Radial Basis Function Networks and Complexity Regularization in Function Learning

Adam Krzyżak, *Senior Member, IEEE,* and Tamás Linder, *Member, IEEE*

***Abstract*— In this paper we apply the method of complexity regularization to derive estimation bounds for nonlinear function estimation using a single hidden layer radial basis function network. Our approach differs from previous complexity regularization neural-network function learning schemes in that we operate with random covering numbers and l_1 metric entropy, making it possible to consider much broader families of activation functions, namely functions of bounded variation. Some constraints previously imposed on the network parameters are also eliminated this way. The network is trained by means of complexity regularization involving empirical risk minimization. Bounds on the expected risk in terms of the sample size are obtained for a large class of loss functions. Rates of convergence to the optimal loss are also derived.**

***Index Terms*— Complexity regularization, convergence rates, function estimation, radial basis functions, random covering numbers.**

I. Introduction

Artificial neural networks have been found effective in learning input–output mappings from noisy examples. In this learning problem an unknown target function is to be inferred from a set of independent observations drawn according to some unknown probability distribution from the input–output space $R^d \times R$. Using this data set the learner tries to determine a function which fits the data in the sense of minimizing some given empirical loss function. The target function may or may not be in the class of functions which are realizable by the learner. In the case when the class of realizable functions consists of some class of artificial neural networks, the above problem has been extensively studied from different viewpoints.

Approximation results (see, e.g., Cybenko [1], Hornik *et al.* [2], Barron [3], and Chen *et al.* [4]) show that virtually any real function of interest in R^d can be appropriately approximated by one-hidden-layer sigmoidal networks. Bounds on the approximation error as a function of the networks size and incremental approximation schemes have been developed by, e.g., Jones [5], Barron [3], and Girosi and Anzellotti [6]. The generalization ability of networks from a finite training set has also been extensively studied through bounding the estimation error by, e.g., White [7], Barron [8], Haussler [9], and Faragó and Lugosi [10]. Barron [11] combined approximation and estimation bounds and obtained the convergence rates for sigmoidal neural networks in function estimation. His results were extended and sharpened by McCaffrey and Gallant [12] and Lee *et al.* [13].

In recent years a special class of artificial neural networks, the radial basis function (RBF) networks have received considerable attention. RBF networks have been shown to be the solution of the regularization problem in function estimation with certain standard smoothness functionals used as stabilizers (see Girosi [14], Girosi *et al.* [15], and the references therein). Universal convergence of RBF nets in function estimation and classification has been proven by Krzyżak *et al.* [16]. Approximation error convergence rates for RBF networks have been studied by Girosi and Anzellotti [6]. In a recent paper Niyogi and Girosi [17] studied the tradeoff between approximation and estimation errors and provided an extensive review of the problem.

In this paper we consider one-hidden-layer RBF networks. We look at the problem of choosing the size of the hidden layer as a function of the available training data by means of complexity regularization. Complexity regularization approach has been applied to model selection by Barron [8], [11] resulting in near optimal choice of sigmoidal network parameters. Our approach here differs from Barron's in that we are using l_1 metric entropy instead of the supremum norm. This allows us to consider a more general class of activation functions, namely the functions of bounded variation, rather than a restricted class of activation functions satisfying a Lipschitz condition. In our complexity regularization approach we are able to choose the network parameters more freely, and no discretization of these parameters is required. For RBF regression estimation with squared error loss, we considerably improve the convergence rate result obtained by Niyogi and Girosi [17].

In Section II the problem is formulated. In Section III two results on the estimation error of complexity regularized RBF nets are presented: one for general loss functions (Theorem 1) and a sharpened version of the first one for the squared loss (Theorem 2). The proofs are given in Section IV. Approximation bounds are combined with the obtained estimation results in Section V yielding convergence rates for function learning with RBF nets.

Manuscript received April 3, 1996; revised November 1, 1997. This work was supported in part by NSERC Grant OGP000270 and Canadian National Networks of Centers of Excellence Grant 293. A. Krzyżak was supported by the Alexander von Humboldt Foundation. T. Linder was supported by the Foundation for Hungarian Higher Education and Research and OTKA Grant F014 174. This paper was presented in part at the 13th International Conference on Pattern Recognition, Vienna, Austria, 1996, and at NIPS'96, Denver, CO, 1996.

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

T. Linder is with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla CA 92093-0407 USA. He is also with the Technical University of Budapest, Hungary.

Publisher Item Identifier S 1045-9227(98)01169-2.

II. Problem Formulation

The task is to predict the value of a real random variable Y upon the observation of an R^d valued random vector X. The accuracy of the predictor $f : R^d \to R$ is measured by the expected risk

$$J(f) = \mathbf{E}L(f(X), Y)$$

where $L : R \times R \to R^+$ is a nonnegative loss function. It will be assumed that there exists a minimizing (measurable) predictor f^* such that

$$J(f^*) = \inf_f J(f).$$

When the probability law governing (X, Y) is known, the optimal predictor f^* can be determined in principle. In the learning model, however, the distribution is only known to be a member of a larger class of distributions. A good predictor f_n is to be determined based on the data $(X_1, Y_1), \cdots, (X_n, Y_n)$ which are independent and identically distributed (i.i.d.) copies of (X, Y). The goal is to make the expected risk $\mathbf{E}J(f_n)$ as small as possible, while f_n is chosen from among a given class $\mathcal{F}$ of candidate functions.

In this paper the set of candidate functions $\mathcal{F}$ will be single-layer feedforward neural networks with RBF activation units. Some of the results, however, will be valid in a more general setting, so that at this point we only specify that $\mathcal{F} = \bigcup_{k=1}^{\infty} \mathcal{F}_k$, where $\mathcal{F}_1, \mathcal{F}_2, \cdots$ is a a sequence of families of candidate functions, typically of increasing complexity. For neural networks, the kth family will be networks with k hidden nodes whose weight parameters satisfy certain constraints. In particular, for RBF's $\mathcal{F}_k$ is the family of networks

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

where $w_0, w_1, \cdots, w_k$ are real numbers called weights, $c_1, \cdots, c_k \in R^d, A_i$ are nonnegative definite $d \times d$ matrices, and x^t denotes the transpose of the column vector x.

The method of empirical minimization is a theoretically attractive tool for choosing the predictor from the training data. It selects an $f \in \mathcal{F}$ which minimizes the empirical risk

$$J_n(f) = \frac{1}{n} \sum_{i=1}^{n} L(f(X_i), Y_i).$$

The well-known problem of overfitting, however, makes it impossible to directly apply empirical minimization in many cases. If $\mathcal{F}$ is rich enough to contain good predictors for a reasonably large class of distributions, the output of empirical minimization will (almost) perfectly fit the data, but it is also bound to have an expected risk much larger than that of the optimal predictor in the class. The method of sieves [18] applied to this problem offers the following remedy: for each data set size n the empirical minimization is carried out over $\mathcal{F}_{k(n)}$, where $k(n)$ is a predetermined function of n. By the appropriate choice of $k(n)$ (which depends on the loss function, the type of network considered, and the family of probability distributions) one can obtain predictors whose expected risk converges to the optimum, i.e.,

$$\mathbf{E}J(f_n) \to J(f^*) \quad \text{as } n \to \infty$$

(see, e.g., [11], [19], [12], and [16]). It is clear that the choice of $k(n)$ (e.g., the number of hidden units for neural-network learning) is determined by the need of balancing between two quantities, the estimation error

$$\mathbf{E}J(f_n) - \inf_{f \in \mathcal{F}_k} J(f)$$

and the approximation error

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

The complexity regularization principle for the learning problem was introduced by Vapnik [20] and fully developed by Barron [8], [11] (see also Lugosi and Zeger [19] and Devroye *et al.* [21]). It enables the learning algorithm to choose $\mathcal{F}_k$ automatically. Complexity regularization penalizes the large candidate classes, which are bound to have small approximation error, in favor of the smaller ones. One form of this method, the minimum description length principle [22] uses as the penalty the length of a binary code describing the class. In a recent work Lugosi and Nobel [23] investigate a novel complexity regularization approach, in which the penalty term is data-dependent.

