Rates of Approximation of Smooth Functions by Gaussian Radial-Basis Networks

Kainen, P.C.

2006

Institute of Computer Science
Academy of Sciences of the Czech Republic

Rates of approximation of smooth functions by Gaussian radial-basis networks

Paul C. Kainen, Věra Kůrková, Marcello Sanguineti

Technical report No. 976

October 2006

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: +420 266 051 111, fax: +420 286 585 789,
e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

Rates of approximation of smooth functions by Gaussian radial-basis networks

Paul C. Kainen[1], Věra Kůrková[2], Marcello Sanguineti[3]

Technical report No. 976

October 2006

Abstract:

Complexity of Gaussian radial-basis-function networks, with varying widths, is bounded above in terms of the rate of decrease of approximation error with increasing number of hidden units. Bounds are explicitly given in terms of norms measuring smoothness (Bessel and Sobolev norms). Estimates are proven using suitable integral representations in the form of networks with continua of hidden units computing scaled Gaussians and translated Bessel potentials.

Keywords:
Gaussian radial-basis functions, rates of approximation, Bessel's potencials.

[1]Department of Mathematics, Georgetown University, Washington, D. C. 20057-1233, USA, kainen@georgetown.edu
[2]Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, Prague 8, Czech Republic, vera@cs.cas.cz
[3]Department of Communications, Computer, and System Sciences (DIST), University of Genoa, Via Opera Pia 13, 16145 Genova, Italy, marcello@dist.unige.it

1 Introduction

Radial-basis function (RBF) networks with Gaussian computational units are known to be able to approximate with an arbitrary accuracy all continuous and all $\mathcal{L}^2$-functions on compact subsets of $\mathbb{R}^d$ [9, 18, 19, 20, 21]. In such approximations, the number n of RBF units plays the role of a measure of model complexity and its size determines the feasibility of network implementation.

Several authors investigated rates of approximation by Gaussian RBF networks with *fixed width.* Girosi and Anzellotti [8] derived an asymptotic upper bound of the form $\mathcal{O}(1/\sqrt{n})$ on approximation error measured by the supremum norm for band-limited functions with continuous derivatives up to the order r with $r > d/2$, where d is the number of variables [8, p. 106]. Using results from statistical learning theory, Girosi [6] extended these bounds to more general classes of kernels. For Gaussians of *varying* widths, Kon, Raphael, and Williams [12, Cor. 3] obtained bounds on a wighted $\mathcal{L}^\infty$-distance from the target function to a linear combination of Gaussians.

In this paper, we also investigate approximation of smooth functions by Gaussian RBF networks with varying widths, but consider Lebesgue measure $\mathcal{L}^2$-distance. We derive upper bounds on rates of approximation in terms of the Bessel and Sobolev norms of the functions to be approximated. Bessel norms are defined in terms of convolution with the Bessel potential kernel, while Sobolev norms use integrals of partial derivatives. Both norms are equivalent but the ratios between them also depend on the number of variables d.

Our estimates hold for all numbers n of hidden units and all degrees $r > d/2$ of Bessel potentials and are of the form $n^{-1/2}$ times the Bessel norm $\|f\|_{L^{1,r}}$ of the function f to be approximated times a factor $k(r,d)$. For a fixed $c > 0$ and the degree $r_d = d/2 + c$, $k(r_d, d)$ decreases to zero exponentially fast. We also derive estimates in terms of $\mathcal{L}^2$ Bessel and Sobolev norms. Our results show that reasonably smooth functions can be approximated quite efficiently by Gaussian radial-basis networks; a preliminary version of our results appeared in [11].

The paper is organized as follows. In Section 2, some concepts, notations and auxiliary results for investigation of approximation by Gaussian RBF networks are introduced. In Section 3, upper bounds on rates of approximation of Bessel potentials by linear combinations of scaled Gaussians are derived in terms of variation norms obtained from integral representations of Bessel potentials and their Fourier transforms. In Section 4, for functions representable as convolutions with Bessel potentials, upper bounds are derived in terms of Bessel potential norms. These bounds are then combined with estimates of variational norms from the previous section to obtain bounds for approximation by Gaussian RBFs in terms of Bessel norms. In Section 5, relationship between Sobolev and Bessel norms is used to obtain bounds in terms of Sobolev norms.

2 Approximation by Gaussian RBF networks

For $\Omega \subseteq \mathbb{R}^d$, $\mathcal{L}^2(\Omega)$ denotes the space of real-valued functions on Ω with norm $\|f\|_{\mathcal{L}^2(\Omega)} = \left(\int |f(x)|^2 dx\right)^{1/2}$. Two functions are identified if they differ only on a set of Lebesgue-measure zero. When $\Omega = \mathbb{R}^d$, we omit it in the notation.

