Hindawi Publishing Corporation
Journal of Inequalities and Applications
Volume 2008, Article ID 640758, 16 pages
doi:10.1155/2008/640758

Research Article

Estimates of the Approximation Error Using Rademacher Complexity: Learning Vector-Valued Functions

Giorgio Gnecco[1, 2] and Marcello Sanguineti[2]

[1] *Department of Mathematics (DIMA), University of Genova, Via Dodecaneso 35, 16146 Genova, Italy*
[2] *Department of Communications, Computer, and System Sciences (DIST), University of Genova, Via Opera Pia 13, 16145 Genova, Italy*

Correspondence should be addressed to Marcello Sanguineti, marcello@dist.unige.it

Received 15 January 2008; Revised 13 August 2008; Accepted 20 October 2008

Recommended by Ulrich Abel

For certain families of multivariable vector-valued functions to be approximated, the accuracy of approximation schemes made up of linear combinations of computational units containing adjustable parameters is investigated. Upper bounds on the approximation error are derived that depend on the Rademacher complexities of the families. The estimates exploit possible relationships among the components of the multivariable vector-valued functions. All such components are approximated simultaneously in such a way to use, for a desired approximation accuracy, less computational units than those required by componentwise approximation. An application to N-stage optimization problems is discussed.

1. Introduction

Various authors have derived upper bounds on the approximation error of certain linear combinations of computational units containing adjustable parameters, called *variable-basis approximation schemes* [1], for various families of functions to be approximated (see, e.g., [2–11] and the references therein). In these schemes, the number of computational units (i.e., the number of basis functions) can be used to measure the *model complexity* [12, 13]; roughly speaking, models with a large complexity make the approximation task computationally inefficient. This typically occurs when the functions to be approximated depend on a large number d of variables, often because of the so-called *curse of dimensionality* [14]. However, experimental results have shown that variable-basis approximation schemes perform successfully in approximation of various high-dimensional mappings and theoretical insights into this have been obtained (see, e.g., [1, 3, 11], and the references therein).

Upper bounds on the approximation error, giving a partial explanation of this efficiency, have been derived using tools from statistical learning theory (SLT) [15]. This approach was first applied by Barron [2] and Girosi [4]. The latter exploited a well-known theorem by Vapnik and Chervonenkis [15], which gives, for a family of real-valued functions, a probabilistic uniform bound on the difference between the expected and empirical risks associated with a learning problem. Such a bound is expressed in terms of a combinatorial parameter, the *VC dimension* of the approximating family. For functions having an integral representation as the convolution $\kappa * \lambda$ of an $\mathcal{L}_1(\mathbb{R}^d)$ function λ with a bounded function $\kappa : \mathbb{R}^d \to \mathbb{R}$, Girosi [4] estimated in terms of VC-dimension the sup-norm error in approximation by linear combinations of $\kappa(\cdot - \mathbf{t}_1), \ldots, \kappa(\cdot - \mathbf{t}_n)$, with the parameters $\mathbf{t}_1, \ldots, \mathbf{t}_n$ varying in $\mathbb{R}^d$; this is a variable-basis approximation scheme [1].

In [8], Girosi's estimate [4] was extended to the approximation of functions for which the representation $\kappa * \lambda$ holds with $\lambda \in \mathcal{L}_p(\mathbb{R}^d)$, $1 < p < \infty$, with the error measured in a weighted essential supremum norm. In [9], Kon and Raphael used Girosi's approach [4] to derive error bounds for approximation in certain Hilbert spaces frequently used in learning theory, called *reproducing kernel Hilbert spaces* (*RKSH*s; see [16, 17], [18, Section III.3], and [19]). In [10], we exploited recent developments of SLT [20] to improve the approximation bounds from [4, 8, 9]. The estimates in [10] were derived in terms of the *Rademacher complexity* [20] of the families of functions to be approximated. Among recent works dealing with learning from the point of view of approximation theory, we cite [18] (on the mathematical foundations of learning), [21] (which uses Rademacher averages, too), [22] (which considers noise in the sampling data), [23], and the references therein. An excellent monograph devoted to this topic is [24].

The above-mentioned estimates were derived for scalar functions. They can be applied separately to every component of a vector mapping but, in doing so, one does not exploit mutual dependencies, similarities, and relationships that may hold among the components themselves. This may happen in various contexts, for example, when one has to approximate the optimal policy functions in N-stage optimization problems [25]. Despite its relevance in a number of applications, in the learning theory community the problem of approximating vector-valued functions has been studied much less than the scalar case. Its importance in learning seems to have been first pointed out in [26, 27]. A framework to study this problem in RKHSs was set down in [28] and further developed in other works; see [29] and the references therein.

In this paper, first we improve an estimate, obtained in [10], of the approximation error for scalar functions. Then, we derive upper bounds on the error of approximating *simultaneously all components* of multivariable vector-valued functions, using a variable-basis approximation scheme in which *all the scalar components share the same adjustable parameters inside the basis functions*. In this way, one may obtain the same approximation accuracy using fewer parameters to be optimized than those required by componentwise (hence scalar) approximation, in which, in general, the adjustable parameters inside the basis functions are different for the different components. We derive our estimates in terms of the Rademacher complexities of the families of functions $\mathbf{f} : X \subset \mathbb{R}^d \to \mathbb{R}^k$, whose components have an integral representation expressed in terms of $\mathcal{L}_1(X)$ functions and a bounded kernel. These components are approximated by linear combinations of functions obtained from the kernel in an adaptive way; this is a variable-basis approximation scheme [1]. We highlight advantages of simultaneous vector function approximation over componentwise approximation and we consider the application to approximate *dynamic programming* for N-stage optimization problems.

The paper is organized as follows. Section 2 describes notations and gives definitions. Section 3 refines some estimates for scalar functions, obtained in [4] and improved in [10]. Section 4 contains our upper bounds on the approximation error, for certain families of vector-valued functions. Section 5 compares the bounds derived for scalar and vector functions. Section 6 discusses the application of the results to approximate dynamic programming.

2. Notations and definitions

By $\mathbb{R}$ and $\mathbb{R}_+$ we denote the sets of real and positive real numbers, respectively, and by $\mathbb{N}$ and $\mathbb{N}_+$ the sets of natural numbers and positive integers, respectively.

For a real normed linear space $(\mathcal{H}, \|\cdot\|)$, $f \in \mathcal{H}$, and $r > 0$, we denote by $B_r(f, \|\cdot\|)$ the closed ball of radius r in the norm $\|\cdot\|$ centered at $f \in \mathcal{H}$, that is,

$$B_r(f, \|\cdot\|) = \{h \in \mathcal{H} \mid \|h - f\| \leq r\}. \tag{2.1}$$

We write $B_r(\|\cdot\|)$ instead of $B_r(0, \|\cdot\|)$. When the norm is clear from the context, we write merely $\mathcal{H}$, $B_r(f)$, and B_r instead of $(\mathcal{H}, \|\cdot\|)$, $B_r(f, \|\cdot\|)$, and $B_r(\|\cdot\|)$, respectively.

For $1 \leq p < \infty$, a positive integer d, and a Lebesgue-measurable set $X \subseteq \mathbb{R}^d$, we denote by $\mathcal{L}_p(X)$ the space of (equivalence classes of) real-valued functions on X that have integrable pth power with respect to the Lebesgue measure, endowed with the standard norm $\|\cdot\|_{p,X}$. By $\mathcal{L}_\infty(X)$ we denote the space of (equivalence classes of) real-valued functions on X that are essentially bounded with respect to the Lebesgue measure, endowed with the essential supremum norm $\|\cdot\|_{\infty,X}$, and by $\mathcal{C}(X)$ the space of continuous functions on X with the supremum norm. Whenever there is no ambiguity, we omit X from the notations.

The *d-dimensional Fourier transform* is defined [30, pages 180, 187] as the operator on $\mathcal{L}_1(\mathbb{R}^d) \cap \mathcal{L}_2(\mathbb{R}^d)$, continuously extended to an operator from $\mathcal{L}_2(\mathbb{R}^d)$ to $\mathcal{L}_2(\mathbb{R}^d)$, such that for every function $f \in \mathcal{L}_1(\mathbb{R}^d) \cap \mathcal{L}_2(\mathbb{R}^d)$ one has

$$f(\mathbf{t}) \longmapsto \widehat{f}(\mathbf{s}) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\langle \mathbf{t}, \mathbf{s}\rangle} f(\mathbf{t}) d\mathbf{t}, \tag{2.2}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product in $\mathbb{R}^d$. For $f \in \mathcal{L}_2(\mathbb{R}^d)$, one has $\|f\|_2 = \|\widehat{f}\|_2$ [30, page 187].