We develop below estimation bounds on the expected risk of complexity regularized neural networks in a framework which extends previous work. The need for such bounds stems from the fact that in [11] the class of activations was restricted to continuous sigmoids satisfying a Lipschitz condition. This restriction excludes activation units with jump discontinuities (e.g., perceptrons). The complexity penalties proposed in this paper make possible to obtain the same good bounds for more general activations. Though the results are mostly specialized to RBF networks, similar statements can be obtained for sigmoidal networks, or other nonlinear estimation schemes.

III. Estimation Bounds Through Complexity Regularization

Let $\mathcal{F}$ be a subset of a space $\mathcal{X}$ of real functions over some set, and let ρ be a pseudometric on $\mathcal{X}$. For $\epsilon > 0$ the *covering number* $N(\epsilon, \mathcal{F}, \rho)$ is defined to be the minimal number of closed ϵ balls whose union cover $\mathcal{F}$. In other words, $N(\epsilon, \mathcal{F}, \rho)$ is the least integer such that there exist $f_1, \cdots, f_N$ with $N = N(\epsilon, \mathcal{F}, \rho)$ satisfying

$$\sup_{f \in \mathcal{F}} \min_{1 \le i \le N} \rho(f, f_i) \le \epsilon.$$

We will mainly be concerned with the case when $\mathcal{F}$ is a family of real functions on R^m, and ρ is given by

$$\rho(f, g) = \frac{1}{n} \sum_{i=1}^{n} |f(z_i) - g(z_i)|$$

for any two functions f and g, where $z_1, \cdots, z_n$ are n given points in R^m. In this case we will use the notation

$N(\epsilon, \mathcal{F}, \rho) = N(\epsilon, \mathcal{F}, z_1^n)$, emphasizing the dependence of the metric ρ on $z_1^n = (z_1, \cdots, z_n)$.

Let us consider the task of predicting the value of a real random variable Y using a function of the R^d valued random vector X. The accuracy $J(f)$ of the prediction is measured by the expected risk

$$J(f) = \mathbf{E}L(f(X), Y)$$

where L is a nonnegative loss function of two real arguments. We assume that there exists a measurable f^* such that $\mathbf{E}L(f^*(X), Y)$ is minimal over all measurable f. The distribution of (X, Y) is assumed to be unknown, but we are given the i.i.d. copies $(X_1, Y_1), \cdots, (X_n, Y_n)$ of (X, Y). Based on this data, we are to pick an f from one of the families of candidate functions $\mathcal{F}_1, \mathcal{F}_2, \cdots$. Let us define the families of functions $\mathcal{H}_k, k = 1, 2, \cdots$ by

$$\mathcal{H}_k = \{L(f(\cdot), \cdot) : f \in \mathcal{F}_k\}.$$

Thus each member of $\mathcal{H}_k$ maps R^{d+1} into R. It will be assumed that for each k we are given a finite, almost sure uniform upper bound on the random covering numbers $N(\epsilon, \mathcal{H}_k, Z_1^n)$, where $Z_1^n = ((X_1, Y_1), \cdots, (X_n, Y_n))$. Denoting this upper bound by $N(\epsilon, \mathcal{H}_k)$, we thus have

$$N\big(\epsilon, \mathcal{H}_k, Z_1^n\big) \le N(\epsilon, \mathcal{H}_k) \quad \text{a.s.} \tag{1}$$

Note that we have suppressed in the notation the possible dependence of this bound on the distribution of (X, Y). Also, we may assume without loss of generality that $N(\epsilon, \mathcal{H}_k)$ is monotone decreasing in ϵ. Finally, assume that $L(f(X), Y)$ is uniformly almost surely bounded by a constant B, i.e.,

$$\mathbf{P}\{L(f(X), Y) \le B\} = 1, \quad f \in \mathcal{F}_k, \quad k = 1, 2, \cdots. \tag{2}$$

We define the complexity penalty of the kth class for n training samples as any nonnegative number Δ_{kn} satisfying

$$\Delta_{kn} \ge \sqrt{128B^2 \frac{\log N(\Delta_{kn}/8, \mathcal{H}_k) + c_k}{n}} \tag{3}$$

where the nonnegative constants c_k satisfy $\sum_{k=1}^{\infty} e^{-c_k} \le 1$. The reason behind defining Δ_{kn} this way will become clear later in the proof of Theorem 1. Note that since $N(\epsilon, \mathcal{H}_k)$ is nonincreasing in ϵ, it is possible to choose such Δ_{kn} for all k and n. We can now define our estimate. Let

$$f_{kn} = \arg\min_{f \in \mathcal{F}_k} J_n(f) = \arg\min_{f \in \mathcal{F}_k} \frac{1}{n} \sum_{i=1}^{n} L(f(X_i), Y_i)$$

that is, f_{kn} minimizes the empirical risk for n training samples over $\mathcal{F}_k$. (We assume the existence of such minimizing function for each k and n.) The penalized empirical risk is defined for each $f \in \mathcal{F}_k$ as

$$\hat{J}_n(f) = J_n(f) + \Delta_{kn}.$$

Our estimate f_n is then defined as the f_{kn} minimizing the penalized empirical risk over all classes

$$f_n = \arg\min_{f_{kn}: k \ge 1} \hat{J}_n(f_{kn}). \tag{4}$$

We have the following theorem for the expected estimation error of the above complexity regularization scheme. The theorem is proved in Section IV.

Theorem 1: For any n and k the complexity regularization estimate (4) satisfies

$$\mathbf{E}J(f_n) - J(f^*) \le \min_{k \ge 1} \left(R_{kn} + \inf_{f \in \mathcal{F}_k} J(f) - J(f^*) \right)$$

where

$$R_{kn} = \min_{u \ge 4\Delta_{kn}} \left(u + 9Be^{-nu^2/(512B^2)} \right).$$

We will now give an explicit choice for Δ_{kn} which works well in typical situations. Since $N(\epsilon, \mathcal{H}_k)$ is an upper bound on the random covering numbers, we can assume without loss of generality that $\log N(\epsilon, \mathcal{H}_n) \ge 1$ for all $\epsilon > 0$ and n. Then

$$\sqrt{128B^2 \frac{\log N(B/\sqrt{n}, \mathcal{H}_k) + c_k}{n}} > \frac{8B}{\sqrt{n}}.$$

Since $N(\epsilon, \mathcal{H}_k)$ is nonincreasing in ϵ, the choice

$$\Delta_{kn} = \sqrt{128B^2 \frac{\log N(B/\sqrt{n}, \mathcal{H}_k) + c_k}{n}} \tag{5}$$

satisfies (3).

For the problems we investigate in this paper, we will find (see Section V) that $N(\epsilon, \mathcal{H}_k) = (A_1/\epsilon)^{A_2 k}$ satisfies condition (1) for some positive constants A_1 and A_2. The c_k may be chosen as $c_k = 2\log k + c_0$ with $c_0 \ge \log(\sum_{k \ge 1} k^{-2})$. Choosing Δ_{kn} as in (5) then gives

$$\begin{aligned}\Delta_{kn} &= \sqrt{128B^2 \frac{kA_2 \log\left(\frac{A_1\sqrt{n}}{B}\right) + 2\log k + c_0}{n}} \\ &= O\left(\sqrt{\frac{k \log n}{n}} \right).\end{aligned}$$

Since $(u + 9Be^{-nu^2/(512B^2)}) \le 2u$ if $u \ge \sqrt{(256B^2 \log n)/n}$, we have obtained that

$$\begin{aligned}R_{kn} &= \min_{u \ge 4\Delta_{kn}} \left(u + 9Be^{-nu^2/(512B^2)} \right) \\ &= O\left(\sqrt{\frac{k \log n}{n}} \right).\end{aligned}$$

Thus we have proved the following corollary of Theorem 1.