For nonzero $f \in \mathcal{L}^2$, $f^o = f/\|f\|_{\mathcal{L}^2}$ denotes the *normalization* of f; for convenience, we put $0^o = 0$. For $F \subset \mathcal{L}^2$, $F|_\Omega$ denotes the set of functions from F restricted to Ω, $\hat{F}$ the set of Fourier transforms of functions in F, and F^o the set of their normalizations. For $n \geq 1$, define

$$\mathrm{span}_n F := \left\{ \sum_{i=1}^n w_i f_i \mid f_i \in F, w_i \in \mathbb{R} \right\}.$$

In this paper, we investigate accuracy measured by $\mathcal{L}^2$-norm with respect to Lebesgue measure λ in approximation by Gaussian radial-basis-function networks.

A *Gaussian radial-basis-function unit* with d inputs computes all *scaled and translated Gaussian functions* on $\mathbb{R}^d$. For $b > 0$, let $\gamma_b : \mathbb{R}^d \to \mathbb{R}$ denote the Gaussian function of *width* b defined by

$$\gamma_b(x) = e^{-b\|x\|^2}.$$

A simple calculation shows that

$$\|\gamma_b\|_{\mathcal{L}^2} = (\pi/2b)^{d/4}. \tag{2.1}$$

Indeed, $\|\gamma_b\|^2_{\mathcal{L}^2} = \int_{\mathbb{R}^d} \left(e^{-b\|x\|^2}\right)^2 dx = \int_{\mathbb{R}^d} e^{-2b\|x\|^2} dx$. Setting $y = (\sqrt{2b})x$ and $dx = (2b)^{-d/2} dy$, we get $\|\gamma_b\|^2_{\mathcal{L}^2} = (2b)^{-d/2} \int_{\mathbb{R}^d} e^{-\|y\|^2} dy = \left(\frac{\pi}{2b}\right)^{d/2}$.

Let

$$G_0 = \{\gamma_b \,|\, b > 0\}$$

denote the *set of the Gaussians centered at* 0 *with varying widths*. For τ_y the *translation operator* defined for any $y \in \mathbb{R}^d$ and any f on $\mathbb{R}^d$ as $(\tau_y f)(x) = f(x-y)$, let

$$G = \left\{\tau_y \gamma_b \,|\, y \in \mathbb{R}^d, b > 0\right\}$$

denote the *set of all translations of the Gaussians with varying widths.*

We investigate rates of approximation by *networks with n Gaussian RBF units and one linear output unit*, which compute functions from the set $\mathrm{span}_n G$.

We utilize properties of the Fourier transform of the Gaussian function. The *d-dimensional Fourier transform* is the operator $\mathcal{F}$ on $\mathcal{L}^2 \cap \mathcal{L}^1$ given by

$$\mathcal{F}(f)(s) = \hat{f}(s) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{ix\cdot s} f(x)\, dx\,, \tag{2.2}$$

where $\cdot$ denotes the Euclidean inner product on $\mathbb{R}^d$.

For every $b > 0$,

$$\widehat{\gamma_b}(x) = (2b)^{-d/2} \gamma_{1/4b}(x) \tag{2.3}$$

(cf.[24, p. 43]). Thus

$$\mathrm{span}_n G_0 = \mathrm{span}_n \widehat{G_0}. \tag{2.4}$$

Plancherel's identity [24, p. 31] asserts that Fourier transform is an isometry on $\mathcal{L}^2$, i.e., for all $f \in \mathcal{L}^2$

$$\|f\|_{\mathcal{L}^2} = \|\hat{f}\|_{\mathcal{L}^2}\,, \tag{2.5}$$

and directly by (2.1) we have

$$\|\gamma_b\|_{\mathcal{L}^2} = \left(\frac{\pi}{2b}\right)^{d/4} = \|\widehat{\gamma_b}\|_{\mathcal{L}^2}. \tag{2.6}$$

In a normed linear space $(\mathcal{X}, \|.\|_\mathcal{X})$, for $f \in \mathcal{X}$ and $A \subset \mathcal{X}$,

$$\|f - A\|_\mathcal{X} = \inf_{g \in A} \|f - g\|_\mathcal{X}$$

denotes the distance from f to A. The following proposition shows that in estimating rates of approximation by linear combinations of scaled Gaussians centered at 0, one can switch between a function and its Fourier transform.

Proposition 1 *For all positive integers* d, n *and all* $f \in \mathcal{L}^2$,
$\|f - \mathrm{span}_n G_0\|_{\mathcal{L}^2} = \|f - \mathrm{span}_n \widehat{G}_0\|_{\mathcal{L}^2} = \|\hat{f} - \mathrm{span}_n \widehat{G}_0\|_{\mathcal{L}^2} = \|\hat{f} - \mathrm{span}_n G_0\|_{\mathcal{L}^2}$.

