Mach Learn
DOI 10.1007/s10994-013-5406-z

# Almost optimal estimates for approximation and learning by radial basis function networks

**Shaobo Lin · Xia Liu · Yuanhua Rong · Zongben Xu**

Received: 29 August 2012 / Accepted: 8 August 2013


**Abstract** This paper quantifies the approximation capability of radial basis function networks (RBFNs) and their applications in machine learning theory. The target is to deduce almost optimal rates of approximation and learning by RBFNs. For approximation, we show that for large classes of functions, the convergence rate of approximation by RBFNs is not slower than that of multivariate algebraic polynomials. For learning, we prove that, using the classical empirical risk minimization, the RBFNs estimator can theoretically realize the almost optimal learning rate. The obtained results underlie the successful application of RBFNs in various machine learning problems.

**Keywords** Learning theory · Approximation theory · Radial basis function networks · Rate of convergence

## 1 Introduction

In physical or biological systems, engineering applications, financial studies, and many other fields, only a finite data set $(x_i, y_i)_{i=1}^m$ be obtained. Learning means synthesizing a function that best represents the relation between the inputs and the corresponding outputs. A learning system is normally developed for defining the function and yielding an estimator. The learning system comprises a hypothesis space, a family of parameterized functions that regulate the forms and properties of the estimator to be found, and a learning strategy or learning algorithm that numerically yields the parameters of the estimator. The central question of learning is and will always be: how well does the synthesized function generalized to reflect the reality that the given samples purport to show us.

The analysis of a learning system can be regarded as studying approximation capability of the hypothesis space and efficiency of the learning strategy. From the point of view

Editor: Paolo Frasconi.

S. Lin · X. Liu · Y. Rong · Z. Xu (✉)
Institute for Information and System Sciences, School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, Shanxi Province, P.R. China
e-mail: zbxu@mail.xjtu.edu.cn

Published online: 18 September 2013

of approximation theory, there are a multitude of ways to choose the hypothesis space, for example, multivariate polynomials, splines, tensor products, radial basis function networks (RBFNs), etc. All of these choices have their own advantages, but have some common disadvantages as well. If the dimension of the considered problem (the number of variables) is large, that is often the case in many applications, a reasonable choice is RBFNs, which is, incidentally, also highly useful in lower dimensional problems.

RBFNs can be formally described as the devices producing input-output mappings depending on some adjustable parameters. The input-output functions take the form of linear combinations of radial functions by units, and can be evaluated in hardware using parallel computation of every units. RBFNs have been extensively used in many fields such as computer graphics (Wendland 2005), adaptive numerical solutions to differential equations (Fedoseyev et al. 2002; Flyer and Wright 2009), machine learning (Caponnetto and DeVito 2007; Cucker and Smale 2001), etc.

A typical issue in RBFN approximation, called the density problem, concerns whether RBFN can approximate an arbitrary function to any desired accuracy by increasing the number of hidden neurons. Under certain assumptions on the activation function, this problem was perfectly resolved in the seminal paper (Park and Sandberg 1991). Similar results can also be found in Park and Sandberg (1993) and Chen and Chen (1995). Another fundamental issue in RBFN approximation is the complexity problem which describes the relationship between the accuracy of approximation and the number of hidden neurons. Generally speaking, the study of complexity problem is more important and difficult than the density problem, since in the former case, we are concerned with not only how many computational units are needed to attain a prescribed accuracy, but also the judgement whether this number can be reduced.

The complexity problem of RBFN approximation were widely studied in Bumann et al. (1995), Buhman (2000), Johnson (1998), Mhaskar (1996), Powell (1990), Schaback (1995, 1996), Wendland (2000) and references therein. More precisely, several important upper (and lower) bound estimations for RBFN approximation have been deduced for some specific activation functions such as Gaussian, thin-plate spline, etc. However, it is still unclear whether these estimations are available for RBFNs with more general activation functions. In this paper, we take an excursion in studying the approximation capability of RBFNs with general activation functions. By imposing activation functions certain restrictions, we show that for non-polynomial target functions, the approximation rate of RBFNs is not slower than that of multivariate algebraic polynomials. Thus, the approximation property of algebraic polynomials automatically provides an upper bound estimation for RBFN approximation. Furthermore, we verify that the established upper bound is almost optimal in the sense that up to a logarithmical factor the upper and lower bounds are asymptotically identical.

According to the well known "bias" and "variance" problem in learning theory (Cucker and Smale 2001) a learning system should reflect a trade-off between the approximation capability and complexity of the hypothesis space. Therefore, from approximation to learning, we should also take account of the price to be paid to get a given accuracy of approximation. Past researches on learning (e.g. Caponnetto and DeVito 2007; Cucker and Smale 2001, 2002; Temlyakov 2008) has been mainly carried out within the theoretical framework of reproducing kernel Hilbert space (RKHS). RKHSs are by definition the Hilbert spaces of functions where point evaluations are continuous linear functionals. This makes the sampling be stable and effective. But, the dimension of RKHS is usually infinite, which implies that the cost of learning by using RKHS method is tremendously high. This observation urges us to search for hypothesis spaces with lower complexities and similar approximation capabilities. Because of their prominent approximation capabilities, RBFNs are natural

alternatives. From the previous work of Maiorov (2006a, 2006b) and Maiorov and Meir (2001), we know that the complexity of RBFN manifold is much lower than that of RKHS. Hence, taking RBFN manifold as the hypothesis space of learning process should be a more reasonable choice. In this paper, using the well known empirical risk minimization rule in the RBFN manifolds, we conclude that such choice is almost optimal in a certain sense.

The rest of paper is organized as follows. In Sect. 2, we present some preliminaries about statistical learning theory and RBFN manifolds. In Sect. 3, we analyze the approximation capacity of RBFNs. In Sect. 4, we derive the almost optimal learning rate of RBFNs. In Sect. 5, we then present all related proofs.

## 2 Preliminaries

In this section, we give a fast review of statistical learning theory and RBFN manifolds.

### 2.1 Statistical learning theory

Let $M \geq 0$, $X \subseteq \mathbb{R}^d$ be the input space and $Y \subseteq [-M, M]$ be the output space. Suppose that the unknown probability measure $\rho$ on $Z := X \times Y$ admits the decomposition

$$\rho(x, y) = \rho_X(x)\rho(y|x).$$

Let $\mathbf{z} = (x_i, y_i)_{i=1}^m$ be a finite random sample of size $m$, $m \in \mathbb{N}$, drawn independently and identically according to the unknown distribution $\rho$. Suppose further that $f : X \to Y$ is a function that one uses to model the correspondence between $X$ and Y, as induced by $\rho$. One natural measurement of the error incurred by using $f$ of this purpose is the generalization error, defined by

$$\mathcal{E}(f) := \int_Z \big(f(x) - y\big)^2 d\rho,$$

which is minimized by the regression function (Cucker and Smale 2001), defined by

$$f_\rho(x) := \int_Y y d\rho(y|x).$$

We do not know this ideal minimizer $f_\rho$, since $\rho$ is unknown, but we have access to random examples from $X \times Y$ sampled according to $\rho$.

Let $L^2_{\rho_X}$ be the Hilbert space of $\rho_X$ square integrable function on $X$, with norm denoted by $\|\cdot\|_\rho$. With the assumption that $f_\rho \in L^2_{\rho_X}$, it is known that, for every $f \in L^2_{\rho_X}$, there holds

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2. \tag{1}$$

The task of the least square regression problem is then to construct functions $f_{\mathbf{z}}$ that approximates $f_\rho$, in the norm $\|\cdot\|_\rho$, using the finite sample $\mathbf{z}$.

