Nonparametric Regression Estimation by Normalized Radial Basis Function Networks

Adam Krzyżak, *Senior Member, IEEE,* and Dominik Schäfer

***Abstract*—This paper establishes weak and strong universal consistency of regression estimates based on normalized radial basis function networks when the network parameters are chosen by empirical risk minimization.**

***Index Terms*—Covering numbers, empirical risk minimization, nonparametric regression estimation, normalized radial basis function networks.**

I. Introduction

THE three mainstream methods of modern neural network modeling are multilayer perceptrons (MLP), radial basis function (RBF) networks and normalized radial basis function (NRBF) networks. These have been proven to be flexible models, applicable for various tasks, such as interpolation, classification, data smoothing, and regression, just to name a few. Accounts of these methods can be found in, e.g., Anthony and Bartlett [1], Barron [2], Cybenko [3], Devroye *et al.* [5], Györfi *et al.* [11] and Hornik *et al.* [13] for multilayer perceptrons; for RBF networks see, e.g., Girosi and Anzellotti [8], Györfi *et al.* [11], Krzyżak *et al.* [15], Krzyżak and Linder [16], Moody and Darken [20], and Park and Sandberg [21], [22].

The present paper is devoted to the study of NRBFs used for regression analysis. To be more precise, we consider **normalized radial basis function (NRBF) networks** $\mathcal{F}(k,\ell,L,R,B)$ with one hidden layer of k nodes ($k \in \mathbb{N}, L \geq \ell \geq 0, R, B > 0$), i.e., the class of functions of the form

$$f(x) = \frac{\sum_{i=1}^{k} w_i K\left((x-c_i)^T A_i (x-c_i)\right)}{\sum_{i=1}^{k} K\left((x-c_i)^T A_i (x-c_i)\right)} =: \frac{\sum_{i=1}^{k} w_i K_{c_i,A_i}(x)}{\sum_{i=1}^{k} K_{c_i,A_i}(x)} \tag{1}$$

for which the following conditions hold.

i) Kernel condition: $K : \mathbb{R}_0^+ \to \mathbb{R}^+$ is a left-continuous, decreasing function, the so-called *kernel*, where $\mathbb{R}_0^+$ are the nonnegative real numbers.

ii) Center condition: $c_1, \ldots, c_k \in \mathbb{R}^d$ are the so-called *center vectors* with $\|c_i\| \leq R$ for all $i = 1, \ldots, k$, where $\|\cdot\|$ is the Euclidean norm.

iii) Receptive field condition: $A_1, \ldots, A_k$ are symmetric, positive definite, real $d \times d$-matrices each of which satisfies the eigenvalue inequalities

$$\ell \leq \lambda_{\min}(A_i) \leq \lambda_{\max}(A_i) \leq L.$$

Here, $\lambda_{\min}(A_i)$ and $\lambda_{\max}(A_i)$ are the minimal and the maximal eigenvalues of A_i, respectively. A_i specifies the shape of the *receptive field* about the center c_i.

iv) Weight condition: $w_1, \ldots, w_k \in \mathbb{R}$ are the *weights* satisfying $|w_i| \leq B$ for all $i = 1, \ldots, k$.

In the remainder of the paper, the kernel is fixed, whereas network parameters $w_i, c_i, A_i, i = 1, \ldots, k$ are learned from the data. Note that $\mathcal{F}$ also depends on K but for the sake of simplicity we suppress it in the notation. Throughout the paper we use the convention $0/0 = 0$. Common choices for the kernel are (we implicitly assume they satisfy i)).

- **Window-type kernels**. These are kernels for which there exists some $\delta > 0$ such that $K(v) \notin (0, \delta)$ for all $v \in \mathbb{R}_0^+$. The classical naive kernel $K(v) = \mathbf{1}_{[0,1]}(v)$ is a member of this class.
- **Non-window-type kernels with bounded support**. These comprise all kernels with support of the form $[0, s]$ which are right-continuous in s. For example, for $K(v) = \max\{1 - v, 0\}$, $K(x^T x)$ is the Epanechnikov kernel.
- **Kernels with unbounded support**, i.e., $K(v) > 0$ for all $v \in \mathbb{R}_0^+$. The most famous example of this class is $K(v) = \exp(-v)$. Then $K(x^T x)$ is the classical Gaussian kernel.

NRBF networks were introduced by Moody and Darken [20] and Specht [25] as modifications of standard RBF networks defined by

$$f(x) = \sum_{i=1}^{k} w_i K((x-c_i)^T A_i (x-c_i)) + w_0. \tag{2}$$

These naturally arise in a variety of problems, such as smoothing splines (cf. Duchon [6]), interpolation using multiquadrics, shifted surface splines or thin-plate splines (see Girosi *et al.* [9]), and regression analysis (see, e.g., Györfi *et al.* [11] and Krzyżak *et al.* [15]). In [15], the authors consider RBF regression estimates and classifiers based on minimizing the empircal risk and prove universal consistency results. The rate of convergence of RBF-based regression estimates using complexity

Manuscript received September 23, 2002; revised August 11, 2004. The work of A. Krzyżak was supported by the NSERC, the FCAR, and the Alexander von Humboldt Foundation. The work of D. Schäfer was supported by the College of Graduates "Parallel and Distributed Systems," University of Stuttgart, Stuttgart, Germany.

A. Krzyżak is with the Department of Computer Science and Software Engineering, Concordia University, Montréal, QC H3G 1M8, Canada (e-mail: krzyzak@cs.concordia.ca).

D. Schäfer is with the Fachbereich Mathematik, Universität Stuttgart, D-70569 Stuttgart, Germany (e-mail: schaefdk@mathematik.uni-stuttgart.de).

Communicated by A. B. Nobel, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2004.842632

regularization is studied in Krzyżak and Linder [16]. Practical and theoretical advantages and disadvantages of the transition from RBFs to NRBFs through normalization are discussed by Shorten and Murray-Smith [24], who also give further references. In particular, radial functions of NRBFs sum to unity at every point of the input space which is highly desirable both from computational and theoretical points of view.