Proof. Using (2.5) and (2.4), respectively, we get $\|f - \mathrm{span}_n G_0\|_{\mathcal{L}^2} = \|\hat{f} - \mathrm{span}_n \hat{G}_0\|_{\mathcal{L}^2} = \|f - \mathrm{span}_n \hat{G}_0\|_{\mathcal{L}^2} = \|\hat{f} - \mathrm{span}_n G_0\|_{\mathcal{L}^2}$. □

To derive our estimates, we use a result on approximation by convex combinations of n elements of a bounded subset of a Hilbert space derived by Maurey [22], Jones [10] and Barron [2, 3]. Let F be a bounded subset of a Hilbert space $(\mathcal{H}, \|.\|_\mathcal{H})$, and $\mathrm{uconv}_n F = \{\frac{1}{n}\sum_{i=1}^n f_i \,|\, f_i \in F\}$ denote the set of n-fold convex combinations of elements of F with all coefficients equal. By Maurey-Jones-Barron's result [3, p. 934], for every function h in $\mathrm{cl\,conv}\,(F\cup -F)$, i.e., in the closure of the symmetric convex hull of F, we have

$$\|h - \mathrm{uconv}_n F\|_{\mathcal{H}} \le n^{-1/2} s_F\,, \tag{2.7}$$

where $s_F = \sup_{f\in F}\|f\|_{\mathcal{H}}$. The bound (2.7) implies an estimate of the distance from $\mathrm{span}_n F$ holding for any function from $\mathcal{H}$. The estimate is formulated in terms of a norm tailored to F, called *F-variation*, which was introduced in [13] as an extension of "variation with respect to half-spaces" defined in [2].

For any bounded subset F of any normed linear space $(\mathcal{X}, \|.\|_{\mathcal{X}})$, F-variation is defined as the *Minkowski functional of the closed convex symmetric hull of F* (where closure is taken with respect to the norm $\|.\|_{\mathcal{X}}$). The variational norm with respect to F in $\mathcal{X}$ is denoted by $\|.\|_{F,\mathcal{X}}$, i.e.,

$$\|h\|_{F,\mathcal{X}} = \inf\left\{c > 0 \,|\, c^{-1}h \in \mathrm{cl\,conv}\,(F\cup -F)\right\}. \tag{2.8}$$

Note that F-variation can be infinite (when the set on the right-hand side is empty) and that it depends on the ambient space norm. When we consider variation with respect to the $\mathcal{L}^2$-norm, we omit $\mathcal{L}^2$ in the notation of variational norm.

Maurey-Jones-Barron's estimate (2.7) implies that for any bounded subset F of a Hilbert space $(\mathcal{H}, \|.\|_{\mathcal{H}})$ and all positive integers n

$$\|h - \mathrm{span}_n F\|_{\mathcal{H}} \le n^{-1/2}\|h\|_{F^o,\mathcal{H}}. \tag{2.9}$$

(see[14]). To apply the upper bound (2.9) to approximation by Gaussian RBFs we take advantage of properties of variational norms given in the remainder of this section.

From the definitions, if ψ is any linear isometry of $(\mathcal{X}, \|.\|_{\mathcal{X}})$, then for any $f \in \mathcal{X}$, $\|f\|_{F,\mathcal{X}} = \|\psi(f)\|_{\psi(F),\mathcal{X}}$. In particular,

$$\|f\|_{G_0^o,\mathcal{X}} = \|\hat{f}\|_{G_0^o,\mathcal{X}} \tag{2.10}$$

Variations with respect to two subsets satisfy the following inequality [16, Proposition 3(iii)].

Lemma 1 *Let F, H be nonempty, nonzero subsets of a normed linear space $(\mathcal{X}, \|.\|_{\mathcal{X}})$ and $s_{H,F} := \sup_{h\in H}\|h\|_F$. Then for every $f \in \mathcal{X}$,*

$$\|f\|_{F,\mathcal{X}} \le s_{H,F}\|f\|_{H,\mathcal{X}}.$$

The next lemma states that the variation of the limit of a sequence of functions is bounded by the limit of variations (see [15, Lemma 7.2]).

Lemma 2 *Let F be a nonempty bounded subset of a normed linear space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, $h \in \mathcal{X}$, $\{h_i\} \subset \mathcal{X}$ be such that $\lim_{i\to\infty}\|h_i - h\|_{\mathcal{X}} = 0$, for all i, $b_i = \|h_i\|_{F,\mathcal{X}} < \infty$ and there exist $\lim_{i\to\infty} b_i = b$. Then $\|h\|_{F,\mathcal{X}} \le b$.*

Variation with respect to a parameterized family of functions can be estimated for functions which can be represented by a suitable integral formula, where integration is respect to the parameter. Let $\Omega \subseteq \mathbb{R}^d$, $\phi : \Omega \times Y \to \mathbb{R}$, and

$$F_\phi := \{\phi(., y) : \Omega \to \mathbb{R} \,|\, y \in Y\}.$$

If for all $y \in Y$, $\phi(., y) \in \mathcal{L}^2(\Omega)$, then we denote by $\Phi : Y \to \mathcal{L}^2$ the mapping defined as $\Phi(y) = \phi(., y) : x \mapsto \phi(x, y)$.