So, the goal of learning is to find the best approximation of the regression function $f_\rho$ within a space $\mathcal{H}$. We need still to address the question of how to find an estimator $f_{\mathbf{z}}$ to $f_\rho$. A popularly used approach is the following empirical risk minimization process. Define the empirical risk of $f \in \mathcal{H}$ by

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^m \big(f(x_i) - y_i\big)^2.$$

We denote by

$$f_{\mathbf{z}} := \arg\min_{f\in\mathcal{H}} \mathcal{E}_{\mathbf{z}}(f),$$

and call it the empirical minimizer. Given a finite ball in a finite dimensional manifold, the problem of finding $f_{\mathbf{z}}$ is numerically executable (Györfy et al. 2002).

### 2.2 Complexity of RBFN manifold

A RBFN can be mathematically expressed as

$$R_{\sigma,N}(x) := \sum_{j=0}^{N} c_j \sigma\big(w_j|x-\theta_j|\big), \quad c_j, w_j \in \mathbb{R},\ \theta_j \in \mathbb{R}^d, \tag{2}$$

where $N \in \mathbb{N}$, $\mathbb{N}$ denotes the set of natural numbers, $|A|$ is the Euclidean norm of vector $A$ and $\sigma$ is the activation function of RBFN. Both coefficients of the linear combinations $c_j$ and parameters of the computation units $w_j$ and $\theta_j$ are adjustable in the process of learning.[1] We denote by $\Phi_{\sigma,N}$ the collection of functions formed as (2). Then, it is well known that $\Phi_{\sigma,N}$ is a nonlinear manifold since the sum of two elements sometimes does not belong to $\Phi_{\sigma,N}$.

Due to the nonlinearity, the complexity of the manifold $\Phi_{\sigma,N}$ can not be measured by the usual dimension of linear space. Thus, some other quantities should be introduced. Three widely used measurements are $\varepsilon$-entropy, Vapnik-Chervonenkis (VC) dimension and pseudo-dimension (Maiorov 2006a; Mendelson and Vershinin 2003). The concept $\varepsilon$-entropy of a set is closely connected to the pseudo-dimension (or VC-dimension), which is stated as follows.

Let $B$ be a Banach space and $V$ a compact set in $B$. The quantity $H_\varepsilon(V,B) = \log_2 \mathcal{N}_\varepsilon(V,B)$, where $\mathcal{N}_\varepsilon(V,B)$ is the number of elements in least $\varepsilon$-net of $V$, is called $\varepsilon$-entropy of $V$ in $B$. The quantity $\mathcal{N}_\varepsilon(V,B)$ is called the $\varepsilon$-covering number of $V$. For any $t \in \mathbb{R}$, define

$$\operatorname{sgn}(t) := \begin{cases} 1, & \text{if } t \geq 0, \\ -1, & \text{if } t < 0. \end{cases}$$

If a vector $\mathbf{t} = (t_1, \ldots, t_n)$ belongs to $\mathbb{R}^n$, then we denote by $\operatorname{sgn}(\mathbf{t})$ the vector $(\operatorname{sgn}(t_1), \ldots, \operatorname{sgn}(t_n))$. Let $\mathbb{B}^d$ be the unit ball in the $d$-dimensional Euclidean space $\mathbb{R}^d$ The VC dimension of a set $V$ over $\mathbb{B}^d$, denoted as $VC\dim(V,\mathbb{B}^d)$, is defined as the maximal natural number $m$ such that there exists a collection $(\mu_1, \ldots, \mu_m)$ in $\mathbb{B}^d$ such that the cardinality of the sgn-vectors set

$$S = \big\{\big(\operatorname{sgn}\big(v(\mu_1)\big), \ldots, \operatorname{sgn}\big(v(\mu_m)\big)\big) : v \in V\big\}$$

equals to $2^m$, that is, the set $S$ coincides with the set of all vertexes of unit cube in $\mathbb{R}^m$. The quantity

$$P\dim\big(V,\mathbb{B}^d\big) := \max_g VC\dim\big(V+g,\mathbb{B}^d\big),$$

is called pseudo-dimension of the set $V$ over $\mathbb{B}^d$, where $g$ runs all functions defined on $\mathbb{B}^d$ and $V+g = \{v+g : v \in V\}$.

Mendelson and Vershinin (2003) (see also Maiorov 2006a) has established the following important relation between the Pseudo-dimension and $\varepsilon$-entropy. This relation together with Lemma 4 below will play a key role in deducing the upper bound of RBFN learning.

[1] In the framework of kernel learning, the coefficients $w_j$'s are assumed to be 1.

**Lemma 1** *Let $V(\mathbb{B}^d)$ be a class of functions which consists of all functions $f \in V$ satisfying $|f(x)| \leq R$ for all $x \in \mathbb{B}^d$. Then,*

$$H_\varepsilon\big(V\big(\mathbb{B}^d\big), L^2\big(\mathbb{B}^d\big)\big) \leq cP \dim\big(V, \mathbb{B}^d\big) \log_2 \frac{R}{\varepsilon},$$

*where $c$ is an absolute positive constant.*

## 3 Approximation by RBFNs

### 3.1 RBFN and polynomial approximation: a comparison

Denote by $L^p(\mathbb{B}^d)$, $(0 < p < \infty)$ the space of real valued and $p$-integrable functions on $\mathbb{B}^d$ endowed with the norm or (quasi-norm)

$$\|f\|_p := \|f\|_{L^p(\mathbb{B}^d)} := \left\{\int_{\mathbb{B}^d} \big|f(x)\big|^p dx\right\}^{1/p} < \infty,$$

$C(\mathbb{B}^d)$ the space of continuous functions with the norm

$$\|f\|_{C(\mathbb{B}^d)} := \max_{x\in\mathbb{B}^d} \big|f(x)\big|.$$

For the sake of simplicity, we denote $\|f\|_\infty := \|f\|_{C(\mathbb{B}^d)}$ and $L^\infty(\mathbb{B}^d) := C(\mathbb{B}^d)$.

Denote by $\mathcal{P}_n := \mathcal{P}_n(\mathbb{B}^d)$ the space of multivariate algebraical polynomials

$$P_n(x) := \sum_{|\mathbf{k}|\leq n} c_{\mathbf{k}} x^{\mathbf{k}}, \quad x \in \mathbb{B}^d,$$

where $x = (x_{(1)}, \ldots, x_{(d)})$, $x^{\mathbf{k}} := x_{(1)}^{k_1} \cdots x_{(d)}^{k_d}$ and $c_{\mathbf{k}} := c_{k_1,k_2,\ldots,k_d} \in \mathbb{R}$. It is obvious that the dimension of the linear space $\mathcal{P}_n$ is $\binom{n+d}{d}$ (see Wendland 2005 for example). For any two sets of functions $W, U \in L^p(\mathbb{B}^d)$, we denote also by

$$E(W, U)_{L^p(\mathbb{B}^d)} := \sup_{f\in W} E(f, U)_{L^p(\mathbb{B}^d)} := \sup_{f\in W} \inf_{g\in U} \|f - g\|_p$$

the distance of $W$ and $U$.

To compare the approximation capabilities between two classes of functions, both the approximation errors and the capacities of these classes should be taken into account. When these classes of functions are parameterized families, their capacities can be measured by the length of parameter vectors (depending on the number of variables e.g., on the degree of an algebraic polynomial, on the number of knots in a spline, on the number of hidden units in a RBFN, etc.). Our first result (Theorem 2) focuses on comparing the approximation capability of RBFNs with that of polynomials. To this end, we should build a convergence rate analysis for RBFN approximation whose target functions are algebraic polynomials.

**Theorem 1** *Let $N$ and $n$ be any natural numbers satisfying $N \geq (2d+5)n\binom{n-1+d}{d}$. If $\sigma : \mathbb{R} \to \mathbb{R}$ is a function with $(n+1)$-times bounded derivatives, and $\sigma^{(s)}(0) \neq 0$ for $s = 0, 1, \ldots, n$, then, for arbitrary $P_n \in \mathcal{P}_n$ and arbitrary $\varepsilon > 0$, there exists an RBFN, $L_N^\sigma$, formed as (2), such that*

$$\big|P_n(x) - L_N^\sigma(x)\big| < \varepsilon. \tag{3}$$

It should be noted that at the first glance, the restrictions to the activation functions in Theorem 1 seem a bit strong, and, further, we can check that the well known Gaussian function does not satisfy the assumptions. This constraint can actually be relaxed and coped with by adding a threshold to the RBFNs. The following Corollary 1 states such a variant.

