

Approximation Error Bounds via Rademacher's Complexity

Giorgio Gnecco

Department of Mathematics (DIMA), University of Genova
Via Dodecaneso 35, 16146 Genova, Italy

and

Department of Communications, Computer, and System Sciences (DIST)
University of Genova, Via Opera Pia 13, 16145 Genova, Italy
giorgio.gnecco@dist.unige.it

Marcello Sanguineti

Department of Communications, Computer, and System Sciences (DIST)
University of Genova, Via Opera Pia 13, 16145 Genova, Italy
marcello@dist.unige.it

Abstract

Approximation properties of some connectionistic models, commonly used to construct approximation schemes for optimization problems with multivariable functions as admissible solutions, are investigated. Such models are made up of linear combinations of computational units with adjustable parameters. The relationship between model complexity (number of computational units) and approximation error is investigated using tools from Statistical Learning Theory, such as Talagrand's inequality, fat-shattering dimension, and Rademacher's complexity. For some families of multivariable functions, estimates of the approximation accuracy of models with certain computational units are derived in dependence of the Rademacher's complexities of the families. The estimates improve previously-available ones, which were expressed in terms of VC dimension and derived by exploiting union-bound techniques. The results are applied to approximation schemes with certain radial-basis-functions as computational units, for which it is shown that the estimates do not exhibit the curse of dimensionality with respect to the number of variables.

Keywords: approximation error, model complexity, curse of dimensionality, Rademacher's complexity, Talagrand's inequality, union bounds, VC dimension.

1 Introduction

Various conditions have been derived, under which linear combinations of computational units containing adjustable parameters are able to approximate with arbitrary accuracy continuous or square-integrable multivariable functions. This approximation capability, also known as “universal approximation property”, is exhibited, e.g., by radial-basis-functions, hinging hyperplanes, free-knot splines, etc. (see, e.g., [1, 2, 3, 4, 5] and the references therein). In these models, belonging to the so-called *variable-basis approximation schemes* [5], the number of computational units (i.e., the basis functions) plays the role of a measure of *model complexity* [6] and is critical for feasibility of implementation: if universality is obtained at the price of a very large model complexity, then approximation is not computationally efficient.

Variable-basis approximation schemes have been widely used to find sub-optimal solutions to optimization problems such as traffic control, routing in telecommunication networks, management of water resources, inventory forecasting, exploration of stochastic graphs, etc. (see [7, 8] and the references therein), in which the admissible solutions depend on a large number d of variables (e.g., the message queues at the nodes of a large-scale telecommunication network and the delays on the network’s links). When, to guarantee a desired degree of accuracy of suboptimal solutions, the model complexity, i.e., the number n of computational units, has to grow fast with d , one may incur the so-called *curse of dimensionality* [9], which makes optimization problems computationally intractable. For a given computational model, *tractability* depends on the family of functions one wants to approximate, the kind of computational units, and the measure used to evaluate the approximation error.

Some insights into characteristic features of the families of functions that can be efficiently approximated by variable-basis functions with certain computational units (thus making the associated problems tractable), can be obtained from inspection of the dependence of the approximation error on model complexity. The purpose of this work is to improve previous results by Girosi [10] and Kon, Raphael, & Williams [11, 12], in which estimates of the approximation error were derived for some families of functions approximated by certain variable-basis approximation schemes.

Girosi [10] was the first to exploit estimation bounds from *Statistical Learning Theory* (SLT) to derive approximation bounds. For functions on $X \subseteq \mathbb{R}^d$ having an integral representation as the convolution $K * \lambda$ of an $\mathcal{L}_1(\mathbb{R}^d)$ function λ with a bounded kernel $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, he estimated the $\mathcal{L}_\infty(\mathbb{R}^d)$ error in approximation by linear combinations of $K(\cdot, \mathbf{t}_1), \dots, K(\cdot, \mathbf{t}_n)$, with $\mathbf{t}_1, \dots, \mathbf{t}_n$ varying in \mathbb{R}^d , which is a variable-basis approximation scheme. As an example, he considered for $r > 0$ the rotation-invariant Bessel potential kernel $\beta_r(\cdot, \mathbf{0})$

(i.e., the function whose Fourier transform is equal to $(2\pi)^{-\frac{d}{2}}(1 + \|\mathbf{s}\|^2)^{-r/2}$), which determines a radial-basis-functions approximation scheme (i.e., the computational units are radially symmetric) formed by linear combinations of $\beta_r(\cdot, \mathbf{t}_1), \dots, \beta_r(\cdot, \mathbf{t}_n)$, with $\mathbf{t}_1, \dots, \mathbf{t}_n$ varying in \mathbb{R}^d .

Kon, Raphael, & Williams [11, 12] recently extended Girosi's results, which require the existence of the \mathcal{L}_1 -norm of the function λ in the convolution $K * \lambda$. In particular, in [11] Girosi's [10] estimate by linear combinations of translates of β_r was extended to the approximation of functions for which the representation $K * \lambda$ holds with $\lambda \in \mathcal{L}_p$ ($1 < p < \infty$), with the error measured in a weighted essential supremum norm. In [12], Kon, Raphael, & Williams used Girosi's [10] approach to derive error bounds for approximation in certain Hilbert spaces, called *Reproducing Kernel Hilbert Spaces* (RKSHs; see [13] and [14, Section III.3]).

The works [10, 11, 12] exploit a main tool from SLT, i.e., the well-known theorem, by Vapnik and Chervonenkis [15], that 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, called *VC dimension*.

In the last years, new estimation bounds for SLT were derived. They improve significantly the previous known bounds like those based on the *VC dimension*. A survey of these new results, which are expressed in terms of a quantity called *Rademacher's complexity*, is [16]. There are several reasons for which these estimates are better than the previous ones. For our purposes, it is enough to recall that classical SLT bounds are obtained by "reducing" a family of functions to a simpler one via ε -coverings, then applying to such a simpler family the *union-bound technique* [16]. The use of union bounds may cause a loss of tightness in the final estimates. Techniques based on Rademacher's complexity, which apply *Talagrand's inequality* [17] instead of union bounds, result in tighter final estimates.

In this paper, we exploit these new tools from SLT in the contexts examined in [10, 11, 12] to improve the approximation bounds for the families of functions considered therein. Taking the hint from Girosi's idea [10, Section 3] of exploiting the structural similarity between an integral representation with a kernel $K(\mathbf{x}, \mathbf{y})$ and the definition of expected risk, we prove for some families F of functions defined on a set X , upper bounds on the error of approximation by linear combinations of functions $K(\cdot, \mathbf{t}_i), \dots, K(\cdot, \mathbf{t}_n)$, with $\mathbf{t}_1, \dots, \mathbf{t}_n$ varying in X , in terms of the Rademacher's complexity of F . Then, we derive an upper bound on Rademacher's complexity in terms of a quantity known as *Dudley's integral* which, in turn, we bound from above in terms of another quantity used in SLT, namely, the *fat-shattering dimension* of the family F . To compare our estimates with those from [10, 11, 12], which are expressed in terms of *VC dimension*, we bound from above the fat-shattering dimension in

terms of the VC dimension.

We apply our estimates to approximation schemes with radial-basis-functions computational units. For Bessel and Gaussian basis functions, we derive upper bounds that do not exhibit the curse of dimensionality with respect to the number d of variables of the functions to be approximated. For Bessel basis functions, the rates are of the same order as those that can be derived by [4, Corollary 8.4], using different techniques. For Gaussian basis functions with varying widths, the error bounds derived in [10] are improved in this paper. Moreover, we obtain such improvements as by-products of our estimates stated in terms of Rademacher's complexity, which might provide even better rates.

The paper is organized as follows. Section 2 describes notations and gives definitions. Section 3 shortly reviews and discusses the literature on approximation error bounds derived by using classical SLT bounds and VC dimension. Section 4 contains our estimates: Section 4.1 describes improvements allowed by exploiting Talagrand's inequality and Rademacher's complexity, and Section 4.2 combines such tools to derive improved upper bounds on the approximation error. Section 5 applies the estimates to some radial-basis-functions schemes with Bessel potentials and Gaussians as computational units. Section 6 is a brief discussion. To make the paper self-contained, the results from SLT, exploited to prove the estimates, are reported in the Appendix.

2 Notations and definitions

By \mathbb{R} and \mathbb{R}_+ we denote the sets of real and positive real numbers, resp., and by \mathbb{N} and \mathbb{N}_+ , the sets of natural numbers and positive integers, resp. For $a \in \mathbb{R}$, $\lceil a \rceil$ is the smallest integer $n \geq a$.

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 $f \in \mathcal{H}$, i.e.,

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

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 $B_r(\|\cdot\|)$, resp.