For a positive integer d, a set $X \subseteq \mathbb{R}^d$, and a family F of functions on X, we denote by $F_{\mathbf{x}} : X \to \mathbb{R}$ a function in F, where $\mathbf{x}$ is a parameter used to identify elements in F(we use the notation $F_{\mathbf{x}}$ since we will consider families F of functions on X having the integral representation $f(\mathbf{x}) = \int_X K_{\mathbf{x}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t}$, where $K : X \times X \to \mathbb{R}$. So, $F_{\mathbf{x}}(\mathbf{t})$ is defined via $K_{\mathbf{x}}(\mathbf{t})$ and the parameter $\mathbf{x}$ used to identify the elements of F is a point $\mathbf{x} \in X$) .

By P_X we denote a probability distribution on X; we write merely P when the set X is clear from the context. For every positive integer n, a *P_X-i.i.d. sequence* is a sequence $\{\mathbf{t}_i\}$ of n points obtained by sampling X independently n times according to P_X (similarly, we define a p_X-i.i.d. sequence when p_X is a probability density).

A *Rademacher random variable* is a random variable taking only the values -1 and $+1$ with equal probability [20]. Let P_X be a probability distribution on $X \subseteq \mathbb{R}^d$, $\{\mathbf{t}_i\}$ a P_X-i.i.d.

sequence, and $\{\varepsilon_i\}$ a sequence of n independent Rademacher random variables. Given a family $F = \{F_{\mathbf{x}}\}$ of functions $f : X \to \mathbb{R}$, the *Rademacher complexity* of F is defined as [20]

$$\mathcal{R}_n(F) \triangleq \mathbb{E}_{\mathbf{t}_1,\ldots,\mathbf{t}_n}\mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_n}\left\{\frac{1}{\sqrt{n}}\sup_{F_{\mathbf{x}}\in F}\left|\sum_{i=1}^{n}\varepsilon_i F_{\mathbf{x}}(\mathbf{t}_i)\right|\right\}. \tag{2.3}$$

Upper bounds on the Rademacher complexities of various families of functions are available; see, for example, [10, 20, 31].

The *VC dimension* of a family $F = \{F_{\mathbf{x}}\}$ of real-valued functions on a set X is the maximum number h of points $\{\mathbf{t}_i\}$ in X that can be separated into two distinct classes in all 2^h possible ways, by using functions of the form $F_{\mathbf{x}}(\mathbf{t}) - \alpha$, where the parameters $\mathbf{x}$ and α vary in X and $\mathbb{R}$, respectively [15].

3. A bound for the approximation of scalar functions

For $r > 0$, the *Bessel potential of order* r is defined as the function $\beta_r : \mathbb{R}^d \to \mathbb{R}$ with the Fourier transform

$$\widehat{\beta}_r(\mathbf{s}) = (2\pi)^{-d/2}\left(1 + \|\mathbf{s}\|^2\right)^{-r/2}. \tag{3.1}$$

We consider the family of functions defined as

$$\mathcal{F}_r^1 \triangleq \left\{f : \mathbb{R}^d \longrightarrow \mathbb{R} \mid f = \beta_r * \lambda,\ \lambda \in \mathcal{L}_1\left(\mathbb{R}^d\right)\right\}, \tag{3.2}$$

where for two functions $g, h : \mathbb{R}^d \to \mathbb{R}$, $(g*h)(x) \triangleq \int_{\mathbb{R}^d} g(y)h(x-y)dy$ is their *convolution*. The space $\mathcal{F}_r^1$ is called *Bessel potential space of order* 1; it is a normed space with the norm $\|\cdot\|_{\mathcal{F}_r^1}$ defined for every $f \in \mathcal{F}_r^1$ as $\|f\|_{\mathcal{F}_r^1} \triangleq \|\lambda\|_1$.

The following result from [10] improves the approximation bound from [4, Proposition 3.1]. We let

$$K_{r,\mathbf{x}}^{\text{Bessel}}(\mathbf{t}) \triangleq \beta_r(\mathbf{x} - \mathbf{t}). \tag{3.3}$$

Theorem 3.1 (see [10]). *There exists an absolute positive constant C such that the following holds. Let r, d be positive integers, $r > d$, and let h_r be the VC dimension of $\{K_{r,\mathbf{x}}^{\text{Bessel}}\}$. For every $f \in \mathcal{F}_r^1$ and every positive integer n, there exist $\mathbf{t}_1, \ldots, \mathbf{t}_n \in \mathbb{R}^d$ and $c_1, \ldots, c_n \in \{-1, 1\}$ such that*

$$\sup_{\mathbf{x}\in\mathbb{R}^d}\left|f(\mathbf{x}) - \frac{\|\lambda\|_1}{n}\sum_{i=1}^{n}c_i\beta_r(\mathbf{x} - \mathbf{t}_i)\right| \leq C\|\lambda\|_1\sqrt{\frac{h_r}{n}}. \tag{3.4}$$

In [4, 10], no upper bound on h_r was given (up to our knowledge, in the literature even the boundedness of h_r had not yet been proven till now). In the following we provide such an upper bound.

Proposition 3.2. *Let r, d be positive integers, $r > d$, and let h_r be the VC dimension of $\{K_{r,\mathbf{x}}^{\text{Bessel}}\}$. Then $h_r \leq d + 2$.*

Proof. Consider the two families of functions $E \triangleq \{K_{r,\mathbf{x}}^{\text{Bessel}}\}$ and $G \triangleq \{G_{\mathbf{x}}\}$, where $G_{\mathbf{x}}(\mathbf{t}) \triangleq e^{-\|\mathbf{x}-\mathbf{t}\|^2}$. By inspection of the proof of [10, Corollary 5.2] we get

$$\beta_r(\mathbf{t}) = \frac{2^{-d/2}}{\Gamma(r/2)} \int_0^{\infty} u^{(r-d)/2-1}\, e^{-\|\mathbf{t}\|^2/4u}\, e^{-u}\, du. \tag{3.5}$$

This integral representation shows that $\beta_r(\mathbf{t})$ is a decreasing function of $\|\mathbf{t}\|$. Let us define the two functions $a, b : [0, +\infty] \to \overline{\mathbb{R}}$ as

$$\begin{gathered} a(\|\mathbf{t}\|) \triangleq \beta_r(\mathbf{t}), \\ b(\|\mathbf{t}\|) \triangleq e^{-\|\mathbf{t}\|^2}, \quad b(+\infty) = 0. \end{gathered} \tag{3.6}$$

Since b is one-to-one, there exists a function $\phi : [0, 1] \to \overline{\mathbb{R}}$ such that $a = \phi \circ b$. Indeed, one can take $\phi = a \circ b^{-1}$. Then $\beta_r(\mathbf{x}-\mathbf{t}) = \phi(e^{-\|\mathbf{x}-\mathbf{t}\|^2})$, that is, $E = \phi(G)$, where ϕ is one-to-one (since it is the composition of two one-to-one functions). Then the VC dimensions $VC(E)$ and $VC(G)$ of E and G are the same. Since G is a subset of the family of functions $L = \{K_{\mathbf{x}}^{\text{Gauss}}\}$ considered in the proof of [10, Corollary 5.2], where it was shown that $VC(L) \leq d + 3$, one has also

$$VC(E) = VC(G) \leq VC(L) \leq d + 3. \tag{3.7}$$

The proof of the bound $VC(L) \leq d + 3$ in [10, Corollary 5.2] can be slightly adapted to find a better upper bound directly on $VC(G)$. Here we report only the modifications that are required. By the same arguments of the proof of [10, Corollary 5.2] it is easy to show that, for every fixed $\mathbf{x}$, each element of the family $\{-\|\mathbf{x} - \mathbf{t}\|^2 + \alpha\}$, where α is a real parameter, can be expressed as a linear combination of the $d + 2$ functions

$$1, t_1, \ldots, t_d, \|\mathbf{t}\|^2. \tag{3.8}$$

Hence, by [32, Theorem 1], $VC(G)$ is at most $d + 2$. □

Combining [10, Theorem 4.5] with the proof technique of Proposition 3.2, we get the following estimate. Recall that a function ψ with bounded variation can be written as the difference of two decreasing functions ψ_1 and ψ_2 [33, Theorem 4, page 331]. In order to state the result we require in addition that ψ_1 and ψ_2 are bounded.