**Corollary 1** *Let $N$ and $n$ be any natural numbers satisfying $N \geq (2d+5)n\binom{n-1+d}{d}$. If $\sigma : \mathbb{R} \to \mathbb{R}$ is a function with $(n+1)$-times bounded derivatives, and there exists at least a point $t_0 \in [0,1)$ such that $\sigma^{(s)}(t_0) \neq 0$ for $s = 0, 1, \ldots, n$, then, for arbitrary $P_n \in \mathcal{P}_n$ and arbitrary $\varepsilon > 0$, there exists an RBFN, $R_N^\sigma$, formed as*

$$\sum_{i=1}^{N} c_k \sigma\big(w_k|x-\theta_k| + t_0\big), \quad c_k, w_k \in \mathbb{R},\ \theta_k \in \mathbb{R}^d$$

*such that*

$$\big|P_n(x) - R_N^\sigma(x)\big| < \varepsilon. \tag{4}$$

There are many activation functions satisfying the assumptions of Corollary 1. For example, the functions having $(n+1)$-th continuous derivatives, which are not algebraic polynomials of degree $n$, satisfy the assumption. As a very special case, the well known Gaussian functions meet the requirement of Corollary 1, and therefore, very commonly used in practice. To further characterize the approximation property of RBFNs , we prove the following Theorem 2, which establishes a relationship between RBFN and algebraic polynomial approximation.

**Theorem 2** *Let $0 < p \leq \infty$, $N$ and $n$ be any natural numbers such that $N \geq (2d+5) \times n\binom{n-1+d}{d}$. If $\sigma : \mathbb{R} \to \mathbb{R}$ is a function with $(n+1)$-times bounded derivatives, and $\sigma^{(s)}(0) \neq 0$ for $s = 0, 1, \ldots, n$, then, for any $f \in L^p(\mathbb{B}^d)$ and arbitrary $\varepsilon > 0$, there holds*

$$E(f, \Phi_{\sigma,N})_{L^p(\mathbb{B}^d)} \leq E(f, \mathcal{P}_n)_{L^p(\mathbb{B}^d)} + \varepsilon. \tag{5}$$

The relationship between RBFN and polynomial approximation has already studied in Maiorov (2003) and Lin et al. (2011a). In Maiorov (2003), Maiorov studied the approximation properties of the radial function manifold $\mathcal{G}_N$ whose elements take the form as

$$G_N(x) := \sum_{j=0}^{N} c_j g_j\big(|x - t_i|\big), \quad c_j \in \mathbb{R},\ g_j \in C(\mathbb{R}),\ t_i \in \mathbb{R}^d. \tag{6}$$

They proved that the approximation capability of $\mathcal{G}_N$ is not worse than that of polynomials of degrees at most $n$ provided $N \geq (2d+5)\binom{n-1+d}{d}$. Conversely, Lin et al. (2011a) deduced that if the target function is radial and $N \sim n^{d-1}$, then the approximation rate of polynomials is also not slower than that of $\mathcal{G}_N$. Noting that the utilized approximants in (6) are linear combinations of different univariate functions, it is difficult to determine the capacity of $\mathcal{G}_N$. Thus, we can not say anything about the comparison between $\mathcal{G}_N$ and $\mathcal{P}_n$, since they are not in the same framework. Furthermore, $\mathcal{G}_N$ is not a parameterized family, which makes it computational infeasible. Differently, since $\Phi_{\sigma,N}$ is parameterized, the capacity of $\Phi_{\sigma,N}$ can be measured by $N$. Theorem 2 shows that if the lengths of parameters of $\Phi_{\sigma,N}$ and $\mathcal{P}_n$ are comparable, i.e., $N = (2d+5)n\binom{n-1+d}{d} \sim n^d \sim \binom{n+d}{d}$, then for arbitrary non-polynomial

function, the approximation rate of RBFN is not slower than that of polynomials. Noting that the comparison is employed into a unified framework, we can draw the conclusion that, as far as the approximation capability is concerned, RBFN is at least not worse than polynomial.

A consensus on RBFN approximation is that it can break the "curse of dimensionality". The results in Barron (1993), Burger and Neubauer (2001), Mhaskar (2004) and Kainen et al. (2012) verified this statement by deducing approximation rates at least $N^{-1/2}$, which is independent of $d$. However, it can be also found in these papers that, to achieve such dimensional-independent approximation rates, the target functions should depend heavily on the activation functions. It was pointed out in Barron et al. (2008, p. 68) that such restrictions may become more and more strong as the dimension $d$ grows. Thus, although the approximation error of RBFNs is independent of the dimension, the applicable target functions become more and more stringent as $d$ grows. Different from these results, the approximation result in this paper is established for arbitrary $p$-times Lebesgue integrable functions. Based on this, the advantages of RBFN can be concluded as following:

(i) For certain classes of target functions, RBFN approximation can break the curse of dimensionality, i.e., it yields an approximation rate at least $N^{-1/2}$.
(ii) For non-polynomial target functions, the approximation capability of RBFN is at least not worse than that of algebraical polynomial.

### 3.2 Almost optimal approximation rate of RBFNs

In this part, we study the approximation rate of RBFN manifolds. At first, we need to characterize the space of functions we wish to approximate. Let $\mathbf{k} = (k_1, k_2, \ldots, k_d)$, $k_i \in \mathbb{N}$, and define the derivative

$$D^{\mathbf{k}} f(x) := \frac{\partial^{|\mathbf{k}|} f}{\partial^{k_1} x_{(1)} \cdots \partial^{k_d} x_{(d)}},$$

where $|\mathbf{k}| := k_1 + \cdots + k_d$. The classical Sobolev class is then defined for any $r \in \mathbb{N}$ by

$$W_p^r := W_p^r\big(\mathbb{B}^d\big) := \Big\{ f : \max_{0 \le |\mathbf{k}| \le r} \big\| D^{\mathbf{k}} f \big\|_p < \infty, r \in \mathbb{N} \Big\}.$$

Based on Theorem 2, the convergence rate of approximation by multivariate algebraical polynomials (DeVore and Lorentz 1993) can easily provide an upper bound of RBFN approximation. We state this as the following Corollary 2.

**Corollary 2** *Let $0 < p \le \infty$, $N$ and $n$ be any natural numbers such that $N \ge (2d+5) \times n\binom{n-1+d}{d}$. If $\sigma : [0,1] \to \mathbb{R}$ is a function with $(n+1)$-times bounded derivatives, and $\sigma^{(s)}(0) \ne 0$ for $s = 0, 1, \ldots, n$, then there exists a constant $C$ depending only on $d$ and $p$ such that*

$$E\big(W_p^r, \Phi_{\sigma,N}\big)_{L^p(\mathbb{B}^d)} \le C N^{-\frac{r}{d}}. \tag{7}$$

We naturally hope to know whether the upper bound given in (7) can be improved. To clarify this, we need to study the lower bound of RBFN approximation. It was proved by Maiorov and Pinkus (1999) (see also Maiorov 2005) that there exists an analytic, strictly increasing and sigmoidal activation function $\sigma$ such that

$$C_1 N^{-\frac{r}{d-1}} \le E\big(W_p^r, \Phi_{\sigma,N}\big)_{C(\mathbb{B}^d)} \le C_2 N^{-\frac{r}{d-1}}, \tag{8}$$

where $C_1$ and $C_2$ are constants depending only on $p$ and $d$. For the thin-plate spline type activation functions, Maiorov (2005) proved that if $p = q = \infty$ or $p = 2$, $1 \le q \le 2$, the upper and lower bounds of approximation by RBFNs are asymptotical identical as $N^{-\frac{r}{d}}$, i.e., there exist constants $C_1$ and $C_2$ depending only on $p$, $q$ and $d$ such that

$$C_1 N^{-\frac{r}{d}} \le E\big(W_p^r, \Phi_{\sigma,N}\big)_{L^q(\mathbb{B}^d)} \le C_2 N^{-\frac{r}{d}}. \tag{9}$$

These assertions show that different activation functions may conduct different approximation rates. So, we turn to ascertain below which activation functions can imply the optimality of the convergence rate in (7). In the following, we focus on two sets of functions:

(i) The class $\Psi_u = \{\psi\}$ which consists of exponential functions of the form $\psi(t) = e^{p(t)}$, where $p(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ is a univariate algebraic polynomial with degree not greater than $u$.