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 p -th power with respect to the Lebesgue measure, endowed with the $\mathcal{L}_p(X)$ -norm $\|\cdot\|_{p,X}$. $\mathcal{L}_\infty(X)$ is the space of (equivalence classes of) real-valued functions on X which are essentially bounded with respect to the Lebesgue measure, endowed with the essential supremum norm $\|\cdot\|_{\infty,X}$. $\mathcal{C}(X)$ is the space of continuous functions on X with the supremum

norm. To simplify the notations, for $r \geq 0$ we let $B_{r,\infty,X} \triangleq B_r(\|\cdot\|_{\infty,X})$. Whenever there is no ambiguity, we omit X from the notations.

For $F \subseteq (\mathcal{H}, \|\cdot\|)$ and $\varepsilon > 0$, $\{f_1, \dots, f_n\} \subseteq F$ is called an ε -net in F , if the family of open balls of radii ε centered at f_i covers F , i.e., if $F \subseteq \bigcup_{i=1}^n B_\varepsilon(f_i, \|\cdot\|)$. A set $\{f_1, \dots, f_n\}$ is called ε -separated if for each distinct pair $i, j \in \{1, \dots, n\}$, one has $\|f_i - f_j\| \geq \varepsilon$. If a set F contains a 2ε -separated subset of size n , then every ε -net in F must contain at least n elements. The ε -covering number $\mathcal{N}(F, \|\cdot\|, \varepsilon)$ of a subset F of $(\mathcal{H}, \|\cdot\|)$ in the metric induced by the norm $\|\cdot\|$ is the cardinality of a minimal ε -net in F , i.e.,

$$\mathcal{N}(F, \|\cdot\|, \varepsilon) \triangleq \min \left\{ n \in \mathbb{N}_+ \mid \exists f_1, \dots, f_n \in F \text{ such that } F \subseteq \bigcup_{i=1}^n B_\varepsilon(f_i) \right\}.$$

We set $\mathcal{N}(F, \|\cdot\|, \varepsilon) = +\infty$ if the set over which the minimum is taken is empty. When the norm is clear from the context, we merely write $\mathcal{N}(F, \varepsilon)$. Note that here we consider covering numbers defined in terms of open balls, as in [18, p. 148], but other Authors (e.g., [19, 20]) use closed balls.

A kernel on $X \subseteq \mathbb{R}^d$ is a symmetric positive-semidefinite function $K : X \times X \rightarrow \mathbb{R}$, i.e., a symmetric function that for all positive integers m , all $(w_1, \dots, w_m) \in \mathbb{R}^m$, and all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in X^m$ satisfies the condition $\sum_{i,j=1}^m w_i w_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$. For every kernel $K : X \times X \rightarrow \mathbb{R}$ and $\mathbf{x} \in X$, we define the function $K_{\mathbf{x}} : X \rightarrow \mathbb{R}$ as

$$K_{\mathbf{x}}(\mathbf{t}) \triangleq K(\mathbf{x}, \mathbf{t}) \quad \forall \mathbf{t} \in X.$$

If there exists a function $k : \mathbb{R}^d \rightarrow \mathbb{R}$ such that the kernel K can be written as $K_{\mathbf{x}}(\mathbf{t}) = k(\mathbf{x} - \mathbf{t})$, then K is called a *convolution kernel*.

For a positive integer d , a set $X \subseteq \mathbb{R}^d$, and a family F of functions on X , we denote by $F_{\mathbf{x}} : X \rightarrow \mathbb{R}$ a function in F , where \mathbf{x} is a parameter used to identify elements in F ¹.

By P_X we denote a (possibly unknown) probability distribution on X , and we write merely P when the set X is clear from the context.

The *expected risk* associated with a function $F_{\mathbf{x}} \in F$ is defined as

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

So, $R(F_{\mathbf{x}}) = \mathbb{E}_{P_X} \{F_{\mathbf{x}}(\mathbf{t})\}$, where \mathbb{E}_{P_X} is the expectation operator; we write merely \mathbb{E} when X and $P_X = P$ are clear from the context.

¹We use the notation $F_{\mathbf{x}}$ since, as it will be clear in Section 3, we shall consider functions on X having the integral representation $f(\mathbf{x}) = \int_X K_{\mathbf{x}}(\mathbf{t}) \lambda(\mathbf{t}) d\mathbf{t}$ for a kernel $K : X \times X \rightarrow \mathbb{R}$ and families F such that $F_{\mathbf{x}}(\mathbf{t})$ is defined via the kernel $K_{\mathbf{x}}(\mathbf{t})$. So the parameter \mathbf{x} used to identify the elements of F will be a point $\mathbf{x} \in X$.

For every positive integer n , we say that a sequence $\{\mathbf{t}_i\}$ of n points obtained by sampling X independently n times according to P_X , is a P_X -i.i.d. sequence.

The *empirical risk* associated with the function $F_{\mathbf{x}}(\mathbf{t}) \in F$ and the sequence $\{\mathbf{t}_i\}$ of samples is defined as

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

Finally, we remark that throughout the paper, C, C_1 , and C_2 denote absolute constants that may take on different values in different formulas.

3 Available approximation bounds derived via classical SLT tools

Previous works deriving approximation bounds via classical SLT tools are the papers by Girosi [10] and its improvements by Kon, Raphael, & Williams [11] and Kon & Raphael [12].

Recall that the *Vapnik-Chervonenkis dimension* (*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 classes Ω_1 and Ω_2 in all 2^h possible ways, by using functions (with argument \mathbf{t}) of the form $F_{\mathbf{x}}(\mathbf{t}) - \alpha$, as the parameters \mathbf{x} and α vary in X and \mathbb{R} , resp. [15]. In particular, if for $\mathbf{t} \in X$

$$F_{\mathbf{x}}(\mathbf{t}) - \alpha \geq 0,$$

then we say that $(F_{\mathbf{x}}, \alpha)$ assigns \mathbf{t} to the class Ω_1 . Similarly, if

$$F_{\mathbf{x}}(\mathbf{t}) - \alpha < 0,$$

then $(F_{\mathbf{x}}, \alpha)$ assigns \mathbf{t} to the class Ω_2 .

Expected risk, empirical risk, and *VC* dimension are related to each other by the following classical result (which here we state for the particular case of a family $F = \{F_{\mathbf{x}}\}$ of functions defined on X , where the parameter \mathbf{x} is a point of X).

Theorem 3.1 (Vapnik & Chervonenkis [15]) *Let F be a family of functions on $X \subseteq \mathbb{R}^d$, $A, B \in \mathbb{R}$ such that for every $\mathbf{x}, \mathbf{t} \in X$ one has $A \leq F_{\mathbf{x}}(\mathbf{t}) \leq B$, h the *VC* dimension of F , P_X a probability distribution on X , and $\{\mathbf{t}_i\}$ a P_X -i.i.d. sequence. For every $0 < \delta < 1$, the following bound holds with probability at least $1 - \delta$:*

$$\sup_{F_{\mathbf{x}} \in F} \left| R(F_{\mathbf{x}}) - R_{\text{emp}}(F_{\mathbf{x}}, \{\mathbf{t}_i\}) \right| \leq (B - A) \sqrt{\frac{h \ln \frac{2en}{h} - \ln \frac{\delta}{4}}{n}}.$$

Let us now consider a family of functions having the integral representation

$$f(\mathbf{x}) = \int_X K_{\mathbf{x}}(\mathbf{t}) \lambda(\mathbf{t}) \, d\mathbf{t},$$

where $K : X \times X \rightarrow \mathbb{R}$ is a kernel and $\lambda \in \mathcal{L}_1(X)$. To derive approximation bounds via Theorem 3.1, Girosi [10] noted that if $\|\lambda\|_1 \neq 0$, then for every $\mathbf{x} \in X$, $f(\mathbf{x})/\|\lambda\|_1$ can be considered as the expected risk relative to the function $F_{\mathbf{x}}(\mathbf{t}) = K_{\mathbf{x}}(\mathbf{t})\text{sgn}(\lambda(\mathbf{t}))$ and the probability density $|\lambda(\mathbf{t})|/\|\lambda\|_1$. By applying Theorem 3.1 and letting $\delta \rightarrow 1$, he obtained the following bound on the $\mathcal{L}_\infty(X)$ error in approximating f by linear combinations of the n functions $K_{\mathbf{x}}(\mathbf{t}_1), \dots, K_{\mathbf{x}}(\mathbf{t}_n)$ centered at suitable points $\mathbf{t}_1, \dots, \mathbf{t}_n \in X$.