Proposition 3.3. *Let d be a positive integer, $X \subset \mathbb{R}^d$ a compact domain, $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ continuous, $\lambda \in \mathcal{L}_1(X)$, and let $f : X \to \mathbb{R}$ have the representation $f(\mathbf{x}) = \int_X K_{\mathbf{x}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t}$, where $K_{\mathbf{x}}(\mathbf{t}) = \psi(\|\mathbf{x} - \mathbf{t}\|)$ and $\psi(z)$ has bounded variation. Suppose that there exist $\tau_1, \tau_2 \geq 0$ and a decomposition of $\psi(z)$ as $\psi(z) = \psi_1(z) - \psi_2(z)$ such that $\psi_1(z)$ and $\psi_2(z)$ are decreasing and $\sup_{t\in\mathbb{R}}|\psi_1(z)| \leq \tau_1$, $\sup_{t\in\mathbb{R}}|\psi_2(z)| \leq \tau_2$. Then there exist two absolute positive constants C_1 and C_2 such that for every positive integer n, there exist $\mathbf{t}_1, \ldots, \mathbf{t}_n \in X$ and $c_1, \ldots, c_n \in \{-1, 1\}$ for which*

$$\sup_{\mathbf{x}\in X} \left| f(\mathbf{x}) - \frac{\|\lambda\|_1}{n} \sum_{i=1}^{n} c_i K(\mathbf{x}, \mathbf{t}_i) \right| \leq \|\lambda\|_1 \left(C_1\sqrt{d+2} + C_2(\tau_1 + \tau_2) \right) \sqrt{\frac{1}{n}}. \tag{3.9}$$

Proof. Setting $K_{1,\mathbf{x}}(\mathbf{t}) \triangleq \psi_1(\|\mathbf{x}-\mathbf{t}\|)$ and $K_{2,\mathbf{x}}(\mathbf{t}) \triangleq \psi_2(\|\mathbf{x}-\mathbf{t}\|)$, we get

$$f(\mathbf{x}) = \int_X K_{\mathbf{x}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t} = \int_X K_{1,\mathbf{x}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t} - \int_X K_{2,\mathbf{x}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t}. \tag{3.10}$$

Since the functions ψ_1 and ψ_2 are decreasing, exploiting similar arguments as in the proof of Proposition 3.2 we conclude that the upper bound $d+2$ holds for the VC dimensions of the families of functions $\{K_{1,\mathbf{x}}\}$ and $\{K_{2,\mathbf{x}}\}$, too. Then the statement follows by [10, Theorem 4.3 and Lemma 4.4] and the subadditivity property of the Rademacher complexity [34, Theorem 12, point 7]. □

4. Bounds for the approximation of vector-valued functions

In [10], the following approximation error bound was obtained in terms of the Rademacher complexity for some quite general families of functions. It improves, at least asymptotically, the bound derived in [4] for the same families of functions.

Theorem 4.1 (see [10]). *Let $X \subset \mathbb{R}^d$ be a compact domain, $K : X \times X \to \mathbb{R}$ continuous, and $\tau > 0$ such that, for all $\mathbf{x}$ and $\mathbf{t}$, one has $|K(\mathbf{x},\mathbf{t})| \le \tau$. Let $\lambda \in \mathcal{L}_1(X)$, let f be a real-valued function on X having the representation $f(\mathbf{x}) = \int_X K(\mathbf{x},\mathbf{t})\lambda(\mathbf{t})d\mathbf{t}$, and $\mathcal{R}_n$ the Rademacher complexity of the family $\{K(\mathbf{x},\cdot)\}$. There exists an absolute positive constant C such that for every positive integer n there exist $\mathbf{t}_1,\ldots,\mathbf{t}_n \in X$ and $c_1,\ldots,c_n \in \{-1,+1\}$ for which*

$$\sup_{\mathbf{x}\in X}\left| f(\mathbf{x}) - \frac{\|\lambda\|_{1,X}}{n}\sum_{i=1}^{n} c_i K(\mathbf{x},\mathbf{t}_i)\right| \le C\|\lambda\|_{1,X}(\mathcal{R}_n+\tau)\sqrt{\frac{1}{n}}. \tag{4.1}$$

Note that the approximation scheme in (4.1) requires n binary parameters and nd real parameters.

Remark 4.2. Inspection of the proof of Theorem 4.1 shows that if $\lambda(\mathbf{t})$ is a nonnegative function, then $c_i = +1$, $i = 1,\ldots,n$. Thus, if $K(\mathbf{x},\mathbf{t})$ is convex (or concave), both $f(\mathbf{x})$ and its sparse approximation $(\|\lambda\|_1/n)\sum_{i=1}^{n} c_i K(\mathbf{x},\mathbf{t}_i)$ are convex (concave, resp.), too. Preservation of convexity (or concavity) is an interesting property of approximators, sometimes exploited in applications [35].

We first derive for vector-valued functions an immediate corollary of Theorem 4.1. Then, we turn our attention to the simultaneous approximation of all the components of a vector-valued function, which allows one to exploit similarities that may hold among them.

Corollary 4.3. *Let $X \subset \mathbb{R}^d$ be a compact domain, $K : X \times X \to \mathbb{R}$ continuous, $\tau > 0$ such that, for all $\mathbf{x}$ and $\mathbf{t}$, one has $|K(\mathbf{x},\mathbf{t})| \le \tau$, and $\mathcal{R}_n$ the Rademacher complexity of the family $\{K(\mathbf{x},\cdot)\}$. Let $\mathbf{f} : X \to \mathbb{R}^k$ be such that each of its components has the representation $f_m(\mathbf{x}) = \int_X K(\mathbf{x},\mathbf{t})\lambda_m(\mathbf{t})d\mathbf{t}$, where $\lambda_m \in \mathcal{L}_1(X)$, $m = 1,\ldots,k$. Then for every $\epsilon > 0$ it is possible to approximate in the supremum norm on X each component f_m of $\mathbf{f}$ with an error at most ϵ by using approximations of the form*

$$\widehat{f}_m(\mathbf{x}) = \frac{\|\lambda_m\|_{1,X}}{n_m}\sum_{i=1}^{n_m} c_{m,i} K(\mathbf{x},\mathbf{t}_{m,i}), \tag{4.2}$$

where, for $i = 1, \dots, n_m$ and $m = 1, \dots, k$, $\mathbf{t}_{m,i} \in X$ and $c_{m,i} \in \{-1, +1\}$, provided that for every $m = 1, \dots, k$ one has

$$n_m \geq \frac{C^2 \left(R_{n_m} + \tau\right)^2 \|\lambda_m\|_{1,X}^2}{\epsilon^2}, \tag{4.3}$$

where C is an absolute positive constant.

Proof. It follows immediately by Theorem 4.1 applied to each of the k components of $\mathbf{f}$, choosing each n_m such that the correspondent approximation error is upper bounded by ϵ. □

To approximate all the k components of $\mathbf{f}$, the approximation scheme (4.2) requires $\overline{n} = \sum_{m=1}^{k} n_m$ binary parameters, and $\overline{n}d$ real parameters.

Let us now consider the case where similarities among the k components of $\mathbf{f}$ are present. We would like to exploit them to obtain a desired approximation accuracy with a number of parameters smaller than the one given by (4.2), by letting some parameters be shared among the k approximators. Assume that, in the expression $f_m(\mathbf{x}) = \int_X K(\mathbf{x}, \mathbf{t}) \lambda_m(\mathbf{t}) d\mathbf{t}$, $m = 1, \dots, k$, for the components of $\mathbf{f}$, each $\lambda_m \in \mathcal{L}_1(X)$ can be written as

$$\lambda_m(\mathbf{t}) = \widetilde{\lambda}_m(\mathbf{t}) \lambda(\mathbf{t}), \tag{4.4}$$

where $\widetilde{\lambda}_m$ are measurable functions taking only the values -1 and $+1$ and $0 \neq \lambda \in \mathcal{L}_1(X)$ is a nonnegative function (there is no loss of generality in assuming that λ is nonnegative). Setting

$$\widetilde{K}_m(\mathbf{x}, \mathbf{t}) \triangleq K(\mathbf{x}, \mathbf{t}) \widetilde{\lambda}_m(\mathbf{t}), \tag{4.5}$$

one has

$$f_m(\mathbf{x}) = \int_X \widetilde{K}_m(\mathbf{x}, \mathbf{t}) \lambda(\mathbf{t}) d\mathbf{t}, \tag{4.6}$$

which can be written as

$$\frac{f_m(\mathbf{x})}{\|\lambda\|_{1,X}} = \int_X \widetilde{K}_m(\mathbf{x}, \mathbf{t}) \frac{\lambda(\mathbf{t})}{\|\lambda\|_{1,X}} d\mathbf{t}. \tag{4.7}$$

As noted in [4], for every $\mathbf{x} \in X$ the representation (4.7) can be regarded as the expected value of the random variable $\operatorname{sgn}(\lambda(\mathbf{t})) \widetilde{K}_m(\mathbf{x}, \mathbf{t})$ with respect to the probability density $|\lambda(\mathbf{t})| / \|\lambda\|_{1,X}$.