(ii) The class $\Phi_u = \{\phi\}$, which consists of all rational functions of degree at most $u$, that is, the functions of the form $\phi(t) = \frac{\alpha(t)}{\beta(t)}$, where $\alpha(\cdot), \beta(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ are univariate algebraic polynomials with degrees not larger than $u$, and $\beta(t) \ne 0$ for all $t \in \mathbb{R}_+$.

Theorem 3 below shows that, when restricted to the above classes of activation functions, the upper bound (7) is almost optimal for RBFNs.

**Theorem 3** *Let $u, N, r \in \mathbb{N}$, $1 \le p, q \le \infty$ and $r/d \ge (\frac{1}{p} - \frac{1}{q})_+$. If $\sigma \in \Phi_u \cup \Psi_u$ satisfying $\sigma^{(s)}(0) \ne 0$ for $s = 0, 1, \ldots, N$, then there exist constants $C_1$ and $C_2$ depending only on $d$, $q$, $u$ and $p$ such that*

$$C_1 (N \log N)^{-\frac{r}{d}} \le E\big(W_p^r, \Phi_{\sigma,N}\big)_{L^q(\mathbb{B}^d)} \le C_2 N^{-\frac{r}{d}}, \tag{10}$$

*where $(t)_+ := \max\{t, 0\}$.*

A series of important estimates for approximating functions in $W_p^r$ by RBFNs were deduced in Bejancu (1997, 2000), Johnson (1998), Lin et al. (2011a, 2011b), Maiorov (2003, 2005), Schaback (1995, 1996) and Xie and Cao (2013). For more details, Shcakback (1995, 1996) gave the upper bound error for approximation by RBFNs with the well-known plate spline activation function when the target function belongs to $W_2^r$. Maiorov (2005) deduced the lower bound of RBFN approximation and also proved that the upper and lower bounds were asymptotically identical. Xie and Cao (2013) deduced an upper bound estimate for the Gaussian RBFN with fixed width. Compared to these work, the novelty of our results stated in Corollary 2 and Theorem 3 is that we focus on a class of activation functions rather than a specific one. It is easy to check that the well known inverse multiquadrics and Wendland functions (Wendland 2005) fulfill the assumptions in Corollary 2, as is the Gaussian function after adding a threshold to the RBFN (see Corollary 1). It can also be found that the established approximation error in (10) depends on the dimension $d$, which differs from the approximation results in Burger and Neubauer (2001), Mhaskar (2004) and Kainen et al. (2012). However, this is not a negative result since the target function is independent of the activation function. The lower bound in (10) also shows that the established approximation rate can not be essentially improved.

## 4 Learning by RBFNs

If we have a particular approximant $f_{\mathbf{z}}$ to $f_\rho$ in hand, the quality of its performance is measured by

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) = \|f_{\mathbf{z}} - f_\rho\|_\rho^2. \tag{11}$$

The error (11) clearly depends on $\mathbf{z}$ and therefore has a stochastic nature. As a result, it is impossible to say something about (11) in general for a fixed $\mathbf{z}$. Instead, we can look at its behavior in statistics as measured by the expected error

$$E_{\rho^m}\big(\|f_{\mathbf{z}} - f_\rho\|_\rho\big) := \int_{Z^m} \|f_{\mathbf{z}} - f_\rho\| d\rho^m,$$

where the expectation is taken over all realizations $\mathbf{z}$ obtained for a fixed $m$, and $\rho^m$ is the $m$ fold tensor product of $\rho$.

It follows from the law of large numbers that by choosing suitable $f_{\mathbf{z}}$, $E_{\rho^m}(\|f_{\mathbf{z}} - f_\rho\|_\rho) \to 0$ as $m \to \infty$. How fast it tends to zero depends at least on three things: (i) the nature of $f_\rho$; (ii) the approximation properties of the hypothesis space $\mathcal{H}$: (iii) how well we do in constructing the estimators $f_{\mathbf{z}}$. Let $\mathcal{M}(\Theta)$ be the class of all Borel measures $\rho$ on $Z$ such that $f_\rho \in \Theta$. Recall that we do not know $\rho$ so that the best we can say about it is that it lies in $\mathcal{M}(\Theta)$. We enter into a competition over all estimators $\mathbb{E}_m : \mathbf{z} \to f_{\mathbf{z}}$ and define

$$e_m(\Theta) := \inf_{\mathbb{E}_m} \sup_{\rho \in \mathcal{M}(\Theta)} E_{\rho^m}\big(\|f_\rho - f_{\mathbf{z}}\|_\rho^2\big).$$

It is easy to see that $e_m(\Theta)$ quantitively measures the quality of $f_{\mathbf{z}}$. If $\Psi = \{f \in W_2^r : \|f\|_\infty \le M\}$ with $r \ge \frac{d}{2}$, then it can be found in DeVore et al. (2006, Eq. (3.26)) that

$$e_m(\Psi) \ge C m^{-\frac{2r}{2r+d}}, \quad m = 1, 2, \ldots, \tag{12}$$

where $C$ is a constant depending only on $M$ and $d$.

Since $y \in [-M, M]$, it is reasonable to set the hypothesis space a subset of $\Phi_{\sigma,N}$ as

$$\mathcal{H}_{\sigma,N}^M := \big\{g : g \in \Phi_{\sigma,N}, \|g\|_\infty \le 2M\big\}. \tag{13}$$

Then, we construct the estimator as

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}_{\sigma,N}^M} \mathcal{E}_{\mathbf{z}}(f). \tag{14}$$

The following Theorem 4 is the main result of this section.

**Theorem 4** *Let $u \in \mathbb{N}$, and $\sigma \in \Phi_u \cup \Psi_u$ satisfy $\sigma^{(s)}(0) \ne 0$ for $s = 0, 1, \ldots, N$. Suppose that $f_\rho \in W_2^r$ and $N = [m^{\frac{d}{d+2r}}]$. If $f_{\mathbf{z}}$ is defined in (14), then there exist constants $C_1$ and $C_2$ depending only on $d, u, r$ and $M$ such that,*

$$C_1 m^{-\frac{2r}{d+2r}} \le e_m(\Psi) \le \sup_{\rho \in \mathcal{M}(\Psi)} E_{\rho^m}\big(\|f_\rho - f_{\mathbf{z}}\|_\rho^2\big) \le C_2 m^{-\frac{2r}{d+2r}} \log^2 m. \tag{15}$$

Due to the nonlinearity, the learning strategy (14) can not be solved easily and may potentially be turned into numerical methods (Györfy et al. 2002). At this point, we do not address the numerical feasibility of our learning strategies. Our main interest is to understand what is the best performance we can expect for the regression problem using RBFN manifold as the hypothesis space. Very fortunately, using the simple empirical risk minimization process in the RBFN manifolds, we can prove that the RBFN manifold is a reasonable choice of

hypothesis space, since it provides a learning rate as (12), which is usually regarded as the baseline of learning rate analysis.