Corollary 1: Assume that $N(\epsilon, \mathcal{H}_k) \le (\frac{A_1}{\epsilon})^{A_2 k}$ for all k. Then the complexity regularized estimate of Theorem 1 gives

$$\begin{aligned}&\mathbf{E}J(f_n) - J(f^*) \\ &\quad \le \min_{k \ge 1} \left(O\left(\sqrt{\frac{k \log n}{n}} \right) + \inf_{f \in \mathcal{F}_k} J(f) - J(f^*) \right).\end{aligned}$$

A. Squared Error Loss

For the special case when

$$L(x, y) = (x - y)^2$$

we can obtain a better upper bound. The estimate will be the same as before, but instead of (3), the complexity penalty Δ_{kn} now has to satisfy

$$\Delta_{kn} \ge C_1 \frac{\log N(\Delta_{kn}/C_2, \mathcal{F}_k) + c_k}{n} \tag{6}$$

where $C_1 = 3499C^4, C_2 = 256C^3$, and $C = \max\{B, 1\}$. Here $N(\epsilon, \mathcal{F}_k)$ is a uniform upper bound on the random l_1 covering numbers $N(\epsilon, \mathcal{F}_k, X_1^n)$. Assume that the class $\mathcal{F} = \bigcup_k \mathcal{F}_k$ is convex, and let $\bar{\mathcal{F}}$ be the closure of $\mathcal{F}$ in $L^2(\mu)$, where μ denotes the distribution of X. Then there is a unique $\bar{f} \in \bar{\mathcal{F}}$ whose squared loss $J(\bar{f})$ achieves $\inf_{f \in \mathcal{F}} J(f)$. We have the following bound on the difference $\mathbf{E}J(f_n) - J(\bar{f})$.

Theorem 2: Assume that $\mathcal{F} = \bigcup_k \mathcal{F}_k$ is a convex set of functions, and consider the squared error loss. Suppose that $|f(x)| \le B$ for all $x \in R^d$ and $f \in \mathcal{F}$. Then complexity regularization estimate with complexity penalty satisfying (6) gives

$$\mathbf{E}J(f_n) - J(\bar{f}) \le 2 \min_{k \ge 1} \left(\Delta_{kn} + \inf_{f \in \mathcal{F}_k} J(f) - J(\bar{f}) \right) + \frac{C_1}{2n}.$$

The proof, which is given in Section IV, uses an idea of Barron [8] and a Bernstein-type uniform probability inequality (Lemma 3 in the Appendix) recently obtained by Lee *et al.* [13]. Note that since $J(\bar{f}) - J(f^*) \ge 0$, we can substitute $J(f^*)$ in place of $J(\bar{f})$ in the statement of the Theorem. However, due to the extra factor of two on the right-hand side, this form of the statement would be weaker.

Just as in the proof of Corollary 1, it is easy to see that when $N(\epsilon, \mathcal{H}_k) = (A_1/\epsilon)^{A_2 k}$, the term Δ_{kn} can be chosen such that $\Delta_{kn} = O(k \log n/n)$. Thus we obtain the following improvement of Corollary 1.

Corollary 2: Assume that $N(\epsilon, \mathcal{H}_k) \le (\frac{A_1}{\epsilon})^{A_2 k}$ for all k. Then the complexity regularized estimate of Theorem 2 for squared error loss gives

$$\begin{aligned} &\mathbf{E}J(f_n) - J(\bar{f}) \\ &\quad \le 2 \min_{k \ge 1} \left(\inf_{f \in \mathcal{F}_k} J(f) - J(\bar{f}) + O\left(\frac{k \log n}{n} \right) \right) + O\left(\frac{1}{n} \right). \end{aligned}$$

IV. Proofs

Proof of Theorem 1: To simplify the proof we will assume that for any k there exists a function minimizing the risk over $\mathcal{F}_k$

$$f_k^* = \arg\min_{f \in \mathcal{F}_k} J(f).$$

Then for any positive ϵ we have

$$\begin{aligned} \mathbf{P}\{J(f_n) - J(f_k^*) \ge \epsilon\} &\le \mathbf{P}\{J(f_n) - \hat{J}_n(f_n) \ge \epsilon/2\} && (7) \\ &\quad + \mathbf{P}\{\hat{J}_n(f_n) - J(f_k^*) \ge \epsilon/2\}. && (8) \end{aligned}$$

Since $\hat{J}_n(f) = J_n(f) + \Delta_{jn}$ for any $f \in \mathcal{F}_j$, the right-hand side of (7) is dealt with as follows:

$$\begin{aligned} &\mathbf{P}\{J(f_n) - \hat{J}_n(f_n) \ge \epsilon/2\} \\ &\quad \le \mathbf{P}\left(\bigcup_{j \ge 1} \{J(f_{jn}) - \hat{J}_n(f_{jn}) \ge \epsilon/2\} \right) \\ &\quad = \mathbf{P}\left(\bigcup_{j \ge 1} \{J(f_{jn}) - J_n(f_{jn}) \ge \epsilon/2 + \Delta_{jn}\} \right) \\ &\quad \le \sum_{j=1}^{\infty} \mathbf{P}\{J(f_{jn}) - J_n(f_{jn}) \ge \epsilon/2 + \Delta_{jn}\} \end{aligned}$$

$$\begin{aligned} &\le \sum_{j=1}^{\infty} \mathbf{P}\left\{ \sup_{f \in \mathcal{F}_j} |J(f) - J_n(f)| \ge \epsilon/2 + \Delta_{jn} \right\} \\ &\overset{(a)}{\le} \sum_{j=1}^{\infty} 8N((\epsilon/2 + \Delta_{jn})/8, \mathcal{H}_j) \\ &\qquad \times \exp\left(-n \left[\frac{\epsilon^2}{512B^2} + \frac{\Delta_{jn}^2}{128B^2} \right] \right) \\ &\overset{(b)}{\le} \sum_{j=1}^{\infty} \frac{8N((\epsilon/2 + \Delta_{jn})/8, \mathcal{H}_j)}{N(\Delta_{jn}/8, \mathcal{H}_j)} \exp\left(-\frac{n\epsilon^2}{512B^2} - c_j \right) \\ &\le 8e^{-n\epsilon^2/(512B^2)} \sum_{j=1}^{\infty} e^{-c_j} \\ &\le 8e^{-n\epsilon^2/(512B^2)} \end{aligned}$$

where in (a) we used Pollard's inequality (see the Appendix) for the class of functions $\mathcal{H}_j$, and in (b) we used the defining inequality (3) for Δ_{jn}. The probability in (8) can be bounded for $\epsilon/4 \ge \Delta_{kn}$ as follows:

$$\begin{aligned} &\mathbf{P}\{\hat{J}_n(f_n) - J(f_k^*) \ge \epsilon/2\} \\ &\quad \le \mathbf{P}\{\hat{J}_n(f_{kn}) - J(f_k^*) \ge \epsilon/2\} \\ &\quad \le \mathbf{P}\{\hat{J}_n(f_k^*) - J(f_k^*) \ge \epsilon/2\} \\ &\quad = \mathbf{P}\{J_n(f_k^*) - J(f_k^*) \ge \epsilon/2 - \Delta_{kn}\} \\ &\quad \le \mathbf{P}\{J_n(f_k^*) - J(f_k^*) \ge \epsilon/4\} \\ &\quad \le e^{-n\epsilon^2/(8B^2)} \end{aligned}$$

where the last inequality follows from Hoeffding's inequality. Thus we have proved that for all $\epsilon \ge 4\Delta_{jn}$

$$\begin{aligned} \mathbf{P}\{J(f_n) - J(f_k^*) \ge \epsilon\} &\le 8e^{-n\epsilon^2/(512B^2)} + e^{-n\epsilon^2/(8B^2)} \\ &\le 9e^{-n\epsilon^2/(512B^2)}. \end{aligned}$$