Theorem 1 *Let d, p be positive integers, $\Omega \subseteq \mathbb{R}^d$, $Y \subseteq \mathbb{R}^p$ be open, $w \in \mathcal{L}^1(Y)$, $\phi : \Omega \times Y \to \mathbb{R}$ be such that $\Phi(Y)$ is a bounded subset of $\mathcal{L}^2(\Omega)$. If w and Φ are continuous on Y except on a closed subset of measure zero and if for almost all $x \in \Omega$,*

$$f(x) = \int w(y)\phi(x, y)dy,$$

then $\|f||_{F_\phi} \le \|w\|_{\mathcal{L}^1(Y)}$.

The theorem can be proved by an argument using Bochner integrals (cf Girosi and Anzellotti [7]) together with the limit property of variational norms given in Lemma 2. It guarantees that if f can be represented as a neural network with a continuum of hidden units computing functions from F_ϕ, then the F_ϕ-variational norm of f is bounded by the $\mathcal{L}^1$-norm of the weight function.

3 Approximation of Bessel potentials by Gaussian RBFs

In this section, we estimate rates of approximation by $\mathrm{span}_n G$ for certain special functions, called Bessel potentials, which are defined by means of their Fourier transforms. For $r > 0$, the *Bessel potential* of order r, denoted by β_r, is the function on $\mathbb{R}^d$ with Fourier transform

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

To estimate G_0^o-variations of β_r and $\hat{\beta}_r$, we use Theorem 1 with representations of these two functions as integrals of scaled Guassians.

For $r > 0$, it is known [23, p. 132] that β_r is non-negative, radial, exponentially decreasing at infinity, analytic except at the origin, and belongs to $\mathcal{L}^1$. It can be expressed by the integral formula (see [17, p. 296] or [23])

$$\beta_r(x) = c_1(r,d) \int_0^\infty e^{-t/(4\pi)}\, t^{-d/2+r/2-1} e^{-(\pi/t)\|x\|^2}\, dt, \tag{3.1}$$

where

$$c_1(r,d) = (2\pi)^{d/2}(4\pi)^{-r/2}/\Gamma(r/2)$$

and $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$ is the Gamma function. The factor $(2\pi)^{d/2}$ occurs since our choice of Fourier transform (2.2) includes the factor $(2\pi)^{-d/2}$. Combining (3.1) with (2.1), we get a representation of the Bessel potential as an integral of normalized scaled Gaussians.

Proposition 2 *For every $r > 0$, d a positive integer, and $x \in \mathbb{R}^d$*

$$\beta_r(x) = \int_0^\infty v_r(t) \gamma^o_{\pi/t}(x)\, dt\,,$$

where $v_r(t) = c_1(r,d)\, 2^{-d/4}\, e^{-t/4\pi}\, t^{-d/4+r/2-1}$.

The next proposition estimates G_0^o-variation of β_r.

Proposition 3 *For d a positive integer and $r > d/2$,*

$$\|\beta_r\|_{G^o} \le \|\beta_r\|_{G_0^o} \le \int_0^\infty v_r(t)\, dt = k(r,d),$$

where $k(r,d) = \frac{(\pi/2)^{d/4}\Gamma(r/2-d/4)}{\Gamma(r/2)}$.

Proof. As $G_0^o \subset G^o$, we get $\|\beta_r\|_{G^o} \le \|\beta_r\|_{G_0^o}$. To estimate $\|\beta_r\|_{G_0^o}$, we apply Theorem 1 with $w = v_r$, $\phi(x,y) = \phi(x,t) = \gamma^o_{\pi/t}(x)$, $Y = (0,\infty)$, $Y_0 = \emptyset$, and $\Omega = \mathbb{R}^d$ to the integral representation from Proposition 2, getting

$$\|\beta_r\|_{G_0^o} \le \int_0^\infty v_r(t)\, dt = c_1(r,d)\, 2^{-d/4} \int_0^\infty e^{-t/(4\pi)}\, t^{-d/4+r/2-1}\, dt\,.$$

To estimate this integral, replace the variable t with $u = t/4\pi$ obtaining $\|\beta_r\|_{G_0^o} \le (4\pi)^{-d/4+r/2}\, c_1(r,d)$ $2^{-d/4} \int_0^\infty u^{-d/4+r/2-1}\, e^{-u}\, du$. Hence, by the definition of the Gamma function, one has

$$\begin{aligned}
\|\beta_r\|_{G_0^o} &\le c_1(r,d)\, 2^{-d/4}(4\pi)^{-d/4+r/2}\Gamma(r/2-d/4) \\
&= (2\pi)^{d/2}(4\pi)^{-r/2}2^{-d/4}(4\pi)^{-d/4+r/2}\Gamma(r/2-d/4)/\Gamma(r/2) \\
&= \frac{(\pi/2)^{d/4}\Gamma(r/2-d/4)}{\Gamma(r/2)} = k(r,d)\,.
\end{aligned}$$

□

The Fourier transform of the Bessel potential can also be expressed as an integral of normalized scaled Gaussians.

Proposition 4 *For every $r > 0$, d a positive integer, and $s \in \mathbb{R}^d$*

$$\hat{\beta}_r(s) = \int_0^\infty w_r(t)\gamma_t^o(s)\, dt\,,$$

where $w_r(t) = (\pi/2t)^{d/4}\, t^{r/2-1}\, e^{-t}/\Gamma(r/2)$.