Optimal (or almost optimal) learning rates of some existing learning methods have been already studied in Caponnetto and DeVito (2007), DeVore et al. (2006), Györfy et al. (2002), Maiorov (2006b), Shivaswamy and Jebara (2007), Zhang et al. (2011) and Zhou and Jetter (2006). For example, Györfy et al. (2002, Chaps. 4–6) proved that the local averaging methods such as partition estimate, Nadaraya-Watson kernel estimate and $k$-nearest neighbor estimate can achieve the optimal learning rate for $W_2^1$. Zhou and Jetter (2006) pointed out that the polynomial estimate can also get the almost optimal learning rate for $W_2^r$.[2] Maiorov verified the almost optimality for feed-forward neural network estimate in Maiorov (2006b). Following these work, we proved, as shown in Theorem 4, that, up to a logarithmical factors, RBFN can also attain the optimal learning rate for $W_2^r$. This result shows that, as far as the theoretical optimality is concerned, the RBFN approach is also one of the most best choices to cope with regression problem. Furthermore, we find that to achieve such an optimal learning rate, only are $m^{\frac{d}{d+2r}}$ neurons sufficient, which is less than that of the kernel methods. This result underlies the successful application and potential advantage of RBFNs in machine learning problems.

## 5 Proofs of theorems

In this section, we provide the proofs of theorems stated in Sects. 3 and 4. To this end, Lemma 2 below, which can be found in Maiorov (2003, Eq. (26)) will play a key role.

**Lemma 2** *Let $L$ and $n$ be any natural numbers satisfying $L \geq (2d+5)\binom{n-1+d}{d}$. Then for any $P_n \in \mathcal{P}_n$, there exists a set of points $\{a_1, \ldots, a_L\} \subset \mathbb{B}^d$ such that*

$$P_n(x) = \sum_{k=1}^{L} \sum_{j=0}^{n} \mathcal{R}(k, j)|x - a_k|^j, \quad x \in \mathbb{B}^d, \tag{16}$$

*where $\mathcal{R}(k, j)$ are constants depending only on $k$ and $j$.*

*Proof of Theorem 1* Since $\sigma(t) \in C^n[0, 1]$ and $\sigma^{(s)}(0) \neq 0$ $(s = 0, 1, \ldots, n)$, then for every $t \in [0, 1]$, $\mu \in (0, 1)$ and $1 \leq m \leq n$, it follows from Taylor's formula (e.g. Xie and Cao 2010) that

$$\sigma(\mu t) = \sigma(0) + \frac{\sigma'(0)}{1!}\mu t + \cdots + \frac{\sigma^{(m)}(0)}{m!}(\mu t)^m + s_m(t), \tag{17}$$

where

$$s_m(t) = \frac{\mu^m}{(m-1)!} \int_0^t \big(\sigma^{(m)}(\mu u) - \sigma^{(m)}(0)\big)(t-u)^{m-1} du \tag{18}$$

and

$$\sigma^{(m)}(\mu u) = \sigma^{(m)}(v)\big|_{v=\mu u}.$$

[2] In Zhou and Jetter (2006), the learning rate was deduced for classification problem by using SVM algorithm associated with the polynomial kernel and the learning rate is indeed not optimal. But we can deduce the almost optimal learning rate for regression problem by using the same method as in Zhou and Jetter (2006).

Hence

$$t^m = \frac{m!}{\mu^m \sigma^{(m)}(0)} \sigma(\mu t) + q_{m-1}(t) + r_m(t), \tag{19}$$

where $q_{m-1}$ is a univariate algebraic polynomial of degree $m-1$ and

$$r_m(t) = -\frac{m!}{\mu^m \sigma^{(m)}(0)} s_m(t).$$

A direct computation shows that

$$\left| r_m(t) \right| \leq M_m \mu, \quad -1 \leq t \leq 1, \tag{20}$$

where $M_m := \max_{-1\leq t\leq 1} \frac{|\sigma^{(m+1)}(t)|}{|\sigma^m(0)|}$.

From (16), for arbitrary $P_n(x) \in \mathcal{P}_n$, there exists a set of points $\{a_1, \ldots, a_L\} \subset \mathbb{B}^d$ such that

$$P_n(x) = \sum_{j=0}^{n} \sum_{i=1}^{L} \mathcal{R}(i, j) |x - a_i|^j. \tag{21}$$

Therefore,

$$\begin{aligned} P_n(x) = &\sum_{i=1}^{L} \mathcal{R}(i, n) |x - a_i|^n + \sum_{i=1}^{L} \mathcal{R}(i, n-1) |x - a_i|^{n-1} \\ &+ \cdots + \sum_{i=1}^{L} \mathcal{R}(i, 1) |x - a_i| + \sum_{i=1}^{L} \mathcal{R}(i, 0). \end{aligned}$$

Since $x, a_i \in \mathbb{B}^d$, we have $|x - a_i| \leq [0, 2]$. For arbitrary $\varepsilon > 0$, if we take $\delta_n \in (0, 1/2]$ such that for all $j = 1, 2, \ldots, L$

$$\sum_{i=1}^{L} \left| \mathcal{R}(i, n) \right| M_n \delta_n \leq \frac{\varepsilon}{n+1}. \tag{22}$$

Then, by (19), there holds $\delta_n |x - a_i| \in [0, 1]$ and

$$|x - a_i|^n = \frac{n!}{\delta_n^n \sigma^{(n)}(0)} \sigma\left(\delta_n |x - a_i|\right) + q_{n-1}\left(|x - a_i|\right) + r_n\left(|x - a_i|\right).$$

Thus,

$$\begin{aligned} P_n(x) = &\sum_{i=1}^{L} \mathcal{R}(i, n) \frac{n!}{\delta_n^n \sigma^{(n)}(0)} \sigma\left(\delta_n |x - a_i|\right) + \sum_{i=1}^{L} \mathcal{R}(i, n) q_{n-1}\left(|x - a_i|\right) \\ &+ \sum_{i=1}^{L} \mathcal{R}(i, n) r_n\left(|x - a_i|\right) + \sum_{i=1}^{L} \mathcal{R}(i, n-1) |x - a_i|^{n-1} \\ &+ \cdots + \sum_{i=1}^{L} \mathcal{R}(i, 1) |x - a_i| + \sum_{i=1}^{L} \mathcal{R}(i, 0). \end{aligned}$$

In other words, for any given $\varepsilon > 0$, there exists $\delta_n \in (0, 1/2]$ such that

$$P_n(x) = \sum_{i=1}^{L} \mathcal{R}(i, n) \frac{n!}{\delta_n^n \sigma^{(n)}(0)} \sigma\left(\delta_n |x - a_i|\right) + P_{n-1}(x) + R_n(x), \tag{23}$$

where

$$R_n(x) := \sum_{i=1}^{L} \mathcal{R}(i,n) r_n\big(|x-a_i|\big),$$

and

$$\begin{aligned} P_{n-1}(x) &:= \sum_{i=1}^{L} \mathcal{R}(i,n) q_{n-1}\big(|x-a_i|\big) + \sum_{i=1}^{L} \mathcal{R}(i,n-1)|x-a_i|^{n-1} \\ &\quad + \cdots + \sum_{i=1}^{L} \mathcal{R}(i,1)|x-a_i| + \sum_{i=1}^{L} \mathcal{R}(i,0) \\ &= \sum_{i=1}^{L} D(i,n-1)|x-a_i|^{n-1} + \sum_{i=1}^{L} D(i,n-2)|x-a_i|^{n-2} \\ &\quad + \cdots + \sum_{i=1}^{L} D(i,1)|x-a_i| + \sum_{i=1}^{L} D(i,0). \end{aligned}$$

Here $D(i,j)$ are constants depending only on $i$ and $j$. It follows from (22) that

$$\big|R_n(x)\big| < \frac{\varepsilon}{n+1}. \tag{24}$$

Furthermore, choose $\delta_{n-1} \in (0, 1/2]$ such that

$$\sum_{i=1}^{L} \big|D(i,n-1)\big| M_{n-1}\delta_{n-1} \le \frac{\varepsilon}{n+1}.$$