Since $J(f_n) \le B$ a.s., for all $u \ge 4\Delta_{kn}$ we obtain

$$\begin{aligned} \mathbf{E}[J(f_n) - J(f_k^*)] &\le u + B\mathbf{P}\{J(f_n) - J(f_k^*) \ge u\} \\ &\le u + 9Be^{-nu^2/(512B^2)} \end{aligned}$$

proving the statement of the theorem. □

Proof of Theorem 2: For the sake of convenience we will again assume that for any k there exists a function minimizing the expected squared loss over $\mathcal{F}_k$

$$f_k^* = \arg\min_{f \in \mathcal{F}_k} J(f).$$

Let $\bar{\mathcal{F}}$ be the $L^2(\mu)$ closure of the convex family of functions $\mathcal{F} = \bigcup_k \mathcal{F}_k$, where μ is the probability measure induced by X, and define $\bar{f}$ as the point in $\bar{\mathcal{F}}$ closest to f^*, where $f^*(x) = \mathbf{E}(Y \mid X = x)$. That is, we have

$$J(\bar{f}) = \inf_{f \in \mathcal{F}} J(f).$$

For any $f \in \bigcup_k \mathcal{F}_k$, let

$$H(f) = J(f) - J(\bar{f})$$

and

$$H_n(f) = J_n(f) - J_n(\bar{f}).$$

Then the estimation error can be written as

$$\begin{aligned}\mathbf{E}J(f_n) - J(\bar{f}) &= \mathbf{E}H(f_n)\\ &= \mathbf{E}[H(f_n) - 2(\hat{J}_n(f_n) - J_n(\bar{f}))]\\ &\quad + 2\mathbf{E}[\hat{J}_n(f_n) - J_n(\bar{f})]. \qquad (9)\end{aligned}$$

The second term on the right-hand side can be bounded as

$$\begin{aligned}\mathbf{E}[\hat{J}_n(f_n) - J_n(\bar{f})] &= \mathbf{E}\left[\inf_{k\ge 1}(J_n(f_{kn}) - J_n(\bar{f}) + \Delta_{kn})\right]\\ &\le \mathbf{E}\left[\inf_{k\ge 1}(J_n(f_k^*) - J_n(\bar{f}) + \Delta_{kn})\right]\\ &\le \inf_{k\ge 1}\mathbf{E}\big[J_n(f_k^*) - J_n(\bar{f}) + \Delta_{kn}\big]\\ &= \inf_{k\ge 1}(J(f_k^*) - J(\bar{f}) + \Delta_{kn}). \qquad (10)\end{aligned}$$

To deal with the first term in (9) let $t > 0$, and consider the probability

$$\begin{aligned}&\mathbf{P}\{H(f_n) - 2[\hat{J}_n(f_n) - J_n(\bar{f})] > t\}\\ &\quad\le \mathbf{P}\left\{\sup_{k\ge 1}(H(f_{kn}) - 2H_n(f_{kn}) - 2\Delta_{kn}) > t\right\}\\ &\quad\le \sum_{k=1}^{\infty}\mathbf{P}\{H(f_{kn}) - 2H_n(f_{kn}) > t + 2\Delta_{kn}\}\\ &\quad= \sum_{k=1}^{\infty}\mathbf{P}\left\{\frac{H(f_{kn}) - H_n(f_{kn})}{t + 2\Delta_{kn} + H(f_{kn})} > \frac{1}{2}\right\}\\ &\quad\le \sum_{k=1}^{\infty}\mathbf{P}\left\{\sup_{f\in\mathcal{F}_k}\frac{H(f) - H_n(f)}{t + 2\Delta_{kn} + H(f)} > \frac{1}{2}\right\}. \qquad (11)\end{aligned}$$

In the key step of the proof a probability inequality by Lee *et al.* [13], described in Lemma 3 in the Appendix, is used. In Lemma 3 we set $\beta = t + \Delta_{kn}, \gamma = \Delta_{kn}$ and $\alpha = 1/2$ to obtain the upper bound

$$\begin{aligned}&\mathbf{P}\left\{\sup_{f\in\mathcal{F}_k}\frac{H(f) - H_n(f)}{t + 2\Delta_{kn} + H(f)} > \frac{1}{2}\right\}\\ &\quad\le 6N(\Delta_{kn}/(128C^3), \mathcal{F}_k)\exp\left(\frac{-3(1/2)^2(t + \Delta_{kn})n}{2624C^4}\right).\end{aligned}$$

It follows from the defining inequality (6) for Δ_{kn} that the above is upper bounded by

$$6\exp\left(-n\frac{3t}{4\cdot 2624C^4} - c_k\right).$$

Since $\sum_k e^{-c_k} \le 1$, we obtain from this and (11) that

$$\begin{aligned}&\mathbf{E}[H(f_n) - 2(\hat{J}_n(f_n) - J_n(\bar{f}))]\\ &\quad\le \int_0^{\infty}\mathbf{P}\{H(f_n) - 2[\hat{J}_n(f_n) - J_n(\bar{f})] > t\}\,dt\\ &\quad\le 6\int_0^{\infty}\sum_{k=1}^{\infty}\exp\left(-n\frac{3t}{4\cdot 2624C^4} - c_k\right)dt\\ &\quad\le 6\int_0^{\infty}\exp\left(-n\frac{3t}{4\cdot 2624C^4}\right)dt\\ &\quad= \frac{8\cdot 2624C^4}{n}.\end{aligned}$$

Finally, this and (10) give

$$\begin{aligned}&\mathbf{E}[J(f_n)] - J(\bar{f})\\ &\quad\le 2\min_{k\ge 1}\left[\Delta_{kn} + \inf_{f\in\mathcal{F}_k}J(f) - J(\bar{f})\right] + \frac{8\cdot 2624C^4}{n}\end{aligned}$$

which completes the proof. □

V. RBF Networks

We will consider RBF networks with one hidden layer. Such a network is characterized by a kernel $K : R^+ \to R$. An RBF net of k nodes is of the form

$$f(x) = \sum_{i=1}^{k} w_i K\big([x - c_i]^t A_i [x - c_i]\big) + w_0 \qquad (12)$$

where $w_0, w_1, \cdots, w_k$ are real numbers called weights, $c_1, \cdots, c_k \in R^d$, and the A_i are nonnegative definite $d \times d$ matrices. The kth candidate class $\mathcal{F}_k$ for the function estimation task is defined as the class of networks with k nodes which satisfy the weight condition $\sum_{i=0}^{k}|w_i| \le b$ for a fixed $b > 0$

$$\mathcal{F}_k = \left\{\sum_{i=1}^{k} w_i K\big([x - c_i]^t A_i [x - c_i]\big) + w_0 : \sum_{i=0}^{k}|w_i| \le b\right\}. \qquad (13)$$

In order to apply Theorem 1 to RBF networks we make the following assumptions on the distribution of (X, Y) and the loss function L:

- Y is bounded almost surely:

$$\mathbf{P}\{|Y| > b\} = 0; \qquad (14)$$

- the loss function satisfies the Lipschitz condition

$$|L(x, y) - L(z, y)| \le M|x - z| \qquad (15)$$

if $|x|, |y|, |z| \le b$. With the above assumptions we obtain the following result for complexity regularized regression estimation using RBF networks. The theorem is proved in Appendix B.

