Nonparametric Estimation and Classification Using Radial Basis Function Nets and Empirical Risk Minimization

Adam Krzyżak, *Member, IEEE,* Tamás Linder, and Gábor Lugosi

***Abstract*— In this paper we study convergence properties of radial basis function (RBF) networks for a large class of basis functions, and review the methods and results related to this topic. We obtain the network parameters through empirical risk minimization. We show the optimal nets to be consistent in the problem of nonlinear function approximation and in nonparametric classification. For the classification problem we consider two approaches: the selection of the RBF classifier via nonlinear function estimation and the direct method of minimizing the empirical error probability. The tools used in the analysis include distribution-free nonasymptotic probability inequalities and covering numbers for classes of functions.**

I. Introduction

In neural network literature much attention has been devoted to multilayer perceptrons (see, e.g., Barron [1], Hornik *et al.* [18], Xu *et al.* [36], and the references therein). Recently, another class of networks, called radial basis function (RBF) networks, has been studied by Broomhead and Lowe [5], Chen *et al.* [6], Moody and Darken [23], Poggio and Girosi [26], Powell [29], and Xu *et al.* [37], [38]. RBF nets have been shown to have universal approximation ability by Hartman *et al.* [16] and Park and Sandberg [24], [25]. Convergence rates for approximations of smooth functions by RBF nets have been studied by Girosi and Anzellotti [14]. In this paper we study generalization abilities of RBF nets (estimation error) and a learning procedure based on empirical risk minimization. We also show convergence of the optimized network in nonlinear functional approximation and classification by using the Vapnik–Chervonenkis approach and covering numbers.

Denote by $\mathcal{F}_k$ the class of RBF networks with one hidden layer and at most k nodes for a fixed kernel K: $\mathbb{R} \to \mathbb{R}$, that is, let $\mathcal{F}_k$ consist of all functions of the form

$$f_\theta(x) = \sum_{i=1}^{k} w_i K([x - c_i]^t A_i [x - c_i]) + w_0 \qquad (1)$$

where $w_0, w_1 \cdots, w_k \in [-b_k, b_k], c_1, \cdots, c_k \in \mathbb{R}^d$, and $A_1, \cdots, A_k \in \mathbb{R}^{d^2}, b_k > 0$ being a parameter of the class (we also allow $b_k = \infty$). The parameter θ is given as $\theta = (w_0, \cdots, w_k, b_1, \cdots, b_k, c_1, \cdots, c_k)$. For a given fixed kernel K, there are three sets of parameters: 1) the $w_i, i = 1, \cdots, k$, which are the weights of the output layer of an RBF net; 2) the center vectors $c_i, i = 1, \cdots, k$; and 3) $A_i, i = 1, \cdots, k$, which are $d \times d$ positive matrices determining the size of the receptive field of the basis functions $K([x - c_i]^t A_i [x - c_i])$. The last two sets constitute the weights of the hidden layer of an RBF net. The problem of determining a specific value $\hat{\theta}$ for θ is called learning or training. The most common choice for K is the Gaussian function, $K(r^2) = e^{-r^2}$ with $A_i^{-1} = \sigma_i^2 I$, but a number of alternatives can also be used [26]. For a specific $K(r^2)$, e.g., a Gaussian $K(r^2) = e^{-r^2}$, the size, shape and orientation of the receptive field of a node are determined by the matrix A_i. When $A_i^{-1} = \sigma_i^2 I$, the shape is a hyperspherical ball with its radius given by σ_i. When $A_i = diag[\sigma_{i1}^2, \cdots, \sigma_{id}^2]$, the shape of the receptive field is an elliptical ball with each axis coinciding with a coordinate axis; the lengths of the axes are determined by $\sigma_{i1}, \cdots, \sigma_{id}$, respectively. When A_i is a nondiagonal but symmetric matrix, we have $A_i = R_i^T D_i R_i$ where D_i is a diagonal matrix which determines the shape and size of the receptive field, and R_i is a rotation matrix which determines the orientation of the receptive field.

In addition to model (1), probabilistic neural networks based on the Bayes–Parzen density estimate have been considered by Specht [30]. The normalized version

$$f_n(x) = \frac{\sum_{i=1}^{n} w_i K([x - c_i]^t A_i [x - c_i])}{\sum_{i=1}^{n} K([x - c_i]^t A_i [x - c_i])} \qquad (2)$$

of (1) has been recently investigated in [23] and [38].

Let us now formulate the problem. Suppose that the random variables X and Y take values in $\mathbb{R}^d$ and $\mathbb{R}$, respectively. To predict the value of Y upon observing X, we need a

Manuscript received March 11, 1994; revised May 27, 1995. This paper was presented in part at the 12th International Conference on Pattern Recognition, Jerusalem, 1994. This research was supported in part by NSERC Grant 0GP000270, Canadian National Networks of Centers of Excellence Grant 293, the Alexander von Humboldt Foundation, the Hungarian National Foundation for Scientific Research Grant F014174, and the Foundation for Hungarian Higher Education and Research.

A. Krzyżak is with the Department of Computer Science, Concordia University, Montreal PQ, Canada H3G 1M8.

T. Linder and G. Lugosi are with the Department of Mathematics and Computer Science, Technical University of Budapest, Hungary.

Publisher Item Identifier S 1045-9227(96)01234-9.

measurable function f: $\mathbb{R}^d \to \mathbb{R}$ such that $f(X)$ is close to Y in some useful sense. If $\boldsymbol{E}|Y|^2 < \infty$, then there exists a measurable function m minimizing the squared L_2 prediction error, that is

$$J^* = \inf_f \boldsymbol{E}|f(X) - Y|^2 = \boldsymbol{E}|m(X) - Y|^2.$$

To estimate m without making any assumption about the distribution of (X,Y), we assume that a training set $D_n = \{X_i, Y_i\}_1^n$ of independent, identically distributed copies of (X,Y) is given, where D_n is independent of (X,Y). To obtain a good predictor we construct an estimate $f_n = f_{\hat{\theta}}$ of m by selecting the parameter vector $\hat{\theta}$ (and thus an estimator $f_{\hat{\theta}}$, depending on D_n) which minimizes the empirical error. In other words, based on the training sequence, we choose an estimator f_n from the class of functions $\mathcal{F}_k$, such that f_n minimizes the empirical L_2 error

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

that is

$$J_n(f_n) \le J_n(f) \text{ for } f \in \mathcal{F}_k.$$

The number of allowable nodes k will be a function of the training set size n, to be specified later. The performance of the estimate f_n is measured by the conditional squared L_2 error

$$J(f_n) = \boldsymbol{E}(|f_n(X) - Y|^2 | D_n).$$

We call a sequence of estimators $\{f_n\}$ strongly consistent for a given distribution of (X,Y), if

$$J(f_n) - J^* \to 0 \text{ almost surely (a.s.) as } n \to \infty.$$

f_n is strongly universally consistent if it is strongly consistent for any distribution of (X,Y) with $\boldsymbol{E}|Y|^2 < \infty$.

Observe that $J^* = \boldsymbol{E}(Y - m(X))^2$, where $m(x) = \boldsymbol{E}(Y|X = x)$ is the regression function, and $J(f_n) - J^* \to 0$ if and only if

$$\boldsymbol{E}((f_n(X) - Y)^2 | D_n) - \boldsymbol{E}(m(X) - Y)^2 = \boldsymbol{E}((f_n(X) - m(X))^2 | D_n) \to 0$$

which is the usual notion of L_2-consistency for regression function estimates.

Estimation of a regression function is in close relationship to pattern recognition. In the classification (pattern recognition) problem Y can take only two values: $Y \in \{-1, 1\}$. A classifier is a binary valued function $g_n(x)$, that can depend on the data D_n, and its error probability $\boldsymbol{P}\{g_n(X) \neq Y | D_n\}$ is to be minimized. The function g^* minimizing the error probability is called the decision, whose error probability $\boldsymbol{P}\{g^*(X) \neq Y\}$ is the Bayes risk. A sequence of classifiers $\{g_n\}$ is called strongly consistent, if

$$\boldsymbol{P}\{g_n(X) \neq Y | D_n\} - L^* \to 0 \text{ almost surely (a.s.) as } n \to \infty.$$

$\{g_n\}$ is strongly universally consistent if it is consistent for any distribution of (X,Y).

It is well known that good estimators of $m(x)$ provide classifiers with small error probability. One can, however, do even better than to derive classifiers from good regression estimators. One of the goals of this paper is to investigate consistency properties of an RBF-estimate of m and of the classifier derived from it, and of an RBF classifier based on the more natural approach of minimizing the empirical error probability.

The idea of empirical risk minimization has extensively been used in the literature. When this minimization is carried out over exceedingly rich (complex) family of candidate functions, the resulting estimate usually overfits the data, i.e., it is not likely to perform well for data statistically independent of the training set. Different measures of complexity of families of functions have been used for different purposes, but they are all related to the cardinality of a finite subset which represents the family in a certain sense. Examples are metric entropy [20], [31], and random covering numbers [27]. Asymptotic properties of the method of empirical risk minimization were studied among others by Vapnik [32] and Haussler [17]. For the candidate functions to approximate closely a large set of target functions, one generally needs to increase the size of the candidate family as the size of the training set increases. A good trade-off, however, should also be maintained between the complexity of the candidate family and the training data size to avoid overfitting. This idea of using candidate classes which grow in a controlled manner with the size of the training data is Grenander's method of sieves [15]. This approach is used in a pattern recognition framework by Devroye [9], and by White [35], and Faragó and Lugosi [13] for neural networks.

In this paper we apply empirical risk minimization together with the method of sieves to establish consistency in regression estimation and pattern recognition using RBF networks. 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 RBF nets.

In Section II, we show that under rather general conditions on the kernel, the family of functions $\cup_{k=1}^{\infty} \mathcal{F}_k$ is dense in $L_p(\mu)$ for any $p > 0$ and probability measure μ on $\mathbb{R}^d$. Section III deals with regression estimation, where in Theorem 2 we prove that the RBF regression estimate based on empirical error minimization is universally consistent. RBF classifiers obtained by empirical error probability minimization are studied in Section IV. Theorem 3 provides a nonasymptotic distribution-free upper bound on the estimation error of RBF classifiers using window kernels, and Theorem 4 deals with the universal consistency of such networks. We then show that there exist smooth unimodal kernels having no nonasymptotic distribution-free upper bounds such as in Theorem 3, in the following sense: for every n and any algorithm which determines the parameters of the RBF net based on a training sequence of length n, there exists a distribution such that $\boldsymbol{P}\{g_n(X) \neq Y | D_n\} - L^* > c$ for a positive universal constant $c > 0$. Finally, in Appendix A we review some results in VC and related theories. These results are fundamental in establishing bounds on the estimation error not only for RBF nets but for any scheme using empirical risk minimization.

