–th Annual Conference on Computational Learning Theory (COLT), 2003.
25. G. Wahba. *Splines Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.
26. T. Zhang. On the dual formulation of regularized linear systems with convex risks. *Machine Learning*, 46, pp. 91–129, 2002.