Theorem 3.2 (Girosi [10, 11]) *Let $X \subseteq \mathbb{R}^d$, $\lambda \in \mathcal{L}_1(X)$, $K : X \times X \rightarrow \mathbb{R}$ be a kernel such that there exists $\tau > 0$ with $|K_{\mathbf{x}}(\mathbf{t})| \leq \tau$ for every $\mathbf{x}, \mathbf{t} \in X$, f a function with the representation $f(\mathbf{x}) = \int_X K_{\mathbf{x}}(\mathbf{t}) \lambda(\mathbf{t}) \, d\mathbf{t}$, and h the VC dimension of the family $F = \{K_{\mathbf{x}}\}$. Then for every positive integer n , there exist $\mathbf{t}_1, \dots, \mathbf{t}_n \in X$ and $c_1, \dots, c_n \in \{-1, 1\}$ such that*

$$\left\| f(\mathbf{x}) - \frac{\|\lambda\|_1}{n} \sum_{i=1}^n c_i K_{\mathbf{x}}(\mathbf{t}_i) \right\|_\infty \leq 4\tau \|\lambda\|_1 \sqrt{\frac{h \ln \frac{2en}{h} + \ln 4}{n}}.$$

Then Girosi [10] applied Theorem 3.2 to $X = \mathbb{R}^d$ and functions of the form $f = \beta_r * \lambda$, where β_r is the *Bessel potential function*, i.e., the inverse Fourier transform of $(2\pi)^{-\frac{d}{2}}(1 + \|\mathbf{s}\|^2)^{-r/2}$.

His result was extended in [11, 12] to the following families of functions.

- In [11, Corollary 3], to the so-called *Sobolev potential spaces* (also called *Sobolev-Liouville spaces*), defined for $p \geq 1$ and $\lambda \in \mathcal{L}_p(\mathbb{R}^d)$ as $\mathcal{L}_p^r \triangleq \{f \in \mathcal{L}_p(\mathbb{R}^d) \mid f = \beta_r * \lambda\}$.
- In [12], to *Reproducing Kernel Hilbert Spaces (RKHSs)*, i.e., Hilbert spaces of functions f on X such that, for every $\mathbf{x} \in X$, the evaluation functional at \mathbf{x} , defined as $\mathcal{F}_{\mathbf{x}}(f) \triangleq f(\mathbf{x})$, is bounded (see [13], [14, Section III.3]). As every RKHS on a set X can be characterized in terms of a kernel, we denote such spaces as \mathcal{H}_K . Two cases in particular were considered:
 - RKHSs that are dense in \mathcal{L}_2 . In this case, for a function $f \in \mathcal{H}_K$, the integral representation $f(\mathbf{x}) = \int_{\mathbb{R}^d} K_{\mathbf{x}}^{1/2}(\mathbf{t})(L_K^{-1/2}f)(\mathbf{t}) \, d\mathbf{t}$ is used in [12, Theorem 4]. Here $L_K^{-1/2}$ is the $(-1/2)$ -power of the operator $L_K : \mathcal{L}_2 \rightarrow \mathcal{L}_2$ defined as $(L_K f)(\mathbf{x}) \triangleq \int_{\mathbb{R}^d} K_{\mathbf{x}}(\mathbf{t})f(\mathbf{t}) \, d\mathbf{t}$, and $K_{\mathbf{x}}^{1/2}(\mathbf{t})$ is the kernel associated with the operator $L_K^{1/2}$;

- RKHSs that are closed subspaces of \mathcal{L}_2 , with the inherited inner product, but that are not dense in \mathcal{L}_2 . In this case, for a function $f \in \mathcal{H}_K$, the integral representation $f(\mathbf{x}) = \int_{\mathbb{R}^d} K_{\mathbf{x}}(\mathbf{t})f(\mathbf{t}) d\mathbf{t}$ is used in [12, Proposition 5].

To improve the results from [10, 11, 12], which are based on the classical SLT theory (VC dimension and Theorem 3.1), we shall exploit recent SLT bounds that improve classical ones.

4 Improved estimates

4.1 Exploiting new tools from SLT

Recall that the main goal of SLT consists in obtaining non-asymptotical, probabilistic, and uniform (i.e., holding simultaneously for every element of a family F) bounds on the difference between the expected risk and the empirical risk associated with an element $F_{\mathbf{x}} \in F$, i.e.:

$$\text{Prob}\left\{\sup_{F_{\mathbf{x}} \in F} |R(F_{\mathbf{x}}) - R_{\text{emp}}(F_{\mathbf{x}}, \{\mathbf{t}_i\})| \leq \varepsilon\right\} \leq \delta(n, q),$$

or, equivalently,

$$\text{Prob}\left\{\sup_{F_{\mathbf{x}} \in F} |R(F_{\mathbf{x}}) - R_{\text{emp}}(F_{\mathbf{x}}, \{\mathbf{t}_i\})| \leq \varepsilon(n, q)\right\} \leq \delta,$$

where n is the number of samples, and q is a parameter dependent on the “complexity” of the family F . In Section 3, the role of q was played by the VC dimension of F ; in the following, it will be played by the Rademacher’s complexity of F , to be defined later on.

The approximation bounds reviewed in Section 3 can be improved avoiding some limitations intrinsic in the procedure in which classical SLT bounds are obtained [16]. We briefly outline such a procedure.

- i) The family F , which might have infinite cardinality, is “approximated” by a “simpler” family F_{ε} , whose finite cardinality is equal to the ε -covering number (in an appropriate metric) of F .
- ii) For each function in F_{ε} a SLT bound is derived, typically via Hoeffding’s inequality or Bernstein’s inequality [21].
- iii) A uniform SLT bound (i.e., a SLT bound for the whole family F_{ε}) is obtained by combining the bounds in ii) by means of the *union-bound technique*, based on the following well-known result from probability theory: for every positive integer k , given k events $\mathcal{A}_1, \dots, \mathcal{A}_k$ one has $P\{\bigcup_{i=1}^k \mathcal{A}_i\} \leq \sum_{i=1}^k P(\mathcal{A}_i)$.

- iv) A uniform SLT bound for the original family F is derived using the relationship between F and F_ε .

One weak point of the procedure outlined above is iii) because, usually, $\sum_{i=1}^k P(\mathcal{A}_i)$ is not a “good approximation” of $P\{\bigcup_{i=1}^k \mathcal{A}_i\}$. Better SLT bounds for the original family F can be derived by using, instead of the union bound technique, a probabilistic inequality known as *Talagrand's inequality* [16, 17]. This approach is a significant improvement over the classical one, because Talagrand's inequality:

- gives a uniform bound without requiring one to apply the union-bound technique;
- holds also for family of functions of infinite cardinality, so there is no need to approximate F with a family F_ε of finite cardinality.

For the sake of completeness, in Theorem 7.1 of the Appendix we give a modified version of Talagrand's inequality, in the form presented in [16].

The right-hand side of Talagrand's inequality (see Theorem 7.1) can be bounded from above in terms of a quantity called Rademacher's complexity (or Rademacher's average) of a family of functions. Recall that a *Rademacher's random variable* is a random variable taking only the values $+1$ and -1 with equal probability [16]. Let P_X be a probability distribution on X , $\{\mathbf{t}_i\}$ a P_X -i.i.d. sequence, and $\{\varepsilon_i\}$ a sequence of n independent Rademacher's random variables. Given a family $F = \{F_{\mathbf{x}}\}$ of functions on X , the *Rademacher's complexity* of F is defined as [16]

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

Using Talagrand's inequality and Rademacher's complexity, the SLT bound reported in Theorem 7.2 of the Appendix was derived. In this section, we shall exploit it to improve the approximation bounds from [10, 11, 12]. We shall proceed as follows.

- We shall express Theorem 7.2, involving the Rademacher's complexity of F , in a form similar to that of Theorem 3.1.
- Proceeding as in [10], for functions having an integral representation as the convolution $K * \lambda$, with $\lambda \in \mathcal{L}_1$ and $K : X \times X \rightarrow \mathbb{R}$ bounded, we shall derive an estimate on the \mathcal{L}_∞ error in approximation by linear combinations of $K(\cdot, \mathbf{t}_1), \dots, K(\cdot, \mathbf{t}_n)$ with $\mathbf{t}_1, \dots, \mathbf{t}_n \in X$.

- In order to compare the results with those reported in [10, 11, 12], which are expressed in terms of VC dimension, we shall bound as follows Rademacher's complexity in terms of VC dimension.
 - First, we shall exploit a result known as *Dudley's integral* (see [16]), which we report in the Appendix as Theorem 7.3. Roughly speaking, it allows one to obtain an upper bound on the Rademacher's complexity of a family F in terms of an integral of the ε -covering number $\mathcal{N}(\varepsilon, F, \mathcal{L}_2(\mu_n))$ of F in the $\mathcal{L}_2(\mu_n)$ measure (μ_n is the empirical measure supported on n samples).
 - Then, we shall find an upper bound on Dudley's integral by using the upper bound from Theorem 7.4 on $\mathcal{N}(\varepsilon, F, \mathcal{L}_2(\mu_n))$, in terms of a quantity called ε -fat-shattering dimension of F [16].
 - Finally, we shall bound from above the ε -fat-shattering dimension by the V_ε dimension and the latter by the VC dimension. To this end, we shall exploit Propositions 7.5 and 7.6.