II. Approximation

The error of the regression estimator can be decomposed into approximation and estimation parts

$$J(f_n) - J^* = (\inf_{f\in\mathcal{F}_k} J(f) - J^*) + (J(f_n) - \inf_{f\in\mathcal{F}_k} J(f)).$$

Since $\boldsymbol{E}|f(X)-Y|^2 - \boldsymbol{E}|m(X)-Y|^2 = \boldsymbol{E}|f(X)-m(X)|^2$ for all f, we have

$$\inf_{f\in\mathcal{F}_k} J(f) - J^* = \inf_{f\in\mathcal{F}_k} \boldsymbol{E}|f(X)-m(X)|^2. \tag{3}$$

In this section we consider the approximation error (3) when $\mathcal{F}_k$ is the family of RBF networks of the form

$$f_\theta(x) = \sum_{i=1}^{k} w_i K((x-c_i)/b_i) + w_0 \tag{4}$$

where K: $\mathbb{R}^d \to \mathbb{R}$ and $\theta = (w_0, \cdots, w_k, b_1, \cdots, b_k, c_1, \cdots, c_k)$ is the vector of parameters $w_0, \cdots, w_k \in \mathbb{R}, b_1, \cdots, b_k \in \mathbb{R}$ and $c_1, \cdots, c_k \in \mathbb{R}^d$. Clearly, when K is radially symmetric, the nets (4) constitute a subset of nets in the form of (1).

In what follows, we will show that $\cup_{k=1}^{\infty} \mathcal{F}_k$ is dense in $L_q(\mu)$ for any $q \in (0,\infty)$, and probability measure μ on $\mathbb{R}^d$, if K is a basis function that satisfies the regularity conditions listed in Theorem 1. We will also show that $\cup_{k=1}^{\infty} \mathcal{F}_k$ is dense in $L_p(\lambda)$ for any $p \in [1,\infty)$, where λ stands for the d-dimensional Lebesgue measure. In fact, any $m \in L_q(\mu) \cap L_p(\lambda)$ can be approximated simultaneously in both norms.

Consequently, for $q = 2$ the approximation error $\inf_{f\in\mathcal{F}_k} J(f) - J^*$ converges to zero if $k \to \infty$. Theorem 1 somewhat generalizes an approximation result of Park and Sandberg [24], [25] who showed that if $K \in L_1(\lambda) \cap L_p(\lambda)$ and $\int K \neq 0$, then the class of RBF nets defined in (4) is dense in $L_p(\lambda)$ for $p \geq 1$.

Theorem 1: Suppose K: $\mathbb{R}^d \to \mathbb{R}$ is bounded and

$$K \in L_1(\lambda) \cap L_p(\lambda) \tag{5}$$

for some $p \in [1,\infty)$, and assume that $\int K(x)\,dx \neq 0$. Let μ be an arbitrary probability measure on $\mathbb{R}^d$ and let $q \in (0,\infty)$. Then the RBF nets in the form (4) are dense in both $L_q(\mu)$ and $L_p(\lambda)$. In particular, if $m \in L_q(\mu) \cap L_p(\lambda)$, then for any ϵ there exists a $\theta = (w_0, \cdots, w_k, b_1, \cdots, b_k, c_1, \cdots, c_k)$ such that

$$\int_{\mathbb{R}^d} |f_\theta(x) - m(x)|^q \mu(dx) < \epsilon \quad \text{and}$$
$$\int_{\mathbb{R}^d} |f_\theta(x) - m(x)|^p\,dx < \epsilon. \tag{6}$$

Proof: We will use the norm notations $||f||_{L_q(\mu)}$ and $||g||_{L_p(\lambda)}$ for the $L_q(\mu)$ and $L_p(\lambda)$ norms, respectively, of any $f \in L_q(\mu)$ and $g \in L_p(\lambda)$. Since $m \in L_q(\mu) \cap L_p(\lambda)$, for any $\delta > 0$ there exists a continuous g supported on a compact set Q such that

$$||m-g||_{L_q(\mu)} < \frac{\delta}{2} \quad \text{and} \quad ||m-g||_{L_p(\lambda)} < \frac{\delta}{2}. \tag{7}$$

Let $\hat{K}(x) = K(x)/\int K(x)\,dx$, $\hat{K}_h(x) = (1/h^d)\hat{K}(x/h)$, and define

$$\sigma_h(x) = \int_{\mathbb{R}^d} g(y)\hat{K}_h(x-y)\,dy.$$

Since g is continuous, we have by [34, Theorem 9.8] that $\lim_{h\to 0} \sigma_h(x) = g(x)$ for all $x \in \mathbb{R}^d$, and we also have $|\sigma_h(x)| \leq B||\hat{K}||_{L_1(\lambda)} = B < \infty$ for all $x \in \mathbb{R}^d$, where $B = \sup_{x\in\mathbb{R}^d} |g(x)|$. Then the dominated convergence theorem implies that $\lim_{h\to 0} ||g - \sigma_h||_{L_q(\mu)} = 0$ and $\lim_{h\to 0} ||g - \sigma_h||_{L_p(\lambda)} = 0$, meaning that we can choose an $h > 0$ such that

$$||g-\sigma_h||_{L_q(\mu)} < \frac{\delta}{4} \quad \text{and} \quad ||g-\sigma_h||_{L_p(\lambda)} < \frac{\delta}{4} \tag{8}$$

are both satisfied. In the rest of the proof, using a probabilistic argument similar to the one by Barron [2], we will demonstrate that there exists an f_θ which approximates σ_h within $\delta/4$ in both $L_q(\mu)$ and $L_p(\lambda)$ norms.

First assume that $g(x) \geq 0$ for all x, and define the probability density function φ by

$$\varphi(x) = Cg(x), \quad C = \left(\int g(x)\,dx\right)^{-1}.$$

Then $\sigma_h(x) = \boldsymbol{E}[\tilde{K}(x,Z)]$, where $\tilde{K}(x,y) = C^{-1}\hat{K}_h(x-y)$ and Z has density φ. For any $r > 0$ let $S_r = \{x \in \mathbb{R}^d : ||x|| \leq r\}$, and define the probability measure $\tilde{\lambda}_r$ by $\tilde{\lambda}_r(B) = \lambda(B \cap S_r)/\lambda(S_r)$. Furthermore, let $\tilde{\mu} = \frac{1}{2}\tilde{\lambda}_r + \frac{1}{2}\mu$, and consider a random variable Y with distribution $\tilde{\mu}$.

Let $Z_1, Z_2, \cdots$ be an i.i.d. sequence independent of Y, with each Z_i having density φ. By the strong law of large numbers, for all x

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} \tilde{K}(x,Z_i) = \sigma_h(x)$$

almost surely. Then by Fubini's theorem

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} \tilde{K}(Y,Z_i) = \sigma_h(Y)$$

almost surely. Let ν denote the measure induced by $(Z_1, Z_2, \cdots)$. It follows after another application of Fubini's theorem that for ν-almost all infinite sequences $z = (z_1, z_2, \cdots)$ the sets $B_z \subset \mathbb{R}^d$ of x's for which

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} \tilde{K}(x,z_i) = \sigma_h(x)$$

fails to hold have $\tilde{\mu}$ measure zero. Thus there exists a sequence $(z_1, z_2, \cdots)$ for which

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} \tilde{K}(x,z_i) = \sigma_h(x)$$

holds for $\tilde{\mu}$-almost all x. In particular, the above convergence holds a.e. $[\lambda]$ on S_r, as well as a.e. $[\mu]$. Since $\sup_{x,y} |\tilde{K}(x,y)| \leq C^{-1} \sup_x |\hat{K}_h(x)| < \infty$, the dominated

convergence theorem implies that for $h_k(x) = (1/k) \Sigma_{i=1}^{k} \tilde{K}(x, z_i)$ we have

$$\lim_{k\to\infty} ||h_k - \sigma_h||_{L_q(\mu)} = 0. \tag{9}$$

To prove convergence in $L_p(\lambda)$, we choose r large enough to ensure that for given $\delta_1 > 0$ the following holds:

$$\sup_{y\in Q} \int_{\mathbb{R}^d - S_r} |\tilde{K}(x, y)|^p \, dx < \&\delta_1 \quad \text{and}$$
$$\int_{\mathbb{R}^d - S_r} |\sigma_h(x)|^p \, dx < \delta_1. \tag{10}$$

Since $K, \sigma_h \in L_p(\lambda)$ and Q is compact, such an r exists. Consider now the decomposition

$$\int_{\mathbb{R}^d} |h_k(x) - \sigma_h(x)|^p \, dx = \int_{S_r} |h_k(x) - \sigma_h(x)|^p \, dx + \int_{\mathbb{R}^d - S_r} |h_k(x) - \sigma_h(x)|^p \, dx.$$

Since $h_k \to \sigma_h$ on S_r a.e. $[\lambda]$, the first integral on the right-hand side converges to zero by dominated convergence. Now the fact that the z_i are in Q implies via (10) that

$$\left(\int_{\mathbb{R}^d - S_r} |h_k(x) - \sigma_h(x)|^p \, dx \right)^{1/p} \le \frac{1}{k} \sum_{i=1}^{k} \left(\int_{\mathbb{R}^d - S_r} |\tilde{K}(x, z_i)|^p \, dx \right)^{1/p} + \left(\int_{\mathbb{R}^d - S_r} |\sigma_h(x)|^p \, dx \right)^{1/p} \le 2\delta_1^{1/p}.$$

Hence by choosing r and δ_1 appropriately, and letting k be large enough, we obtain $||h_k - \sigma_h||_{L_p(\lambda)} \le \delta/4$. Since the h_k are RBF nets in the form of (4), this and (9) imply the existence of an f_θ such that

$$||\sigma_h - f_\theta||_{L_q(\mu)} < \frac{\delta}{4} \quad \text{and} \quad ||\sigma_h - f_\theta||_{L_p(\lambda)} < \frac{\delta}{4}. \tag{11}$$

To prove (11) for general g we use the decomposition $g(x) = g^+(x) - g^-(x)$, where g^+ and g^- denote the positive and negative parts of g, respectively. Then

$$\sigma_h(x) = \sigma_h^{(1)}(x) - \sigma_h^{(2)}(x) = \int_{\mathbb{R}^d} g^+(y)\hat{K}(x - y) \, dy - \int_{\mathbb{R}^d} g^-(y)\hat{K}(x - y) \, dy.$$

Now $\sigma_h^{(1)}$ and $\sigma_h^{(2)}$ are approximated as above by $f_{\theta^{(1)}}$ and $f_{\theta^{(2)}}$, respectively, and for $f_\theta = f_{\theta^{(1)}} + f_{\theta^{(2)}}$ we obtain (11).