In the following, we study the application of NRBFs in regression analysis: Suppose that the random variables X and Y take values in $\mathbb{R}^d$ and $\mathbb{R}$, respectively. The task is to find a measurable function $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(X)$ is a good approximation of Y in the mean-squared-error sense. In particular, if $\boldsymbol{E}|Y|^2 < \infty$, we aim to find a measurable function r minimizing the L_2-risk, that is,

$$J^* = \inf_{f:\mathbb{R}^d \to \mathbb{R} \text{ measurable}} \boldsymbol{E}|f(X) - Y|^2 = \boldsymbol{E}|r(X) - Y|^2.$$

The solution of this minimization problem is given by the regression function $r(x) = \boldsymbol{E}[Y \mid X = x]$. The regression function, however, can only be computed if the distribution of (X, Y) happens to be known. Otherwise, we have to rely on estimates of r.

In order to estimate r without making any assumptions about the distribution of (X, Y), we assume that a training set

$$D_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

of independent and identically distributed (i.i.d.) copies of (X, Y) is given, where D_n is independent of (X, Y). The method of **empirical risk minimization** obtains an estimate $\hat{f}_n$ of r by selecting the parameter vector which minimizes the residual sum of squares over a suitable class $\mathcal{F}_n$ of functions. In other words, based on the training sequence, we choose an estimator $\hat{f}_n \in \mathcal{F}_n$, such that $\hat{f}_n$ minimizes the *empirical* L_2-risk (mean residual sum of squares)

$$J_n(f) = \frac{1}{n} \sum_{j=1}^{n} |f(X_j) - Y_j|^2$$

that is,

$$J_n(\hat{f}_n) \le J_n(f), \qquad \text{for all } f \in \mathcal{F}_n.$$

The performance of the regression estimator $\hat{f}_n$ is measured by

$$J(\hat{f}_n) = \boldsymbol{E}[|\hat{f}_n(X) - Y|^2 \mid D_n].$$

In this framework a regression estimator $\hat{f}_n$ is called **strongly (weakly) consistent** if it asymptotically attains the minimal L_2-risk J^* almost surely (in probability), i.e., if

$$J(\hat{f}_n) - J^* \to 0 \quad \text{converges with probability 1 (in probability) as } n \to \infty. \quad (3)$$

Observe that $J(\hat{f}_n) - J^* \to 0$ if and only if

$$\boldsymbol{E}[|\hat{f}_n(X) - Y|^2 \mid D_n] - \boldsymbol{E}|r(X) - Y|^2 = \boldsymbol{E}[|\hat{f}_n(X) - r(X)|^2 \mid D_n] \to 0$$

which is the usual notion of L_2-consistency for regression function estimates (cf. Györfi *et al.* [11]).

The idea of empirical risk minimization has been extensively used in literature. When the minimization is carried out over exceedingly rich (complex) families $\mathcal{F}_n$ of candidate functions, the resulting estimate usually overfits the data, i.e., it is not likely to perform well for new data that is independent of the training set. Different measures of complexity of $\mathcal{F}_n$ have been used for different purposes, but they are all related to the cardinality of a finite subset representing the whole family in a certain sense. Examples are metric entropy (Kolmogorov and Tihomirov [14]), Vapnik–Chervonenkis (VC)-dimension (Vapnik and Chervonenkis [26]), and random covering numbers (Pollard [23]). Based on these measures, asymptotic properties of the method of empirical risk minimization were studied among others by Vapnik [27] and Haussler [12]. The class $\mathcal{F}_n$ of candidate functions should clearly allow the statistician to find good approximations for a multitude of target functions. Therefore, one generally needs to increase the size of the candidate family as the size of the training set increases. However, a good tradeoff should also be maintained between the complexity of the candidate family and the training data size to avoid overfitting. The idea of using nested candidate classes which grow in a *controlled* manner with the size of the training data is Grenander's method of sieves [10]. This approach was successfully applied to pattern recognition by Devroye *et al.* [5], to regression estimation by Györfi *et al.* [11] and Lugosi and Zeger [19], and by Faragó and Lugosi [7] and White [28] in the neural network framework.

In this paper, we apply empirical risk minimization to obtain consistent regression estimates using NRBF networks. First, consistency results for regression estimates based on NRBF networks were obtained by Krzyżak and Niemann [17] and Xu *et al.* [29] under the special assumption that the centers are placed at the data points, the covariance matrices are chosen according to some specified rules and only the output weights are learned by minimizing the residual sum of squares. Much more flexibility is gained by learning all these parameters from the data. The present paper establishes weak and strong universal consistency of the regression estimates derived from NRBF networks when *all* network parameters (including the centers) are learned by empirical risk minimization. In doing so, we demonstrate how to apply the tools of the trade (covering numbers, VC-dimensions, and their connections with each other) to feedforward NRBF nets. In Section II, we present the main results. Auxiliary lemmas are presented in Section III, and the proofs of the main results can be found in Section IV.

II. Consistent Estimation Algorithms

We first consider the case of window-type kernels and kernels of unbounded support. Let the parameters $k_n \in \mathbb{N}, L_n, R_n$ and $B_n > 0$ tend to ∞ as $n \to \infty$. Note that the $\mathcal{F}(k_n, 0, L_n, R_n, B_n)$ are not nested as n increases. We therefore consider the nested models

$$\mathcal{F}_n := \bigcup_{k=1}^{k_n} \mathcal{F}(k, 0, L_n, R_n, B_n).$$

The condition on the minimal admissible eigenvalue is not used here.

Suppose i.i.d. observations $D_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are available. Let the estimate $\hat{f}_n \in \mathcal{F}_n$ of the regression function $r(\cdot) = \boldsymbol{E}[Y \mid X = \cdot]$ be chosen by empirical risk minimization, i.e.,

$$\hat{f}_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2. \tag{4}$$

This estimation strategy yields consistent estimators when a window type kernel or a kernel with unbounded support is used.

Theorem 1: Suppose $k_n, L_n, R_n, B_n \to \infty \quad (n \to \infty)$ in such a way that

a) for window type kernel K

$$\frac{B_n^4 k_n}{n} \log B_n^2 k_n \to 0;$$

b) for kernel K with unbounded support

$$\frac{B_n^4 k_n}{n} \log \frac{B_n^2}{K(4R_n^2 L_n)} \to 0.$$

Then the estimate $\hat{f}_n$ defined by (4) is weakly consistent for any distribution of (X, Y) with $\boldsymbol{E}Y^2 < \infty$ and $\|X\| \le Q < \infty$ almost surely. If, in addition

$$B_n^4 \le \frac{n}{(1+\beta)\log n}$$

for some $\beta > 0$ and all sufficiently large n, then $\hat{f}_n$ is strongly consistent.

In case a) one may choose $L_n = R_n = \infty$ for all $n \ge 1$, which means that learning is unrestricted in choosing the best parameters $c_i, A_i, i = 1, \ldots, k$ and, as shall be seen in the proof, the condition $\|X\| \le Q < \infty$ almost surely can be skipped.