4.2 Approximation bounds

From Theorem 7.2 we obtain immediately the next result, which is a reformulation of the SLT bound from Theorem 7.2 in a form similar to that of Theorem 3.1.

Theorem 4.1 *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, and F a family of $[0, 1]$ -valued functions with Rademacher's complexity \mathcal{R}_n . There exists an absolute constant C such that for all $0 < \delta < 1$, with probability at least $1 - \delta$ one has:*

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

Note that in Theorem 4.1 the constraint on n , which appeared in the formulation of Theorem 7.2, is implicitly contained in the second side of equation (1).

Since in the following we will have to consider classes of $[-\tau, \tau]$ -valued functions ($\tau > 0$), the following slight modification of Theorem 4.1 is more useful in order to proceed.

Theorem 4.2 *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, and F a family of $[-\tau, \tau]$ -valued functions ($\tau > 0$) with Rademacher's complexity \mathcal{R}_n . There exists an absolute constant C such that for all $0 < \delta < 1$, with probability at least $1 - \delta$ one has:*

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

Proof. The case of a class of $[-\tau, \tau]$ -valued functions can be reconducted to that of Theorem 4.1 by considering a class $F^\tau = \{F_{\mathbf{x}}^\tau\}$ of translated and scaled functions defined as $F_{\mathbf{x}}^\tau \triangleq \frac{K_{\mathbf{x}}(\mathbf{t}) + \tau}{2\tau}$ for $\mathbf{x} \in X$, which is a class of $[0, 1]$ -valued functions. The relationship between the Rademacher's complexities of F and F^τ can be obtained by [16, Theorem 15], which gives some simple structural results for Rademacher's complexity. Indeed, as an immediate application of that theorem, we get

$$\mathcal{R}_n(F^\tau) = \frac{1}{2\tau} \mathcal{R}_n(\{K_{\mathbf{x}}(\mathbf{t}) + \tau\}) \leq \frac{1}{2\tau} (\mathcal{R}_n(\{K_{\mathbf{x}}(\mathbf{t})\}) + \tau) = \frac{1}{2\tau} (\mathcal{R}_n(F) + \tau).$$

It turns out that the absolute constant C is the same as in Theorem 4.1. \square

As it will be shown later, the next theorem improves the bound on the \mathcal{L}_∞ approximation error derived in [10] (reported here as Theorem 3.2). The proof exploits ideas from the proof of [10, Proposition 3.1] and properties of Rademacher's complexity.

Theorem 4.3 *Let $X \subseteq \mathbb{R}^d$, $\lambda \in \mathcal{L}_1(X)$, f be a 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's complexity of the family $\{K_{\mathbf{x}}\}$. For every positive integer n there exist $\mathbf{t}_1, \dots, \mathbf{t}_n \in X$, $c_1, \dots, c_n \in \{-1, 1\}$, and an absolute constant C such that: if there exists $\tau > 0$ such that for all \mathbf{x} and \mathbf{t} , one has $|K_{\mathbf{x}}(\mathbf{t})| \leq \tau$, then*

$$\left\| f(\mathbf{x}) - \frac{\|\lambda\|_1}{n} \sum_{i=1}^n c_i K_{\mathbf{x}}(\mathbf{t}_i) \right\|_\infty \leq C \|\lambda\|_1 (\mathcal{R}_n + \tau) \sqrt{\frac{1}{n}}.$$

Proof. Without loss of generality we assume $\|\lambda\|_1 \neq 0$, as the case $\|\lambda\|_1 = 0$ is trivial. For every $\mathbf{x} \in X$, $\frac{f(\mathbf{x})}{\|\lambda\|_1} = \int_X K_{\mathbf{x}}(\mathbf{t}) \frac{\lambda(\mathbf{t})}{\|\lambda\|_1} d\mathbf{t}$ can be considered as the expected risk relative to the function $F_{\mathbf{x}}(\mathbf{t}) = K_{\mathbf{x}}(\mathbf{t}) \text{sign}(\lambda(\mathbf{t}))$, where $\text{sign}(z) = -1$ for $z < 0$ and $\text{sign}(z) = 1$ for $z \geq 0$, and the probability density $|\lambda(\mathbf{t})|/\|\lambda\|_1$. Note that λ can assume both positive and negative values but the Rademacher's complexities of $F = \{K_{\mathbf{x}}(\mathbf{t})\}$ and $F_\lambda = \{K_{\mathbf{x}}(\mathbf{t}) \text{sign}(\lambda(\mathbf{t}))\}$ are the same. Indeed, changing the sign of $K_{\mathbf{x}}(\mathbf{t}_i)$ in the definition of the Rademacher's complexity of $F = \{K_{\mathbf{x}}(\mathbf{t})\}$ is equivalent to changing the sign of each ε_i , and $\{\varepsilon_i\}$ are independent and symmetrically distributed around 0.

Then by Theorem 4.2, for every $\delta > 0$ and every sequence $\{\mathbf{t}_i\}$ obtained sampling X n times independently according to $\frac{|\lambda(\mathbf{t})|}{\|\lambda\|_1}$, we get with probability at least $1 - \delta$

$$\frac{1}{2\tau} \left\| \frac{f(\mathbf{x})}{\|\lambda\|_1} - \frac{1}{n} \sum_{i=1}^n \text{sign}(\lambda(\mathbf{t}_i)) K_{\mathbf{x}}(\mathbf{t}_i) \right\|_{\infty} \leq C \sqrt{\frac{1}{n} \max \left\{ \left(\frac{\mathcal{R}_n + \tau}{2\tau} \right)^2, \ln \frac{1}{\delta} \right\}}.$$

By letting $\delta \rightarrow 1$ and multiplying both sides of this inequality by $\|\lambda\|_1$, we conclude that there exist $\mathbf{t}_1, \dots, \mathbf{t}_n \in X$ and $c_1, \dots, c_n \in \{-1, 1\}$ such that $\left\| f(\mathbf{x}) - \frac{\|\lambda\|_1}{n} \sum_{i=1}^n c_i K_{\mathbf{x}}(\mathbf{t}_i) \right\|_{\infty} \leq C \|\lambda\|_1 (\mathcal{R}_n + \tau) \sqrt{\frac{1}{n}}$. \square

The computation of Rademacher's complexity is not always easy; however, an easy situation occurs, e.g., when F is the unit ball of a RKHS, as in this case the Rademacher's complexity can be expressed in terms of the sum of the eigenvalues of the operator associated with the kernel [16]. However, to compare our results with those from [10, 11, 12], we shall estimate the Rademacher's complexity in terms of the VC dimension, by means of Propositions 7.5 and 7.6.

The following lemma gives, for a family F of $[0, 1]$ -valued functions, an upper bound on the Rademacher's complexity of F in terms of its VC dimension. The lemma is analogous to [16, Corollary 5], which refers to the case of Boolean functions. As intermediate steps, the proof of the lemma estimates Rademacher's complexity in terms of ε -fat-shattering dimension, V_{ε} dimension and V dimension.

Recall that for $\varepsilon > 0$, a class F of real-valued functions on X ε -shatters a set $S \subseteq X$ if there exists a function g such that for every $T \subseteq S$, there is $f_T \in F$ satisfying the following: for every $\mathbf{t} \in S \setminus T$, one has $f_T(\mathbf{t}) \leq g(\mathbf{t}) - \varepsilon$ and for every $\mathbf{t} \in T$, $f_T(\mathbf{t}) \geq g(\mathbf{t}) + \varepsilon$. The ε -fat-shattering dimension of F , denoted by $\text{fat}_{\varepsilon}(F)$, is the maximal cardinality of an ε -shattered subset of X [16, 22]. A scaled version of the VC dimension, called here V_{ε} dimension, is defined by considering, in the definition of an ε -shattered set, only constant functions g , i.e., $g(\mathbf{t}) = \alpha \in \mathbb{R}$ [22]. Finally, for $\varepsilon = 0$ we obtain the definition of V dimension [22]².