It is easy to see how similarities among the k components of f can be modeled in this way, for example, by assuming that there exists $\zeta > 0$ such that $K(\mathbf{x}, \mathbf{t}) \cong 0$ if $\|\mathbf{x} - \mathbf{t}\| > \zeta$. Indeed, if there exists $\mathbf{x}' \in X$ such that $\lambda(\mathbf{t}) \cong 0$ on a neighborhood $\mathcal{N}'$ of $\mathbf{x}'$ of radius larger than ζ, all components of f will assume very small values on another neighborhood $\mathcal{N}'' \subset \mathcal{N}'$ of $\mathbf{x}'$.

Exploiting the particular nature of the functions $\widetilde{\lambda}_m(\mathbf{t})$, which can take only the values −1 and +1, the following proposition shows that the Rademacher complexities of the families $\{\widetilde{K}_m(\mathbf{x},\cdot)\}$ and $\{K(\mathbf{x},\cdot)\}$ are the same.

Proposition 4.4. *Let $\mathcal{R}_n$ and $\mathcal{R}_{m,n}$ be the Rademacher complexities of the families $\{K(\mathbf{x},\cdot)\}$ and $\{\widetilde{K}_m(\mathbf{x},\cdot)\}$, respectively. Then for every $m = 1,\ldots,k$, $\mathcal{R}_{m,n} = \mathcal{R}_n$.*

Proof. It follows directly from the definition of the Rademacher complexity, as each $\widetilde{\lambda}_m(\mathbf{t})$ can take only the values −1 and +1 and the Rademacher random variables ϵ_i are independent and symmetrically distributed around 0. □

To derive an extension of Theorem 4.1 to families of vector-valued functions, we will exploit the following estimate from [10]. Recall that for a family $F = \{F_\mathbf{x}\}$ of functions on X and a probability distribution P_X on X, the *expected risk* associate with a function $F_\mathbf{x} \in F$ is defined as

$$R(F_\mathbf{x}) \triangleq \int_X F_\mathbf{x}(\mathbf{t})dP_X(\mathbf{t}). \tag{4.8}$$

So, $R(F_\mathbf{x}) = \mathbb{E}_{P_X}\{F_\mathbf{x}(\mathbf{t})\}$, where $\mathbb{E}_{P_X}$ is the expectation operator. The *empirical risk* associate with the function $F_\mathbf{x}(\mathbf{t}) \in F$ and the sequence $\{\mathbf{t}_i\}$ of samples is defined as

$$R_{\text{emp}}(F_\mathbf{x},\{\mathbf{t}_i\}) \triangleq \frac{1}{n}\sum_{i=1}^{n} F_\mathbf{x}(\mathbf{t}_i). \tag{4.9}$$

Theorem 4.5 (see [10]). *Let P_X be a probability distribution on $X \subseteq \mathbb{R}^d$, $\{\mathbf{t}_i\}$ a P_X-i.i.d. sequence of n points in X, $\tau > 0$, and F a family of $[-\tau,\tau]$-valued continuous functions with Rademacher complexity $\mathcal{R}_n$. Then there exists an absolute constant C such that for all $0 < \delta < 1$, with probability at least $1-\delta$, one has*

$$\sup_{F_\mathbf{x}\in F}\left|R(F_\mathbf{x}) - R_{\text{emp}}(F_\mathbf{x},\{\mathbf{t}_i\})\right| \le 2\tau C\sqrt{\frac{1}{n}\max\left\{\left(\frac{\mathcal{R}_n+\tau}{2\tau}\right)^2, \ln\frac{1}{\delta}\right\}}. \tag{4.10}$$

In the next theorem (Theorem 4.7), we will improve the estimate from Theorem 4.1 for approximation of vector-valued functions (for $k = 1$ and in the limit $\delta \to 0$ one gets Theorem 4.1). To derive such an extension, we will need the density result stated in the following lemma. For $X \subset \mathbb{R}^d$ compact, $K : X \times X \to \mathbb{R}$ continuous, and $\lambda \in \mathcal{L}_1(X)$, we let

$$\begin{aligned} F_\lambda &\triangleq \left\{f : X \longrightarrow \mathbb{R} \mid f(\mathbf{x}) = \int_X K(\mathbf{x},\mathbf{t})\lambda_{\text{bound}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t},\ \|\lambda_{\text{bound}}\|_{\infty,X} \le 1\right\}, \\ G_\lambda &\triangleq \left\{g : X \longrightarrow \mathbb{R} \mid g(\mathbf{x}) = \int_X K(\mathbf{x},\mathbf{t})\lambda_{\text{bin}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t},\ \lambda_{\text{bin}} : X \longrightarrow \{\pm 1\}\ \text{measurable}\right\}. \end{aligned} \tag{4.11}$$

Lemma 4.6. *Let $X \subset \mathbb{R}^d$ be compact, $K : X\times X \to \mathbb{R}$ continuous, and $\lambda \in \mathcal{L}_1(X)$. Then G_λ is dense in F_λ in the supremum norm.*

Proof. Given an element of F_λ, say $\tilde{f}_{\text{bound}}(\mathbf{x}) = \int_X K(\mathbf{x},\mathbf{t})\tilde{\lambda}_{\text{bound}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t}$, and an arbitrary $\epsilon > 0$, let us find an element of G_λ, say $\tilde{g}_{\text{bin}}(\mathbf{x}) = \int_X K(\mathbf{x},\mathbf{t})\tilde{\lambda}_{\text{bin}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t}$, such that $\sup_{\mathbf{x}\in X}|\tilde{f}_{\text{bound}}(\mathbf{x}) - \tilde{g}_{\text{bin}}(\mathbf{x})| < \epsilon$.

As $K(\mathbf{x},\mathbf{t})$ is uniformly continuous on $X \times X$, for every $\eta > 0$ it is possible to find a partition $\{P_i : i = 1,\ldots,N(\eta)\}$ of X such that for every $i, j \in \{1,\ldots,N(\eta)\}$ one has

$$\left|K_{\max(i,j)} - K_{\min(i,j)}\right| \leq \eta, \tag{4.12}$$

where $K_{\max(i,j)} \triangleq \max_{(\mathbf{x},\mathbf{t})\in P_i\times P_j} K(\mathbf{x},\mathbf{t})$ and $K_{\min(i,j)} \triangleq \min_{(\mathbf{x},\mathbf{t})\in P_i\times P_j} K(\mathbf{x},\mathbf{t})$. Let $\mathbf{x} \in P_i$. As

$$\begin{aligned}
\int_{P_i} K(\mathbf{x},\mathbf{t})\tilde{\lambda}_{\text{bound}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t} = {} & \int_{P_j} \left(K(\mathbf{x},\mathbf{t}) - K_{\min(i,j)}\right)\tilde{\lambda}_{\text{bound}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t} \\
& + K_{\min(i,j)}\int_{P_j} \left(\tilde{\lambda}_{\text{bound}}(\mathbf{t}) - \tilde{\lambda}_{\text{bin}}(\mathbf{t})\right)\lambda(\mathbf{t})d\mathbf{t} \\
& + \int_{P_j} \left(K_{\min(i,j)} - K(\mathbf{x},\mathbf{t})\right)\tilde{\lambda}_{\text{bin}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t} \\
& + \int_{P_j} K(\mathbf{x},\mathbf{t})\tilde{\lambda}_{\text{bin}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t},
\end{aligned} \tag{4.13}$$

we get

$$\begin{aligned}
\sup_{\mathbf{x}\in P_i} & \left| \int_{P_j} K(\mathbf{x},\mathbf{t})\tilde{\lambda}_{\text{bound}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t} - \int_{P_j} K(\mathbf{x},\mathbf{t})\tilde{\lambda}_{\text{bin}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t} \right| \\
& \leq \sup_{\mathbf{x}\in P_i} \left| \int_{P_j} \left(K(\mathbf{x},\mathbf{t}) - K_{\min(i,j)}\right)\tilde{\lambda}_{\text{bound}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t} \right| \\
& \quad + \sup_{\mathbf{x}\in P_i} \left| K_{\min(i,j)}\int_{P_j} \left(\tilde{\lambda}_{\text{bound}}(\mathbf{t}) - \tilde{\lambda}_{\text{bin}}(\mathbf{t})\right)\lambda(\mathbf{t})d\mathbf{t} \right| \\
& \quad + \sup_{\mathbf{x}\in P_i} \left| \int_{P_j} \left(K_{\min(i,j)} - K(\mathbf{x},\mathbf{t})\right)\tilde{\lambda}_{\text{bin}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t} \right|.
\end{aligned} \tag{4.14}$$