A similar process as (21)–(24) then yields

$$P_{n-1}(x) = \sum_{i=1}^{L} D(i,n-1) \frac{(n-1)!}{\delta_{n-1}^{n-1}\sigma^{n-1}(0)} \sigma\big(\delta_{n-1}|x-a_i|\big) + P_{n-2}(x) + R_{n-1}(x),$$

where

$$\begin{aligned} P_{n-2}(x) &:= \sum_{i=1}^{L} D(i,n-1) q_{n-2}\big(|x-a_i|\big) + \sum_{i=1}^{L} D(i,n-2)|x-a_i|^{n-2} \\ &\quad + \cdots + \sum_{i=1}^{L} D(i,1)|x-a_i| + \sum_{i=1}^{L} D(i,0) \\ &= \sum_{i=1}^{L} E(i,n-2)|x-a_i|^{n-2} + \sum_{i=1}^{L} E(i,n-3)|x-a_i|^{n-3} \\ &\quad + \cdots + \sum_{i=1}^{L} E(i,1)|x-a_i| + \sum_{i=1}^{L} E(i,0), \end{aligned}$$

$E(i,j)$ are constants depending only on $i$ and $j$, and

$$R_{n-1}(x) := \sum_{i=1}^{L} D(i,n-1) r_{n-1}\big(|x-a_i|\big).$$

Similarly, we obtain

$$\left|R_{n-1}(x)\right| < \frac{\varepsilon}{n+1}. \tag{25}$$

After repeating the above method $n+1$ times, we then finally obtain

$$\begin{aligned} P_n(x) &= \sum_{i=1}^{L} A_{i,n}\sigma\left(\delta_n|x-a_i|\right) + \sum_{i=1}^{L} A_{i,n-1}\sigma\left(\delta_{n-1}|x-a_i|\right) \\ &\quad + \cdots + \sum_{i=1}^{L} A_{i,0}\sigma\left(\delta_0|x-a_i|\right) + R(x) \\ &= \sum_{j=0}^{n}\sum_{i=1}^{L} A_{i,j}\sigma\left(\delta_j|x-a_i|\right) + R(x), \end{aligned}$$

where

$$R(x) := \sum_{j=0}^{n} R_j(x)$$

and $A_{i,j}$ are constants depending on $i$, $j$, $\varepsilon$, and $\sigma$. From (24) and (25) it is easy to deduce that

$$\left|R(x)\right| < (n+1)\frac{\varepsilon}{n+1} = \varepsilon.$$

Thus, with the RBFN $L_N^\sigma$

$$L_N^\sigma = \sum_{k=1}^{Ln} c_k\sigma\left(w_k|x-\theta_k|\right)$$

there hold the following

$$\left|P_n(x) - L_N^\sigma(x)\right| < \varepsilon, \quad x \in \mathbb{B}^d.$$

This implies Theorem 1. □

*Proof of Theorem 2* It is obvious that there exists a $P_n \in \mathcal{P}_n$ such that

$$\|f - P_n\|_p < E(f, P_n)_{L^p(\mathbb{B}^d)} + \frac{\varepsilon}{2}.$$

On the other hand, Theorem 1 shows that for $P_n \in \mathcal{P}_n$ there exists an $L_N^\sigma$ formed as (2) such that

$$\left\|P_n - L_N^\sigma\right\|_p < \frac{\varepsilon}{2}.$$

Combining these two inequalities, the estimation (5) then directly follows. This finishes the proof of Theorem 2. □

To prove Theorem 3, we need the following Lemma 3 which can be found in Maiorov and Meir (2001).

**Lemma 3** *Let* $\frac{r}{d} > (\frac{1}{p} - \frac{1}{q})_+$. *Assume* $\sigma$ *is one of the following types*: (i) *A piecewise polynomial function*; (ii) *A rational function*; (iii) *A Gaussian*. *Then for any* $1 \le p, q \le \infty$, *there exists an absolute constant* $C$ *such that*

$$E\left(W_p^r, \Phi_{\sigma,N}\right)_q \ge C(N\log N)^{-\frac{r}{d}}. \tag{26}$$

*Proof of Theorem 3* It follows from Corollary 2 that the second inequality of (10) holds. On the other hand, since $\sigma \in \Phi_u$, we obtain from Lemma 3 that the first inequality of (10) also holds. This arrives to Theorem 3. □

In order to prove the upper bound of Theorem 4, we need the following two lemmas, which can be found in Maiorov (2006b) and Zhou and Jetter (2006), respectively.

**Lemma 4** *For any natural $N$ and any positive number $\varepsilon$, the following inequality holds*

(i) $\mathcal{H}_\varepsilon(\mathcal{H}^M_{\sigma,N}, L^2(\mathbb{B}^d)) \le cd^u N \log^2 \frac{2M}{\varepsilon}$, *if* $\sigma \in \Psi_u$;
(ii) $\mathcal{H}_\varepsilon(\mathcal{H}^M_{\sigma,N}, L^2(\mathbb{B}^d)) \le cd^u N \log N \log \frac{2M}{\varepsilon}$, *if* $\sigma \in \Phi_u$.

**Lemma 5** *Let $\mathcal{G}$ be a set of functions on $Z$ such that, for some $c \ge 0$, $|g - E(g)| \le B$ almost everywhere and $E(g^2) \le cE(g)$ for each $g \in \mathcal{G}$. Then, for every $\varepsilon > 0$,*

$$\mathrm{Prob}_{z\in Z^m}\left\{\sup_{f\in\mathcal{G}} \frac{E(g) - \frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{E(g)+\varepsilon}} \le \sqrt{\varepsilon}\right\} \le \mathcal{N}_\varepsilon\big(\mathcal{G}, C\big(\mathbb{B}^d\big)\big) \exp\left\{-\frac{m\varepsilon}{2c + \frac{2B}{3}}\right\}.$$

*Proof of Theorem 4* For simplicity, we only prove (15) for $\sigma \in \Psi_u$. The case $\sigma \in \Phi_u$ can be similarly justified. From (1), it follows that

$$\|f_{\mathbf{z}} - f_\rho\|_\rho^2 = \big\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) - \big(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\big)\big\} + \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho) := \mathcal{S}_1 + \mathcal{S}_2.$$

Therefore,

$$E_{\rho^m}\big(\|f_{\mathbf{z}} - f_\rho\|_\rho^2\big) \le E_{\rho^m}\big(\big\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) - \big(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\big)\big\}\big) + E_{\rho^m}\big(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\big).$$

Now we use Lemmas 4 and 5 to estimate $\mathcal{S}_1$. Set

$$\mathcal{F}_{2M} := \big\{\big(f(x) - y\big)^2 - \big(f_\rho(x) - y\big)^2 : f \in \mathcal{H}^M_{\sigma,N}\big\}.$$

Then for any fixed $g \in \mathcal{F}_{2M}$, there exists $f \in \mathcal{H}^M_{\sigma,N}$ such that $g(z) = (f(x) - y)^2 - (f_\rho(x) - y)^2$. Therefore,

$$E_{\rho^m}(g) = \mathcal{E}(f) - \mathcal{E}(f_\rho) \ge 0, \qquad \frac{1}{m}\sum_{i=1}^m g(z_i) = \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_\rho).$$

Since $|f(x)| \le 2M$ and $|f_\rho(x)| \le M$ almost everywhere, we deduce that

$$\big|g(\mathbf{z})\big| = \big|\big(f(x) - f_\rho(x)\big)\big(\big(f(x) - y\big) + \big(f_\rho(x) - y\big)\big)\big| \le 15M^2.$$

It then follows that $|g(\mathbf{z}) - E(g)| \le 30M^2$ almost every where and

$$E_{\rho^m}\big(g^2\big) \le 30M^2\|f - f_\rho\|_\rho^2 = 30M^2 E_{\rho^m}(g).$$