Finally, from (7), (8), and (11) we conclude that

$$||m - f_\theta||_{L_q(\mu)} < \frac{\delta}{4} \quad \text{and} \quad ||m - f_\theta||_{L_p(\lambda)} < \frac{\delta}{4}$$

which proves (6) after choosing a suitable δ as a function of ϵ. Note that the above proof also establishes the first statement of the theorem, namely that $\{f_\theta : \theta \in \Theta\}$ is dense in both $L_q(\mu)$ and $L_p(\lambda)$. □

For smooth classes of function, one can obtain rate-of-approximation results as in [2]. One of the underlying methods here is the investigation of the properties of the best nth order convex approximation from a set of functions, when the target function is assumed to lie in the closure of the convex hull of the set. Such results are given by Barron [2], Darken *et al.* [8], and Girosi and Anzellotti [14]. The very important question of incremental (i.e., recursive) approximations is also dealt with by Jones [19], as well as in the above cited papers.

III. Regression Estimation

In this section we consider regression estimation using RBF networks of the form

$$f_\theta(x) = \sum_{i=1}^{k} w_i K([x - c_i]^t A_i [x - c_i]) + w_0. \tag{12}$$

Here $\theta = (w_0, \cdots, w_k, c_1, \cdots, c_k, A_1, \cdots, A_k)$ is the vector of parameters, where $w_0, \cdots, w_k \in \mathbb{R}, c_1, \cdots, c_k \in \mathbb{R}^d$, and $A_1, \cdots, A_k$ are positive $d \times d$ matrices. The scalar basis function $K : [0, \infty) \to \mathbb{R}$ is a monotone decreasing, left-continuous, bounded function. Given the training set $D_n = ((X_1, Y_1), \cdots, (X_n, Y_n))$ consisting of n i.i.d. copies of (X, Y), our estimate of the regression function $m(x) = \boldsymbol{E}(Y|X = x)$ is the RBF f_θ which minimizes the empirical L_2 error

$$J_n(f_\theta) = \frac{1}{n} \sum_{j=1}^{n} |f_\theta(X_j) - Y_j|^2. \tag{13}$$

To be more specific, for each n we fix Θ_n as the set of parameters defined by

$$\Theta_n = \left\{ \theta = (w_0, \cdots, w_{k_n}, c_1, \cdots, c_{k_n}, A_1, \cdots, A_{k_n}) : \sum_{i=0}^{k_n} |w_i|^2 \le b_n \right\}$$

and we choose our regression estimator as an $f_\theta, \theta \in \Theta_n$ satisfying

$$J_n(f_\theta) = \min_{\hat{\theta}\in\Theta_n} J_n(f_{\hat{\theta}}). \tag{14}$$

Thus the optimal f_θ is sought among the RBF's consisting of at most k_n neurons satisfying the weight constraint $\Sigma_{i=0}^{k_n} |w_i|^2 \le b_n$. If we assume that $|K(r)| \le 1$ for all $r \ge 0$, then these constraints and the Cauchy–Schwarz inequality imply that for any $\theta \in \Theta_n$ and $x \in \mathbb{R}^d$

$$|f_\theta(x)|^2 = \left| \sum_{i=1}^{k} w_i K([x - c_i]^t A_i [x - c_i]) + w_0 \right|^2 \le b_n(k_n + 1). \tag{15}$$

In what follows we will denote by f_n, for convenience, the empirically optimal RBF net in (14). With a slight abuse of notation we put $\mathcal{F}_n = \{f_\theta : \theta \in \Theta_n\}$ thus making explicit the dependence on n of $\mathcal{F}_k, k = k_n$. We have the following consistency result for the regression estimate f_n:

Theorem 2: Consider a family of RBF nets defined by (12), with $k \geq 1$ arbitrary, such that the defining kernel $K:[0,\infty) \to \mathbb{R}$ is integrable with respect to the Lebesgue measure. Suppose furthermore that K is monotone decreasing, left-continuous, and bounded. If $k_n, b_n \to \infty$ and $k_n^3 b_n^2 \log(k_n^3 b_n^2)/n \to 0$ as $n \to \infty$, then the RBF net f_n minimizing the empirical L_2 error over $\mathcal{F}_n = \{f_\theta : \theta \in \Theta_n\}$ satisfies

$$J(f_n) - J^* \to 0 \quad \text{as} \quad n \to \infty$$

in probability for any distribution of (X,Y) satisfying $\boldsymbol{E}|Y|^2 < \infty$. If in addition $k_n/n^\delta \to \infty$ as $n \to \infty$ for some $\delta \geq 0$, then f_n is strongly universally consistent, i.e., the above convergence holds almost surely.

Proof: We begin with the usual decomposition of the error into approximation and estimation parts

$$J(f_n) - J^* = (\inf_{f \in \mathcal{F}_n} J(f) - J^*) + (J(f_n) - \inf_{f \in \mathcal{F}_n} J(f)).$$

Approximation Error: Since $\boldsymbol{E}|f(X) - Y|^2 - \boldsymbol{E}|m(X) - Y|^2 = \boldsymbol{E}|f(X) - Y|^2$ for all f, we have

$$\inf_{f \in \mathcal{F}_n} J(f) - J^* = \inf_{f \in \mathcal{F}_n} \boldsymbol{E}|f(X) - m(X)|^2.$$

But $\cup_{n=1}^\infty \mathcal{F}_n$ is just the set of functions of the form (12) since $k_n, b_n \to \infty$ as $n \to \infty$. Thus the right-hand side of the above equation tends to zero by Theorem 1, implying that the approximation error $\inf_{f \in \mathcal{F}} J(f) - J^*$ converges to zero.

Estimation Error, Y Unbounded: To deal with the estimation error $J(f_n) - \inf_{f \in \mathcal{F}_n} J(f)$ we will use the well-known inequality

$$\begin{aligned} J(f_n) &- \inf_{f \in \mathcal{F}_n} J(f) \\ &\leq 2 \sup_{f \in \mathcal{F}_n} |J_n(f) - \boldsymbol{E}J_n(f)| \\ &= 2 \sup_{f \in \mathcal{F}_n} \left| \frac{1}{n} \sum_{j=1}^n |f(X_j) - Y_j|^2 - \boldsymbol{E}|f(X) - Y|^2 \right|. \end{aligned} \tag{16}$$

To prove that this supremum converges to zero, we will use nonasymptotic uniform large deviation inequalities involving suprema over classes of functions.

First we restate a result by Lugosi and Zeger [21] asserting that if the right-hand side of (16) converges to zero either in probability or almost surely for all distributions such that $|Y|$ is bounded, then the estimation error converges to zero in the same sense for any distribution such that $\boldsymbol{E}|Y|^2 < \infty$. For the sake of completeness, we include the proof in Appendix B.

Lemma 1 (Lugosi and Zeger [21]): If

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

in probability (almost surely) for every distribution of (X,Y) such that Y is bounded with probability one, then

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

in probability (almost surely) for every distribution of (X,Y) such that $\boldsymbol{E}|Y|^2 < \infty$.

Estimation Error, Y Bounded: By the above result, we may assume that $\boldsymbol{P}\{|Y| \leq L\} = 1$ for some positive L. If $|y| \leq L$, then by (15) the functions $h(x,y) = (f(x) - y)^2, f \in \mathcal{F}_n$, are bounded above

$$\begin{aligned} h(x,y) &\leq 4 \max\{|f(x)|^2, |y|^2\} \leq 4 \max\{b_n(k_n+1), L^2\} \\ &\leq 5 b_n k_n \end{aligned} \tag{17}$$

when $b_n(k_n+1) \geq L^2$ (i.e., when n is large enough). Thus, for such n, the supremum in (16) is bounded above by

$$\sup_{h \in \mathcal{H}_n} \left| \frac{1}{n} \sum_{i=1}^n h(X_i, Y_i) - \boldsymbol{E}h(X,Y) \right| \tag{18}$$

where $\mathcal{H}_n = \{h(x,y) : h(x,y) = (f(x) - y)^2, f \in \mathcal{F}_n\}$, and for all $h \in \mathcal{H}_n$ we have $|h(x,y)| \leq 5 b_n k_n$ for all $(x,y) \in \mathbb{R}^d \times [-L, L]$. We will now use the notion of covering numbers (Definition 1 in Appendix A) and Lemma 3 by Pollard in Appendix A. We apply Lemma 3 with $m = d+1, \mathcal{F} = \mathcal{H}_n, Z_1^n = ((X_1,Y_1), \cdots, (X_n,Y_n))$, and $B = 5 b_n k_n$ to obtain

$$\begin{aligned} &\boldsymbol{P}\{J(f_n) - \inf_{f \in \mathcal{F}_n} J(f) > \epsilon\} \\ &\quad \leq \boldsymbol{P}\left\{ \sup_{h \in \mathcal{H}_n} \left| \frac{1}{n} \sum_{i=1}^n h(X_i, Y_i) - Eh(X,Y) \right| > \epsilon/2 \right\} \\ &\quad \leq 8 \boldsymbol{E} N(\epsilon/16, \mathcal{H}_n, Z_1^n) e^{-n\epsilon^2/512(5k_n b_n)^2}. \end{aligned} \tag{19}$$

Bounding the Covering Number: In the remaining part of the proof we derive an upper bound on $N(\epsilon, \mathcal{H}_n, z_1^n)$, which will imply consistency through the above inequality. Let f_1 and f_2 be two real functions on $\mathbb{R}^d$ satisfying $|f_i(x)|^2 \leq b_k(k_n+1), i = 1,2$, for all $x \in \mathbb{R}^d$. Then for $h_1(x,y) = (f_1(x) - y)^2$ and $h_2(x,y) = (f_2(x) - y)^2$, and any $z_1^n = ((x_1,y_1), \cdots, (x_n,y_n))$ with $|y_i| \leq L, i = 1, \cdots, n$, we have