As to the first and third terms in the right-hand side of (4.14), by Hölder's inequality and (4.12) we get

$$\begin{aligned}
\sup_{\mathbf{x}\in P_i} \left| \int_{P_j} \left(K(\mathbf{x},\mathbf{t}) - K_{\min(i,j)}\right)\tilde{\lambda}_{\text{bound}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t} \right| & \leq \sup_{\mathbf{x}\in P_i} \left\|K(\mathbf{x},\cdot) - K_{\min(i,j)}\right\|_{\infty,P_j} \left\|\tilde{\lambda}_{\text{bound}}\lambda\right\|_{1,P_j} \\
& \leq \eta\|\lambda\|_{1,P_j}, \\
\sup_{\mathbf{x}\in P_i} \left| \int_{P_j} \left(K_{\min(i,j)} - K(\mathbf{x},\mathbf{t})\right)\tilde{\lambda}_{\text{bin}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t} \right| & \leq \sup_{\mathbf{x}\in P_i} \left\|K(\mathbf{x},\cdot) - K_{\min(i,j)}\right\|_{\infty,P_j} \left\|\tilde{\lambda}_{\text{bin}}\lambda\right\|_{1,P_j} \\
& \leq \eta\|\lambda\|_{1,P_j}.
\end{aligned} \tag{4.15}$$

Now we consider the second term in the right-hand side of (4.14). Let $c \triangleq \int_{P_j} \tilde{\lambda}_{\text{bound}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t}$ and divide P_j into two sets A_j and B_j, such that $\int_{A_j} |\lambda(\mathbf{t})|d\mathbf{t} = (\|\lambda\|_{1,P_j} + c)/2$ (such sets exist by [33, Theorem 6, page 300], as $\|\widetilde{\lambda}_{\text{bound}}\|_\infty \le 1$ implies $c \le \|\lambda\|_{1,P_j}$). Choosing $\widetilde{\lambda}_{\text{bin}}$ such that $\widetilde{\lambda}_{\text{bin}}(\mathbf{t}) = \operatorname{sgn}(\lambda(\mathbf{t}))$, for all $\mathbf{t} \in A_j$ and $\widetilde{\lambda}_{\text{bin}}(\mathbf{t}) = -\operatorname{sgn}(\lambda(\mathbf{t}))$, for all $\mathbf{t} \in B_j$, simple computations give

$$\int_{P_j} \left(\widetilde{\lambda}_{\text{bound}}(\mathbf{t}) - \widetilde{\lambda}_{\text{bin}}(\mathbf{t})\right)\lambda(\mathbf{t})d\mathbf{t} = 0. \tag{4.16}$$

Summing up, taking $\eta = \epsilon/2$ for every $\epsilon > 0$ we get

$$\begin{aligned}\sup_{\mathbf{x}\in X}&\left|\widetilde{f}_{\text{bound}}(\mathbf{x}) - \widetilde{g}_{\text{bin}}(\mathbf{x})\right| \\ &= \sup_{\mathbf{x}\in X}\left|\sum_{P_j}\left(\int_{P_j} K(\mathbf{x},\mathbf{t})\widetilde{\lambda}_{\text{bound}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t} - \int_{P_j} K(\mathbf{x},\mathbf{t})\widetilde{\lambda}_{\text{bin}}(\mathbf{t})\lambda(\mathbf{t})d\mathbf{t}\right)\right| \le \epsilon\|\lambda\|_1.\end{aligned} \tag{4.17}$$

The statement follows by letting $\epsilon \to 0$ and considering correspondingly refined partitions. □

Now we can prove our main result.

Theorem 4.7. *Let $X \subset \mathbb{R}^d$ be a compact domain, $K : X \times X \to \mathbb{R}$ continuous, $\tau > 0$ such that, for all $\mathbf{x}$ and $\mathbf{t}$, one has $|K(\mathbf{x},\mathbf{t})| \le \tau$, and $\mathcal{R}_n$ be the Rademacher complexity of the family $\{K(\mathbf{x},\cdot)\}$. Let $0 \ne \lambda \in \mathcal{L}_1(X)$ be a nonnegative function on X and $\mathbf{f} : X \to \mathbb{R}^k$ such that each of its components has the representation $f_m(\mathbf{x}) = \int_X \widetilde{K}_m(\mathbf{x},\mathbf{t})\lambda(\mathbf{t})d\mathbf{t}$. For $m = 1,\ldots,k$, set $\widetilde{K}_m(\mathbf{x},\mathbf{t}) = K(\mathbf{x},\mathbf{t})\widetilde{\lambda}_m(\mathbf{t})$, where each $\widetilde{\lambda}_m$ is a measurable function such that $\|\widetilde{\lambda}_m\|_{\infty,X} \le 1$. Then there exists an absolute constant C such that for every positive integer n there exist $\mathbf{t}_1,\ldots,\mathbf{t}_n \in X$ and $c_{m,1},\ldots,c_{m,n} \in \{-1,+1\}$, $m = 1,\ldots,k$, for which one has simultaneously for all $m = 1,\ldots,k$*

$$\sup_{\mathbf{x}\in X}\left|f_m(\mathbf{x}) - \frac{\|\lambda\|_{1,X}}{n}\sum_{i=1}^{n} c_{m,i}K(\mathbf{x},\mathbf{t}_i)\right| \le 2\tau C\|\lambda\|_{1,X}\sqrt{\frac{1}{n}\max\left\{\left(\frac{\mathcal{R}_n+\tau}{2\tau}\right)^2, \ln k\right\}}. \tag{4.18}$$

Proof. For every $\mathbf{x} \in X$, $f_m(\mathbf{x})/\|\lambda\|_{1,X} = \int_X \widetilde{K}_m(\mathbf{x},\mathbf{t})(\lambda(\mathbf{t})/\|\lambda\|_{1,X})d\mathbf{t}$ can be considered as the expected risk of the function $\operatorname{sgn}(\lambda(\cdot))\widetilde{K}_m(\mathbf{x},\cdot)$ with respect to the probability density $|\lambda(\mathbf{t})|/\|\lambda\|_{1,X}$. Combining Proposition 4.4, Theorem 4.5, and Lemma 4.6, for every positive integer l and $m = 1,\ldots,k$, it is possible to find an approximation $\widetilde{\lambda}^{l}_{\text{bin},m}(\mathbf{t})$ of $\widetilde{\lambda}_m(\mathbf{t})$ such that for every $\delta_m > 0$ and every $\lambda(\mathbf{t})/\|\lambda\|_{1,X}$-i.i.d. sequence $\{\mathbf{t}_i\}$, we get with probability at least $1-\delta_m$

$$\sup_{\mathbf{x}\in X}\left|\frac{f_m(\mathbf{x})}{\|\lambda\|_{1,X}} - \frac{1}{n}\sum_{i=1}^{n}\widetilde{\lambda}^{l}_{\text{bin},m}(\mathbf{t}_i)K(\mathbf{x},\mathbf{t}_i)\right| \le 2\tau C\sqrt{\frac{1}{n}\max\left\{\left(\frac{\mathcal{R}_n+\tau}{2\tau}\right)^2, \ln\frac{1}{\delta_m}\right\}} + \frac{1}{l}. \tag{4.19}$$

For simplicity and without any loss of generality, we take $\delta_m = \delta < 1/k$, for all $m = 1,\ldots,k$. By standard probability arguments we get with probability at least $1 - k\delta$

$$\sup_{\mathbf{x}\in X}\left|\frac{f_m(\mathbf{x})}{\|\lambda\|_{1,X}} - \frac{1}{n}\sum_{i=1}^{n}\widetilde{\lambda}^{l}_{\text{bin},m}(\mathbf{t}_i)K(\mathbf{x},\mathbf{t}_i)\right| \le 2\tau C\sqrt{\frac{1}{n}\max\left\{\left(\frac{\mathcal{R}_n+\tau}{2\tau}\right)^2, \ln\frac{1}{\delta}\right\}} + \frac{1}{l} \tag{4.20}$$

simultaneously for all components of $\mathbf{f}$. Then, there exists a choice $\mathbf{t}_i^{l,\delta}$ of $\mathbf{t}_i$, $i = 1, \ldots, n$, for which the above-written upper bound holds.