Proof. First we show that $\hat{\beta}_r(s) = I/\Gamma(r/2)$, where

$$I = \int_0^\infty t^{r/2-1}\, e^{-t}\, e^{-t\|s\|^2}\, dt.$$

Indeed, putting $u = t(1 + \|s\|^2)$, so $dt = du(1 + \|s\|^2)^{-1}$, we get

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

By (2.1) $\|\gamma_t\|_{\mathcal{L}^2} = (\pi/2t)^{d/4}$, so $\hat{\beta}_r(s) = \int_0^\infty (\pi/2t)^{d/4}\, t^{r/2-1}\, e^{-t}/\Gamma(r/2)\, \gamma_t^o(s)dt$. □

The next proposition gives an upper bound on G_0^o-variation of $\hat{\beta}_r$.

Proposition 5 *For d a positive integer and $r > d/2$,*

$$\|\hat{\beta}_r\|_{G^o} \le \|\hat{\beta}_r\|_{G_0^o} \le \int_0^\infty w_r(t)dt = k(r, d),$$

where $k(r, d) = \frac{(\pi/2)^{d/4}\Gamma(r/2-d/4)}{\Gamma(r/2)}$.

Proof. A straightforward calculation shows that the $\mathcal{L}^1$-norm of the weighting function w_r is the same as the $\mathcal{L}^1$-norm of the weighting function v_r and the upper bound follows from Theorem 1 as in Proposition 3 but with $\phi(x, y) = \phi(x, t) = \gamma_t^o(x)$. □

Because the Fourier transform is an isometry on $\mathcal{L}^2$, by (2.10) the functions β_r and $\hat{\beta}_r$ have the same variation with respect to G_0^o. Propositions 3 and 5 give the same upper bound $k(r, d)$ on this number. If for some fixed $c > 0$, $r_d = d/2 + c$, then $k(r_d, d) \to 0$ exponentially fast as $d \to \infty$.

An application of (2.9) with Propositions 3 or 5 shows the following result:

Theorem 2 *For d, n positive integers and $r > d/2$*
$\|\beta_r - \mathrm{span}_n\, G_0\|_{\mathcal{L}^2} = \|\hat{\beta}_r - \mathrm{span}_n\, G_0\|_{\mathcal{L}^2} \le k(r, d)\, n^{-1/2}$.

As above, for $c > 0$ and d large enough, The theorem shows that the Bessel potential of order $r_d = d/2 + c$ can be well-approximated by a network with just one Gaussian unit; hence, β_{r_d} is close in $\mathcal{L}^2$-norm to a multiple of some d-dimensional Gaussian centered at the origin.

4 Approximation of smooth functions by Gaussian RBFs

In this section we estimate rates of approximation by Gaussian RBF for functions in the Bessel potential spaces. To obtain the estimates we first derive upper bounds on variation with respect to the set of translated Bessel potentials and then combine them with the estimates of G_0-variation of Bessel potentials from the previous section.

Let $h * g$ denote the *convolution* of two functions h and g,

$$(h * g)(x) = \int_{\mathbb{R}^d} h(y)g(x - y)dy.$$

For d a positive integer, $r > d/2$, and $q \in [1, \infty]$, the *Bessel potential space* (with respect to $\mathbb{R}^d$) [23, pp.134-136] denoted by $(L^{q,r}, \|.\|_{L^{q,r}})$ is defined as

$$L^{q,r} := \{f \mid f = w * \beta_r,\ w \in \mathcal{L}^q\}$$

and

$$\|f\|_{L^{q,r}} := \|w\|_{\mathcal{L}^q} \quad \text{for} \quad f = w * \beta_r.$$

Since the Fourier transform (2.2) of a convolution is $(2\pi)^{d/2}$ times the product of the transforms, we have $\hat{w} = (2\pi)^{-d/2}\hat{f}/\hat{\beta}_r$. Thus $w = (2\pi)^{-d/2}(\hat{f}/\hat{\beta}_r)\check{}$ is uniquely determined by f and so the Bessel potential norm is well-defined.

For τ_y the translation operator $(\tau_y f)(x) = f(x-y)$ let

$$G_{\beta_r} = \{\tau_y \beta_r \,|\, y \in \mathbb{R}^d\}$$

denote the *set of translates of the Bessel potential of order* r. For $r > d/2$, β_r belongs to $\mathcal{L}^2$; since translation does not change the $\mathcal{L}^2$-norm, $G_{\beta_r} \subset \mathcal{L}^2$.