Non-window-type kernels with bounded support $[0, s]$ require more care. In the case of window-type kernels and kernels of unbounded support, we are in a position to control the size of the denominator in (1) by assuming $K(v) \notin (0, \delta)$ and the receptive field condition iii) in Section I, respectively. For non-window-type kernels with bounded support, the denominator can attain very small values, the ratio in (1) thus becoming unstable. Nonetheless, with the following slightly modified estimation procedure we obtain weak and strong consistency.

For $\delta > 0$ and $f \in \mathcal{F}(k, \ell, L, R, B)$ let $f^{(\delta)}$ be the NRBF with the same parameters as f, except for the kernel $K(v)$, which is replaced by the window-type kernel

$$K^{(\delta)}(v) = \begin{cases} \delta, & \text{if } K(v) \in (0, \delta] \text{ or } v = s \\ K(v), & \text{otherwise.} \end{cases}$$

Now fix positive sequences $\delta_n, \ell_n \to 0 \ (n \to \infty)$ and choose the estimate $\hat{f}_n(\cdot)$ from

$$\mathcal{F}_n := \bigcup_{k=1}^{k_n} \mathcal{F}(k, \ell_n, L_n, R_n, B_n)$$

such as to satisfy

$$\hat{f}_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f^{(\delta_n)}(X_i) - Y_i|^2. \tag{5}$$

For suitable choice of the sequences δ_n and ℓ_n, the following consistency result holds.

Theorem 2: Let the kernel K be a non-window-type kernel with bounded support $[0, s]$. Suppose $k_n, L_n, R_n, B_n \to \infty$ and $\ell_n, \delta_n \to 0 \ (n \to \infty)$ in such a way that

$$\frac{B_n^4 k_n}{n} \log \frac{B_n^2 k_n}{\delta_n} \to 0 \tag{6}$$

$$\frac{B_n^4 k_n}{\ell_n^{d/2}} \left(s^{d/2} - (K^{-1}(\delta_n))^{d/2} \right) \to 0 \tag{7}$$

with $K^{-1}(\delta) := \sup\{x : K(x) \ge \delta\}, \sup \emptyset := 0$. Then the estimate $\hat{f}_n$ defined by (5) is weakly consistent for any distribution of (X, Y) where X has a bounded density with respect to (w.r.t.) the Lebesgue measure and $\boldsymbol{E}Y^2 < \infty$. If in addition

$$B_n^4 \le \frac{n}{(1+\beta)\log n} \tag{8}$$

for some $\beta > 0$ and all sufficiently large n, then $\hat{f}_n$ is strongly consistent.

Examples: For the naive kernel NRBF with $K(v) = \mathbf{1}_{[0,1]}(v)$

$$B_n^4 k_n = O(n^q), \qquad \text{for } q \in (0, 1)$$

suffices for weak consistency. For Gaussian kernel NRBFs, $K(v) = \exp(-v)$

$$B_n^4 k_n R_n^2 L_n = O(n^q), \qquad \text{for } q \in (0, 1)$$

will do. Epanechnikov kernel NRBFs, $K(v) = \max\{1 - v, 0\}$, have $s = 1$ and

$$s^{d/2} - (K^{-1}(\delta_n))^{d/2} = 1 - (1 - \delta_n)^{d/2} \le \max\{1, d/2\}\delta_n$$

and weak consistency holds if, e.g.,

$$\frac{B_n^4 k_n}{\ell_n^{d/2}} = O(n^q) \text{ and } \delta_n \sim n^{-\epsilon}, \qquad \text{for } q \in (0, 1),\ \epsilon > q.$$

It is interesting to compare these results with the results in Krzyżak *et al.* [15, Theorem 2] and the results in Györfi *et al.* [11, Theorem 17.1] for (nonnormalized) RBFs. There, an RBF with the same number k_n of nodes and the sum of the squared weights not exceeding $k_n B_n^2$ requires $k_n^5 B_n^4 \log(k_n^5 B_n^4)/n \to 0$ for weak consistency, whereas the NRBFs considered in this paper maintain consistency for larger number k_n of nodes, basically for $k_n B_n^4 \log(k_n B_n^2)/n \to 0$.

The proofs of the theorems will rely on upper bounds on the covering number of the class of NRBFs and on a denseness result given in Section III.

III. Tools for the Proofs of the Theorems

A. Covering Numbers

Let $\mathcal{F}$ be a class of real-valued functions on $\mathbb{R}^d$. Let $x_1^n = (x_1, \ldots, x_n) \in \mathbb{R}^{dn}$, $\epsilon > 0$, and let

$$\|f\|_{L(\nu_n)} = \frac{1}{n}\sum_{i=1}^{n} |f(x_i)|.$$

We say a class $\mathcal{G}$ of real-valued functions on $\mathbb{R}^d$ is an ϵ-cover of $\mathcal{F}$ with respect to $\|\cdot\|_{L(\nu_n)}$ on x_1^n, if for each $f \in \mathcal{F}$ there exists a $g \in \mathcal{G}$ such that

$$\frac{1}{n}\sum_{j=1}^{n} |f(x_j) - g(x_j)| \leq \epsilon.$$

The **covering number** $N(\epsilon, \mathcal{F}, x_1^n)$ is the smallest integer m such that an ϵ-cover $\mathcal{G}$ of $\mathcal{F}$ with respect to $\|\cdot\|_{L(\nu_n)}$ on x_1^n exists with cardinality $|\mathcal{G}| = m$.

Covering numbers are a useful tool to obtain uniform tail inequalities for the deviation of the average of i.i.d. random variables from their mean. Several useful results may be found, e.g., in Pollard [23].

For RBF and NRBF networks, the following class of functions $\mathbb{R}^d \to \mathbb{R}$ plays a crucial role:

$$\mathcal{K} := \Big\{ K((\cdot - c)^T A(\cdot - c)) : c \in \mathbb{R}^d, \\ A \in \mathbb{R}_{d\times d} \text{ symmetric, positive definite} \Big\}.$$

Concerning its covering number, Krzyżak *et al.* [15] prove the following.