Lemma 4.4 *Let F be a family of $[0, 1]$ -valued functions with finite VC dimension. There exists an absolute constant C such that for every positive integer*

²The terminology is not uniform in the literature. For example, in [22] the ε -fat-shattering dimension is called P_{γ} dimension (the latter name comes from the fact that it is a generalization of the pseudo-dimension P , introduced in [23]). Moreover, the definition of V dimension given in [22] is very similar, but not identical, to the definition of VC dimension for a family of real-valued functions. However, the upper bound $V(F) \leq VC(F)$ stated in Proposition 7.5 follows easily from their definitions.

n

$$\mathcal{R}_n(F) \leq C\sqrt{VC(F)}$$

Proof. Propositions 7.5 and 7.6 imply that for a family F of $[0, 1]$ -valued functions with finite VC dimension, for $\varepsilon \leq 1$ the ε -fat-shattering dimension can be bounded from above as follows in terms of VC dimension:

$$\text{fat}_\varepsilon(F) \leq \left(2\left\lceil\frac{1}{2\varepsilon}\right\rceil - 1\right)V_{\frac{\varepsilon}{2}}(F) \leq \left(\frac{2}{\varepsilon} - 1\right)V(F) \leq \left(\frac{2}{\varepsilon} - 1\right)VC(F), \quad (3)$$

which is a polynomial in $1/\varepsilon$ of degree lower than 2.

As for $\varepsilon \geq 1$ $\mathcal{N}(\varepsilon, F, \mathcal{L}_2(\mu_n)) = 1$ for every family of $[0, 1]$ -valued functions, one has $\int_0^\infty (\ln \mathcal{N}(\varepsilon, F, \mathcal{L}_2(\mu_n)))^{1/2} d\varepsilon = \int_0^1 (\ln \mathcal{N}(\varepsilon, F, \mathcal{L}_2(\mu_n)))^{1/2} d\varepsilon$. By (3) and Theorem 7.4, there exist absolute constants C, C_1 , and C_2 such that

$$\begin{aligned} \int_0^1 (\ln \mathcal{N}(\varepsilon, F, \mathcal{L}_2(\mu_n)))^{1/2} d\varepsilon &\leq \int_0^1 \left(\ln\left(\frac{2}{\varepsilon}\right)^{C_1 \text{fat}_{C_2\varepsilon}(F)}\right)^{1/2} d\varepsilon \\ &\leq \int_0^1 \left(\ln\left(\frac{2}{\varepsilon}\right)^{C_1\left(\frac{2}{\varepsilon}-1\right)VC(F)}\right)^{1/2} d\varepsilon \\ &= \int_0^1 \left(C_1\left(\frac{2}{\varepsilon}-1\right)VC(F)\right)^{\frac{1}{2}} \left(\ln\left(\frac{2}{\varepsilon}\right)\right)^{1/2} d\varepsilon. \end{aligned}$$

Note that in the statement of Theorem 7.4 the value of the positive constant C_2 is not given, however the second inequality above would be true also for $C_2\varepsilon > 1$, since it is easy to check that, for a class F of $[0, 1]$ -valued functions, $\text{fat}_{C_2\varepsilon}(F) = 0$ if $C_2\varepsilon > 1$.

By standard integrability criteria

$$\int_0^1 \left(C_1\left(\frac{2}{\varepsilon}-1\right)VC(F)\right)^{1/2} \left(\ln\left(\frac{2}{\varepsilon}\right)\right)^{1/2} d\varepsilon \leq C(VC(F))^{1/2}.$$

Putting all together, we get

$$\int_0^\infty (\ln \mathcal{N}(\varepsilon, F, \mathcal{L}_2(\mu_n)))^{1/2} d\varepsilon \leq C(VC(F))^{\frac{1}{2}}.$$

We conclude by Theorem 7.3 and by performing the average with respect to $\{\mathbf{t}_i\}$. This last step is due to the fact that Rademacher's complexity is the average, with respect to $\{\mathbf{t}_i\}$, of the first side of Dudley's integral formula. \square

The next theorem improves the bound on the $\mathcal{L}_\infty(X)$ approximation error derived in [10] (reported here as Theorem 3.2), using VC dimension as in Theorem 3.2, instead of Rademacher's complexity as in Theorem 4.3.

Theorem 4.5 *Let $X \subseteq \mathbb{R}^d$, $\lambda \in \mathcal{L}_1(X)$, f be a function on X having the representation $f(\mathbf{x}) = \int_X K_{\mathbf{x}}(\mathbf{t}) \lambda(\mathbf{t}) d\mathbf{t}$, and h the VC dimension of the family $\{K_{\mathbf{x}}\}$. For every positive integer n , there exist $\mathbf{t}_1, \dots, \mathbf{t}_n \in X$, $c_1, \dots, c_n \in \{-1, 1\}$, and an absolute constant C such that:*

i) *if for all \mathbf{x} and \mathbf{t} one has $0 \leq K_{\mathbf{x}}(\mathbf{t}) \leq 1$, then*

$$\left\| f(\mathbf{x}) - \frac{\|\lambda\|_1}{n} \sum_{i=1}^n c_i K_{\mathbf{x}}(\mathbf{t}_i) \right\|_{\infty} \leq C \|\lambda\|_1 \sqrt{\frac{h}{n}};$$

ii) *if there exists $\tau > 0$ such that for all \mathbf{x} and \mathbf{t} one has $|K_{\mathbf{x}}(\mathbf{t})| \leq \tau$, then*

$$\left\| f(\mathbf{x}) - \frac{\|\lambda\|_1}{n} \sum_{i=1}^n c_i K_{\mathbf{x}}(\mathbf{t}_i) \right\|_{\infty} \leq 2\tau C \|\lambda\|_1 \sqrt{\frac{h}{n}}.$$

i) When λ assumes both positive and negative values, the Rademacher's complexities of $F = \{K_{\mathbf{x}}(\mathbf{t})\}$ and $F_{\lambda} = \{K_{\mathbf{x}}(\mathbf{t}) \text{sign}(\lambda(\mathbf{t}))\}$, where $\text{sign}(z) = -1$ for $z < 0$ and $\text{sign}(z) = 1$ for $z \geq 0$, are the same. So, the statement follows from Theorem 4.3 i) and Lemma 4.4.

ii) The case $-\tau \leq K_{\mathbf{x}}(\mathbf{t}) \leq \tau$ can be reduced to i) one still by considering the class F^{τ} of translated and scaled functions defined as $F^{\tau} \triangleq \left\{ \frac{K_{\mathbf{x}}(\mathbf{t}) + \tau}{2\tau} \right\}$, which is a class of $[0, 1]$ -valued functions. It easily follows from the definition of VC dimension that $VC(F)$ and $VC(F^{\tau})$ are equal. Indeed, the effect of a translation τ is a corresponding translation of the constant α appearing in the definition of VC dimension, while the subsequent scaling does not change the associated classification rule. \square

For a discussion on how the absolute constant C in Theorem 4.5 can be estimated, see Section 6. However, even without knowing C , one can see that Theorem 4.5 improves Theorem 3.2, at least for sufficiently large values of n , thanks to the absence of the extra factor $\ln((2en)/h)$. Further improvements may be obtained by evaluating C and directly using the fat-shattering dimension instead of the VC dimension. Here we have chosen to use the VC dimension for two reasons: i) to allow an immediate comparison with the results reported in [10, 11, 12]; ii) from the practical point of view, estimates of the VC dimension for these cases were already available. Note also that basically, the reason for which Theorem 4.5 improves Theorem 3.2 is that it has been derived exploiting bounds based on Talagrand's inequality, instead of the union-bound technique.

5 Application to radial-basis-functions approximation schemes

In this section we derive upper bounds on the approximation error by certain radial-basis-functions approximation schemes, i.e., approximation schemes having for a positive integer n the form

$$\sum_{i=1}^n c_i k(\mathbf{x} - \mathbf{t}_i),$$

where $c_1, \dots, c_n \in \mathbb{R}$, $\mathbf{t}_1, \dots, \mathbf{t}_n \in \mathbb{R}^d$, and $k : \mathbb{R} \rightarrow \mathbb{R}$ is a radial function. In particular, we consider the following basis functions: the Bessel potentials and the Gaussian. Throughout this section, we set $X = \mathbb{R}^d$.

The d -dimensional Fourier transform is defined as the operator from $\mathcal{L}_2(\mathbb{R}^d)$ to $\mathcal{L}_2(\mathbb{R}^d)$ such that

$$f(\mathbf{x}) \mapsto \hat{f}(\mathbf{s}) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\langle \mathbf{x}, \mathbf{s} \rangle} f(\mathbf{x}) d\mathbf{x},$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product in \mathbb{R}^d .

For $r > 0$, $\beta_r(\mathbf{x} - \mathbf{t})$ is the *Bessel potential of order r* , defined as the function $\beta_r : \mathbb{R}^d \rightarrow \mathbb{R}$ with the Fourier transform

$$\hat{\beta}_r(\mathbf{s}) = (2\pi)^{-\frac{d}{2}} (1 + \|\mathbf{s}\|^2)^{-r/2}.$$

For every $r > 0$, $\beta_r \in \mathcal{L}_1$, β_r is non-negative, radially-decreasing with exponential decay at infinity, and analytic except at the origin ([24], [25, p. 132]). If $r > d/2$, then $\beta_r \in \mathcal{L}_2$. For applications of Bessel potentials see, e.g., [26].