The $\mathcal{L}^2$-norm of β_r can be calculated by switching to $\hat{\beta}_r$ and using Plancherel's equality (2.5). For every $r > d/2$

$$\|\beta_r\|_{\mathcal{L}^2} = \|\hat{\beta}_r\|_{\mathcal{L}^2} = \lambda(r,d) := \pi^{d/4}\left(\frac{\Gamma(r-d/2)}{\Gamma(r)}\right)^{1/2}. \tag{4.1}$$

Indeed, using radial symmetry $\|\hat{\beta}_r\|^2_{\mathcal{L}^2} = \int_{\mathbb{R}^d}(1+\|x\|^2)^{-r}dx = \omega_d I$, where $\omega_d := 2\pi^{d/2}/\Gamma(d/2)$ is the area of the unit sphere in $\mathbb{R}^d$ [5, p. 303] and $I = \int_0^\infty (1+\rho^2)^{-r}\rho^{d-1}d\rho$. Substituting $\sigma = \rho^2$, one gets $d\rho = (1/2)\sigma^{-1/2}d\sigma$; hence,

$$I = (1/2)\int_0^\infty \frac{\sigma^{d/2-1}}{(1+\sigma)^r}d\sigma = \frac{\Gamma(d/2)\Gamma(r-d/2)}{2\Gamma(r)}$$

(see [4, p. 60] for the last equality).

As all elements of G_{β_r} have the same $\mathcal{L}^2$-norm equal to $\lambda(r,d)$,

$$\|.\|_{G^o_{\beta_r}} = \lambda(r,d)\|.\|_{G_{\beta_r}}. \tag{4.2}$$

Functions in the Bessel potential space are convolutions with β_r which are integral formulas. Thus we get the following upper bound:

Proposition 6 *Let d be a positive integer, $r > d/2$, $w : \mathbb{R}^d \to \mathbb{R}$ continuous except on a closed set Z_0 of measure zero, $w \in \mathcal{L}^1$, and $f = w * \beta_r$. Then*
$\|f\|_{G_{\beta_r}} \le \|w\|_{\mathcal{L}^1} = \|f\|_{L^{1,r}}$.

Proof. The bounds follow from Theorem 1, applied to the integral formula $f(x) = \int w(y)\beta_r(x-y)dy = \int w(y)\lambda(r,d)\beta^o_r(x-y)dy$ combined with (4.2). Take $Y = \mathbb{R}^d$, $Y_0 = Z_0$, let $\phi(x,y) = \beta^o_r(x-y)$, and let $w(y)\lambda(r,d)$ be the weight function. The condition $r > d/2$ is needed to ensure that $G_{\beta_r} \subset \mathcal{L}^2$. □

For $h : U \to \mathbb{R}$, U a topological space, let $\operatorname{supp} h = \operatorname{cl}\{u \in U \,|\, h(u) \neq 0\}$.

Proposition 7 *Let d be a positive integer, $r > d/2$, $w \in \mathcal{L}^2$ continuous except on a closed set of measure zero, $\lambda(\operatorname{supp} w) = a < \infty$, and $f = w * \beta_r$. Then*

$$\|f\|_{G_{\beta_r}} \le a^{1/2}\|f\|_{L^{2,r}}.$$

Proof. By the Cauchy-Schwartz inequality, $\|w\|_{\mathcal{L}^1} \le a^{1/2}\|w\|_{\mathcal{L}^2} = a^{1/2}\|f\|_{L^{2,r}}$. But by Proposition 6, $\|f\|_{G_{\beta_r}} \le \|w\|_{\mathcal{L}^1}$. □

These estimates of variations give an upper bound on rates of approximation by linear combinations of n translates of the Bessel potential β_r.

Theorem 3 *Let d, n be positive integers, $r > d/2$, w continuous except on a closed set of measure zero, $f = w * \beta_r$, and $\lambda(r,d) = \pi^{d/4}\left(\frac{\Gamma(r-d/2)}{\Gamma(r)}\right)^{1/2}$.*
(i) For $w \in \mathcal{L}^1$,

$$\|f - \mathrm{span}_n G_{\beta_r}\|_{\mathcal{L}^2} \leq (\lambda(r,d)\|f\|_{L^{1,r}})\, n^{-1/2}.$$

(ii) For $w \in \mathcal{L}^2$ with $a = \lambda(\mathrm{supp}\, w) < \infty$,

$$\|f - \mathrm{span}_n G_{\beta_r}\|_{\mathcal{L}^2} \leq \left(a^{1/2}\lambda(r,d)\|f\|_{L^{2,r}}\right) n^{-1/2}.$$

Proof. (i) By Proposition 6, (4.2), and (2.9).
(ii) As in Proposition 7, $w \in \mathcal{L}^2$ and $\mathrm{supp}(w) = a < \infty$ implies $w \in \mathcal{L}^1$; the rest follows from Proposition 7. □

Composing estimates of variations with respect to sets of translated Bessel potentials and Gaussians, we get an upper bound on rates of approximation by networks with n Gaussian RBF units for functions from Bessel spaces.