As X is compact, by letting $l \to \infty$ and $\delta \to 1/k$ it is possible to extract a subsequence denoted—with a little abuse—by $\{\mathbf{t}_i^{l,\delta}\}$, such that, for each $i = 1, \ldots, n$, $\mathbf{t}_i^{l,\delta} \to \mathbf{t}_i^* \in X$. Similarly, since for every l one has a finite number of binary parameters $c_{m,i}^{l,\delta} = \widetilde{\lambda}_{\text{bin},m}^{l}(\mathbf{t}_i^{l,\delta})$, one can also extract a subsequence denoted—with a little abuse—by $\{c_{m,i}^{l,\delta}\}$, such that, for every $i = 1, \ldots, n$ and $m = 1, \ldots, k$, $c_{m,i}^{l,\delta} \to c_{m,i}^* \in \{-1, +1\}$ as $l \to \infty$ and $\delta \to 1/k$. Then

$$\begin{aligned}\sup_{\mathbf{x}\in X}\left|f_m(\mathbf{x}) - \frac{\|\lambda\|_{1,X}}{n}\sum_{i=1}^{n} c_{m,i}^* K(\mathbf{x}, \mathbf{t}_i^*)\right| &\leq \sup_{\mathbf{x}\in X}\left|f_m(\mathbf{x}) - \frac{\|\lambda\|_{1,X}}{n}\sum_{i=1}^{n} c_{m,i}^{l,\delta} K(\mathbf{x}, \mathbf{t}_i^{l,\delta})\right| \\ &\quad + \sup_{\mathbf{x}\in X}\left|\frac{\|\lambda\|_{1,X}}{n}\sum_{i=1}^{n}\left[c_{m,i}^{l,\delta} K(\mathbf{x}, \mathbf{t}_i^{l,\delta}) - c_{m,i}^* K(\mathbf{x}, \mathbf{t}_i^*)\right]\right|.\end{aligned} \tag{4.21}$$

As $K(\mathbf{x}, \mathbf{t})$ is uniformly continuous on $X \times X$, in the limit we get

$$\sup_{\mathbf{x}\in X}\left|f_m(\mathbf{x}) - \frac{\|\lambda\|_{1,X}}{n}\sum_{i=1}^{n} c_{m,i}^* K(\mathbf{x}, \mathbf{t}_i^*)\right| \leq 2\tau C\|\lambda\|_{1,X}\sqrt{\frac{1}{n}\max\left\{\left(\frac{\mathcal{R}_n + \tau}{2\tau}\right)^2, \ln k\right\}}. \tag{4.22}$$

□

Remark 4.8. Note that Theorem 4.7 does not require knowledge of the Rademacher complexities of the families $\{\widetilde{K}_m(\mathbf{x}, \cdot)\}$, which might not be related in a simple way to that of $\{K(\mathbf{x}, \cdot)\}$ (differently from Proposition 4.4, where a very special structure of $\widetilde{\lambda}_m(\mathbf{t})$ has been used).

Remark 4.9. If $((\mathcal{R}_n + \tau)/2\tau)^2 \geq \ln k$ (in particular, this holds for the scalar case $k = 1$), one simply gets

$$\sup_{\mathbf{x}\in X}\left|f_m(\mathbf{x}) - \frac{\|\lambda\|_{1,X}}{n}\sum_{i=1}^{n} c_{m,i} K(\mathbf{x}, \mathbf{t}_i)\right| \leq C\|\lambda\|_{1,X}(\mathcal{R}_n + \tau)\sqrt{\frac{1}{n}},$$

which has the same form of the bound given in Theorem 4.1 but holds simultaneously for all the components of $\mathbf{f}$.

Corollary 4.10. *Let $X \subset \mathbb{R}^d$ be a compact domain, $K : X \times X \to \mathbb{R}$ continuous, $\tau > 0$ such that, for all $\mathbf{x}$ and $\mathbf{t}$, one has $|K(\mathbf{x}, \mathbf{t})| \leq \tau$, and let $\mathcal{R}_n$ be the Rademacher complexity of the family $\{K(\mathbf{x}, \cdot)\}$. Let $0 \neq \lambda \in \mathcal{L}_1(X)$ be a nonnegative function on X and $\mathbf{f} : X \to \mathbb{R}^k$ such that every component has the representation $f_m(\mathbf{x}) = \int_X \widetilde{K}_m(\mathbf{x}, \mathbf{t})\lambda(\mathbf{t})d\mathbf{t}$. For $m = 1, \ldots, k$, set $\widetilde{K}_m(\mathbf{x}, \mathbf{t}) = K(\mathbf{x}, \mathbf{t})\widetilde{\lambda}_m(\mathbf{t})$, where each $\widetilde{\lambda}_m$ is a measurable function such that $\|\widetilde{\lambda}_m\|_{\infty,X} \leq 1$. Then for each $\epsilon > 0$ it is possible to approximate in the supremum norm on X each component f_m of $\mathbf{f}$ with an error at most ϵ by using approximations of the form*

$$\widehat{f}_m(\mathbf{x}) = \frac{\|\lambda\|_{1,X}}{\widetilde{n}}\sum_{i=1}^{\widetilde{n}} c_{m,i} K(\mathbf{x}, \mathbf{t}_i), \tag{4.23}$$

where $\mathbf{t}_i \in X$ *and* $c_{m,i} \in \{-1,+1\}$, *for* $i = 1,\ldots,\tilde{n}$ *and* $m = 1,\ldots,k$, *provided that*

$$\tilde{n} \geq \frac{C^2\|\lambda\|_{1,X}^2}{\epsilon^2}\max\left\{(R_n+\tau)^2,\ 4\tau^2\ln k\right\}, \tag{4.24}$$

where C *is an absolute constant.*

Proof. It follows immediately by applying Theorem 4.7, by choosing n such that the approximation error for each of the k components of $\mathbf{f}$ is upper bounded by ϵ. □

5. Comparison between the scalar and the vector approximation schemes

To approximate in the supremum norm on X all the k components of a function $\mathbf{f} : X \subset \mathbb{R}^d \to \mathbb{R}^k$ with an error at most $\epsilon > 0$, the vector approximation scheme (4.23) requires

(i) $\tilde{n}k$ binary parameters,

(ii) $\tilde{n}d$ real parameters,

where $\tilde{n} \geq (C^2\|\lambda\|_{1,X}^2/\epsilon^2)\max\{(R_n+\tau)^2,\ 4\tau^2\ln k\}$.

To the same end, the scalar approximation scheme (4.2) requires

(i) $\overline{n}$ binary parameters,

(ii) $\overline{n}\,d$ real parameters,

where $\overline{n} = \sum_{m=1}^{k} n_m$ and $n_m \geq C^2(R_{n_m}+\tau)^2\|\lambda_m\|_{1,X}^2/\epsilon^2$.

In contrast to (4.2), in (4.23) the real parameters $\mathbf{t}_i$ are the same for all the components. In the following, we compare the lower bounds on the numbers of real parameters given by (4.2) and (4.23). To this end, we focus on a case in which the dependence of the Rademacher complexity $\mathcal{R}_n$ on n can be estimated from above.