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n |h_1(x_i,y_i) - h_2(x_i,y_i)| \\ &\quad = \frac{1}{n} \sum_{i=1}^n |(f_1(x_i) - y_i)^2 - (f_2(x_i) - y_i)^2| \\ &\quad \leq \frac{1}{n} \sum_{i=1}^n 2|f_1(x_i) - f_2(x_i)| \\ &\qquad \cdot \max\{|f_1(x_i) - L|, |f_2(x_i) - L|\} \\ &\quad \leq 4\sqrt{b_n k_n} \frac{1}{n} \sum_{i=1}^n |f_1(x_i) - f_2(x_i)| \end{aligned} \tag{20}$$

since $|f(x) - L| \leq 2\sqrt{b_n(k_n+1)}$ for n large enough. Since $|f|^2 \leq b_n(k_n+1)$ for all $f \in \mathcal{F}_n$, the functions $f_1, \cdots, f_l, l = N(\epsilon, \mathcal{F}_n, x_1^n)$, in the definition of the covering number can be chosen so that they also satisfy $|f_i|^2 \leq b_n(k_n+1), i = 1, \cdots, l$, where $x_1^n = (x_1, \cdots, x_n)$. Combining this with (20) we conclude that for n large enough

$$N(\epsilon, \mathcal{H}_n, z_1^n) \leq N\left(\frac{\epsilon}{4\sqrt{k_n b_n}}, \mathcal{F}_n, x_1^n \right). \tag{21}$$

Next we estimate $N(\epsilon, \mathcal{F}_n, x_1^n)$ by building up $\mathcal{F}_n$ from a relatively simpler class of functions. For this, our tool will be

Lemma 4 in Appendix A. Let us define the classes of functions $\mathcal{G}$ and $\mathcal{G}_n$ by

$$\mathcal{G} = \{K([x-c]^t A[x-c]): c \in \mathbb{R}^d\},$$
$$\mathcal{G}_n = \{w \cdot g: g \in \mathcal{G}, |w| \le \sqrt{b_n}\}.$$

Then by Lemma 4 (2), for any $x_1^n \in \mathbb{R}^{nd}$ we have

$$N(\epsilon, \mathcal{G}_n, x_1^n) \le \frac{2\sqrt{b_n}}{\epsilon} N\left(\frac{\epsilon}{2\sqrt{b_n}}, \mathcal{G}, x_1^n\right).$$

Since

$$\mathcal{F}_n = \left\{ f = \sum_{i=1}^{k_n} w_i g_i + w_0: g \in \mathcal{G}, \sum_{i=0}^{k_n} |w_i|^2 \le b_n \right\}$$

by applying Lemma 4 (1) $k_n + 1$ times we obtain

$$\begin{aligned} N(\epsilon, \mathcal{F}_n, x_1^n) &\le \frac{2\sqrt{b_n}(k_n+1)}{\epsilon} (N(\epsilon/(k_n+1), \mathcal{G}_n, x_1^n))^{k_n} \\ &\le \frac{2\sqrt{b_n}(k_n+1)}{\epsilon} \left(\frac{2\sqrt{b_n}(k_n+1)}{\epsilon} \right. \\ &\quad \left. \cdot N\left(\frac{\epsilon}{2\sqrt{b_n}(k_n+1)}, \mathcal{G}, x_1^n\right)\right)^{k_n}. \end{aligned} \tag{22}$$

In the last step, we estimate $N(\epsilon, \mathcal{G})$. We will need the concept of VC dimension of a class of sets as given in Definition 2 in Appendix A and the connection between covering numbers and VC dimension given by Lemma 5 in Appendix A. Using an argument by Pollard [27] we will show that the collection of graphs of functions in $\mathcal{G}$ has finite VC dimension, which will result in a suitable bound on $N(\epsilon, \mathcal{G}, x_1^n)$ via Lemma 5.

Since K is left continuous and monotone decreasing we have $K([x-c]^t A[x-c]) \ge t$ iff $[x-c]^t A[x-c] \le \varphi(t)$, where $\varphi(t) = \max\{y: K(y) \ge t\}$. Equivalently, (x,t) must satisfy

$$x^t Ax - x^t(Ac + A^t c) + c^t Ac - \varphi(t) \le 0.$$

Consider now the set of real functions on $\mathbb{R}^{d+1}$ that are given for any $(x,s) \in \mathbb{R}^d \times \mathbb{R}$ by

$$g_{A,\alpha,\beta,\gamma}(x,s) = x^t Ax + x^t \alpha + \beta s + \gamma$$

where A ranges over all $d \times d$ matrices, and $\alpha \in \mathbb{R}^d, \beta, \gamma \in \mathbb{R}$ are arbitrary. The collection $\{g_{A,\alpha,\beta,\gamma}\}$ is a (d^2+d+1)-dimensional vector space of functions, thus the class of sets of the form $\{(x,s): g_{A,\alpha,\beta,\gamma}(x,s) \le 0\}$ has VC dimension at most d^2+d+2 by Lemma 6 in Appendix A. Clearly, if for a given a collection of points $\{(x_i,t_i)\}$ a set $\{(x,t): g(x) \ge t\}, g \in \mathcal{G}$ picks out the points $(x_{i_1}, t_{i_1}), \cdots, (x_{i_l}, t_{i_l})$, then there exist A, α, β, γ such that $\{(x,s): g_{A,\alpha,\beta,\gamma}(x,s) \ge 0\}$ picks out exactly $(x_{i_1}, \varphi(t_{i_1})), \cdots, (x_{i_l}, \varphi(t_{i_l}))$. This shows that $V_{\mathcal{G}} \le d^2+d+2$. Thus Lemma 5 gives

$$N(\epsilon, \mathcal{G}, x_1^n) \le 2\left(\frac{2e}{\epsilon}\right)^{2(d^2+d+2)}$$

from which, upon substitution into (22), we obtain

$$N(\epsilon, \mathcal{F}_n, x_1^n) \le \left(\frac{4e\sqrt{b_n}(k_n+1)}{\epsilon}\right)^{2(k_n+1)(d^2+d+3)}.$$

In view of (21) and (19) this implies that with appropriate constants C_1, C_2, and C_3 we have

$$\begin{aligned} &\boldsymbol{P}\{J(f_n) - \inf_{f \in \mathcal{F}_n} J(f) > \epsilon\} \\ &\quad \le \left(\frac{C_1 b_n k_n^{3/2}}{\epsilon}\right)^{C_2 k_n} e^{-n\epsilon^2/C_3(k_n b_n)^2} \\ &\quad = \exp\left(-\frac{n}{(k_n b_n)^2}\left[\epsilon^2/C_3 - \frac{C_2 k_n^3 b_n^2}{n}\right.\right. \\ &\qquad \left.\left. \cdot \log \frac{C_1 b_n k_n^{3/2}}{\epsilon}\right]\right). \end{aligned} \tag{23}$$

Since $k_n, b_n \to \infty$ as $n \to \infty$, the above upper bound tends to zero for any $\epsilon > 0$ if

$$\frac{k_n^3 b_n^2 \log k_n^3 b_n^2}{n} \to 0 \quad \text{as} \quad n \to \infty$$

and the universal consistency is proved. It is easy to check that if we additionally have $k_n/n^\delta \to \infty$ for some $\delta > 0$, then the upper bound (23) is summable in n for any $\epsilon > 0$, and the strong universal consistency follows by the Borel–Cantelli lemma. □

IV. Classification

In the classification (pattern recognition) problem, based upon the observation of a random vector $X \in \mathbb{R}^d$, one has to guess the value of a corresponding label Y, where Y is a random variable taking its values from $\{-1, 1\}$. The decision is now a function $g: \mathbb{R}^d \to \{-1, 1\}$, whose goodness is measured by the error probability $L(g) = \boldsymbol{P}\{g(X) \ne Y\}$. It is well known that the decision function that minimizes the error probability is given by

$$g^*(x) = \begin{cases} -1 & \text{if } m(x) \le 0 \\ 1 & \text{otherwise} \end{cases}$$

where $m(x) = \boldsymbol{E}(Y|X=x)$, g^* is called the Bayes decision, and its error probability $L^* = \boldsymbol{P}\{g^*(X) \ne Y\}$ is the Bayes risk.

When the joint distribution of (X,Y) is unknown (as is typical in practical situations), a good decision has to be learned from a training sequence

$$D_n = ((X_1, Y_1), \cdots, (X_n, Y_n))$$

which consists of n independent copies of the $\mathbb{R}^d \times \{-1,1\}$-valued pair (X,Y). Then formally, a decision rule g_n is a function $g_n: \mathbb{R}^d \times (\mathbb{R}^d \times \{-1,1\})^n \to \{-1,1\}$, whose error probability is given by

$$L(g_n) = \boldsymbol{P}\{g_n(X, D_n) \ne Y | D_n\}.$$

Note that $L(g_n)$ is a random variable, as it depends on the (random) training sequence D_n. For notational simplicity, we will write $g_n(x)$ instead of $g_n(x, D_n)$.

A sequence of classifiers $\{g_n\}$ is called strongly consistent if

$$\boldsymbol{P}\{g_n(X) \ne Y | D_n\} - L^* \to 0$$
almost surely (a.s.) as $n \to \infty$

and $\{g_n\}$ is strongly universally consistent if it is consistent for any distribution of (X,Y).

It is intuitively clear that pattern recognition is closely related to regression function estimation. This is seen by observing that the function m defining the optimal decision g^* is just the regression function $\boldsymbol{E}(Y|X=x)$. Thus, having a good estimate $f_n(x)$ of the regression function m, we expect a good performance of the decision rule

$$g_n(x) = \begin{cases} -1 & \text{if } f_n(x) \le 0 \\ 1 & \text{otherwise.} \end{cases} \tag{24}$$

Indeed, we have the well-known inequality

$$\begin{aligned} &\boldsymbol{P}\{g_n(X) \neq Y|X=x, D_n\} - \boldsymbol{P}\{g^*(X) \neq Y|X=x\} \\ &\quad \le |f_n(x) - m(x)| \end{aligned} \tag{25}$$

(see, e.g., [10]) and in particular

$$\begin{aligned} &\boldsymbol{P}\{g_n(X) \neq Y|D_n\} - \boldsymbol{P}\{g^*(X) \neq Y\} \\ &\quad \le (\boldsymbol{E}((f_n(X) - m(X))^2|D_n))^{1/2}. \end{aligned}$$

Therefore, any strongly consistent estimate f_n of the regression function m leads to a strongly consistent classification rule g_n via (24). For example, if f_n is an RBF-estimate of m based on minimizing the empirical L_2 error $J_n(f_\theta)$, then according to the consistency theorem discussed in the previous section, g_n is a strongly universally consistent classification rule. That is, for any distribution of (X,Y), it is guaranteed that the error probability of the RBF-classifier gets arbitrarily close to that of the best possible classifier if the training sequence D_n is long enough.