We shall consider the family \mathcal{F}_r^1 of functions defined as

$$\mathcal{F}_r^1 \triangleq \{f : \mathbb{R}^d \rightarrow \mathbb{R} \mid f = \beta_r * \lambda, \lambda \in \mathcal{L}_1\}, \quad (4)$$

where for two functions $g, h : \mathbb{R}^d \rightarrow \mathbb{R}$, $(g * h)(x) \triangleq \int_{\mathbb{R}^d} g(y)h(x - y)dy$ is their *convolution*.

For every integer $p \in [1, \infty)$, if $f \in \mathcal{L}_1$ and $g \in \mathcal{L}_p$ then $f * g \in \mathcal{L}_p$ and $\|f * g\|_p \leq \|f\|_1 \|g\|_p$ [27, Section IV.4]. So, for every $r > 0$ one has $\mathcal{F}_r^1 \subset \mathcal{L}_1$. The space \mathcal{F}_r^1 is called *Sobolev potential space* (or *Sobolev-Liouville 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 $\|f\|_{\mathcal{F}_r^1} \triangleq \|\lambda\|_1$.

Let $K_{r, \mathbf{x}}^{\text{Bessel}}(\mathbf{t}) \triangleq \beta_r(\mathbf{x} - \mathbf{t})$. The following estimate improves [10, Proposition 3.1].

Corollary 5.1 *Let d be a positive integer, $r > d$ and h_r 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 $\mathbf{t}_1, \dots, \mathbf{t}_n \in \mathbb{R}^d$, $c_1, \dots, c_n \in \{-1, 1\}$, and an absolute constant C such that*

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \left| f(\mathbf{x}) - \frac{\|\lambda\|_1}{n} \sum_{i=1}^n c_i \beta_r(\mathbf{x} - \mathbf{t}_i) \right| \leq 2\tau C \|\lambda\|_1 \sqrt{\frac{h_r}{n}}.$$

where $\tau > 0$ is such that $|\beta_r(\mathbf{z})| \leq \tau$ for every $\mathbf{z} \in \mathbb{R}^d$.

Proof. Proceeding as in the proof of [10, Proposition 3.1], we note that, for $r > d$, one has $\mathcal{L}_r^1 \subset \mathcal{C}$ [25], where \mathcal{C} denotes the space of continuous functions on \mathbb{R}^d . So the essential supremum can be replaced by a supremum. Since $r > d$, it can be shown that there exists $\tau > 0$ such that, for all $\mathbf{z} \in \mathbb{R}^d$, one has $|\beta_r(\mathbf{z})| \leq \tau$. The statement follows by Theorem 4.5 ii). \square

Corollary 5.1 gives a bound on the error of approximation by linear combinations of translates of the Bessel potentials of order r . Unfortunately, as noted in [10], both the analytic expression of β_r and the VC dimension of the family $\{K_{r,\mathbf{x}}^{\text{Bessel}}\}$ are unknown.

In [10], the integral representation from [25, p. 132] of the kernel β_r in terms of Gaussians of different widths was used to derive a bound on the error of approximation of functions from \mathcal{F}_r^1 by a linear combination of Gaussians with different widths and centers [10, Proposition 3.2]. Instead of using [25, p. 132], in the following we derive directly an integral representation of β_r in terms of Gaussians. Then, we improve the bound [10, Proposition 3.2] by exploiting such a representation and Corollary 5.1.

Recall that *Gaussian radial-basis-function* computational units compute scaled and translated Gaussian functions on \mathbb{R}^d . Let $\gamma, \gamma_b : \mathbb{R}^d \rightarrow \mathbb{R}$ denote the Gaussian function and the Gaussian function scaled by $b > 0$, resp., i.e.,

$$\gamma(\mathbf{x}) = e^{-\|\mathbf{x}\|^2} \quad \text{and} \quad \gamma_b(\mathbf{x}) = e^{-b\|\mathbf{x}\|^2}.$$

Corollary 5.2 *Let d be a positive integer and $r > d$. For every $f \in \mathcal{F}_r^1$ and every positive integer n , there exist $\mathbf{t}_1, \dots, \mathbf{t}_n \in \mathbb{R}^d$, $b_1, \dots, b_n \in \mathbb{R}$, $c_1, \dots, c_n \in \{-1, +1\}$, and an absolute constant C such that*

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \left| f(\mathbf{x}) - \frac{\|\lambda\|_1}{n} \sum_{i=1}^n c_i e^{-\frac{\|\mathbf{x}_i - \mathbf{t}_i\|^2}{b_i}} \right| \leq C \|\lambda\|_1 \sqrt{\frac{d+3}{n}}. \quad (5)$$

Proof. The Fourier transform of the Gaussian function is a scaled Gaussian multiplied by a scalar: for every $b > 0$,

$$\widehat{\gamma}_b(\mathbf{s}) = (2b)^{-d/2} \gamma_{1/4b}(\mathbf{s}). \quad (6)$$

For every positive integer d and every $r > 0$, one has

$$\hat{\beta}_r(\mathbf{s}) = \frac{1}{\Gamma(r/2)} \int_0^\infty u^{r/2-1} e^{-u(1+\|\mathbf{s}\|^2)} du, \quad (7)$$

where $\Gamma(z) \triangleq \int_0^\infty t^{z-1} e^{-t} dt$ is the Gamma function. To see this, let $I \triangleq \int_0^\infty u^{r/2-1} e^{-u(1+\|\mathbf{s}\|^2)} dt$ and $v \triangleq u(1 + \|\mathbf{s}\|^2)$. Then, $du = dv(1 + \|\mathbf{s}\|^2)^{-1}$ and $I = (1 + \|\mathbf{s}\|^2)^{-r/2} \int_0^\infty v^{r/2-1} e^{-v} dv = \hat{\beta}_r(\mathbf{s})\Gamma(r/2)$.

By (6) with $b = 1$ and (7) we get

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

Thus, every $f = \beta_r * \lambda$ can be written as

$$f(\mathbf{x}) = \frac{2^{-d/2}}{\Gamma(r/2)} \int_0^\infty \int_{\mathbb{R}^d} e^{-\frac{\|\mathbf{x}-\mathbf{t}\|^2}{4u}} \Lambda(u, \mathbf{t}) d\mathbf{t} du,$$

where $\Lambda(u, \mathbf{t}) \triangleq e^{-u} u^{\frac{r-d}{2}-1} \lambda(\mathbf{t})$ is integrable since $r > d$. So, we can apply Theorem 4.5 i) with the kernel $K_{\mathbf{x}}^{\text{Gauss}}(\mathbf{t}, u) \triangleq e^{-\frac{\|\mathbf{x}-\mathbf{t}\|^2}{4u}}$. While using Theorem 4.5 i), the essential supremum can be replaced by a supremum (as in the proof of Corollary 5.1). Moreover, we can also estimate from above the VC dimension of the family $\{K_{\mathbf{x}}^{\text{Gauss}}\}$ as follows. As in the proof of [11, Proposition 5] we note that, since the exponential function is one-to-one, it is sufficient to bound from above the VC dimension of the family $\left\{-\frac{\|\mathbf{x}-\mathbf{t}\|^2}{4u}\right\}$, where \mathbf{x} plays the role of a parameter, and $(\mathbf{t}, u) \in \mathbb{R}^d \times [0, \infty)$ is the variable. By definition, the VC dimension of $\left\{-\frac{\|\mathbf{x}-\mathbf{t}\|^2}{4u}\right\}$ is equal to the VC dimension of the class of binary classifiers associated with $\left\{-\frac{\|\mathbf{x}-\mathbf{t}\|^2}{4u} + \alpha, \alpha \in \mathbb{R}\right\}$. As

$$-\frac{\|\mathbf{x}-\mathbf{t}\|^2}{4u} = -\frac{\|\mathbf{t}\|^2}{4u} + \frac{\mathbf{x} \cdot \mathbf{t}}{2u} - \frac{\|\mathbf{x}\|^2}{4u},$$

each element of $\left\{-\frac{\|\mathbf{x}-\mathbf{t}\|^2}{4u} + \alpha\right\}$, with parameters \mathbf{x} and α , is a function of (\mathbf{t}, u) that can be expressed as a linear combination of the $d + 3$ functions

$$1, \frac{1}{u}, \frac{t_1}{u}, \dots, \frac{t_d}{u}, \frac{\|\mathbf{t}\|^2}{u}.$$

Hence, by [28, Theorem 1] the VC dimension of the family $\{K_{\mathbf{x}}^{\text{Gauss}}\}$ is at most $d + 3$. \square

The bound from Corollary 5.2 depends substantially on the square root \sqrt{d} of the number of variables of functions in \mathcal{F}_r^1 . So, for functions having the

integral representation $\beta_r * \lambda$ for a function λ whose \mathcal{L}_1 -norm does not grow “too fast” with d , the bound does not exhibit the curse of dimensionality.