Theorem 4 *Let d, n be positive integers, $r > d/2$, w continuous except on a closed set of measure zero, $f = w * \beta_r$, and $k(r,d) = \frac{(\pi/2)^{d/4}\Gamma(r/2-d/4)}{\Gamma(r/2)}$.*
(i) For $w \in \mathcal{L}^1$,

$$\|f - \mathrm{span}_n G\|_{\mathcal{L}^2} \leq (k(r,d)\,\|f\|_{L^{1,r}})\, n^{-1/2}.$$

(ii) For $w \in \mathcal{L}^2$ and $\lambda(\mathrm{supp}\, w) = a < \infty$,

$$\|f - \mathrm{span}_n G\|_{\mathcal{L}^2} \leq \left(k(r,d)\, a^{1/2}\,\|f\|_{L^{2,r}}\right) n^{-1/2}.$$

Proof. By (2.9), $\|f - \mathrm{span}_n G\|_{\mathcal{L}^2} \leq \|f\|_{G^o} n^{-1/2}$. By Lemma 1 with $\mathcal{X} = \mathcal{L}^2, F = G^o, H = G_{\beta_r}$, using Proposition 3 and the fact that G^o is closed under translations, we have

$$\|f\|_{G^o} \leq \sup\{\|\tau_y(\beta_r)\|_{G^o} \,|\, y \in \mathbb{R}^d\}\|f\|_{G_{\beta_r}} \leq k(r,d)\|f\|_{G_{\beta_r}}.$$

Then the statements follow from the upper bounds on $\|f\|_{G_{\beta_r}}$ given in Propositions 6 and 7, resp. □

5 Upper bounds in terms of Sobolev norms

In this section, bounds on approximation by Gaussian RBFs are given in terms of Sobolev norms. Two norms are *equivalent* if each is bounded by a multiple of the other. The equivalence between Sobolev and Bessel potential norms is well-known (e.g., [24] or [1, p. 252]). As constants of equivalence were not readily available, we derive one of them here in a special case.

Let r be a positive integer and let $W^{2,r}$ denote the *Sobolev space* of functions with t-th order partials in $\mathcal{L}^2$ for $0 \leq t \leq r$ [24] with corresponding norm

$$\|f\|_{W^{2,r}} = \left(\sum_{|\alpha|\leq r} \|D^\alpha f\|^2_{\mathcal{L}^2}\right)^{1/2},$$

where α denotes a multi-index (i.e., a vector of non-negative integers), D^α denotes the corresponding partial derivative operator, and $|\alpha| = \alpha_1 + \cdots + \alpha_d$.

Let d and r be positive integers, $r > d/2$, $w \in \mathcal{L}^2$, and $f = w * \beta_r$. Then

$$\|f\|_{L^{2,r}} \leq (2\pi)^{-d/2}(r!)^{1/2}\|f\|_{W^{2,r}} \tag{5.1}$$

Indeed, since $f = w * \beta_r$, $\hat{f} = (2\pi)^{d/2}\hat{w}\hat{\beta}_r$ and so

$$\|f\|_{L^{2,r}} = (2\pi)^{-d/2}\|\hat{f}/\hat{\beta}_r\|_{\mathcal{L}^2} = (2\pi)^{-d/2}\left(\int_{\mathbb{R}^d}|\hat{f}(s)|^2(1+|s|^2)^r ds\right)^{1/2}.$$

Let $\binom{r}{\sigma}$ denote the multinomial coefficient $r!/\sigma_1!\dots\sigma_t!$. Note that $(1+|s|^2)^r = \sum_{|\sigma|=r}\binom{r}{\sigma}|u^{2\sigma}|$, for $u \in \mathbb{R}^{d+1}$ defined by $u_j = s_j$, $j = 1,\dots,d$, $u_{d+1} = 1$, for $\sigma = (\sigma_1,\dots\sigma_{d+1}) \in \mathbb{N}^{d+1}$ a multi-index of length $d+1$, and $|u^{2\sigma}| = |u_1^{2\sigma_1}\cdots u_{d+1}^{2\sigma_{d+1}}|$. Hence, we have

$$\int_{\mathbb{R}^d}|\hat{f}(s)|^2(1+|s|^2)^r ds \le c(r,d)\int_{\mathbb{R}^d}|\hat{f}(s)|^2\sum_{|\alpha|\le r}|s^{2\alpha}|ds,$$

where $c(r,d) = \max\left\{\binom{r}{\sigma} \,\middle|\, |\sigma| = r\right\}$. It follows from basic properties of the Fourier transform that the integral on the right-hand side is the square of the Sobolev norm of f; see, e.g., [24, p. 162]. Clearly, $c(r,d) \le r!$, and equality holds if and only if $r \le d$. This establishes (5.1).

Thus the larger the dimension, the more the magnitudes of these two equivalent norms differ. We can now estimate the rate of approximation by scaled and translated Gaussians in terms of the Sobolev norm of the function to be approximated.