While consistency is an extremely important property, it gives little information about the finite-sample behavior of $L(g_n)$. The intuitive reason why we can do much better than basing our decision on an L_2-consistent RBF regression estimate is that the empirical L_2 error has only a vague relationship with the error probability. For the classification problem, it is more natural to minimize the number of errors committed by the corresponding classifiers on the training sequence.

Let $K: \mathbb{R}^d \to \mathbb{R}$ be a kernel function. Consider RBF networks given by

$$f_\theta(x) = \sum_{i=1}^{k} w_i K(A_i[x - c_i]) + w_0 \tag{26}$$

where $\theta = (w_0, \cdots, w_k, c_1, \cdots, c_k, A_1, \cdots, A_k)$ is the vector of parameters, $w_0, \cdots, w_k \in \mathbb{R}, c_1, \cdots, c_k \in \mathbb{R}^d$, and $A_1, \cdots, A_k$ are nonsingular $d \times d$ matrices. Let $\{k_n\}$ be a sequence of positive integers. Define $\mathcal{F}_n$ as the set of RBF networks in the form of (26) with $k = k_n$. Given an f_θ as above, we define the classifier $g_\theta: \mathbb{R}^d \to \{-1, 1\}$ as

$$g_\theta(x) = \begin{cases} -1 & \text{if } f_\theta(x) \le 0 \\ 1 & \text{otherwise.} \end{cases} \tag{27}$$

Let $\mathcal{G}_n$ be the class of classifiers based of the class of functions $\mathcal{F}_n$. To every classifier $g \in \mathcal{G}_n$, assign the empirical error probability

$$\hat{L}_n(g) = \frac{1}{n} \sum_{i=1}^{n} I_{\{g(X_i) \neq Y_i\}}$$

i.e., the normalized number of errors committed by g in classifying D_n. It is a natural choice to pick a classifier g_n from $\mathcal{G}_n$ by minimizing the empirical error probability

$$\hat{L}_n(g_n) \le \hat{L}_n(\phi) \text{ for } \phi \in \mathcal{G}_n.$$

In the sequel we investigate the behavior of the error probability $L(g_n) = \boldsymbol{P}\{g_n(X) \neq Y|D_n\}$ of the selected classifier.

The distance $L(g_n) - L^*$ between the error probability of the selected rule and the Bayes risk may be decomposed into a random and a deterministic component

$$L(g_n) - L^* = (L(g_n) - \inf_{g \in \mathcal{G}_n} L(g)) + (\inf_{g \in \mathcal{G}_n} L(g) - L^*)$$

where the first term on the right-hand side is called the estimation error and the second is called the approximation error.

We begin by investigating the estimation error. The estimation error measures the distance of the error probability of the selected classifier from that of the best k_n-node RBF classifier. The size of the estimation error is an interesting quantity in itself, as it tells us how far we are from the best classifier realizable by a network with complexity k_n. Assume first that $K: \mathbb{R}^d \to \{0,1\}$, i.e., $K(x) = I_{\hat{C}}(x)$ is the indicator function of some subset $\hat{C}$ of $\mathbb{R}^d$. We have the following:

Theorem 3: Assume that K is an indicator function. Assume that the class of sets

$$\mathcal{C}_1 = \{\{x \in \mathbb{R}^d: K(A[x-c]) > 0\}: c \in \mathbb{R}^d, A \text{ invertible}\} \tag{28}$$

has a finite VC dimension $V_{\mathcal{C}_1}$. Then for every n, k_n and $\epsilon > 0$

$$\begin{aligned} &\boldsymbol{P}\{L(g_n) - \inf_{g \in G_n} L(g) > \epsilon\} \\ &\quad \le 4\left(\frac{4en}{V_{\mathcal{C}_1}+1}\right)^{k_n(V_{\mathcal{C}_1}+2)} e^{-n\epsilon^2/32} \\ &\quad = 4\exp\left(-n\left[\frac{\epsilon^2}{32} - \frac{C_2 k_n \log(C_1 n)}{n}\right]\right) \end{aligned}$$

for some constants C_1 and C_2 depending only on $V_{\mathcal{C}_1}$.

The importance of the theorem above lies in the fact that it gives a distribution-free, nonasymptotic bound for the error probability of the selected classification rule. By a simple bounding argument it follows from Theorem 3 that

$$\boldsymbol{E}L(g_n) - \inf_{g \in G_n} L(g) \le c\sqrt{\frac{k_n \log n}{n}}$$

where the constant c depends on the class $\mathcal{C}_1$ only. This inequality tells us that no matter what the distribution of (X,Y) is, we are always within $O(\sqrt{(k_n \log n)/n})$ of the best error probability achievable by a k_n-node RBF. We emphasize that this is not true for classifiers based on minimizing the empirical L_2-error.

Examples: The quantity $V_{\mathcal{C}_1}$ depends only on the set $\hat{C} \in \mathbb{R}^d$ defining the indicator K.

1) When $\hat{C}$ is the unit sphere, $\mathcal{C}_1$ is just the family of d-dimensional ellipsoids. It is well known (see, e.g., Pollard [27]) that $V_{\mathcal{C}_1} < \infty$ in this case (in fact, it is not hard to see that $V_{\mathcal{C}} \le d^2 + d + 2$).

2) Suppose $\hat{C}$ is a convex polytope, i.e., the intersection of l ($l > d$) halfspaces. Then $\mathcal{C}_1$ is a collection of polytopes of l faces, and it follows from, e.g., Pollard [27] that $V_{\mathcal{C}_1} \leq l(d+1)$.

Proof of Theorem 3: We will start with the observation that

$$L(g_n) - \inf_{g \in \mathcal{G}_n} L(g) \leq 2 \sup_{g \in \mathcal{G}_n} |\hat{L}_n(g) - \boldsymbol{E}\hat{L}_n(g)| \tag{29}$$

(see, e.g., Devroye [9]). Let $\mathcal{C}_n$ denote the collection of subsets of sets $\mathbb{R}^d \times \{-1, 1\}$ in the form

$$\{(x, y): g(x) \neq y\}; \qquad g \in \mathcal{G}_n. \tag{30}$$

Then the above supremum is just

$$\sup_{g \in \mathcal{G}_n} |\hat{L}_n(g) - \boldsymbol{E}\hat{L}_n(g)| = \sup_{C \in \mathcal{C}_n} \left| \frac{1}{n} \sum_{j=1}^{n} I_C(X_j, Y_j) - \boldsymbol{E} I_C(X, Y) \right|. \tag{31}$$

We estimate the right-hand side by the Vapnik–Chervonenkis inequality (Lemma 7 in Appendix A). In our case $\mathcal{C} = \mathcal{C}_n$ and $Z_j = (X_j, Y_j), j = 1, \cdots, n$. It is not hard to see that

$$S(n, \mathcal{C}_n) = S(n, \hat{\mathcal{C}}_n)$$

where $\hat{\mathcal{C}}_n$ is the collection of subsets of $\mathbb{R}^d$ of the form $\{x: g(x) = 1\}, g \in \mathcal{G}_n$. The classifier g has a feedforward architecture with $k_n + 1$ computational nodes all having binary outputs. Thus the shatter coefficient $S(n, \hat{\mathcal{C}}_n)$ has an upper bound (Baum and Haussler [3, Theorem 1])

$$S(n, \hat{\mathcal{C}}_n) \leq \left(\frac{(k_n + 1)en}{V_n} \right)^{V_n}$$

where $V_n = \sum_{i=1}^{k_n+1} V_i$, the sum of the VC dimensions of the classes of indicators at each node. In our case the first k_n nodes are equipped with the class $\mathcal{C}_1$ [defined in (28)], and have VC dimension $V_{\mathcal{C}_1}$. The $(k_n+1)^{th}$ node is associated via (27) with the class of k_n-dimensional linear threshold functions which has VC dimension $k_n + 1$ (see Cover [7] and Wenocur and Dudley [33]). Thus we have

$$V_n = k_n(V_{\mathcal{C}_1} + 1) + 1$$

and we can write

$$S(\mathcal{C}_n, 2n) \leq \left(\frac{4en}{V_{\mathcal{C}_1} + 1} \right)^{k_n(V_{\mathcal{C}_1}+2)}.$$

Combining (29), (30), and Lemma 7, we obtain the desired inequality. □

Now, with a strong upper bound on the estimation error in hand, it is easy to obtain conditions on k_n for strong universal consistency of RBF classifiers based on the minimization of the empirical error probability. The next theorem states that if $k_n \to \infty$ as $n \to \infty$ not too rapidly, then the sequence g_n is strongly universally consistent. Faragó and Lugosi [13] proved a similar result for sigmoidal neural networks trained by minimizing the empirical error probability.

Theorem 4: Let K be an indicator such that $V_{\mathcal{C}_1} < \infty$ for $\mathcal{C}_1$ defined in (28). Suppose that the set of RBF networks given by (26), k being arbitrary, is dense in $L_1(\mu)$ on balls $\{x \in \mathbb{R}^d: \|x\| \leq B\}$ for any probability measure μ on $\mathbb{R}^d$. If $k_n \to \infty$ and $n^{-1}(k_n \log n) \to 0$ as $n \to \infty$, then the sequence of classifiers g_n minimizing the empirical error probability is strongly universally consistent.

Remark: Note that the approximation result of Theorem 1 applies when K is an indicator with $\int K < \infty$, thus the denseness condition in Theorem 4 holds in this case.