Recall that a *reproducing kernel Hilbert space (RKHS)* is a Hilbert space $\mathscr{H}$ formed by functions defined on a nonempty set X such that for every $\mathbf{x} \in X$ the evaluation functional $\mathcal{F}_\mathbf{x}$, defined for any $f \in \mathscr{H}$ as $\mathcal{F}_\mathbf{x}(f) = f(\mathbf{x})$, is bounded [16, 19]. RKHSs can be characterized in terms of *kernels*, which for $X \subseteq \mathbb{R}^d$ are *symmetric positive-semidefinite* functions $K : X \times X \to \mathbb{R}$, that is, symmetric functions such that for all positive integers m, all $(w_1,\ldots,w_m) \in \mathbb{R}^m$, and all $(\mathbf{x}_1,\ldots,\mathbf{x}_m) \in X^m$ satisfy the condition $\sum_{i,j=1}^{m} w_i w_j K(\mathbf{x}_i,\mathbf{x}_j) \geq 0$. A kernel is called *positive-definite* if the previous inequality is strict for all $(w_1,\ldots,w_m) \neq (0,\ldots,0)$. If the kernel K is also continuous, then it is called a *Mercer kernel*. For every kernel $K : X \times X \to \mathbb{R}$ and $\mathbf{x} \in X$, we define the function $K_\mathbf{x} : X \to \mathbb{R}$ as

$$K_\mathbf{x}(\cdot) \triangleq K(\mathbf{x},\cdot). \tag{5.1}$$

If there exists a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ such that the kernel K can be written as $K_\mathbf{x}(\mathbf{t}) = \kappa(\mathbf{x}-\mathbf{t})$, then K is called a *convolution kernel* or a *translation-invariant kernel*. By the Riesz representation theorem [36, page 200], for every $\mathbf{x} \in X$ there exists a unique element $K_\mathbf{x} \in \mathscr{H}$, called the *representer* of $\mathbf{x}$, such that $\mathcal{F}_{\mathbf{x}}(f) = \langle f, K_{\mathbf{x}} \rangle$, for all $f \in \mathcal{H}$ (this property is called the *reproducing property*). It is easy to check that the function $K : X \times X$ defined for all $\mathbf{x}, \mathbf{y} \in X$ as $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the ordinary scalar product in $\mathbb{R}^d$ restricted to X, is a kernel. On the other hand, every kernel $K : X \times X \to \mathbb{R}$ generates an RKHS $\mathcal{H}_K(X)$ that is the completion of the linear span of the set $\{K_{\mathbf{x}} : \mathbf{x} \in X\}$, with the inner product defined as $\langle K_{\mathbf{x}}, K_{\mathbf{y}} \rangle_K = K(\mathbf{x}, \mathbf{y})$; we denote by $\|\cdot\|_K$ the induced norm (see, e.g., [16] and [19, page 81]).

For a compact domain $X \subset \mathbb{R}^d$, a positive-definite, continuous kernel $K : X \times X \to \mathbb{R}$, and a probability measure μ on X, we define the integral operator $T_K : \mathcal{L}_2(\mu) \to \mathcal{L}_2(\mu)$ as

$$(T_K f)(\mathbf{x}) = \int_X K(\mathbf{x}, \mathbf{t}) f(\mathbf{t}) d\mu(\mathbf{t}). \tag{5.2}$$

The following proposition gives an upper bound on the Rademacher complexity of the family $G_K = \{K(\mathbf{x}, \cdot)\}$ associate with the kernel K defining T_K. By $\mathrm{Tr}(T_K)$, we denote the *trace* of T_K, that is, the sum of its eigenvalues.

Proposition 5.1. *Let $X \subset \mathbb{R}^d$ be a compact domain, $K : X \times X \to \mathbb{R}$ a positive-definite Mercer kernel, μ a probability measure on X, and $T_K : \mathcal{L}_2(\mu) \to \mathcal{L}_2(\mu)$ the integral operator defined by* (5.2). *Let $\mathcal{H}_K(X)$ be the RKHS associate with the kernel K, $G_K = \{K(\mathbf{x}, \cdot)\}$, and $s_K \triangleq \sup_{\mathbf{x} \in X} \|K(\mathbf{x}, \cdot)\|_K = \max_{\mathbf{x} \in X} \sqrt{K(\mathbf{x}, \mathbf{x})}$. If the largest eigenvalue $l_1(T_K)$ of T_K satisfies $l_1(T_K) \geq 1/n$, then*

$$\mathcal{R}_n(G_K) \leq s_K \sqrt{\mathrm{Tr}(T_K)}.$$

Proof. Let B_K be the unit ball of $\mathcal{H}_K(X)$ centered at 0. By the definition of s_K, $G_K \subseteq s_K B_K$. Then, the statement follows by [20, Theorem 15, parts 1 and 3] and [20, Theorem 16]. □

Let us assume that the required approximation error $\epsilon > 0$ is sufficiently small, so that the assumption $l_1(T_K) \geq 1/n$ of Proposition 5.1 is satisfied with n equal to each n_m in Corollary 4.3 and to $\tilde{n}$ in Corollary 4.10. Then, taking for simplicity of comparison $\|\lambda_m\|_{1,X} = \|\lambda\|_{1,X}$ (which holds, e.g., when the functions $\tilde{\lambda}_m$ take values in $\{\pm 1\}$) and using the upper bound on the Rademacher complexity given in Proposition 5.1, the lower bounds on $\overline{n}$ and $\tilde{n}$ given by Corollary 4.3 and Corollary 4.10 become, respectively,

(i) $\overline{n} \geq k C^2 \left(s_K \sqrt{\mathrm{Tr}(T_K)} + \tau\right)^2 \|\lambda\|_{1,X}^2 / \epsilon^2$;

(ii) $\tilde{n} \geq (C^2 \|\lambda\|_{1,X}^2 / \epsilon^2) \max\{\left(s_K \sqrt{\mathrm{Tr}(T_K)} + \tau\right)^2, 4\tau^2 \ln k\}$.

Inspection of the respective proofs shows that all the absolute positive constants C above are equal. Thus, for a Mercer kernel satisfying the assumptions of Proposition 5.1, the lower bound obtained by applying Corollary 4.10 and Proposition 5.1 is an improvement over the one obtained by applying Corollary 4.3 and Proposition 5.1, at least for a large number k of components.

6. Application to N-stage optimization problems

Let us consider the solution, by the well-known dynamic programming (DP) algorithm [14], of the following *finite-horizon, discrete-time dynamic optimization problem*, modeled as in [37]: given $\mathbf{x}_0 \in X$, find $\mathbf{x}_1, \dots, \mathbf{x}_N \in X$ such that

$$J^o(\mathbf{x}_0) = \sup \left\{ \sum_{t=0}^{N-1} \beta^t h(\mathbf{x}_t, \mathbf{x}_{t+1}) + \beta^N h_N(\mathbf{x}_N) \right\}, \\ \text{where } (\mathbf{x}_t, \mathbf{x}_{t+1}) \in D, \quad t = 0, 1, \dots, N-1. \tag{6.1}$$

The vector $\mathbf{x}_t \in X \subseteq \mathbb{R}^d$ represents the state of a dynamical system, X is the set to which the state vector belongs (*state space*), $D \subseteq X \times X$ is the graph of a *correspondence* that models the transition from one stage to the following one, $h : D \to \mathbb{R}$ is a *transition reward*, $h_N : X \to \mathbb{R}$ is the *final reward* associate with the final stage, $0 < \beta \leq 1$ is a *discount factor*, and J^o is the so-called *value function*. Using an economic terminology, $\beta^t h(\mathbf{x}_t, \mathbf{x}_{t+1})$ and $\beta^N h_N(\mathbf{x}_N)$ are actualized values of transition and final rewards, respectively. For simplicity of notations and without loss of generality, we assume that X, h, and D do not depend on the time t.

Dynamic programming considers the following N subproblems:

$$J_N^o(\mathbf{x}_N) = h_N(\mathbf{x}_N); \tag{6.2}$$

given $\mathbf{x}_i \in X$, find $\mathbf{x}_{i+1}, \dots, \mathbf{x}_N \in X$ such that

$$J_i^o(\mathbf{x}_i) = \sup \left\{ \sum_{t=i}^{N-1} \beta^{t-i} h(\mathbf{x}_t, \mathbf{x}_{t+1}) + \beta^{N-i} h_N(\mathbf{x}_N) \right\}, \quad i = N-1, \dots, 0, \\ \text{where } (\mathbf{x}_t, \mathbf{x}_{t+1}) \in D, \quad t = i, \dots, N-1. \tag{6.3}$$

Under suitable conditions on the problem formulation [38], Theorem 4.1 can be exploited to find sparse approximations for both the ith stage value functions $J_i^o : X \to \mathbb{R}$ and each component of the optimal ith stage policy functions $\mathbf{g}_i^o : X \to X$, $i = N-1, \dots, 0$. For the latter, since each component is dealt with separately, in general one expects that the parameters $\mathbf{t}_1, \dots, \mathbf{t}_n \in X$ and $c_1, \dots, c_n \in \{-1, +1\}$ for which the bound given in Theorem 4.1 holds will be different from one component to the other. In order to reduce the number of parameters needed to obtain a desired accuracy of approximation of the optimal policy functions, it is useful to have at one's disposal an approximation error bound for which some of the parameters are common to the approximators of *all the components* of each optimal policy function. If similarities (correlations) among the components of each optimal policy function to be approximated exist and can be modeled as in Theorem 4.7, the latter gives such an upper bound. In dynamic optimization problems, these similarities can arise in several ways (although their analysis is problem-dependent).