Lemma 1: For any $x_1^n = (x_1, \ldots, x_n) \in \mathbb{R}^{dn}$ and for any $\sup_x K(x)/4 > \epsilon > 0$ there exists a subset $\mathcal{K}(\epsilon) \subseteq \mathcal{K}$ being an ϵ-cover of $\mathcal{K}$ with respect to $\|\cdot\|_{L(\nu_n)}$ on x_1^n with cardinality $2\,(4e/\epsilon)^{2(d^2+d+2)}$. Thus,

$$N(\epsilon, \mathcal{K}, x_1^n) \leq 2\left(\frac{4e}{\epsilon}\right)^{2(d^2+d+2)}. \tag{9}$$

Lemma 1 will enable us to obtain the following bounds on the covering numbers of the NRBF networks under consideration in this paper.

Lemma 2: Pick $\sup_x K(x)/4 > \epsilon > 0$ and $x_1^n = (x_1, \ldots, x_n) \in \mathbb{R}^{dn}$.

a) Let K be a window-type kernel that does not attain values in $(0, \delta)$. Then

$$N(\epsilon, \mathcal{F}(k, \ell, L, R, B), x_1^n) \\ \leq 2^k \cdot \left(\frac{4B}{\epsilon} + 1\right)^k \cdot \left(\frac{16eBk}{\epsilon\delta}\right)^{2k(d^2+d+2)}.$$

b) Let K be a kernel with unbounded support. If $\|x_i\| \leq Q$ $(i = 1, \ldots, n)$, we have

$$N(\epsilon, \mathcal{F}(k, \ell, L, R, B), x_1^n) \\ \leq 2^k \cdot \left(\frac{4B}{\epsilon} + 1\right)^k \cdot \left(\frac{16eB}{\epsilon \cdot K((R+Q)^2 L)}\right)^{2k(d^2+d+2)}.$$

Proof: First, consider *case a)*. Let

$$f(x) = \frac{\sum_{i=1}^{k} w_i K_i(x)}{\sum_{i=1}^{k} K_i(x)} \in \mathcal{F}(k, \ell, L, R, B)$$

where we use the abbreviation $K_i(x) = K_{c_i, A_i}(x)$. The ϵ-cover of $\mathcal{F}(k, \ell, L, R, B)$ we are about to construct consists of functions of the form

$$g(x) = \frac{\sum_{i=1}^{k} \tilde{w}_i \tilde{K}_i(x)}{\sum_{i=1}^{k} \tilde{K}_i(x)}$$

with $\tilde{K}_i(x) = K_{\tilde{c}_i, \tilde{A}_i}(x)$. Note that $\tilde{w}_i, \tilde{c}_i$ and $\tilde{A}_i$ are not required to satisfy the weight, center, and receptive field conditions, respectively. Put $M_i(x) := \max\{\delta, K_i(x)\}$ and $\tilde{M}_i(x) := \max\{\delta, \tilde{K}_i(x)\}$. Then

$$\begin{aligned} &|f(x) - g(x)| \\ &= \left| \sum_{i=1}^{k} \left(w_i \frac{K_i(x)}{K_i(x) + \sum_{j\neq i} K_j(x)} - \tilde{w}_i \frac{\tilde{K}_i(x)}{\tilde{K}_i(x) + \sum_{j\neq i} \tilde{K}_j(x)} \right) \right| \\ &= \left| \sum_{i=1}^{k} \left(w_i \frac{K_i(x)}{M_i(x) + \sum_{j\neq i} K_j(x)} - \tilde{w}_i \frac{\tilde{K}_i(x)}{\tilde{M}_i(x) + \sum_{j\neq i} \tilde{K}_j(x)} \right) \right| \\ &\leq |I| + |II| + |III| \end{aligned}$$

with

$$\begin{aligned} I &:= \frac{\sum_{i=1}^{k} (w_i - \tilde{w}_i) K_i(x)}{\sum_{i=1}^{k} K_i(x)} \\ II &:= \sum_{i=1}^{k} \tilde{w}_i \frac{K_i(x) - \tilde{K}_i(x)}{M_i(x) + \sum_{j\neq i} K_j(x)} \\ III &:= \sum_{i=1}^{k} \tilde{w}_i \tilde{K}_i(x) \left(\frac{1}{M_i(x) + \sum_{j\neq i} K_j(x)} \right. \\ &\qquad \left. - \frac{1}{\tilde{M}_i(x) + \sum_{j\neq i} \tilde{K}_j(x)} \right) \\ &= \sum_{i=1}^{k} \frac{\tilde{w}_i \tilde{K}_i(x)}{\tilde{M}_i(x) + \sum_{j\neq i} \tilde{K}_j(x)} \\ &\qquad \cdot \frac{(\tilde{M}_i(x) - M_i(x)) + \sum_{j\neq i} (\tilde{K}_j(x) - K_j(x))}{M_i(x) + \sum_{j\neq i} K_j(x)}. \end{aligned}$$

Clearly

$$|I| \leq \max_{i=1,\ldots,k} |w_i - \tilde{w}_i|$$

and because of $M_i(x) \geq \delta$

$$|II| \leq \frac{B}{\delta} \sum_{i=1}^{k} |K_i(x) - \tilde{K}_i(x)|.$$

Note that by the Lipschitz continuity of $\max\{\delta,\cdot\}$ we have $|\tilde{M}_i(x) - M_i(x)| \le |\tilde{K}_i(x) - K_i(x)|$ so that

$$|III| \le \sum_{i=1}^{k} \frac{|\tilde{w}_i|\tilde{K}_i(x)}{\sum_{j=1}^{k}\tilde{K}_j(x)} \cdot \frac{\sum_{j=1}^{k}|\tilde{K}_j(x) - K_j(x)|}{\delta}$$
$$\le \frac{B}{\delta}\sum_{j=1}^{k}|\tilde{K}_j(x) - K_j(x)|.$$

Hence,

$$|f(x) - g(x)| \le \max_{i=1,\dots,k}|w_i - \tilde{w}_i| + \frac{2B}{\delta}\sum_{i=1}^{k}|K_{c_i,A_i}(x) - K_{\tilde{c}_i,\tilde{A}_i}(x)|$$

and we have

$$|f(x) - g(x)| \le \epsilon$$

if only

$$\max_{i=1,\dots,k}|w_i - \tilde{w}_i| \le \frac{\epsilon}{2}$$

and, for all i

$$|K_{c_i,A_i}(x) - K_{\tilde{c}_i,\tilde{A}_i}(x)| \le \frac{\epsilon\delta}{4Bk}.$$

To achieve this, we can choose the w_i from an equidistant $\epsilon/2$-cover of $[-B,B]$ and the $K_{\tilde{c}_i,\tilde{A}_i}, i = 1,\dots,k$ from the cover $\mathcal{K}(\epsilon\delta/(4Bk))$ from Lemma 1. This yields

$$N(\epsilon,\mathcal{F}(k,\ell,L,R,B),x_1^n) \le \left(\frac{4B}{\epsilon}+1\right)^k \cdot \prod_{i=1}^{k}\left(2\cdot\left(\frac{16eBk}{\epsilon\delta}\right)^{2(d^2+d+2)}\right)$$
$$= 2^k \cdot \left(\frac{4B}{\epsilon}+1\right)^k \cdot \left(\frac{16eBk}{\epsilon\delta}\right)^{2k(d^2+d+2)}$$

the assertion.