Corollary 5.2 improves [10, Proposition 3.2], as the multiplicative term $\sqrt{d+3}$ in the numerator replaces the term $\sqrt{(d+1) \ln \frac{2en}{d+1} + \ln 4}$ therein.

Note that various Authors derived upper bounds of the order $\mathcal{O}\left(\sqrt{\frac{1}{n}}\right)$ for approximation schemes formed by linear combinations of various kinds of computational units: Gaussian radial basis functions [4, 29], sigmoidal neural networks [1, 30], hinging hyperplanes [2], sines and cosines with variable frequencies and phases [3], etc. So the estimate from [10, Proposition 3.2] is worse than the above-mentioned ones, whereas our bound from Corollary 5.2 has the same order.

6 Discussion

Relationships with other works. In [31], estimates of the \mathcal{L}_∞ -approximation error were derived for certain variable-basis approximation schemes, using techniques different from those exploited in this paper.

A measure-theoretic formulation. To derive our estimates by applying Talagrand’s inequality (here in the form of Theorem 7.1), we have considered the representation $\frac{f(\mathbf{x})}{\|\lambda\|_1} = \int_X K_{\mathbf{x}}(\mathbf{t}) \frac{\lambda(\mathbf{t})}{\|\lambda\|_1} d\mathbf{t}$ as the expected risk relative to the function $F_{\mathbf{x}}(\mathbf{t}) = K_{\mathbf{x}}(\mathbf{t}) \text{sign}(\lambda(\mathbf{t}))$, and the probability density $|\lambda(\mathbf{t})|/\|\lambda\|_1$. More generally, one can use the following measure-theoretic approach, which allows one to deal with the case in which one cannot define a probability density function in the classical sense.

Let

$$f(\mathbf{x}) = \int_X K_{\mathbf{x}}(\mathbf{t}) d\mu,$$

where μ is a real signed measure with bounded total variation $|\mu|$ [32, Sections 6.1 and 6.2]. Define the positive and negative variations of μ as $\mu^+ \triangleq \frac{1}{2}(|\mu| + \mu)$ and $\mu^- \triangleq \frac{1}{2}(|\mu| - \mu)$, resp., which are two positive and bounded measures [32, Section 6.4]. Then, $\mu = \mu^+ - \mu^-$ and $|\mu| = \mu^+ + \mu^-$. By the Hahn Decomposition Theorem [32, Section 6.14], there exist two disjoint sets A, B such that $A \cup B = X$, $d\mu^+ = d|\mu|$ on A , $d\mu^- = d|\mu|$ on B and

$$f(\mathbf{x}) = \int_X K_{\mathbf{x}}(\mathbf{t}) d\mu = \int_A K_{\mathbf{x}} d\mu^+ - \int_B K_{\mathbf{x}}(\mathbf{t}) d\mu^-.$$

Then we get

$$\begin{aligned} f(\mathbf{x}) &= \int_A K_{\mathbf{x}}(\mathbf{t}) d|\mu| - \int_B K_{\mathbf{x}}(\mathbf{t}) d|\mu| = \int_X K_{\mathbf{x}}(\mathbf{t}) [I_A(\mathbf{t}) - I_B(\mathbf{t})] d|\mu| \\ &= \frac{\int_X K_{\mathbf{x}}(\mathbf{t}) [I_A(\mathbf{t}) - I_B(\mathbf{t})] d|\mu|}{\int_X d|\mu|} \int_X d|\mu|, \end{aligned}$$

where I_A and I_B are the indicator functions of the sets A and B , resp., and we have multiplied and divided by $\int_X d|\mu|$ to get a probability measure.

Now one can apply Theorem 7.2 to the family of functions $F^{A,B} = \{K_{\mathbf{x}}^{A,B}\}$ with $K_{\mathbf{x}}^{A,B} \triangleq K_{\mathbf{x}}(\mathbf{t}) [I_A(\mathbf{t}) - I_B(\mathbf{t})]$. As the function $I_A(\mathbf{t}) - I_B(\mathbf{t})$ can take only the values $+1$ and -1 , the Rademacher's complexity of the family $F^{A,B} = \{K_{\mathbf{x}}^{A,B}\}$ is equal to that of the family $F = \{K_{\mathbf{x}}\}$.

Localized Rademacher's complexity. In [16], an improvement of classical SLT bounds was derived in terms of a quantity called *localized Rademacher's complexity*. We do not report here the definition of such a complexity but we recall that it can be used to improve SLT results as it allows one to bound the probability

$$\text{Prob}\left\{\exists F_{\mathbf{x}} \in F \mid R_{\text{emp}}(F_{\mathbf{x}}, \{\mathbf{t}_i\}) \leq \varepsilon, R(F_{\mathbf{x}}) \geq 2\varepsilon\right\}.$$

Indeed, in SLT this quantity is more interesting than

$$\text{Prob}\left\{\sup_{F_{\mathbf{x}} \in F} |R(F_{\mathbf{x}}) - R_{\text{emp}}(F_{\mathbf{x}}, \{\mathbf{t}_i\})| \leq \varepsilon\right\} \quad (8)$$

since, in order to obtain an approximation of the minimum of the expected risk, it is sufficient having a small estimation error when the observed empirical risk is small.

However, to derive approximation bounds from SLT bounds, the localized Rademacher's complexity cannot be fruitfully used (at least not in a straightforward way), because in this context what one needs is a bound of the form (8). Indeed, following Girosi's [10] approach, (8) gives a bound on the \mathcal{L}_{∞} error for the function one wants to approximate, while the bound based on the localized Rademacher's complexity does not.

Computations of the constants. The absolute constant C in Theorems 4.3 and 4.5 can be computed by using the values (or upper bounds on the values) of:

- the absolute constant C in Talagrand's inequality (Theorem 7.1), required to compute the absolute constant C in Theorem 7.2 or equivalently the absolute constant C in Theorem 4.2;
- the absolute constant C in Theorem 7.3;

- the absolute constants C_1 and C_2 in Theorem 7.4;
- the absolute constant C_2 in the proof of Lemma 4.4.

As regards the constant in Theorem 7.3, the value $C = 12$ has been derived in [33]. For a value of the constant in the Talagrand's inequality (in a different form than the one reported in Theorem 7.1), see [34].

Improvements of other estimates. The estimates in [11, 12] were derived using similar techniques as in [10]. So, it turns out that our Theorems 4.3 and 4.5 can be applied also to improve the bounds in [11, 12] (in particular, [11, Corollary 3] and [12, Theorem 4 and Proposition 5]).

7 Appendix

Theorem 7.1 (Talagrand's inequality [16]) *Let P_X be a probability distribution on $X \subseteq \mathbb{R}^d$, $\{\mathbf{t}_i\}$ be a P_X -i.i.d. sequence of n points in X , and $F \subseteq B_{1,\infty,X}$. There exists an absolute constant $C \geq 1$ such that for every $0 < \delta < 1$, every $0 < \varepsilon < 1$, and every positive integer $n \geq \left\lceil \frac{4C^2}{\varepsilon^2} \ln \frac{1}{\delta} \right\rceil$, the following bound holds with probability at least $1 - \delta$:*

$$\sup_{F_{\mathbf{x}} \in F} |R(F_{\mathbf{x}}) - R_{\text{emp}}(F_{\mathbf{x}}, \{\mathbf{t}_i\})| \leq 2\mathbb{E}_{\mathbf{t}_1, \dots, \mathbf{t}_n} \left\{ \sup_{F_{\mathbf{x}} \in F} |R(F_{\mathbf{x}}) - R_{\text{emp}}(F_{\mathbf{x}}, \{\mathbf{t}_i\})| \right\} + \frac{3\varepsilon}{4}.$$

Theorem 7.2 (SLT bound via Rademacher's complexity [16]) *Let P_X be a probability distribution on $X \subseteq \mathbb{R}^d$, $\{\mathbf{t}_i\}$ be a P_X -i.i.d. sequence of n points in X , and $F \subseteq B_{1,\infty,X}$. There exists an absolute constant C such that for every $0 < \delta < 1$, every $0 < \varepsilon < 1$, and every positive integer $n \geq \left\lceil \frac{C}{\varepsilon^2} \max \left\{ \mathcal{R}_n^2(F), \ln \frac{1}{\delta} \right\} \right\rceil$, the following bound holds with probability at least $1 - \delta$:*

$$\sup_{F_{\mathbf{x}} \in F} |R(F_{\mathbf{x}}) - R_{\text{emp}}(F_{\mathbf{x}}, \{\mathbf{t}_i\})| \leq \varepsilon.$$