Theorem 3: Let K be of bounded variation, and assume that $\sup_x |K(x)| \le 1$. Then with assumptions (14) and (15) the estimate satisfies

$$\begin{aligned}&\mathbf{E}J(f_n) - J(f^*)\\ &\quad\le \min_{k\ge 1}\left(O\left(\sqrt{\frac{k\log n}{n}}\right) + \inf_{f\in\mathcal{F}_k}J(f) - J(f^*)\right).\end{aligned}$$

A. L^p Losses

Condition (15) on the loss function is satisfied for the pth power of the L^p loss for $1 \le p < \infty$. In this case $L(x, y) = |x - y|^p, J(f) = \mathbf{E}|f(X) - Y|^p$, and (15) holds with $M = p(2b)^{p-1}$. Let μ denote the probability measure induced by X. Then by the triangle inequality, we have

$$(J(f))^{1/p} - (J(f^*))^{1/p} \le \|f - f^*\|_{L^p(\mu)} \qquad (16)$$

where $\|f-g\|_{L^p(\mu)}$ denotes the $L^p(\mu)$ norm $(\int |f-f^*|^p d\mu)^{1/p}$. Define $\bar{\mathcal{F}}$ to be the closure in $L^p(\mu)$ of the convex hull of the functions $\hat{b}K([x-c]^tA[x-c])$ and the constant function $h(x)=1, x\in R^d$, where $|\hat{b}|\le b, c\in R^d$, and A varies over all nonnegative $d\times d$ matrices. That is, $\bar{\mathcal{F}}$ is the closure of $\mathcal{F}=\bigcup_k \mathcal{F}_k$, where $\mathcal{F}_k$ is given in (13). Let $g\in\bar{\mathcal{F}}$ be arbitrary. If, as in Theorem 3, we assume that $|K|$ is uniformly bounded, then by [24, Corollary 1], we have for $1\le p\le 2$ that

$$\inf_{f\in\mathcal{F}_k}\|f-g\|_{L^p(\mu)} = O(1/\sqrt{k}) \tag{17}$$

where $\mathcal{F}_k$ is given in (13). The constant in the $O(1/\sqrt{k})$ term depends on b and p, but not on g. The approximation error $\inf_{f\in\mathcal{F}_k} J(f)-J(f^*)$ can be dealt with using this result if the optimal f^* happens to be in $\bar{\mathcal{F}}$. In this case, $\inf_{f\in\mathcal{F}_k} J(f)-J(f^*)\to 0$ as $k\to 0$, and we have that

$$\inf_{f\in\mathcal{F}_k} J(f)-J(f^*) = O\left(\left(\inf_{f\in\mathcal{F}_k} J(f)\right)^{1/p} - (J(f^*))^{1/p}\right).$$

Thus by (16) and (17) we obtain

$$\inf_{f\in\mathcal{F}_k} J(f)-J(f^*) = O(1/\sqrt{k})$$

for all $1\le p\le 2$. Values of p close to one are of great importance for robust neural-network regression (see, e.g., [25]). For $1\le p\le 2$, Theorem 3 gives the following convergence rate for complexity regularized RBF L^p regression estimation:

$$\mathbf{E}J(f_n)-J(f^*) \le \min_{k\ge 1}\left[O\left(\sqrt{\frac{k\log n}{n}}\right)+O\left(\sqrt{\frac{1}{k}}\right)\right] = O\left(\left(\frac{\log n}{n}\right)^{1/4}\right).$$

For $p=1$, i.e., for L^1 regression estimation, this rate is known to be optimal within the logarithmic factor.

For squared error loss $J(f)=\mathbf{E}(f(X)-Y)^2$ we have $f^*(x)=\mathbf{E}(Y\mid X=x)$, and therefore

$$J(f)-J(f^*) = \mathbf{E}(f(X)-f(X))^2 = \|f-f^*\|^2_{L^2(\mu)}.$$

If $f^*\in\bar{\mathcal{F}}$, then by specializing (17) to $p=2$ (and also by an earlier result of Jones [5] and Barron [3]) we obtain

$$\inf_{f\in\mathcal{F}_k} J(f)-J(f^*) = O(1/k). \tag{18}$$

It is easy to check that the class $\bigcup_k \mathcal{F}_k$ is convex if the $\mathcal{F}_k$ are the collections of RBF nets defined in (13). This and Lemma 4 in the Appendix imply that the conditions of Corollary 2 are satisfied, and we can get rid of the square root in Theorem 3.

Theorem 4: Let $\sup_x|K(x)|\le 1$ and assume that K is of bounded variation. Suppose furthermore that $|Y|$ is a bounded random variable, and let $L(x,y)=(x-y)^2$. Then the complexity regularization RBF squared regression estimate satisfies

$$\mathbf{E}J(f_n)-\inf_{f\in\mathcal{F}} J(f) \le 2\min_{k\ge 1}\left(\inf_{f\in\mathcal{F}_k} J(f)-\inf_{f\in\mathcal{F}} J(f) + O\left(\frac{k\log n}{n}\right)\right)+O\left(\frac{1}{n}\right).$$

If $f^*\in\bar{\mathcal{F}}$, this result and (18) give

$$\mathbf{E}J(f_n)-\inf_{f\in\mathcal{F}} J(f) \le \min_{k\ge 1}\left[O\left(\frac{k\log n}{n}\right)+O\left(\frac{1}{k}\right)\right] = O\left(\left(\frac{\log n}{n}\right)^{1/2}\right).$$

This result sharpens and extends the main result (Theorem 3.1) of Niyogi and Girosi [17] where the weaker $O(\sqrt{\frac{k\log n}{n}})+O(\frac{1}{k})$ convergence rate was obtained (in a PAC-like formulation) for the squared loss of Gaussian RBF network regression estimation. Note that our result is valid for a very large class of RBF schemes, including the Gaussian RBF networks considered in [17]. Also, our rate is the same obtained by Barron [11] for sigmoidal networks. Our result, however, differs from Barron's. First, due to the technique of l_1 covering, we can allow the basis functions to have discontinuities, as long as they are of bounded variation. In [11] the sigmoids are required to be continuous and satisfy a Lipschitz condition because covering in supremum norm is used to obtain bounds on the estimation error. For the same reason, the network parameters are discretized in [11], while we allow the minimization in the definition of the estimate f_n to be carried out over a continuum of the parameter values w_i, c_i, and A_i. Third, our only restriction on the parameters is the requirement that $\sum_{i=0}^k|w_i|\le b$. The location parameters c_i and the matrices A_i determining the receptive field size are varied freely, while in [3] the parameter $a\in R^d$ for the sigmoidal unit $\phi(a^tx+b)$ must have a bounded l_1 norm which depends on the sample size n and on the rate at which $\phi(z)$ approaches its limit as $|z|\to\infty$. Extending Barron's result, McCaffrey and Gallant [12] eliminated the need to discretize the parameters and obtained a convergence rate which is better for small dimensions and smooth regression functions. This result also assumes a continuous activation function, namely the so called cosine squasher.