Now we apply Lemma 5 with $B = c = 30M^2$ to the set of functions $\mathcal{F}_{2M}$, which yields

$$\sup_{f\in B_{2M}} \frac{\{\mathcal{E}(f) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_\rho)\}}{\sqrt{\{\mathcal{E}(f) - \mathcal{E}(f_\rho)\} + \varepsilon}} = \sup_{g\in\mathcal{F}_{2M}} \frac{E_{\rho^m}(g) - \frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{E_{\rho^m}(g) + \varepsilon}} \le \sqrt{\varepsilon}$$

with confidence at least

$$1 - \mathcal{N}_\varepsilon\big(\mathcal{F}_{2M}, C\big(\mathbb{B}^d\big)\big) \exp\left\{-\frac{m\varepsilon}{80M^2}\right\}.$$

Observe that for any $g_1, g_2 \in \mathcal{F}_{2M}$ there exist $f_1, f_2 \in \mathcal{H}_{\sigma,N}^M$ such that

$$g_j(z) = \big(f_j(x) - y\big)^2 - \big(f_\rho(x) - y\big)^2, \quad j = 1, 2.$$

It is obvious that

$$\big|g_1(\mathbf{z}) - g_2(\mathbf{z})\big| = \big|\big(f_1(x) - y\big)^2 - \big(f_2(x) - y\big)^2\big| \le 6M\|f_1 - f_2\|_\infty.$$

We see that for any $\varepsilon > 0$, an $(\frac{\varepsilon}{6M})$-covering of $\mathcal{H}_{\sigma,N}^M$ provides an $\varepsilon$-covering of $\mathcal{F}_{2M}$. Therefore,

$$\mathcal{N}_\varepsilon\big(\mathcal{F}_{2M}, C\big(\mathbb{B}^d\big)\big) \le \mathcal{N}_{\frac{\varepsilon}{6M}}\big(\mathcal{H}_{\sigma,N}^M, C\big(\mathbb{B}^d\big)\big).$$

Thus, the confidence is

$$1 - \mathcal{N}_\varepsilon\big(\mathcal{F}_{2M}, C\big(\mathbb{B}^d\big)\big)\exp\left\{-\frac{m\varepsilon}{80M^2}\right\} \ge 1 - \mathcal{N}_{\frac{\varepsilon}{6M}}\big(\mathcal{H}_{\sigma,N}^M, C\big(\mathbb{B}^d\big)\big)\exp\left\{-\frac{m\varepsilon}{80M^2}\right\}.$$

Since

$$\mathcal{N}_{\frac{\varepsilon}{6M}}\big(\mathcal{H}_{\sigma,N}^M, C\big(\mathbb{B}^d\big)\big) \le \mathcal{N}_{\frac{\varepsilon}{6M}}\big(\mathcal{H}_{\sigma,N}^M, L^2\big(\mathbb{B}^d\big)\big),$$

it follows from Lemma 4 that

$$\mathcal{N}_{\frac{\varepsilon}{6M}}\big(\mathcal{H}_{\sigma,N}^M, C\big(\mathbb{B}^d\big)\big) \le \exp\left\{cd^u N \log^2 \frac{12M^2}{\varepsilon}\right\}.$$

Thus, we have

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\{\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_\mathbf{z}(f_\mathbf{z}) - \mathcal{E}_\mathbf{z}(f_\rho)\}}{\sqrt{\{\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho)\} + \varepsilon}} \le \sqrt{\varepsilon}\right\}\\ &\ge 1 - \exp\left\{cd^u N \log^2 \frac{12M^2}{\varepsilon} - \frac{m\varepsilon}{80M^2}\right\}.\end{aligned}$$

Since

$$\sqrt{\varepsilon}\sqrt{\big\{\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho)\big\} + \varepsilon} \le \frac{1}{2}\big\{\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho)\big\} + \varepsilon,$$

we conclude that with confidence at least

$$1 - \exp\left\{cd^u N \log^2 \frac{12M^2}{\varepsilon} - \frac{m\varepsilon}{80M^2}\right\}$$

there holds

$$\big(\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho)\big) - \big(\mathcal{E}_\mathbf{z}(f_\mathbf{z}) - \mathcal{E}_\mathbf{z}(f_\rho)\big) \le \frac{1}{2}\big(\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho)\big) + \varepsilon.$$

Hence,

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in Z^m}\big\{\big\{\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho)\big\} - 2\big\{\mathcal{E}_\mathbf{z}(f_\mathbf{z}) - \mathcal{E}_\mathbf{z}(f_\rho)\big\} \le \varepsilon\big\}\\ &\ge 1 - \exp\left\{cd^u N \log^2 \frac{24M^2}{\varepsilon} - \frac{m\varepsilon}{160M^2}\right\}.\end{aligned}$$

Set

$$\mathcal{T} := \big\{\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho)\big\} - 2\big\{\mathcal{E}_\mathbf{z}(f_\mathbf{z}) - \mathcal{E}_\mathbf{z}(f_\rho)\big\}.$$

Then

$$\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho) = \mathcal{T} + 2\mathcal{S}_2. \tag{27}$$

For arbitrary $\mu \geq \frac{24M^2}{m}$, there holds

$$\begin{aligned}
E_{\rho^m}(\mathcal{T}) &= \int_0^\infty \mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\left\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)\right\} - 2\left\{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\right\} > \varepsilon\right\} d\varepsilon \\
&\leq \mu + \int_\mu^\infty \exp\left\{cd^u N \log^2 \frac{24M^2}{\varepsilon} - \frac{m\varepsilon}{160M^2}\right\} d\varepsilon \\
&\leq \mu + \int_\mu^\infty \exp\left\{cd^u N \log m \log \frac{24M^2}{\varepsilon} - \frac{m\varepsilon}{160M^2}\right\} d\varepsilon \\
&\leq \mu + \exp\left\{-\frac{m\mu}{160M^2}\right\} \int_\mu^\infty \left(\frac{24M^2}{\varepsilon}\right)^{cd^u N \log m} d\varepsilon \\
&\leq \mu + \exp\left\{-\frac{m\mu}{160M^2}\right\} \left(\frac{24M^2}{\mu}\right)^{cd^u N \log m} \mu \\
&\leq \mu + \exp\left\{-\frac{m\mu}{160M^2}\right\} m^{cd^u N \log m} \mu.
\end{aligned}$$

By setting $\mu = 160cd^u M^2 \frac{N \log^2 m}{m}$, we obtain

$$E_{\rho^m}(\mathcal{T}) \leq \frac{320cd^u M^2 N \log^2 m}{m}. \tag{28}$$

Now, we turn to estimate $E_{\rho^m}(\mathcal{S}_2)$. Note first that

$$\begin{aligned}
E_{\rho^m}(\mathcal{S}_2) &= E_{\rho^m}\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\right) = E_{\rho^m}\left(\frac{1}{m}\sum_{i=1}^m \left(f_{\mathbf{z}}(x_i) - y_i\right)^2 - \frac{1}{m}\sum_{i=1}^m \left(f_\rho(x_i) - y_i\right)^2\right) \\
&= E_{\rho^m}\left(\inf_{f\in\mathcal{H}^M_{\sigma,N}}\left(\frac{1}{m}\sum_{i=1}^m \left(f(x_i) - y_i\right)^2 - \frac{1}{m}\sum_{i=1}^m \left(f_\rho(x_i) - y_i\right)^2\right)\right) \\
&\leq \inf_{f\in\mathcal{H}^M_{\sigma,N}} \left(E_{\rho^m}\left(\left(f(x) - y\right)^2\right) - E_{\rho^m}\left(\left(f_\rho(x) - y\right)^2\right)\right) \\
&= \inf_{f\in\mathcal{H}^M_{\sigma,N}} \int_{\mathbb{B}^d} \left(f(x) - f_\rho(x)\right)^2 d\rho.
\end{aligned}$$