For *part b)* we observe that for the denominator in (1) we have the following inequality:

$$\inf_{\|x\|\le Q}\sum_{i=1}^{k}K_{c_i,A_i}(x) \ge k\cdot\inf_{0\le v\le(R+Q)^2L}K(v) = k\cdot K((R+Q)^2L) > 0. \quad (10)$$

Indeed, for fixed $\|x\| \le Q, \|c\| \le R$ and fixed symmetric, positive-definite $A \in \mathbb{R}_{d\times d}$ satisfying the above receptive field condition, we can find an orthonormal basis $\{x_1,\dots,x_d\}$ of $\mathbb{R}^d$ consisting of eigenvectors of A, $Ax_i = \lambda_i(A)x_i$. The expansion $x - c = \sum_{i=1}^{d}\langle x-c, x_i\rangle x_i$ yields

$$0 \le (x-c)^T A(x-c) = \sum_{i=1}^{d}\langle x-c,x_i\rangle^2\lambda_i(A)$$
$$\le \langle x-c, \sum_{i=1}^{d}\langle x-c,x_i\rangle x_i\rangle\lambda_{\max}(A) = \|x-c\|^2\lambda_{\max}(A).$$

By the center and receptive field condition, this may be continued to obtain

$$0 \le (x-c)^T A(x-c) \le \|x-c\|^2\lambda_{\max}(A) \le (R+Q)^2L.$$

Hence,

$$K_{c,A}(x) \ge \inf_{0\le v\le(R+Q)^2L}K(v)$$

which proves (10).

To finish the proof of part b) we can argue along the lines of part a) with $\delta = k\cdot K((R+Q)^2L)$. □

B. A Denseness Result

The following lemma shows that the class of all NRBFs of the form (1) is dense in the space $L_q(\mu)$ of functions whose qth moment w.r.t. the distribution μ of X ($q \ge 1$) exists. This result holds for all kernels satisfying the kernel condition i) in Section I.

Lemma 3: For any probability measure μ on $\mathbb{R}^d$, any function $g \in L_q(\mu)$ ($q > 0$), any $\epsilon > 0$, and any left-continuous, decreasing kernel K we can find an NRBF

$$f(x) = \frac{\sum_{i=1}^{k} w_i K(\|x-c_i\|^2/h)}{\sum_{i=1}^{k} K(\|x-c_i\|^2/h)} \quad (11)$$

with $k \in \mathbb{N}, w_i \in \mathbb{R}, c_i \in \mathbb{R}^d, h > 0$ such that

$$\|f-g\|_{L_q(\mu)} := \left(\int |f(x)-g(x)|^q\mu(dx)\right)^{1/q} < \epsilon. \quad (12)$$

Proof: Pick a continuous function $\tilde{g}$ with bounded support satisfying

$$\|g-\tilde{g}\|_{L_q(\mu)} \le \frac{\epsilon}{3} \quad (13)$$

(for existence of such a function see, e. g., [11, Theorem A1]). Without loss of generality we may assume $\tilde{g}(\cdot) \ge 0$, otherwise, add $\inf_x \tilde{g}(x)$ to g, $\tilde{g}$ and consider modified weights $w_i + \inf_x \tilde{g}(x)$ in (11). For $h > 0$ we define

$$\sigma_h(x) := \frac{\int \mathbf{1}_{[-1/h,1/h]^d}(z)\tilde{g}(z)K(\|x-z\|^2/h)dz}{\int \mathbf{1}_{[-1/h,1/h]^d}(z)K(\|x-z\|^2/h)dz}$$

with the convention $0/0 = 0$, where $\mathbf{1}_A(\cdot)$ is an indicator function of set A. Substitution $y := (x-z)/\sqrt{h}$ yields

$$\sigma_h(x) = \frac{h^{d/2}\int \mathbf{1}_{[-1/h,1/h]^d}(x-\sqrt{h}y)\tilde{g}(x-\sqrt{h}y)K(\|y\|^2)dy}{h^{d/2}\int \mathbf{1}_{[-1/h,1/h]^d}(x-\sqrt{h}y)K(\|y\|^2)dy}$$
$$\to \frac{\tilde{g}(x)\int K(\|y\|^2)dy}{\int K(\|y\|^2)dy} = \tilde{g}(x)$$

by the Lebesgue dominated convergence theorem as $h \searrow 0$. Since $|\sigma_h(x)| \le \|\tilde{g}\|_\infty < \infty$, convergence holds in $L_q(\mu)$ and we can choose an $h > 0$ such that

$$\|\sigma_h - \tilde{g}\|_{L_q(\mu)} \le \frac{\epsilon}{3}. \quad (14)$$

Now we approximate σ_h by an NRBF of the form (11). To this end, we sample Z_i i.i.d. from the uniform distribution on

$[-1/h, 1/h]^d$ and derive a uniform strong law of large numbers for

$$\frac{\sum_{i=1}^{k} \tilde{g}(Z_i) K(\|x - Z_i\|^2/h)}{\sum_{i=1}^{k} K(\|x - Z_i\|^2/h)}. \tag{15}$$

If $\{h_1, \ldots, h_\ell\}$ is an ϵ-cover with respect to $\|\cdot\|_{L(\nu_n)}$ on z_1^n of

$$\mathcal{K}_1 := \left\{ K(\|x - \cdot\|^2/h) : x \in \mathbb{R}^d \right\}$$

then

$$\{\tilde{g} \cdot \mathbf{1}_{[-1/h,1/h]^d} \cdot h_1, \ldots, \tilde{g} \cdot \mathbf{1}_{[-1/h,1/h]^d} \cdot h_\ell\}$$

constitutes an $(\|\tilde{g}\|_\infty \epsilon)$-cover of

$$\mathcal{K}_{\tilde{g}} := \left\{ f_x(\cdot) := \tilde{g}(\cdot) \mathbf{1}_{[-1/h,1/h]^d}(\cdot) K(\|x - \cdot\|^2/h) : x \in \mathbb{R}^d \right\}.$$