Theorem 7.3 (Dudley's integral [16]) *Let F be a family of functions on $X \subseteq \mathbb{R}^d$, μ_n be the empirical measure supported on n samples $\{\mathbf{t}_i\} \subset X$, and $\mathcal{N}(\varepsilon, F, \mathcal{L}_2(\mu_n))$ be the ε -covering number of F with respect to $\mathcal{L}_2(\mu_n)$. There exists an absolute constant C such that for every positive integer n the following holds:*

$$\mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n} \left\{ \frac{1}{\sqrt{n}} \sup_{F_{\mathbf{x}} \in F} \left| \sum_{i=1}^n \varepsilon_i F_{\mathbf{x}}(\mathbf{t}_i) \right| \right\} \leq C \int_0^\infty \left(\ln \mathcal{N}(\varepsilon, F, \mathcal{L}_2(\mu_n)) \right)^{1/2} d\varepsilon.$$

Theorem 7.4 (Bound on ε -covering numbers [16, 35]) *Let $X \subseteq \mathbb{R}^d$, $F \subseteq B_{1,\infty,X}$, μ a probability measure on X , $\mathcal{N}(\varepsilon, F, \mathcal{L}_2(\mu))$ the ε -covering number of F with respect to $\mathcal{L}_2(\mu)$, and $\text{fat}_\varepsilon(F)$ the ε -fat-shattering dimension of F . There exist absolute constants C_1 and C_2 such that for every $0 < \varepsilon < 1$ the following holds:*

$$\mathcal{N}(\varepsilon, F, \mathcal{L}_2(\mu)) \leq \left(\frac{2}{\varepsilon}\right)^{C_1 \text{fat}_{C_2\varepsilon}(F)}.$$

Proposition 7.5 (V_ε , V and VC dimensions [22]) *For every family F of bounded real-valued functions and every $\varepsilon > 0$,*

$$V_\varepsilon(F) \leq V(F) \leq VC(F).$$

Proposition 7.6 (V_ε and ε -fat-shattering dimensions [22]) *For every family F of bounded real-valued functions and every $\varepsilon > 0$,*

$$V_\varepsilon(F) \leq \text{fat}_\varepsilon(F) \leq \left(2 \left\lceil \frac{1}{2\varepsilon} \right\rceil - 1\right) V_{\frac{\varepsilon}{2}}(F).$$

Remark. The bound $V(F) \leq VC(F)$ actually does not appear in [22], but it follows easily from the (slightly different) definitions of V dimension and VC dimension for a family of real-valued functions. In fact, in the definition of VC dimension the parameter α can vary with \mathbf{x} , but this does not hold in the definition of V dimension.

Remark. Propositions 7.5 and 7.6 were derived in [22] for $[0, 1]$ -valued functions but, for every $a > 0$, they hold for classes of $[-a, a]$ -valued functions. To see this, given one such class F , first add to each of its functions the value a , so that the interval where they take values becomes $[0, 2a]$. The VC , VC_ε , and fat_ε -shattering dimensions of the new class are the same as those of F (just imagine to translate by a the quantities used in the definitions). Then, one goes from $[0, 2a]$ to $[0, 1]$ by scaling the function values. In doing so, from the definitions one can see that the VC dimension does not change, whereas the ε -dependent dimensions are modified as follows: the value ε for a class F of $[0, 2a]$ -valued functions corresponds to the value $\varepsilon/(2a)$ for the same class of functions scaled in $[0, 1]$.

8 Acknowledgement

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] A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. on Information Theory*, **39** (1993), 930–945.
- [2] L. Breiman, Hinging hyperplanes for regression, classification, and function approximation, *IEEE Trans. on Information Theory*, **39** (1993), 993–1013.
- [3] L. K. Jones, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, *Annals of Statistics*, **20** (1992), 608–613.
- [4] F. Girosi and G. Anzellotti, Rates of convergence for Radial Basis Functions and neural networks, in *Artificial Neural Networks for Speech and Vision*, R. J. Mammone, editor, 97–113, Chapman & Hall, London, 1993.
- [5] V. Kůrková and M. Sanguineti, Comparison of worst case errors in linear and neural network approximation, *IEEE Trans. on Information Theory*, **48** (2002), 264–275.
- [6] V. Kůrková and M. Sanguineti, Learning with generalization capability by kernel methods of bounded complexity, *J. of Complexity*, **21** (2005), 350–367.
- [7] R. Zoppoli, M. Sanguineti, and T. Parisini, Approximating networks and extended Ritz method for the solution of functional optimization problems, *J. of Optimization Theory and Applications*, **112** (2002), 403–440.
- [8] V. Kůrková and M. Sanguineti, Error estimates for approximate optimization by the extended Ritz method, *SIAM J. on Optimization*, **15** (2005), 461–487.
- [9] R. Bellman, *Dynamic Programming*, Princeton University Press, 1957.
- [10] F. Girosi, Approximation error bounds that use VC-bounds, in *Proc. Int. Conf. on Artificial Neural Networks* (Paris, 1995), 295–302.
- [11] M. A. Kon, L. A. Raphael, and D. A. Williams, Extending Girosi’s approximation estimates for functions in Sobolev spaces via Statistical Learning Theory, *J. of Analysis and Applications*, **3**(2) (2005), 67–90.
- [12] M. A. Kon and L. A. Raphael, Approximating functions in reproducing kernel Hilbert spaces via Statistical Learning Theory, in *Wavelets and Splines*, G. Chen and M. J. Lai, editors, 271–286, Nashboro Press, Nashville, TN, 2006.

- [13] N. Aronszajn, Theory of reproducing kernels, *Trans. of AMS*, **68** (1950), 337–404 .
- [14] F. Cucker and S. Smale, On the mathematical foundations of learning, *Bulletin of AMS*, **39** (2001), 1–49.
- [15] V. P. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.
- [16] S. Mendelson, A few notes on Statistical Learning Theory, in *Advanced Lectures on Machine Learning - LNCS 2600*, Mendelson, S. and Smola, A., editors, 1–40, Springer, 2003.
- [17] M. Talagrand, Sharper bounds for Gaussian and empirical processes, *Annals of Probability*, **22** (1994), 20–76.
- [18] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, Cambridge, UK, 1999.
- [19] B. Carl, I. Kyrezi, and A. Pajor, Metric entropy of convex hulls in Banach spaces, *J. London Math. Soc.*, **60** (1999), 871–896.
- [20] M. Vidyasagar, *A Theory of Learning and Generalization*, Springer, Berlin, 1997.
- [21] W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. of the American Statistical Association*, **53** (1963), 13–30.
- [22] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, Scale sensitive dimensions, uniform convergence and learnability, *J. of the ACM*, **4** (1997), 615–631.
- [23] D. Pollard, *Empirical Processes: Theory and Applications*, Volume 2 of NSF-CBMS Regional Conf. Series in Probability and Statistics, Institute of Mathematical Statistics and American Statistical Association, 1990.
- [24] N. Aronszajn and K. T. Smith, Theory of Bessel potentials - Part I, *Ann. Inst. Fourier*, **11** (1961), 385–475 (Also: Technical Report 22 - Office of Naval Research, Lawrence, Kansas, 1959).
- [25] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
- [26] L. P. Castro, Strongly elliptic operators for a plane wave diffraction problem in Bessel potential spaces, *J. of Inequalities in Pure and Applied Mathematics*, **3** (2002), 2, Article 25.

- [27] H. Brezis, *Analyse Fonctionnelle - Théorie et Applications*. Masson, Paris, 1983.
- [28] E. Sontag, VC dimension of neural networks, in *Neural Networks and Machine Learning*, C. Bishop, editor, 69–95, Springer-Verlag, Berlin, 1998.
- [29] H. N. Mhaskar and C. A. Micchelli, Dimension independent bounds on the degree of approximation by neural networks, *IBM J. of Research and Development*, **38** (1994), 277–284.
- [30] A. R. Barron, Neural net approximation, in *Proc. 7th Yale Workshop on Adaptive and Learning Systems*, K. S.Narendra, editor, 69–72, Yale University Press, 1992.
- [31] H. N. Mhaskar, Neural networks for optimal approximation of smooth and analytic functions, *Neural Computation*, **8** (1996), 164–177.
- [32] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, Singapore, 1987 (III Ed.).
- [33] [http://www.cs.berkeley.edu/~bartlett/courses/281b-sp06/lecture 24.ps](http://www.cs.berkeley.edu/~bartlett/courses/281b-sp06/lecture%2024.ps).
- [34] <http://ocw.mit.edu/OcwWeb/Mathematics/18-465Spring-2004/LectureNotes/index.htm>.
- [35] S. Mendelson and R. Vershynin, Entropy and the combinatorial dimension, *Inventiones Mathematicae*, **152** (2003), 37–55.

Received: March 15, 2007