The above convergence rate results hold in the case when there exists an f^* minimizing the risk which is a member of the $L^p(\mu)$ closure of $\mathcal{F}=\cup\mathcal{F}_k$, where

$$\mathcal{F}_k = \left\{\sum_{i=1}^k w_iK\big([x-c_i]^tA_i[x-c_i]\big)+w_0 : \sum_{i=0}^k|w_i|\le b\right\}. \tag{19}$$

In other words, f^* should be such that for all $\epsilon>0$ there exists a k and a member f of $\mathcal{F}_k$ with $\|f-f^*\|_{L^p(\mu)}<\epsilon$. The precise characterization of $\bar{\mathcal{F}}$ seems to be difficult. However, based on the work of Girosi and Anzellotti [6] we can describe a large class of functions that is *contained* in $\bar{\mathcal{F}}$.

Let $H(x,z)$ be a bounded, real, and measurable function of two variables $x\in R^d$ and $z\in R^n$. Suppose that λ is a signed measure on R^n with finite total variation $\|\lambda\|$ (see, e.g., Royden [26]). If $g(x)$ is defined as

$$g(x)=\int_{R^n} H(x,z)\lambda(dz)$$

then $g\in L^p(\mu)$ for any probability measure μ on R^d. One can reasonably expect that g can be approximated well by functions $f(x)$ of the form

$$f(x)=\sum_{i=1}^{k} w_i H(x,z_i)$$

where $z_1,\cdots,z_k \in R^n$ and $\sum_{i=1}^k |w_i| \le \|\lambda\|$. The case $n=d$ and $H(x,z)=G(x-z)$ is investigated in [6], where a detailed description of function spaces arising from the different choices of the basis function G is given and approximation by convex combinations of translates and dilates of a Gaussian function is considered. In general we can prove the following.

Lemma 1: Let

$$g(x)=\int_{R^n} H(x,z)\lambda(dz), \tag{20}$$

where $H(x,z)$ and λ are as above. Define for each $k\ge 1$ the class of functions

$$\mathcal{G}_k=\left\{f(x)=\sum_{i=1}^{k} w_i H(x,z_i) : \sum_{i=0}^{k} |w_i| \le \|\lambda\|\right\}.$$

Then for any probability measure μ on R^d and for any $1\le p<\infty$, the function g can be approximated in $L^p(\mu)$ arbitrarily closely by members of $\mathcal{G}=\cup\mathcal{G}_k$, i.e.,

$$\inf_{f\in\mathcal{G}_k} \|f-g\|_{L^p(\mu)} \to 0 \qquad \text{as } k\to\infty.$$

In other words, $g\in\bar{\mathcal{G}}$.

To prove this lemma one need only slightly adapt the proof of Theorem 8.2 in [6], which is based on the notion of vector-valued integration and proves convergence in $L^2(R^d)$. A more elementary, probabilistic proof can be based on the proof of Theorem 1 of [16]. It is worth mentioning that in general, the closure of $\cup G_k$ is richer than the class of functions having representation as in (20).

To apply the lemma for RBF networks considered in this paper, let $n=d^2+d$, $z=(A,c)$, and $H(x,z)=K([x-c]^tA[x-c])$. Then we obtain that $\bar{\mathcal{F}}$ contains all the functions g with the integral representation

$$g(x)=\int_{R^{d^2+d}} K\big([x-c]^tA[x-c]\big)\lambda(dc\,dA)$$

for which $\|\lambda\|\le b$, where b is the constraint on the weights as in (19). One important example of functions g obtainable in this manner is given by Girosi [14]. He uses the Gaussian basis function

$$H(x,z)=H(x,c,\sigma)=\exp\bigg(-\frac{\|x-c\|^2}{\sigma}\bigg)$$

where $c\in R^d$, $\sigma>0$, and $t=(c,\sigma)$. Using results from Stein [27] he shows that members of the Bessel potential space of order $2m>d$ have integral representation in the form of (20) with this $H(x,z)$, and that they can be approximated by functions of the form

$$f(x)=\sum_{i=1}^{k} w_i \exp\bigg(-\frac{\|x-c_i\|^2}{\sigma_i}\bigg) \tag{21}$$

in $L^2(R^d)$ as well as in supremum norm, and thus in $L^p(\mu)$. The space of functions thus obtained includes the Sobolev space $H^{2m,1}$ of functions whose weak derivatives up to order $2m$ are in $L^1(R^d)$. Note that the class of RBF networks considered in our Theorem 3 and Theorem 4 contain (21) as a special case.

VI. Conclusion

In this paper we applied complexity regularization to obtain estimation bounds in nonlinear function estimation for a large class of loss functions. This approach has been used to obtain the rates of convergence for radial basis nets. The network parameters were learned by minimizing the penalized empirical risk and the analysis involved the random covering numbers and metric entropy. The rates obtained in this paper for radial basis networks substantially improve the existing results by extending the class of functions to functions of bounded variation and by improving on the rates of convergence. An interesting open problem is to obtain the lower bounds on the rates of convergence of radial basis networks.

Appendix A

Lemma 2 ([28]): Let $\mathcal{F}$ be a class of real functions on R^m with $|f(z)|\le B$ for all $f\in\mathcal{F}$, $z\in R^m$, and let $Z_1^n=(Z_1,\cdots,Z_n)$ be R^m valued i.i.d. random variables. Then for any $\epsilon>0$

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

The next result is a probability inequality by Lee *et al.* [13]. In a sense, it provides a sharpening of Pollard's inequality for the uniform deviation of the squared error loss. As in the proof of Theorem 2, let $\mathcal{F}=\bigcup_k\mathcal{F}_k$, where the $\mathcal{F}_k$ are families of real functions on R^d which have uniform upper bounds $N(\epsilon,\mathcal{F}_k)$ on their l_1 random covering numbers. Let X be an R^d valued random vector and let Y be a real random variable. Denote by $\bar{\mathcal{F}}$ the closure of $\mathcal{F}$ in $L^2(\mu)$, where μ is the probability measure induced by X, and let $\bar{f}\in\mathcal{F}$ be the function closest to $f^*(x)=\mathbf{E}[Y\mid X=x]$ in $L^2(\mu)$ norm, that is,

$$J(\bar{f})=\mathbf{E}|\bar{f}(X)-Y|^2=\inf_{f\in\mathcal{F}} J(f).$$

Let $(X_1,Y_1),\cdots,(X_n,Y_n)$ be i.i.d. copies of (X,Y), and for any $f\in\mathcal{F}$ define

$$H(f)=J(f)-J(\bar{f}) \quad\text{and}\quad H_n(f)=J_n(f)-J_n(\bar{f})$$

where $J_n(f)=n^{-1}\sum_{i=1}^n (f(X_i)-Y_i)^2$. Then the following holds.

Lemma 3 ([13, Th. 3]): Assume that $\mathcal{F}$ is convex and $|f(X)|\le B$ for all $f\in\mathcal{F}$ and $x\in R^d$. Suppose furthermore that $\mathbf{P}\{|Y|>B\}=0$. Let $C=\max\{B,1\}$, and let $\beta,\gamma>0$ and $0<\alpha\le 1/2$. Then for any n and k we have

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}_k}\frac{H(f)-H_n(f)}{\beta+\gamma+H(f)}\ge\alpha\right\} \le 6N\Big(\frac{\alpha\gamma}{128C^3},\mathcal{F}_k\Big)e^{-3\alpha^2\beta n/(2624C^4)}.$$