Thus,

$$N(\delta, \mathcal{K}_{\tilde{g}}, z_1^n) \le N(\delta/\|\tilde{g}\|_\infty, \mathcal{K}_1, z_1^n).$$

For the right-hand side, we clearly have

$$N(\delta/\|\tilde{g}\|_\infty, \mathcal{K}_1, z_1^n) \le N(\delta/\|\tilde{g}\|_\infty, \mathcal{K}, z_1^n)$$

with

$$\mathcal{K} := \Big\{ K_{c,A}(\cdot) : c \in \mathbb{R}^d, \; A \in \mathbb{R}_{d \times d} \text{ symmetric, positive definite} \Big\}.$$

The covering number of $\mathcal{K}$ can be bounded for any $\frac{\sup_x K(x)}{4} > \delta > 0$ by

$$N(\delta, \mathcal{K}, z_1^n) \le 2 \left(\frac{4e}{\delta} \right)^{2(d^2+d+2)}$$

using (9). Hence,

$$N(\delta/\|\tilde{g}\|_\infty, \mathcal{K}_1, z_1^n) \le 2 \left(\frac{4e\|\tilde{g}\|_\infty}{\delta} \right)^{2(d^2+d+2)}.$$

The result of Pollard [23] (also see [11, Theorem 9.1]) allows us to establish

$$\boldsymbol{P}\left(\sup_{f_x \in \mathcal{K}_{\tilde{g}}} \left| \frac{1}{k} \sum_{i=1}^{k} f_x(Z_i) - \boldsymbol{E} f_x(Z_i) \right| > \delta \right) \le 16 \left(\frac{32e\|\tilde{g}\|_\infty}{\delta} \right)^{2(d^2+d+2)} \exp\left(-\frac{k\delta^2}{128\|\tilde{g}\|_\infty^2 \|K\|_\infty^2} \right).$$

Using

$$\boldsymbol{E} f_x(Z_i) = \left(\frac{h}{2} \right)^d \int \mathbf{1}_{[-1/h,1/h]^d}(z) \tilde{g}(z) K(\|x - z\|^2/h) dz$$

and applying the Borel–Cantelli lemma we find

$$\sup_x \left| \frac{1}{k} \sum_{i=1}^{k} \tilde{g}(Z_i) K(\|x - Z_i\|^2/h) - \left(\frac{h}{2} \right)^d \int \mathbf{1}_{[-1/h,1/h]^d}(z) \tilde{g}(z) K(\|x - z\|^2/h) dz \right| \to 0$$

(this takes care of the numerator in (15)) and

$$\sup_x \left| \frac{1}{k} \sum_{i=1}^{k} K(\|x - Z_i\|^2/h) - \left(\frac{h}{2} \right)^d \int \mathbf{1}_{[-1/h,1/h]^d}(z) K(\|x - z\|^2/h) dz \right| \to 0$$

(this takes care of the denominator in (15)) both a.s. as $k \to \infty$. This implies the existence of a set Ω' of $\boldsymbol{P}$-measure 1 in the underlying probability space $(\Omega, \mathcal{A}, \boldsymbol{P})$ such that for any $\omega \in \Omega'$

$$\frac{\sum_{i=1}^{k} \tilde{g}(Z_i(\omega)) K(\|x - Z_i(\omega)\|^2/h)}{\sum_{i=1}^{k} K(\|x - Z_i(\omega)\|^2/h)} \to \frac{\int \mathbf{1}_{[-1/h,1/h]^d}(z) \tilde{g}(z) K(\|x - z\|^2/h) dz}{\int \mathbf{1}_{[-1/h,1/h]^d}(z) K(\|x - z\|^2/h) dz} = \sigma_h(x)$$

for all x as $k \to \infty$. Choose some $\omega \in \Omega'$ and set

$$c_i := Z_i(\omega) \quad \text{and} \quad w_i := \tilde{g}(c_i)$$

to obtain

$$\frac{\sum_{i=1}^{k} w_i K(\|x - c_i\|^2/h)}{\sum_{i=1}^{k} K(\|x - c_i\|^2/h)} \to \sigma_h(x)$$

for all x as $k \to \infty$. Since $w_i \le \|\tilde{g}\|_\infty$, and by the Lebesgue bounded convergence theorem convergence holds in $L_q(\mu)$, thus, we may choose k large enough such that

$$\left\| \sum_{i=1}^{k} w_i K(\|\cdot - c_i\|^2/h) \Big/ \sum_{i=1}^{k} K(\|\cdot - c_i\|^2/h) - \sigma_h(\cdot) \right\|_{L_q(\mu)} \le \frac{\epsilon}{3}. \tag{16}$$

Equations (13), (14), and (16) plus the triangle inequality yield the assertion. □

IV. Proofs of the Theorems

With the tools of the preceding Section, we can move on to the proofs of Theorem 1 and 2.

Proof of Theorem 1: Consider the decomposition

$$J(\hat{f}_n) - J^* = \left(\inf_{f \in \mathcal{F}_n} J(f) - J^* \right) + \left(J(\hat{f}_n) - \inf_{f \in \mathcal{F}_n} J(f) \right) =: A_n + E_n$$

into **approximation error** A_n and **estimation error** E_n. μ denotes the distribution of (X, Y).

1) *Approximation error* A_n: Choose $\sup_x K(x)/4 > \epsilon > 0$. According to Lemma 3, taking $q = 2$, there exists an NRBF $\tilde{f}$ of the form (11) such that

$$\inf_{f \in \mathcal{F}_n} J(f) - J^* = \inf_{f \in \mathcal{F}_n} \|f - r\|_{L_2(\mu)}^2 \le \|\tilde{f} - r\|_{L_2(\mu)}^2 \le \epsilon^2$$

where r is the regression function. Since $k_n, L_n, R_n, B_n \to \infty$, we have $f \in \mathcal{F}_n$ for all $n \geq n_0$ (n_0 sufficiently large). This proves

$$0 \leq \limsup_{n\to\infty} A_n = \limsup_{n\to\infty} \inf_{f\in\mathcal{F}_n} J(f) - J^* \leq \epsilon^2;$$

and from ϵ being arbitrary

$$\lim_{n\to\infty} A_n = 0.$$

Hence, it suffices to analyze the convergence properties of the estimation error E_n.