Proof of Theorem 4: The consistency of g_n is proved by decomposing $L(g_n) - L^*$ in the usual way

$$L(g_n) - L^* = (L(g_n) - \inf_{g \in \mathcal{G}_n} L(g)) + (\inf_{g \in \mathcal{G}_n} L(g) - L^*).$$

For the first term on the right-hand side we invoke Theorem 3. Since $n^{-1}(k_n \log n) \to 0$ as $n \to \infty$, the right-hand side of the inequality of Theorem 3 is summable in n for any $\epsilon > 0$. Therefore the estimation error $L(g_n) - \inf_{g \in \mathcal{G}_n} L(g)$ converges to zero with probability one by the Borel–Cantelli theorem.

To bound the approximation error, recall (25). Clearly

$$\inf_{g \in \mathcal{G}_n} L(g) - L^* \leq \inf_{f \in \mathcal{F}_n} \int_{\mathbb{R}^d} \min\{|f(x) - m(x)|, 2\} \mu(dx) \tag{32}$$

where μ is the measure induced by X. Now the denseness condition and the fact that $k_n \to \infty$ as $n \to \infty$ imply that for any $B > 0$

$$\lim_{n \to \infty} \inf_{f \in \mathcal{F}_n} \int_{\|x\| \leq B} |f(x) - m(x)| \mu(dx) = 0. \tag{33}$$

Since $\lim_{B \to \infty} \boldsymbol{P}\{\|X\| > B\} = 0$, (32) and (33) imply that $\inf_{g \in \mathcal{G}_n} L(g) - L^* \to 0$ as $n \to \infty$. The proof of the theorem is complete. □

A natural question immediately arises: What other kernels provide the same behavior of the estimation error as Theorem 3? For example, it is easy to see that if K is a weighted sum of finitely many indicators, then a result similar to Theorem 3 still holds. The key question is whether the VC dimension of the class of sets of the form

$$\left\{ x: \sum_{i=1}^{k} w_i K(A_i[x - c_i]) + w_0 > 0 \right\}$$

is finite, and how it depends on k. If the VC dimension is infinite, then as Blumer *et al.* [4] point out, there is no distribution-free upper bound on the estimation error. In fact, if the VC dimension is infinite, then for every n, and for any training method, there exists a distribution such that the error probability of the resulting classifier is at least a universal constant away from the optimal error probability $\inf_{g \in \mathcal{G}_n} L(g)$ in the class. Bounding the VC dimension of such classes is a challenging problem. One would suspect that for "nice" unimodal kernels the situation should not be dramatically different from when K is an indicator of (say) a ball. It may come as a surprise that for some "nice," smooth kernels this is not the case. Our counterexample is based on the work of Macintyre and Sontag. We show that there exists a symmetric, unimodal, continuous one-dimensional kernel,

with the property $K(x) \leq K(y)$ if $|x| \geq |y|$, such that the VC *dimension* corresponding to the class $\mathcal{G}_n$ is infinite if $k_n \geq 2$. A finite set S is said to be shattered by the class of sets $\mathcal{C}$ if every $B \subset S$ is of the form $B = S \cap C$ for some $C \in \mathcal{C}$. Thus the VC dimension of $\mathcal{C}$ is infinite iff for any n there exists a set of n elements shattered by $\mathcal{C}$. The construction of our example relies on the following lemma (C^∞ denotes the space of functions having derivatives of arbitrary order).

Lemma 2 [22]: There exists a bounded, monotone increasing C^∞ function r: $\mathbb{R} \to \mathbb{R}$, such that the class of sets

$$\mathcal{A} = \{A_a = \{x \in \mathbb{R}: r(ax) + r(-ax) > 0\}: a > 0\}$$

has infinite VC dimension. Further, the the points $x_1, \cdots, x_n$ that are shattered by $\mathcal{A}$ can all be chosen in the interval (0, 1).

Based on the construction of $r(x)$ in the above result, we now show that there exists a "nice" kernel such that the VC dimension of the corresponding class of RBF classifiers $\mathcal{G}_n$ is infinite for every n. For nets using such a kernel, by a result of Blumer *et al.* [4], for every n and for any method of selecting the parameters of the net, there exists a distribution such that the error probability of the resulting classifier is larger than $\inf_{g \in G_n} L(g) + c$, where c is a universal constant. Interestingly, this does not mean that strong universal consistency is not achievable with RBF nets based on such kernels. For example, the kernel in the counterexample below satisfies the conditions of Theorem 2, therefore the classification rule based on the regression function estimate obtained by minimizing the empirical squared L_2 error remains strongly universally consistent. This makes the point that consistency is a very weak property when it comes to studying finite-sample behavior of classifiers. Theorem 3 shows that minimizing the empirical error probability in an RBF class based on a window kernel has a desirable finite-sample property, which many other algorithms and kernels fail to share.

Theorem 5: There exists a continuous kernel K: $\mathbb{R} \to \mathbb{R}$, which is unimodal, bounded, monotone increasing for $x < 0$, and decreasing for $x > 0$, such that the shatter coefficient of the class of sets

$$\mathcal{A}_K = \left\{ \left\{ x: K\left(\frac{h-x}{h}\right) + K\left(\frac{-h-x}{h}\right) > 0 \right\}: h > 0 \right\}$$

equals 2^n for all n. Thus, the VC dimension corresponding to $\mathcal{G}_n$ is infinite whenever $k_n \geq 2$.

Proof: The pathological function r of Lemma 2 is constructed as follows: Let $\alpha(x) = (1 - x^2)^{-1}$ and define

$$\beta(x) = \begin{cases} \int_0^x \alpha(t)\, dt & \text{if } x \geq 0 \\ -\int_x^0 \alpha(t)\, dt & \text{otherwise} \end{cases}$$

and

$$\hat{r}(x) = 3\beta(x) + \alpha(x)\cos(x).$$

Then $\hat{r}$ is bounded, infinitely differentiable, and monotone increasing. Furthermore

$$\hat{r}(ax) + \hat{r}(-ax) = 2\alpha(ax)\cos(ax).$$

Thus $\hat{r}(ax) + \hat{r}(-ax) > 0$ iff $\cos(ax) > 0$. It is not hard to see that for any n there exist $x_1, \cdots, x_n > 1$ shattered by sets of the form $\{x: \cos(ax) > 0\}$. Now r is defined as

$$r(x) = 3\beta(x) + \alpha(x)e^{-1/x^2}\cos(1/x).$$

Clearly, r satisfies the same conditions as $\hat{r}$, and

$$r(ax) + r(-ax) = 2\alpha(ax)e^{-1/(ax)^2}\cos(1/(ax))$$

which is positive iff $\cos(1/(ax)) > 0$. Thus $r(x) + r(-x) > 0$ iff $\hat{r}(1/x) + \hat{r}(-1/x) > 0$. But all the shattered points for $\hat{r}$ were greater than one, therefore their reciprocals shattered by the sets $\{x: \cos(1/(ax)) > 0\}$ are in (0, 1). To use the above construction of Macintyre and Sontag, we make the simple observation that the shattered x_i can be chosen so that

$$\cos(a_j x_i) > 0 \text{ if } x_i \in S_j \tag{34}$$

and

$$\cos(a_j x_i) < 0 \text{ if } x_i \in \{x_1, \cdots, x_n\} - S_j \tag{35}$$

for some $a_1, \cdots, a_N > 0, N = 2^n$, where $S_1, \cdots, S_N$ are all the subsets of $\{x_1, \cdots, x_n\}$. Define the kernel K as

$$K(x) = r(-|x| + 1).$$

Then $K(x) = g(|x|)$ with $g(t) = r(-t - 1), t \geq 0$, and K is monotone decreasing, bounded and continuous. In fact, g (and thus K) is infinitely differentiable everywhere except at zero. Now for $|x| < h$

$$K\left(\frac{h-x}{h}\right) + K\left(\frac{-h-x}{h}\right) = r\left(\frac{x}{h}\right) + r\left(\frac{-x}{h}\right). \tag{36}$$

Thus the theorem will be proved if we can show that there exist $\{x_1, \cdots, x_n\}$ shattered by the sets $\{x: r(x/h_j) + r(-x/h_j) > 0\}, j = 1, \cdots, N, N = 2^n$, such that

$$\max_{1 \leq i \leq n} x_i \leq \min_{1 \leq j \leq N} h_j. \tag{37}$$

Let the x_i's and a_j's be as in (34) and (35). The value of $\cos(a_j x_i)$ is determined by b_{ij} in the unique representation of x_i

$$x_i = k_{ij}\frac{2\pi}{a_j} + b_{ij}, \quad k_{i,j} \geq 0 \text{ integer}, \quad 0 \leq b_{ij} < \frac{2\pi}{a_j}.$$

By the continuity of $\cos(x)$ we can choose the a_j's satisfying (34) and (35) to be rational: $a_j = p_j/q_j$, for some integers $p_j, q_j, j = 1, \cdots, N$. Define $Z = 2\pi \prod_{j=1}^N q_j$. Then $M_j = a_j Z/(2\pi)$ is an integer for all j, and for any positive integer L we have

$$x_i + LZ = (k_{ij} + LM_j)\frac{2\pi}{a_j} + b_{ij}$$

which implies $\cos(a_j(x_i + LZ)) = \cos(a_j x_i)$ for all i, j. *This means that if L is chosen large enough, the points* $y_i = x_i + LZ$ satisfy $\max_i(1/y_i) < \min_j a_j$, and therefore $\{x: K((a_j - x)/a_j) + K((-a_j - x)/a_j) > 0\}, j = 1, \cdots, N$ shatter $\{1/y_1, \cdots, 1/y_n\}$ by (36) and (37). □

V. Conclusions

In this paper, we have established convergence properties of general RBF nets. Upon imposing mild conditions on the basis functions they have been seen to approximate arbitrary functions in $L_p(\mu)$ norm, $p>0$. We have proved the generalization ability of RBF nets in nonlinear function estimation for a large class of basis functions with parameters learned through empirical risk minimization. We have considered two approaches to classification based on RBF networks. In one approach, we base the classification on nonlinear function estimation and show the resulting network to be consistent. We obtain better results in training the network by minimizing the empirical error probability directly, provided that the RBF net uses a window kernel. We give a counterexample of RBF nets in which these better properties cannot be achieved by any method of training the network. It remains an open problem to characterize kernels which will share these convergence properties with the window kernels.

Appendix A
Review of Some Results Concerning VC Dimension and Covering Numbers

In what follows, we list some important definitions and results that are crucial in dealing with empirical risk minimization. First, we give the definition of covering numbers for classes of functions.