For example, applying the dynamic programming algorithm at stage $N-1$, for each $\mathbf{x}_{N-1} \in X$ one has to find

$$\mathbf{x}_N^o = \mathbf{g}_{N-1}^o(\mathbf{x}_{N-1}) = \arg\max_{\mathbf{x}_N \in X} \left[h(\mathbf{x}_{N-1}, \mathbf{x}_N) + \beta h_N(\mathbf{x}_N) \right]. \tag{6.4}$$

Similarities among the components of $\mathbf{g}^o_{N-1}$ may be present, depending on the properties of h_{N-1} and h_N, which are given in the problem formulation.

Another case in which there can be similarities among the components of the optimal policy function may occur in the case of *infinite-horizon, discrete-time dynamic optimization problems* modeled as follows: given $\mathbf{x}_0 \in X$ and $0 < \beta < 1$, find $\mathbf{x}_1, \mathbf{x}_2, \ldots, \in X$ such that

$$J^o(\mathbf{x}_0) = \sup\left\{\sum_{t=0}^{\infty}\beta^t h(\mathbf{x}_t, \mathbf{x}_{t+1})\right\}, \qquad \text{where } (\mathbf{x}_t, \mathbf{x}_{t+1}) \in D, \quad t = 0, 1, \ldots. \tag{6.5}$$

It is known that under suitable hypotheses an optimal policy has one or more stationary points, that is, there exists $\mathbf{x}_0 \in X$ such that $\mathbf{g}^o(\mathbf{x}_0) = \mathbf{x}_0$. In such a case, if $\mathbf{g}^o$ is Lipschitz continuous (which holds under some conditions; see [37]) and its Lipschitz constant is "sufficiently small," then one has $\mathbf{g}^o(\mathbf{x}) \cong \mathbf{x}$ on a "sufficiently large" neighborhood of $\mathbf{x}_0$.

Acknowledgment

The authors were partially supported by a PRIN grant from the Italian Ministry for University and Research, project "Models and Algorithms for Robust Network Optimization."

References

[1] V. Kůrková and M. Sanguineti, "Comparison of worst case errors in linear and neural network approximation," *IEEE Transactions on Information Theory*, vol. 48, no. 1, pp. 264–275, 2002.

[2] A. R. Barron, "Neural net approximation," in *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems*, K. Narendra, Ed., pp. 69–72, Yale University Press, New Haven, Conn, USA, May 1992.

[3] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[4] F. Girosi, "Approximation error bounds that use VC-bounds," in *Proceedings of the International Conference on Artificial Neural Networks (ICANN '95)*, pp. 295–302, Paris, France, October 1995.

[5] Y. Makovoz, "Uniform approximation by neural networks," *Journal of Approximation Theory*, vol. 95, no. 2, pp. 215–228, 1998.

[6] V. E. Maiorov, "On best approximation by ridge functions," *Journal of Approximation Theory*, vol. 99, no. 1, pp. 68–94, 1999.

[7] A. Pinkus, "Approximation theory of the MLP model in neural networks," *Acta Numerica*, vol. 8, no. 1, pp. 143–195, 1999.

[8] M. A. Kon, L. A. Raphael, and D. A. Williams, "Extending Girosi's approximation estimates for functions in Sobolev spaces via statistical learning theory," *Journal of Analysis and Applications*, vol. 3, no. 2, pp. 67–90, 2005.

[9] M. A. Kon and L. A. Raphael, "Approximating functions in reproducing kernel Hilbert spaces via statistical learning theory," in *Wavelets and Splines: Athens 2005*, G. Chen and M.-J. Lai, Eds., Modern Methods in Mathematics, pp. 271–286, Nashboro Press, Brentwood, Tenn, USA, 2006.

[10] G. Gnecco and M. Sanguineti, "Approximation error bounds via Rademacher's complexity," *Applied Mathematical Sciences*, vol. 2, no. 1–4, pp. 153–176, 2008.

[11] V. Kůrková and M. Sanguineti, "Geometric upper bounds on rates of variable-basis approximation," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5681–5688, 2008.

[12] V. Kůrková and M. Sanguineti, "Learning with generalization capability by kernel methods of bounded complexity," *Journal of Complexity*, vol. 21, no. 3, pp. 350–367, 2005.

[13] V. Kůrková and M. Sanguineti, "Approximate minimization of the regularized expected error over kernel models," *Mathematics of Operations Research*, vol. 33, no. 3, pp. 747–756, 2008.

[14] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, USA, 1957.

[15] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

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

[17] J. Buescu and A. C. Paixão, "Inequalities for differentiable reproducing kernels and an application to positive integral operators," *Journal of Inequalities and Applications*, vol. 2006, Article ID 53743, 9 pages, 2006.

[18] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2002.

[19] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Function*, vol. 100 of *Graduate Texts in Mathematics*, Springer, New York, NY, USA, 1984.

[20] S. Mendelson, "A few notes on statistical learning theory," in *Advanced Lectures on Machine Learning*, S. Mendelson and A. Smola, Eds., vol. 2600 of *Lecture Notes in Computer Science*, pp. 1–40, Springer, Berlin, Germany, 2003.

[21] G. Blanchard, O. Bousquet, and L. Zwald, "Statistical properties of kernel principal component analysis," *Machine Learning*, vol. 66, no. 2-3, pp. 259–294, 2007.

[22] S. Smale and D.-X. Zhou, "Shannon sampling and function reconstruction from point values," *Bulletin of the American Mathematical Society*, vol. 41, no. 3, pp. 279–305, 2004.

[23] S. Smale and D.-X. Zhou, "Learning theory estimates via integral operators and their approximations," *Constructive Approximation*, vol. 26, no. 2, pp. 153–172, 2007.

[24] F. Cucker and D.-X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK, 2007.

[25] D. P. Bertsekas, *Dynamic Programming and Optimal Control. Vol. II*, Athena Scientific, Belmont, Mass, USA, 2nd edition, 2001.

[26] F. A. Mussa-Ivaldi, "From basis functions to basis fields: vector field approximation from sparse data," *Biological Cybernetics*, vol. 67, no. 6, pp. 479–489, 1992.

[27] F. A. Mussa-Ivaldi and F. Gandolfo, "Networks that approximate vector-valued mappings," in *Proceedings of IEEE International Conference on Neural Networks (ICNN '93)*, vol. 3, pp. 1973–1978, San Francisco, Calif, USA, March-April 1993.

[28] C. A. Micchelli and M. Pontil, "On learning vector-valued functions," *Neural Computation*, vol. 17, no. 1, pp. 177–204, 2005.

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

[30] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, NY, USA, 3rd edition, 1987.

[31] C. S. Ong, X. Mary, S. Canu, and A. J. Smola, "Learning with non positive kernels," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, pp. 639–646, Banff, Canada, July 2004.

[32] E. Sontag, "VC dimension of neural networks," in *Neural Networks and Machine Learning*, C. Bishop, Ed., pp. 69–95, Springer, Berlin, Germany, 1998.

[33] A. N. Kolmogorov and S. V. Fomīn, *Introductory Real Analysis*, Dover, New York, NY, USA, 1975.

[34] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

[35] O. L. Mangasarian, J. B. Rosen, and M. E. Thompson, "Convex kernel underestimation of functions with multiple local minima," *Computational Optimization and Applications*, vol. 34, no. 1, pp. 35–45, 2006.

[36] A. Friedman, *Foundations of Modern Analysis*, Dover, New York, NY, USA, 1982.

[37] L. Montrucchio, "Lipschitz continuous policy functions for strongly concave optimization problems," *Journal of Mathematical Economics*, vol. 16, no. 3, pp. 259–273, 1987.

[38] G. Gnecco and M. Sanguineti, "Suboptimal solutions to dynamic optimization problems via approximations of the policy functions," *Journal of Optimization Theory and Applications*, to appear.