Appendix B

Proof of Theorem 3: We only have to show that the conditions of Theorem 1 and Corollary 1 hold. First consider condition (2). Since $|K| \leq 1$, we have $|f| \leq b$ for all $f \in \mathcal{F}_k$ and all k. Combining this with $|Y| \leq b$ a.s. and the Lipschitz condition (15), we obtain

$$L(f(X), Y) \leq 2Mb.$$

Thus (2) holds with $B = 2Mb$. In the rest of the proof we will prove that for appropriate positive constants A_1 and A_2, $(\frac{A_1}{\epsilon})^{A_2 k}$ is an a.s. uniform upper bound on $N(\epsilon, \mathcal{H}_k, Z_1^n)$ for each k and n. First we consider the connection between $N(\epsilon, \mathcal{H}_k, Z_1^n)$ and $N(\epsilon, \mathcal{F}_k, X_1^n)$, where $Z_1^n = ((X_1, Y_1), \cdots, (X_n, Y_n))$. For any $z_1^n = ((x_1, y_1), \cdots, (x_n, y_n))$ with $\max_i |y_i| \leq b$, and for any f_1, f_2 with $|f_1|, |f_2| \leq b$ we have

$$\frac{1}{n}\sum_{i=1}^{n} |L(f_1(x_i), y_i) - L(f_2(x_i), y_i)| \leq \frac{M}{n}\sum_{i=1}^{n} |f_1(x_i) - f_2(x_i)|.$$

It follows that with probability 1 we have

$$N\big(\epsilon, \mathcal{H}_k, Z_1^n\big) \leq N\big(\epsilon/M, \mathcal{F}_k, X_1^n\big)$$

so that

$$N(\epsilon, \mathcal{H}_k) \leq N(\epsilon/M, \mathcal{F}_k). \tag{22}$$

The next lemma determines $N(\epsilon, \mathcal{F}_k)$.

Lemma 4: Assume that $|K(x)| \leq 1$ for all $x \in R^d$, and suppose that K has total variation $V < \infty$. Then a uniform upper bound $N(\epsilon, \mathcal{F}_k)$ on the random covering numbers is given by

$$(e^2(d^2+d+3))^{2(k+1)}\left(\frac{2e(b+\epsilon)}{\epsilon}\right)^{k+1} \times \left(\frac{Ve(b+\epsilon)}{\epsilon}\right)^{2(k+1)(d^2+d+2)}.$$

The proof of the lemma is given below. It is immediate that the lemma implies

$$N(\epsilon, \mathcal{F}_k) \leq \left(\frac{\hat{A}_1}{\epsilon}\right)^{\hat{A}_2 k}$$

for some constants $\hat{A}_1$ and $\hat{A}_2$. Hence by (22) we have

$$N(\epsilon, \mathcal{H}_k) = \left(\frac{A_1}{\epsilon}\right)^{A_2 k}$$

with $A_1 = M\hat{A}_1$ and $A_2 = \hat{A}_2$. Now Corollary 1 gives the desired result. □.

Definition 1: Let $\mathcal{C}$ be a collection of subsets of 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

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

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.

Proof: Since K is of bounded variation it can be decomposed as the difference of two monotone increasing functions: $K = K_1 - K_2$. Let $\mathcal{G}$ be the collection of functions $[x - c]^t A[x - c]$ parameterized by $c \in R^d$ and the nonnegative definite matrix A. Also, let $\hat{\mathcal{F}}_i = \{K_i(g(\cdot)) : g \in \mathcal{G}\}, i = 1, 2$, and let $\mathcal{F} = \{K(g(\cdot)) : g \in \mathcal{G}\}$. Then by a lemma of Pollard [29] concerning the covering number of sums of families of functions, we have

$$N(\epsilon, \mathcal{F}) \leq N(\epsilon/2, \hat{\mathcal{F}}_1)N(\epsilon/2, \hat{\mathcal{F}}_2) \tag{23}$$

because $\mathcal{F} \subset \{f_1 - f_2 : f_1 \in \hat{\mathcal{F}}_1, f_2 \in \hat{\mathcal{F}}_2\}$. Since $\mathcal{G}$ spans a $d^2 + d + 1$-dimensional vector space, by Pollard [28] the collection of sets $\mathcal{G}^+ = \{\{(x,t) : g(x) - t \geq 0\} : g \in \mathcal{G}\}$ has VC dimension $V_{\mathcal{G}^+} \leq d^2 + d + 2$. Since K_i is monotone, it follows from Noland and Pollard [30] that $V_{\hat{\mathcal{F}}_i^+} \leq d^2 + d + 2$, where the families of sets $\hat{\mathcal{F}}_i^+$ are defined just as $\mathcal{G}^+$ with $\hat{\mathcal{F}}_i$ in place of $\mathcal{G}$. Let V_1 and V_2 be the total variations of K_1 and K_2, respectively. Then $V = V_1 + V_2$ and $0 \leq K_i(x) + \alpha_i \leq V_i, x \in R, i = 1, 2$, for suitably chosen constants α_1 and α_2. A result of Haussler [31] states that if $0 \leq f(x) \leq B$ for all $f \in \mathcal{F}$ and x, then

$$N(\epsilon, \mathcal{F}) \leq e(V_{\mathcal{F}} + 1)\left(\frac{2eB}{\epsilon}\right)^{V_{\mathcal{F}}}.$$

It follows that

$$N(\epsilon, \hat{\mathcal{F}}_i) \leq e(d^2 + d + 3)\left(\frac{2eV}{\epsilon}\right)^{d^2+d+2}$$

and since $V_1 \cdot V_2 \leq V^2/4$, this and (23) implies that

$$N(\epsilon, \mathcal{F}) \leq e^2(d^2 + d + 3)^2\left(\frac{eV}{\epsilon}\right)^{2(d^2+d+2)}.$$

Since $\mathcal{F}_k$ is defined as

$$\mathcal{F}_k = \left\{\sum_{i=1}^{k} w_i f_i + w_0 : \sum_{i=0}^{k} w_i \leq b, f_i \in \mathcal{F}\right\}$$

using Lemma 5 below with $\epsilon = 2\delta$ and $B = 1$, we obtain

$$\begin{aligned} N(\epsilon, \mathcal{F}_k) &\leq \left(\frac{2e(b+\epsilon)}{\epsilon}\right)^{k+1}[N(\epsilon/(b+2\delta), \mathcal{F})]^{k+1} \\ &\leq (e^2(d^2+d+3))^{2(k+1)}\left(\frac{2e(b+\epsilon)}{\epsilon}\right)^{k+1} \\ &\quad \times \left(\frac{Ve(b+\epsilon)}{\epsilon}\right)^{2(k+1)(d^2+d+2)}. \end{aligned}$$

□

Lemma 5: Let $\mathcal{G}_1, \cdots, \mathcal{G}_k$ be classes of real functions over the same domain, and define $\mathcal{F}$ as

$$\mathcal{F} = \left\{ \sum_{i=1}^{k} w_i f_i : (w_1, \cdots, w_k) \in R^k : \right. \\ \left. \sum_{i=1}^{k} |w_i| \le b, f_i \in \mathcal{G}_i, i = 1. \cdots, n \right\}.$$

Let $N(\epsilon, \mathcal{G}_i)$ be the covering number of $\mathcal{G}_i$ with respect to a norm $||\cdot||$ over the linear space spanned by the $\mathcal{G}_i$, and assume that $||f|| \le B$ for all $f \in \mathcal{G}_i, i = 1, \cdots, k$. Then we have for any $\epsilon, \delta > 0$

$$N(\epsilon + \delta, \mathcal{F}) \le \left(\frac{Beb}{\delta} \right)^k \prod_{i=1}^{k} N(\epsilon/(b + 2\delta), \mathcal{G}_i)$$