Definition 1: Let $\mathcal{F}$ be a class of real functions on $\mathbb{R}^m$. The covering number $N(\epsilon, \mathcal{F}, z_1^n)$ is defined for any $\epsilon \geq 0$ and $z_1^n = (z_1, \cdots, z_n) \in \mathbb{R}^{nm}$ as the smallest integer l such that there exist functions $g_1, \cdots, g_l$: $\mathbb{R}^m \to \mathbb{R}$ satisfying for each $f \in \mathcal{F}$

$$\min_{1 \leq i \leq l} \frac{1}{n} \sum_{j=1}^{n} |f(z_j) - g_i(z_j)| \leq \epsilon.$$

When $Z_1^n = (Z_1, \cdots, Z_n)$ is an n-vector of R^m-valued random variables, the covering number $N(\epsilon, \mathcal{F}, Z_1^n)$ is a random variable with expectation $\boldsymbol{E}N(\epsilon, \mathcal{F}, z_1^n)$. The next result, by Pollard, is our main tool in Section III.

Lemma 3 [27]: Let $\mathcal{F}$ be a class of real functions on $\mathbb{R}^m$ with $|f(z)| \leq B$ for all $f \in \mathcal{F}, z \in \mathbb{R}^m$, and let $Z_1^n = (Z_1, \cdots, Z_n)$ be $\mathbb{R}^m$ valued i.i.d. random variables. Then for any $\epsilon > 0$

$$\boldsymbol{P}\left\{\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - \boldsymbol{E}f(Z_1) \right| > \epsilon \right\} \leq 8\boldsymbol{E}(N(\epsilon/8, \mathcal{F}, Z_1^n))e^{-n\epsilon^2/128B^2}.$$

The following lemma is often useful when $\mathcal{F}$ is built up from relatively simpler classes.

Lemma 4 [28]: Let $\mathcal{F}$ and $\mathcal{G}$ be two families of real functions on $\mathbb{R}^m$ with $|f(z)| \leq B_1$ and $|g(z)| \leq B_2$ for all $z \in \mathbb{R}^m, f \in \mathcal{F}$ and $g \in \mathcal{G}$.

1) If $\mathcal{F} \oplus \mathcal{G}$ denotes the set of functions $\{f+g: f \in \mathcal{F}, g \in \mathcal{G}\}$, then for any $z_1^n \in \mathbb{R}^{nm}$ and $\epsilon, \delta > 0$ we have

$$N(\epsilon + \delta, \mathcal{F} \oplus \mathcal{G}, z_1^n) \leq N(\epsilon, \mathcal{F}, z_1^n) N(\delta, \mathcal{G}, z_1^n).$$

2) If $\mathcal{F} \odot \mathcal{G}$ denotes the set of functions $\{f \cdot g: f \in \mathcal{F}, g \in \mathcal{G}\}$, then for any $z_1^n \in \mathbb{R}^{nm}$ and $\epsilon > 0$ we have

$$N(\epsilon, \mathcal{F} \odot \mathcal{G}, z_1^n) \leq N(\epsilon/(2B_2), \mathcal{F}, z_1^n) N(\epsilon/(2B_1), \mathcal{G}, z_1^n).$$

The notions of shatter coefficient and VC dimension are used in this paper for bounding covering numbers in Section III, as well as directly in Section IV through the celebrated Vapnik–Chervonenkis inequality (Lemma 7).

Definition 2: Let $\mathcal{C}$ be a collection of subsets of $\mathbb{R}^m$. The nth shatter coefficient $S(n, \mathcal{C})$ of $\mathcal{C}$ is defined as the maximum number of distinct subsets $\mathcal{C}$ can pick from a finite set of n elements

$$S(n, \mathcal{C}) = \max_{\substack{S \subset \mathbb{R}^m \\ |S| = n}} |\{S \cap C: C \in \mathcal{C}\}|.$$

The VC dimension of $\mathcal{C}$ (denoted by $V_{\mathcal{C}}$) is the largest n satisfying $S(n, \mathcal{C}) = 2^n$. By definition $V_{\mathcal{C}} = \infty$ if $S(n, \mathcal{C}) = 2^n$ for all n.

If $\mathcal{C} = \{\{z: g(z) = 1\}; g \in \mathcal{G}\}$, for a class $\mathcal{G}$ of indicators, then $V_{\mathcal{G}} = V_{\mathcal{C}}$ by definition.

A connection between covering numbers and VC dimension is given by the following:

Lemma 5 [17]: Let $\mathcal{F}$ be a collection of real functions on $\mathbb{R}^m$ with $|f(z)| \leq B$ for all $z \in \mathbb{R}^m$ and $f \in \mathcal{F}$. Suppose that the family of sets

$$\{(z,t) \in \mathbb{R}^{m+1}: f(z) \geq t\}, \qquad f \in \mathcal{F}$$

has finite VC dimension $V_{\mathcal{F}}$. Then for any $z_1^n \in \mathbb{R}^{nm}$ and $\epsilon > 0$ we have

$$N(\epsilon, \mathcal{F}, z_1^n) \leq 2\left(\frac{2eB}{\epsilon}\right)^{2V_{\mathcal{F}}}.$$

The VC dimension can often be upper-bounded by means of the following very useful result.

Lemma 6 [27]: Let $\mathcal{F}$ be a d-dimensional vector space of real functions on $\mathbb{R}^m$. Then the class of sets of the form $\{x: f(x) \geq 0\}, f \in \mathcal{F}$, has VC dimension no greater than $d+1$.

The following is the fundamental VC inequality.

Lemma 7 [31]: Let $\mathcal{C}$ be a class of sets in $\mathbb{R}^m$, and let $Z_1, \cdots, Z_n$ be a sequence of $\mathbb{R}^m$-valued i.i.d. random variables. Then for any ϵ and n

$$\boldsymbol{P}\left\{\sup_{C \in \mathcal{C}} \left| \frac{1}{n} \sum_{j=1}^{n} I_C(Z_j) - \boldsymbol{P}\{Z_1 \in C\} \right| > \epsilon \right\} \leq 4s(2n, \mathcal{C})e^{-n\epsilon^2/8}$$

where $s(n, \mathcal{C})$ is the nth shatter coefficient of $\mathcal{C}$.

Appendix B
Proof of Lemma 1

The proof is given for L_p norms loss functions ($1 \leq p < \infty$), thus $J(f) = (\boldsymbol{E}|f(X) - Y|^p)^{1/p}$. Let $L > 0$ be an arbitrary fixed number and introduce the following truncated random variables:

$$Y_L = \begin{cases} Y & \text{if } |Y| \leq L \\ L \operatorname{sgn}(Y) & \text{otherwise} \end{cases}$$

and

$$Y_{j,L} = \begin{cases} Y_j & \text{if } |Y_j| \le L \\ L \operatorname{sgn}(Y_j) & \text{otherwise} \end{cases}$$

for $j = 1, \cdots, n$, where $\operatorname{sgn}(x) = 2I_{\{x \ge 0\}} - 1$. Further, let $\hat{m}_n$ be a function in $\mathcal{F}_n$ that minimizes the empirical error based on the truncated variables

$$\left(\frac{1}{n}\sum_{j=1}^{n}|\hat{m}_n(X_j) - Y_{j,L}|^p\right)^{1/p} \le \left(\frac{1}{n}\sum_{j=1}^{n}|f(X_j) - Y_{j,L}|^p\right)^{1/p} \quad \text{for every } f \in \mathcal{F}_n.$$

Also, denote by f^* a function that minimizes $J(f)$ over $\mathcal{F}_n$.

Observe that the triangle inequality implies

$$\begin{aligned} J(m_n) &= (\boldsymbol{E}(|m_n(X) - Y|^p|D_n))^{1/p} \\ &\le (\boldsymbol{E}(|m_n(X) - Y_L|^p|D_n))^{1/p} + (\boldsymbol{E}|Y_L - Y|^p)^{1/p} \end{aligned}$$

and similarly

$$\begin{aligned} &\inf_{f\in\mathcal{F}_n}(\boldsymbol{E}|Y_L - f(X)|^p)^{1/p} \\ &\quad \le (\boldsymbol{E}|Y_L - f^*(X)|^p)^{1/p} \\ &\quad \le (\boldsymbol{E}|Y - f^*(X)|^p)^{1/p} + (\boldsymbol{E}|Y_L - Y|^p)^{1/p} \\ &\quad = \inf_{f\in\mathcal{F}_n}(\boldsymbol{E}|Y - f(X)|^p)^{1/p} + (\boldsymbol{E}|Y_L - Y|^p)^{1/p}. \end{aligned}$$

Combining the two inequalities above, we obtain

$$\begin{aligned} &J(m_n) - \inf_{f\in\mathcal{F}_n} J(f) \\ &\quad = (\boldsymbol{E}(|m_n(X) - Y|^p|D_n))^{1/p} \\ &\qquad - \inf_{f\in\mathcal{F}_n}(\boldsymbol{E}|Y - f(X)|^p)^{1/p} \\ &\quad \le (\boldsymbol{E}(|m_n(X) - Y_L|^p|D_n))^{1/p} \\ &\qquad - \inf_{f\in\mathcal{F}_n}(\boldsymbol{E}|Y_L - f(X)|^p)^{1/p} + 2(\boldsymbol{E}|Y_L - Y|^p)^{1/p}. \end{aligned} \tag{38}$$