Theorem 5 *Let d, n, r be positive integers, $r > d/2$, w continuous except on a closed set of measure zero, and $f = w * \beta_r$.*
For $w \in \mathcal{L}^2$ and $\lambda(\operatorname{supp} w) = a < \infty$,

$$\|f - \operatorname{span}_n G\|_{\mathcal{L}^2} \le \left(\left(\frac{\pi}{8}\right)^{d/4}\frac{\Gamma(r/2 - d/4)}{\Gamma(r/2)}a^{1/2}(r!)^{1/2}\right)\|f\|_{W^{2,r}}n^{-1/2}.$$

Proof. Using Theorem 4.4(ii) and (5.1), the $\mathcal{L}^2$-distance from f to $\operatorname{span}_n G$ is at most $(k(r,d)\,a^{1/2}$ $(2\pi)^{-d/2}(r!)^{1/2}\|f\|_{W^{2,r}})n^{-1/2}$, and the result follows. □

Acknowledgements

Collaboration of V. K. and M. S. partially supported by 2004-2006 Scientific Agreement among University of Genoa, National Research Council of Italy, and Academy of Sciences of the Czech Republic, V. K. partially supported by GA ČR grant 201/05/0557 and by the Institutional Research Plan AV0Z10300504; M. S. by PRIN Grant from the Italian Ministry for University and Research, project New Techniques for the Identification and Adaptive Control of Industrial Systems; collaboration of P. C. K. with V. K. and M. S. partially supported by Georgetown University.

Bibliography

[1] Adams, R. A., Fournier, J. J. F.: Sobolev Spaces. Academic Press, Amsterdam (2003)

[2] Barron, A. R.: Neural net approximation. Proc. 7th Yale Workshop on Adaptive and Learning Systems, K. Narendra, Ed., Yale University Press (1992) 69–72

[3] Barron, A. R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory **39** (1993) 930–945

[4] Carlson, B. C.: Special Functions of Applied Mathematics, Academic Press, New York (1977)

[5] Courant, R.: Differential and Integral Calculus, vol. 2, Wiley, New York, 1936 (1964 edition, transl. E. J. McShane)

[6] Girosi, F.: Approximation error bounds that use VC-bounds. In Proceedings of the International Conference on Neural Networks, Paris (1995) 295-302

[7] Girosi, F., Anzellotti, G.: Rates of convergence for radial basis functions and neural networks. Memo AI-Lab, MIT, 1995.

[8] Girosi, F., Anzellotti, G.: Rates of convergence for radial basis functions and neural networks. In Artificial Neural Networks for Speech and Vision, R. J. Mammone (Ed.), Chapman & Hall, London (1993) 97-113

[9] Hartman, E. J., Keeler, J. D., Kowalski, J. M.: Layered neural networks with Gaussian hidden units as universal approximations. Neural Computation **2** (1990) 210–215

[10] Jones, L. K.: 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

[11] Kainen, P. C., Kůrková, V., Sanguineti, M.: Estimates of approximation rates by Gaussian radial-basis functions. In *Adaptive and Natural Computing Algorithms - ICANNGA'07* (Eds. B. Beliczynski, A. Dzielinski, M. Iwanowski, B. Ribeiro), Part II, LNCS 4432 (pp.11-18). Berlin, Heidelberg: Springer-Verlag, 2007, to appear.

[12] Kon, M. A., Raphael, L. A., Williams, D. A.: Extending Girosi's approximation estimates for functions in Sobolev spaces via statistical learning theory. J. of Analysis and Applications **3** (2005) 67-90

[13] Kůrková, V.: Dimension–independent rates of approximation by neural networks. In Computer–Intensive Methods in Control and Signal Processing: Curse of Dimensionality, K. Warwick and M. Kárný, Eds., Birkhäuser, Boston (1997) 261–270

[14] Kůrková, V.: High-dimensional approximation and optimization by neural networks. Chapter 4 in Advances in Learning Theory: Methods, Models and Applications, J. Suykens et al., Eds., IOS Press, Amsterdam (2003) 69–88

[15] Kůrková, V., Kainen, P. C., Kreinovich, V.: Estimates of the number of hidden units and variation with respect to half-spaces. Neural Networks **10** (1997) 1061–1068

[16] Kůrková, V., Sanguineti, M.: Comparison of worst case errors in linear and neural network approximation. IEEE Trans. on Information Theory **48** (2002) 264–275.

[17] Martínez, C., Sanz, M.: The Theory of Fractional Powers of Operators. Elsevier, Amsterdam (2001)

[18] Mhaskar, H. N.: Versatile Gaussian networks. Proc. IEEE Workshop of Nonlinear Image Processing (1995) 70-73

[19] Mhaskar, H. N., Micchelli, C. A.: Approximation by superposition of a sigmoidal function and radial basis functions. Advances in Applied Mathematics **13** (1992) 350-373

[20] Park, J., Sandberg, I. W.: Universal approximation using radial–basis–function networks. Neural Computation **3** (1991) 246-257

[21] Park, J., Sandberg, I.: Approximation and radial basis function networks. Neural Computation **5** (1993) 305-316

[22] Pisier, G.: Remarques sur un resultat non publié de B. Maurey. In Seminaire d'Analyse Fonctionelle, vol. I(12). École Polytechnique, Centre de Mathématiques, Palaiseau 1980-1981

[23] Stein, E. M.: Singular Integrals and Differentiability Properties of Functions. Princeton University Press, Princeton, NJ (1970)

[24] Strichartz, R.: A Guide to Distribution Theory and Fourier Transforms. World Scientific, NJ (2003)