2) *Estimation error* E_n: First note that (cf. Devroye [4] and Haussler [12])

$$\begin{aligned} J(\hat{f}_n) &- \inf_{f\in\mathcal{F}_n} J(f) \\ &\leq 2 \sup_{f\in\mathcal{F}_n} \left| \frac{1}{n}\sum_{i=1}^n |f(X_i)-Y_i|^2 - \boldsymbol{E}|f(X)-Y|^2 \right| \\ &= 2 \sup_{h\in\mathcal{H}_n} \left| \frac{1}{n}\sum_{i=1}^n h(X_i,Y_i) - \boldsymbol{E}h(X,Y) \right| \end{aligned}$$

with

$$\mathcal{H}_n := \left\{ h : h(x,y) = |f(x)-y|^2, f \in \mathcal{F}_n \right\}.$$

Theorem 1 in Lugosi and Zeger [19] (see also Lemma 1 in Krzyżak *et al.* [15]) asserts that

$$J(\hat{f}_n) - \inf_{f\in\mathcal{F}_n} J(f) \to 0$$

in probability (almost surely) for *every* distribution of (X,Y) with $\boldsymbol{E}Y^2 < \infty$, if only

$$\sup_{f\in\mathcal{F}_n} \left| \frac{1}{n}\sum_{i=1}^n |f(X_i)-Y_i|^2 - \boldsymbol{E}|f(X)-Y|^2 \right| \to 0 \qquad (17)$$

in probability (almost surely) for *every* distribution of (X,Y) with Y being *bounded* with probability one. Thus, without loss of generality, we may assume that $|Y_i| \leq M < \infty$ when proving

$$\sup_{h\in\mathcal{H}_n} \left| \frac{1}{n}\sum_{i=1}^n h(X_i,Y_i) - \boldsymbol{E}h(X,Y) \right| \to 0$$

in probability (almost surely). To this end, observe that any $h \in \mathcal{H}_n$ is bounded in sup-norm by

$$||h||_\infty \leq \sup_{f\in\mathcal{F}_n} (||f||_\infty + M)^2 = (B_n + M)^2$$

and the result of Pollard [23, Lemma 33] (also see [11, Theorem 9.1]) yields

$$\begin{aligned} \boldsymbol{P}&\left(J(\hat{f}_n) - \inf_{f\in\mathcal{F}_n} J(f) > \epsilon \right) \\ &\leq \boldsymbol{P}\left(\sup_{h\in\mathcal{H}_n} \left| \frac{1}{n}\sum_{i=1}^n h(X_i,Y_i) - \boldsymbol{E}h(X,Y) \right| > \frac{\epsilon}{2} \right) \\ &\leq 8 \cdot \boldsymbol{E}N(\epsilon/16, \mathcal{H}_n, Z_1^n) \cdot \exp\left(-\frac{n\epsilon^2}{512(B_n+M)^4} \right) \qquad (18) \end{aligned}$$

with $Z_1^n := \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$. Observing that for $h_j(x,y) = |f_j(x)-y|^2$ ($f_j \in \mathcal{F}_n, j = 1,2$)

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n |h_1(X_i,Y_i) - h_2(X_i,Y_i)| \\ &\quad = \frac{1}{n}\sum_{i=1}^n \left| |f_1(X_i)-Y_i|^2 - |f_2(X_i)-Y_i|^2 \right| \\ &\quad \leq \frac{1}{n}\sum_{i=1}^n |f_1(X_i)+f_2(X_i)-2Y_i| \cdot |f_1(X_i)-f_2(X_i)| \\ &\quad \leq 2(B_n+M)\frac{1}{n}\sum_{i=1}^n |f_1(X_i)-f_2(X_i)| \end{aligned}$$

we bound the covering number by

$$N(\epsilon/16, \mathcal{H}_n, Z_1^n) \leq N(\epsilon/(32(B_n+M)), \mathcal{F}_n, X_1^n).$$

From the definition of $\mathcal{F}_n$

$$\begin{aligned} N(\epsilon/16, \mathcal{H}_n, Z_1^n) &\\ &\leq \sum_{k=1}^{k_n} N(\epsilon/(32(B_n+M)), \mathcal{F}(k,0,L_n,R_n,B_n), X_1^n). \end{aligned}$$

We now apply Lemma 2 for the different kernel classes. First, for *part a)*

$$\begin{aligned} N(\epsilon/16, \mathcal{H}_n, Z_1^n) &\\ &\leq \sum_{k=1}^{k_n} 2^k \cdot \left(\frac{128B_n(B_n+M)}{\epsilon} + 1 \right)^k \\ &\quad \cdot \left(\frac{512eB_n(B_n+M)k}{\epsilon\delta} \right)^{2k(d^2+d+1)} \\ &\leq k_n \cdot \left(\frac{1024}{\epsilon} \cdot B_n^2 \right)^{k_n} \left(\frac{1024eB_n^2k_n}{\epsilon\delta} \right)^{2k_n(d^2+d+1)} \qquad (19) \end{aligned}$$

for sufficiently large n. Equations (18) and (19) yield

$$\begin{aligned} &\boldsymbol{P}\left(J(\hat{f}_n) - \inf_{f\in\mathcal{F}_n} J(f) > \epsilon \right) \\ &\leq 8 \cdot \left(\frac{1024}{\epsilon} \cdot B_n^2 k_n \right)^{k_n} \cdot \left(\frac{1024eB_n^2k_n}{\epsilon\delta} \right)^{2k_n(d^2+d+1)} \\ &\quad \cdot \exp\left(-\frac{n\epsilon^2}{8192B_n^4} \right) \\ &\leq C_1 \exp\left(-\frac{n}{B_n^4} \left(C_2 - C_3 \frac{B_n^4 k_n}{n} \left(\log(B_n^2k_n) + \log\frac{B_n^2k_n}{\delta} \right) \right) \right) \end{aligned}$$

with suitable constants $C_j = C_j(d,\epsilon)$. $C_2 = \epsilon^2/8192 > 0$; therefore, the right-hand side of this chain of inequalities tends to zero if

$$\frac{B_n^4 k_n}{n} \log\frac{B_n^2 k_n}{\delta} \to 0 \qquad (20)$$

as $n \to \infty$, which proves the statement on weak consistency. We obtain strong consistency by the Borel–Cantelli lemma

if the right-hand side is summable in n, which is fulfilled if additionally

$$B_n^4 \leq \frac{n}{(1+\beta)\log n}$$

for some $\beta > 0$ and all sufficiently large n.