Now, we bound the difference on the right-hand side of the inequality

$$\begin{aligned} &(\boldsymbol{E}(|m_n(X) - Y_L|^p|D_n))^{1/p} - \inf_{f\in\mathcal{F}_n}(\boldsymbol{E}|Y_L - f(X)|^p)^{1/p} \\ &\quad = (\boldsymbol{E}(|m_n(X) - Y_L|^p|D_n))^{1/p} \\ &\qquad - \left(\frac{1}{n}\sum_{j=1}^{n}|m_n(X_j) - Y_{j,L}|^p\right)^{1/p} \\ &\qquad + \left(\frac{1}{n}\sum_{j=1}^{n}|m_n(X_j) - Y_{j,L}|^p\right)^{1/p} \\ &\qquad - \inf_{f\in\mathcal{F}_n}(\boldsymbol{E}|Y_L - f(X)|^p)^{1/p} \\ &\quad \le \sup_{f\in\mathcal{F}_n}\left|(\boldsymbol{E}|f(X) - Y_L|^p)^{1/p}\right. \\ &\qquad \left. - \left(\frac{1}{n}\sum_{j=1}^{n}|f(X_j) - Y_{j,L}|^p\right)^{1/p}\right| \\ &\qquad + \left(\frac{1}{n}\sum_{j=1}^{n}|Y_j - Y_{j,L}|^p\right)^{1/p} \\ &\qquad + \left(\frac{1}{n}\sum_{j=1}^{n}|m_n(X_j) - Y_j|^p\right)^{1/p} \\ &\qquad - \inf_{f\in\mathcal{F}_n}(\boldsymbol{E}|Y_L - f(X)|^p)^{1/p} \end{aligned} \tag{39}$$

$$\begin{aligned} &\le \sup_{f\in\mathcal{F}_n}\left|(\boldsymbol{E}|f(X) - Y_L|^p)^{1/p}\right. \\ &\qquad \left. - \left(\frac{1}{n}\sum_{j=1}^{n}|f(X_j) - Y_{j,L}|^p\right)^{1/p}\right| \\ &\qquad + \left(\frac{1}{n}\sum_{j=1}^{n}|Y_j - Y_{j,L}|^p\right)^{1/p} \\ &\qquad + \left(\frac{1}{n}\sum_{j=1}^{n}|\hat{m}_n(X_j) - Y_j|^p\right)^{1/p} \\ &\qquad - \inf_{f\in\mathcal{F}_n}(\boldsymbol{E}|Y_L - f(X)|^p)^{1/p} \end{aligned} \tag{40}$$

$$\begin{aligned} &\le \sup_{f\in\mathcal{F}_n}\left|(\boldsymbol{E}|f(X) - Y_L|^p)^{1/p}\right. \\ &\qquad \left. - \left(\frac{1}{n}\sum_{j=1}^{n}|f(X_j) - Y_{j,L}|^p\right)^{1/p}\right| \\ &\qquad + 2\left(\frac{1}{n}\sum_{j=1}^{n}|Y_j - Y_{j,L}|^p\right)^{1/p} \\ &\qquad + \left(\frac{1}{n}\sum_{j=1}^{n}|\hat{m}_n(X_j) - Y_{j,L}|^p\right)^{1/p} \\ &\qquad - \inf_{f\in\mathcal{F}_n}(\boldsymbol{E}|Y_L - f(X)|^p)^{1/p} \end{aligned} \tag{41}$$

$$\begin{aligned} &\le 2\sup_{f\in\mathcal{F}_n}\left|(\boldsymbol{E}|f(X) - Y_L|^p)^{1/p}\right. \\ &\qquad \left. - \left(\frac{1}{n}\sum_{j=1}^{n}|f(X_j) - Y_{j,L}|^p\right)^{1/p}\right| \\ &\qquad + 2\left(\frac{1}{n}\sum_{j=1}^{n}|Y_j - Y_{j,L}|^p\right)^{1/p} \end{aligned} \tag{42}$$

where (39) and (41) follow from the triangle inequality,

while (40) exploits the defining optimality property of m_n. Combining this with (38) and using the strong law of large numbers, we get

$$\limsup_{n\to\infty}(J(m_n) - \inf_{f\in\mathcal{F}_n} J(f))$$
$$\leq 2\cdot\limsup_{n\to\infty}\left(\sup_{f\in\mathcal{F}_n}\left|(\boldsymbol{E}|f(X)-Y|^p)^{1/p} - \left(\frac{1}{n}\sum_{j=1}^{n}|f(X_j)-Y_j|^p\right)^{1/p}\right|\right)$$
$$+ 4(\boldsymbol{E}|Y_L - Y|^p)^{1/p} \text{ a.s.}$$

The first term of the right-hand side is zero almost surely by the conditions of the theorem, while the second term can be made arbitrarily small by appropriate choice of L. □

REFERENCES

[1] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," in *Proc. 4th Annu. Wkshp. Computa. Learning Theory,* vol. 4, 1991, pp. 243–249.

[2] ——, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory,* vol. 39, pp. 930–945, May 1993.

[3] E. B. Baum and D. Haussler, "What size net gives valid generalization?" *Neural Computa.,* vol. 1, pp. 151–160, 1989.

[4] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the Vapnik–Chervonenkis dimension," *J. ACM,* vol. 36, pp. 929–965, 1989.

[5] D. S. Broomhead and D. Lowe, "Multivariable functional interpolation and adaptive networks," *Complex Syst.,* vol. 2, pp. 321–323, 1988.

[6] S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," *IEEE Trans. Neural Networks,* vol. 2, pp. 302–309, 1991.

[7] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. Electron. Comput.,* vol. 14, pp. 326–334, 1965.

[8] C. Darken, M. Donahue, L. Gurvits, and E. Sontag, "Rate of approximation results motivated by robust neural network learning," in *Proc. 6th Annu. Wkshp. Computa. Learning Theory,* 1993, pp. 303–309.

[9] L. Devroye, "Automatic pattern recognition: A study of the probability of error," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. PAMI-10, pp. 530–543, 1988.

[10] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L1 View.* New York: Wiley, 1985.

[11] L. Devroye and A. Krzyżak, "An equivalence theorem for L_1 convergence of the kernel regression estimate," *J. Statist. Planning Inference,* vol. 23, pp. 71–82, 1989.

[12] N. Dunford and J. T. Schwartz, *Linear Operators, Part I. General Theory.* New York: Interscience, 1957.

[13] A. Faragó and G. Lugosi, "Strong universal consistency of neural network classifiers," *IEEE Trans. Inform. Theory,* vol. 39, pp. 1146–1151, July 1993.

[14] F. Girosi and G. Anze llotti, "Rates of convergence for radial basis functions and neural networks," in *Artificial Neural Networks for Speech and Vision,* R. J. Mammone, Ed. London: Chapman and Hall, 1993, pp. 97–114.

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

[16] E. J. Hartman, J. D. Keeler, and J. M. Kowalski, "Layered neural networks with Gaussian hidden units as universal approximations," *Neural Computa.,* vol. 2, pp. 210–215, 1990.

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

[18] K. Hornik, S. Stinchocombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks,* vol. 2, pp. 359–366, 1989.

[19] L. K. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training," *Annals Statist.,* vol. 20, pp. 608–613, Mar. 1992.

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

[21] G. Lugosi and K. Zieger, "Nonparametric estimation via empirical risk minimization," *IEEE Trans. Inform. Theory,* vol. 41, ppl 677–687, May 1995.

[22] A. Macintyre and E. D. Sontag, "Finiteneness results for sigmoidal 'neural' networks," in *Proc. 25th Annu. ACM Symp. Theory Computing,* 1993, pp. 325–334.

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

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

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

[26] T. Poggio and F. Girosi, "A theory of networks for approximation and learning," *Proc. IEEE,* vol. 78, pp. 1481–1497, 1990.

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

[28] ——, *Empirical Processes: Theory and Applications,* NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, 1990.

[29] M. J. D. Powell, "Radial basis functions for multivariable interpolation: A review," in *Algorithms for Approximation.* Oxford, U.K.: Clarendon, 1987, pp. 143–167.

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

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

[32] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data.* New York: Springer-Verlag, 1982.

[33] R. S. Wenocur and R. M. Dudley, "Some special Vapnik–Chervonenkis classes," *Discrete Math.,* vol. 33, pp. 313–318, 1981.

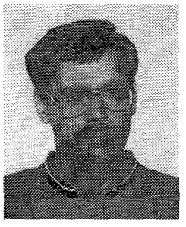
[34] R. L. Wheeden and A. Zygmund, *Measure and Integral.* New York: Marcel Dekker, 1977.

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

[36] L. Xu, S. Klasa, and A. L. Yuille, "Recent advances on techniques static feedforward networks with supervised learning," *Int. J. Neural Syst.,* vol. 3, pp. 253–290, 1992.

[37] L. Xu, A. Krzyżak, and E. Oja, "Rival penalized competitive learning for clustering analysis, RBF net, and curve detection," *IEEE Trans. Neural Networks,* vol. 4, pp. 636–649, July 1993.

[38] 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 Networks,* vol. 7, pp. 609–628, 1994.



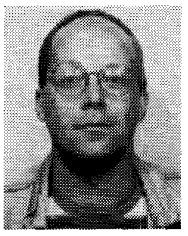
Adam Krzyżak (M'86) was born in Szczecin, Poland, on May 1, 1953. He received the M.Sc. and Ph.D. degrees in computer engineering from the Technical University of Wrocław, Poland, in 1977 and 1980, respectively.

In 1980 he became an Assistant Professor in the Institute of Engineering Cybernetics, Technical University of Wrocław, Poland. From November 1982 to July 1983 he was a Postdoctorate Fellow receiving the International Scientific Exchange Award in the School of Computer Science, McGill University, Montreal, PQ, Canada. Since August 1983, he has been with the Department of Computer Science, Concordia University, Montreal, where he is currently an Associate Professor. He has published about 100 papers in the areas of pattern recognition, image processing, computer vision, and identification and estimation of control systems, as well as on various applications of probability theory and statistics.

Dr. Krzyżak won the Vineberg Memorial Fellowship at Technion-Israel Institute of Technology in 1991, and in 1992 he won the Humboldt Research Fellowship in Germany. He is an Associate Editor of the *Pattern Recognition Journal, The International Journal of Applied Software Technology,* and Coeditor of the book, *Computer Vision and Pattern Recognition* (Singapore: World Scientific, 1989). He has served on the program committees of Vision Interface'88, Vision Interface'94, Vision Interface'95, and the 1995 International Conference on Document Processing and Applications.

Tamás Linder was born in Budapest, Hungary, in 1964. He received the M.S. degree in electrical engineering from the Technical University of Budapest in 1988, and the Ph.D. degree from the Hungarian Academy of Sciences in electrical engineering in 1992.

He was a Post-Doctoral Fellow at the University of Hawaii, Honolulu, in 1993, and a Visiting Fulbright Scholar at the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign in 1993–1994. He is currently an Associate Professor at the Department of Mathematics and Computer Science, Technical University of Budapest. His research interests include information theory, source coding and vector quantization, rate-distortion theory, machine learning, and statistical pattern recognition.



Gábor Lugosi was born on July 13, 1964, in Budapest, Hungary. He received the bachelor's degree in electrical engineering from the Technical University of Budapest in 1987, and the Ph.D. degree from the Hugarian Academy of Sciences in 1991.

He has been with the Department of Mathematics and Computer Science, Faculty of Electrical Engineering, at the Technical University of Budapest. His main research interests include pattern recognition, information theory, and nonparametric statistics.