From Theorem 3, the definition of $\mathcal{H}^M_{\sigma,N}$, and the well known Sobolev embedding theorem, it follows that when $r \geq \frac{d}{2}$, there holds

$$\inf_{f\in\mathcal{H}^M_{\sigma,N}} \left\{\int_{\mathbb{B}^d} \left(f(x) - f_\rho(x)\right)^2 d\rho\right\}^{\frac{1}{2}} \leq E_{\rho^m}\left(W_2^r, \mathcal{H}^M_{\sigma,N}\right)_{C(\mathbb{B}^d)} \leq CN^{-\frac{r}{d}}.$$

Therefore, we have

$$E_{\rho^m}(\mathcal{S}_2) \leq CN^{-\frac{2r}{d}}. \tag{29}$$

If we setting $N = m^{\frac{d}{2r+d}}$, then (27), (28) and (29) imply that there exists a constant $C$ depending only on $M$, $d$ and $u$ such that

$$E_{\rho^m}\left(\|f_{\mathbf{z}} - f_\rho\|_\rho^2\right) \leq Cm^{-\frac{2r}{2r+d}} \log^2 m,$$

which arrives to the upper bound of (15). However it follows from (12) that there exists a constant $C$ depending only on $M$ and $d$ that satisfies

$$E_{\rho^m}\left(\|f_{\mathbf{z}} - f_\rho\|_\rho^2\right) \geq Cm^{-\frac{2r}{2r+d}},$$

which gives a lower bound estimation of learning. With this, the proof of Theorem 4 is completed. □

**Acknowledgements** Three anonymous referees, to which we feel much indebted and are grateful, have carefully read the paper and have provided us with numerous constructive suggestions. As a result, the overall quality of the paper has been noticeably enhanced.

The research was supported by the National 973 Programming (2013CB329404), the Key Program of National Natural Science Foundation of China (Grant No. 11131006), and the National Natural Science Foundations of China (Grants No. 61075054).

## References

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

Barron, A., Cohen, A., Dahmen, W., & Devore, R. (2008). Approximation and learning by greedy algorithms. *The Annals of Statistics*, *36*, 64–94.

Bejancu, A. (1997). *The uniform convergence of multivariate natural splines* (DAMTP Technical Report). University of Cambridge.

Bejancu, A. (2000). *On the accuracy of surface spline approximation and interpolation to bump functions* (DAMTP Technical Report). University of Cambridge.

Buhman, M. (2000). Radial basis functions. *Acta Numerica*, *9*, 1–38.

Bumann, M., Dyn, N., & Levin, D. (1995). On quasi-interpolation by radial basis functions with scattered centres. *Constructive Approximation*, *11*, 239–254.

Burger, M., & Neubauer, A. (2001). Error bounds for approximation with neural networks. *Journal of Approximation Theory*, *112*, 235–250.

Caponnetto, A., & DeVito, E. (2007). Optimal rates for the regularized least squares algorithm. *Foundations of Computational Mathematics*, *7*, 331–368.

Chen, T., & Chen, H. (1995). Approximation capability to functions of several variables, nonlinear functionals and operators by radial basis function neural networks. *IEEE Transactions on Neural Networks*, *6*, 904–910.

Cucker, F., & Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, *39*, 1–49.

Cucker, F., & Smale, S. (2002). Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, *2*, 413–428.

DeVore, R., & Lorentz, G. (1993). *Constructive approximation*. Berlin: Springer.

DeVore, R., Kerkyacharian, G., Picard, D., & Temlyakov, V. (2006). Approximation methods for supervised learning. *Foundations of Computational Mathematics*, *6*, 3–58.

Fedoseyev, A., Friedman, M., & Kansa, E. (2002). Improved multiquadric method for elliptic partial differential equations via PDE collocation on the boundary. *Computers and Mathematics*, *43*, 439–455.

Flyer, N., & Wright, G. (2009). A radial basis function method for the shallow water equations on a sphere. *Proceedings of the Royal Society A*, *465*, 1949–1976.

Györfy, L., Kohler, M., Krzyzak, A., & Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Berlin: Springer.

Johnson, M. (1998). A bound on the approximation order of surface splines. *Constructive Approximation*, *14*, 429–438.

Kainen, P., Kurková, V., & Sanguineti, M. (2012). Dependence of computational models on input dimension: tractability of approximation and optimization tasks. *IEEE Transactions on Information Theory*, *58*, 1203–1214.

Lin, S., Cao, F., & Xu, Z. (2011a). The essential rate of approximation for radial function manifold. *Science China Mathematics*, *54*, 1985–1994.

Lin, S., Cao, F., & Xu, Z. (2011b). Essential rate for approximation by spherical neural networks. *Neural Networks*, *24*, 752–758.

Maiorov, V. (2003). On best approximation of classes by radial functions. *Journal of Approximation Theory*, *120*, 36–70.

Maiorov, V. (2005). On lower bounds in radial basis approximation. *Advances in Computational Mathematics*, *22*, 103–113.

Maiorov, V. (2006a). Pseudo-dimension and entropy of manifolds formed by affine invariant dictionary. *Advances in Computational Mathematics*, *25*, 435–450.

Maiorov, V. (2006b). Approximation by neural networks and learning theory. *Journal of Complexity*, *22*, 102–117.
Maiorov, V., & Meir, R. (2001). Lower bounds for multivariate approximation by affine-invariant dictionaries. *IEEE Transactions on Information Theory*, *47*, 1569–1575.
Maiorov, V., & Pinkus, A. (1999). Lower bounds for approximation by MLP neural networks. *Neurocomputing*, *25*, 81–91.
Mendelson, S., & Vershinin, R. (2003). Entropy and the combinatorial dimension. *Inventiones Mathematicae*, *125*, 37–55.
Mhaskar, H. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation*, *8*, 164–177.
Mhaskar, H. (2004). On the tractability of multivariate integration and approximation by neural networks. *Journal of Complexity*, *20*, 561–590.
Park, J., & Sandberg, I. (1991). Universal approximation using radial-basis-function networks. *Neural Computation*, *3*, 246–257.
Park, J., & Sandberg, I. (1993). Approximation and radial-basis-function networks. *Neural Computation*, *5*, 305–316.
Powell, M. (1990). The theory of radial basis approximation. In W. A. Light (Ed.), *Advances in numerical analysis* (Vol. 2, pp. 105–210). Oxford: Oxford University Press.
Schaback, R. (1995). Error estimates and conditions numbers for radial basis functions interpolations. *Advances in Computational Mathematics*, *3*, 251–264.
Schaback, R. (1996). Approximation by radial basis functions with finitely many centers. *Constructive Approximation*, *12*, 331–340.
Shivaswamy, P., & Jebara, T. (2007). Ellipsoidal machines. *Journal of Machine Learning Research—Proceedings Track*, *2*, 484–491.
Temlyakov, V. (2008). Approximation in learning theory. *Constructive Approximation*, *27*, 33–74.
Wendland, H. (2000). Optimal approximation orders on $L_p$ for radial basis functions. *East Journal on Approximations*, *6*, 87–102.
Wendland, H. (2005). *Scattered data approximation*. New York: Cambridge University Press.
Xie, T., & Cao, F. (2010). The errors in simultaneous approximation by feed-forward neural networks. *Neurocomputing*, *73*, 903–907.
Xie, T., & Cao, F. (2013). The rate of approximation of Gaussian radial basis neural networks in continuous function space. *Acta Mathematica Sinica. English Series*, *29*, 295–302.
Zhang, Y., Cao, F., & Xu, Z. (2011). Optimal rate of the regularized regression learning algorithm. *International Journal of Computer Mathematics*, *88*, 1471–1483.
Zhou, D. X., & Jetter, K. (2006). Approximation with polynomial kernels and SVM classifiers. *Advances in Computational Mathematics*, *25*, 323–344.