For *part b)*, under the assumption of μ having bounded support, say $\|X\| < Q < \infty$, it suffices to prove (17) for every distribution such that $\|X\| < Q$ and Y is bounded with probability one. Here, we exploit $K((R_n+Q)^2 L_n) \geq K(4R_n^2 L_n)$ for sufficiently large n, and we may replace δ from (19) onwards by $k_n K(4R_n^2 L_n)$. □

Proof of Theorem 2: The argument for the *approximation error* of the estimate is the same as in the proof of Theorem 1. Therefore, we need only consider the *estimation error*. From (20), it follows that (6)

$$\frac{B_n^4 k_n}{n} \log \frac{B_n^2 k_n}{\delta_n} \to 0$$

is sufficient for

$$J\left(\left(\hat{f}_n\right)^{\delta_n}\right) - \inf_{f \in \mathcal{F}_n} J\left(f^{(\delta_n)}\right) \to 0$$

in probability. This holds almost surely if in addition (8) is satisfied for some $\beta > 0$ and all sufficiently large n. Hence, to verify that the estimation error vanishes, it suffices to prove that almost surely

$$\left| J\Big((\hat{f}_n)^{(\delta_n)}\Big) - J\Big(\hat{f}_n\Big) \right| \to 0$$

and

$$\left| \inf_{f \in \mathcal{F}_n} J\left(f^{(\delta_n)}\right) - \inf_{f \in \mathcal{F}_n} J(f) \right| \to 0.$$

One can show that by assumption (7)

$$\left| J\left((\hat{f}_n)^{(\delta_n)}\right) - J\left(\hat{f}_n\right) \right| \leq 8\sqrt{\|g\|_\infty c_d} \left(\frac{B_n^4 k_n}{\ell_n^{d/2}} \left(s^{d/2} - K^{-1}(\delta_n)^{d/2} \right) \right)^{1/2} \to 0$$

where g is a density of X and c_d is the volume of the unit sphere in $\mathbb{R}^d$. The details of the proof can be found in Krzyżak and Schäfer [18] and are omitted. □

V. Conclusion

In this paper, the weak and strong universal consistency of the regression estimate based on normalized radial basis function networks has been investigated for a large class of radial kernels and for network parameters learned by empirical risk minimization. An important issue of the rates of convergence has not been dealt with and is left for future research.

References

[1] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

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

[3] G. Cybenko, "Approximations by superpositions of sigmoidal functions," *Math. Contr., Signals, Syst.*, vol. 2, pp. 303–314, 1989.

[4] L. Devroye, "Automatic pattern recognition: A study of the probability of error," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 4, pp. 530–543, Jul. 1988.

[5] L. Devroye, L. Györfi, and G. Lugosi, *Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[6] J. Duchon, "Sur l'erreur d'interpolation des fonctions de plusieurs variables par les D^m-splines," *RAIRO Anal. Numér.*, vol. 12, no. 4, pp. 325–334, 1978.

[7] A. Faragó and G. Lugosi, "Strong universal consistency of neural network classifiers," *IEEE Trans. Inf. heory*, vol. 39, no. 4, pp. 1146–1151, Jul. 1993.

[8] 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, Ed. London, U.K.: Chapman and Hall, 1993, pp. 97–113.

[9] F. Girosi, M. Jones, and T. Poggio, "Regularization theory and neural network architectures," *Neural Comput.*, vol. 7, pp. 219–267, 1995.

[10] U. Grenander, *Abstract Inference*. New York: Wiley, 1981.

[11] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer-Verlag, 2002.

[12] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Inform. Comput.*, vol. 100, pp. 78–150, 1992.

[13] K. Hornik, S. Stinchcombe, and H. White, "Multilayer feed-forward networks are universal approximators," *Neural Netw.*, vol. 2, pp. 359–366, 1989.

[14] A. N. Kolmogorov and V. M. Tihomirov, "ϵ-entropy and ϵ-capacity of sets in function spaces," *Transl. Amer. Math. Soc.*, vol. 17, pp. 277–364, 1961.

[15] A. Krzyżak, T. Linder, and G. Lugosi, "Nonparametric estimation and classification using radial basis function nets and empirical risk minimization," *IEEE Trans. Neural Netw.*, vol. 7, no. 2, pp. 475–487, Mar. 1996.

[16] A. Krzyżak and T. Linder, "Radial basis function networks and complexity regularization in function learning," *IEEE Trans. Neural Netw.*, vol. 9, no. 2, pp. 247–256, Mar. 1998.

[17] A. Krzyżak and H. Niemann, "Convergence and rates of convergence of radial basis functions networks in function learning," *Nonlin. Anal.*, vol. 47, pp. 281–292, 2001.

[18] A. Krzyżak and D. Schäfer, "Nonparametric Regression Estimation by Normalized Radial Basis Function Networks," Universität Stuttgart, Mathematisches Institut A, Stuttgart, Germany, Preprint 2002-15, 2002.

[19] G. Lugosi and K. Zeger, "Nonparametric estimation via empirical risk minimization," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 677–687, May 1995.

[20] J. Moody and J. Darken, "Fast learning in networks of locally-tuned processing units," *Neural Comput.*, vol. 1, pp. 281–294, 1989.

[21] J. Park and I. W. Sandberg, "Universal approximation using radial-basis-function networks," *Neural Comput.*, vol. 3, pp. 246–257, 1991.

[22] ——, "Approximation and radial-basis-function networks," *Neural Comput.*, vol. 5, pp. 305–316, 1993.

[23] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984.

[24] R. Shorten and R. Murray-Smith, "Side effects of normalizing radial basis function networks," *Int. J. Neural Syst.*, vol. 7, pp. 167–179, 1996.

[25] D. F. Specht, "Probabilistic neural networks," *Neural Netw.*, vol. 3, pp. 109–118, 1990.

[26] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. its Applic.*, vol. 16, pp. 264–280, 1971.

[27] V. N. Vapnik, *Estimation of Dependences based on Empirical Data*, 2nd ed. New York: Springer-Verlag, 1999.

[28] H. White, "Connectionist nonparametric regression: Multilayer feedforward networks that can learn arbitrary mappings," *Neural Netw.*, vol. 3, pp. 535–549, 1990.

[29] L. Xu, A. Krzyżak, and A. L. Yuille, "On radial basis function nets and kernel regression: Approximation ability, convergence rate and receptive field size," *Neural Netw.*, vol. 7, pp. 609–628, 1994.