Proof of Lemma 5: Let

$$S_b = \left\{ w \in R^k : \sum_{i=1}^{k} |w_i| \le b \right\}$$

and assume that $S_{b,\delta}$ is a finite subset of R^k with the covering property $\max_{w \in S_b} \min_{x \in S_{b,\delta}} ||w - x||_1 \le \delta$, where $||y||_1$ denotes the l_1 norm of any $y \in R^k$. Also, let the $\mathcal{G}_i(\epsilon)$ be the minimal covers for the $\mathcal{G}_i$, that is, each $\mathcal{G}_i(\epsilon)$ has cardinality $N(\mathcal{G}_i, \epsilon)$ and $\min_{g \in \mathcal{G}_i(\epsilon)} ||f - g|| \le \epsilon$ for all $f \in \mathcal{G}_i$. Let $f \in \mathcal{F}$ given by $f = \sum_{i=1}^{k} w_i f_i$, and choose $x \in S_{b,\delta}$ and $\hat{f}_i \in \mathcal{G}_i(\epsilon)$ with $||w - x||_1 \le \delta$ and $||f_i - \hat{f}_i|| \le \epsilon$, $i = 1, \cdots, k$. Since $||f_i|| \le B$ for all i, we have

$$\begin{aligned} &\left\| f - \sum_{i=1}^{k} x_i \hat{f}_i \right\| \\ &\quad \le \left\| \sum_{i=1}^{k} w_i f_i - \sum_{i=1}^{k} x_i f_i \right\| + \left\| \sum_{i=1}^{k} x_i f_i - \sum_{i=1}^{k} x_i \hat{f}_i \right\| \\ &\quad \le \sum_{i=1}^{k} |w_i - x_i| \, ||f_i|| + \sum_{i=1}^{k} |x_i| \, ||f_i - \hat{f}_i|| \\ &\quad \le \delta B + \epsilon b_\delta \end{aligned}$$

where $b_\delta = \max_{x \in S_{b,\delta}} ||x||_1$. It follows that a set of functions of cardinality:

$$|S_{b,\delta}| \cdot \prod_{i=1}^{k} N(\epsilon, \mathcal{G}_i)$$

will $(\delta B + \epsilon b_\delta)$-cover $\mathcal{F}$. Thus we need only to bound the cardinality of $S_{b,\delta}$. The obvious choice for $S_{b,\delta}$ is a rectangular grid with edge length $2\delta/k$. Define $S_{b,\delta}$ as the points on this grid whose l_1 Voronoi regions intersect S_b. These Voronoi regions (and the associated grid points) are certainly contained in $S_{b+2\delta}$. Since the volume of $S_{b+2\delta}$ is $(2(b + 2\delta))^k/k!$, the cardinality of $S_{b+2\delta}$ is upper bounded by

$$\frac{(2(b+2\delta))^k}{k!} \left(\frac{2\delta}{k} \right)^{-k} = \frac{1}{k!} \left(\frac{k(b+2\delta)}{\delta} \right)^k \le \left(\frac{e(b+2\delta)}{\delta} \right)^k.$$

Since we have $b_\delta \le b + 2\delta$, this implies

$$N(\delta B + \epsilon(b + 2\delta), \mathcal{F}) \le \left(\frac{e(b + 2\delta)}{\delta} \right)^k \prod_{i=1}^{k} N(\epsilon, \mathcal{G}_i)$$

or

$$N(\delta + \epsilon, \mathcal{F}) \le \left(\frac{Be(b + 2\delta/B)}{\delta} \right)^k \prod_{i=1}^{k} N(\epsilon/(b + 2\delta/B), \mathcal{G}_i).$$

□

REFERENCES

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

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

[3] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory*, vol. 39, pp. 930–944, 1993.

[4] T. Chen, H. Chen, and R. Liu, "Approximation capability in $C(\bar{R}^n)$ by multilayer feedforward networks and related problems," *IEEE Trans. Neural Networks*, vol. 6, pp. 25–30, Jan. 1995.

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

[6] 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: Chapman and Hall, 1993, pp. 97–113.

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

[8] A. R. Barron, "Complexity regularization with application to artificial neural networks," in *Nonparametric Functional Estimation and Related Topics*, G. Roussas, Ed., NATO ASI Series. Dordrecht, The Netherlands: Kluwer, 1991, pp. 561–576.

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

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

[11] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, pp. 115–133, 1994.

[12] D. F. McCaffrey and A. R. Gallant, "Convergence rates for single hiddden layer feedforward networks," *Neural Networks*, vol. 7, no. 1, pp. 147–158, 1994.

[13] W. S. Lee, P. L. Bartlett, and R. C. Williamson, "Efficient agnostic learning of neural networks with bounded fan-in," *IEEE Trans. Inform. Theory*, vol. 42, pp. 2118–2132, Nov. 1996.

[14] F. Girosi, "Regularization theory, radial basis functions and networks," in *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, V. Cherkassky, J. H. Friedman, and H. Wechsler, Eds. Berlin: Springer-Verlag, 1992, pp. 166–187.

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

[16] A. Krzyżak, T. Linder, and G. Lugosi, "Nonparametric estimation and classification using radial basis function nets and empirical risk minimization," *IEEE Trans. Neural Networks*, vol. 7, pp. 475–487, Mar. 1996.

[17] P. Niyogi and F. Girosi, "On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions," *Neural Computa.*, vol. 8, pp. 819–842, 1996.

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

[19] G. Lugosi and K. Zeger, "Nonparametric estimation via empirical risk minimization," *IEEE Trans. Inform. Theory*, vol. 41, pp. 677–678, 1995.

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

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

[22] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Ann. Statist.*, vol. 11, pp. 416–431, 1983.

[23] G. Lugosi and A. Nobel, “Adaptive model selection using empirical complexities,” Dept. Statist., Univ. North Carolina, Chapel Hill, Tech. Rep. 2346, 1996.

[24] 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.

[25] W. J. Rey, *Introduction to Robust and Quasi-Robust Statistical Methods*. Berlin: Springer-Verlag, 1983.

[26] H. L. Royden, *Real Analysis*. New York: Macmillan, 1968.

[27] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*. Princeton, NJ: Princeton Univ. Press, 1970.

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

[29] ——, *Empirical Processes: Theory and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics. Hayward, CA: Inst. Math. Statist., 1990.

[30] D. Nolan and D. Pollard, “U-processes: Rates of convergence,” *Ann. Statist.*, vol. 15, pp. 780–799, 1987.

[31] D. Haussler, “Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension,” *J. Combinatorial Theory Series A*, vol. 69, pp. 217–232, 1995.

Adam Krzyżak (M’84–SM’96) 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, Quebec, Canada. Since August 1983, he has been with the Department of Computer Science, Concordia University, Montreal, where he is currently an Associate Professor. In 1991 he held Vineberg Memorial Fellowship at Technion-Israel Institute of Technology and in 1992 Humboldt Research Fellowship at the University of Erlangen-Nürnberg, Germany. He visited the University of California Irvine, Information Systems Laboratory at Stanford University and Riken Frontiers Research Laboratory, Japan. He has published more than 100 papers in the areas of pattern recognition, image processing, computer vision, identification, and nonparametric estimation.

Dr. Krzyżak is an Associate Editor of the *Pattern Recognition Journal* and *International Journal of Applied Software Technology* and coeditor of the book, *Computer Vision and Pattern Recognition* (Singapore: World, 1989). He has served on the program committees of Vision Interface’88, Vision Interface’94, Vision Interface’95, and 1995 International Conference on Document Processing and Applications, Montreal, Canada.

Tamás Linder (S’92–M’93) was born in Budapest, Hungary, in 1964. He received the M.S. degree from the Technical University of Budapest in 1988, and the Ph.D. degree from the Hungarian Academy of Sciences in 1992, both in electrical engineering.

He was a Postdoctoral Fellow at the University of Hawaii in 1992, and a Fulbright Scholar at the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign in 1993–1994. He has been an Associate Professor of Electrical Engineering at the Technical University of Budapest since 1994 and is currently visiting the Department of Electrical and Computer Engineering, University of California at San Diego. His research interests include communications and information theory, vector quantization, rate-distortion theory, and machine learning.