Nonparametric Estimation via Empirical Risk Minimization

Gábor Lugosi and Kenneth Zeger, *Senior Member, IEEE*

***Abstract*— A general notion of universal consistency of nonparametric estimators is introduced that applies to regression estimation, conditional median estimation, curve fitting, pattern recognition, and learning concepts. General methods for proving consistency of estimators based on minimizing the empirical error are shown. In particular, distribution-free almost sure consistency of neural network estimates and generalized linear estimators is established.**

***Index Terms*— Regression estimation, nonparametric estimation, consistency, pattern recognition, neural networks, series methods, sieves.**

I. Introduction

LET the random variables X and Y take their values from $\mathbb{R}^d$ and $\mathbb{R}$, respectively. Denote the measure of X on $\mathbb{R}^d$ by μ, and the measure of (X,Y) on $\mathbb{R}^d \times \mathbb{R}$ by ν. We are interested in predicting the value of Y from X, that is, in a measurable function $m : \mathbb{R}^d \to \mathbb{R}$ such that $m(X)$ approximates Y well. One can show that if $\mathbf{E}|Y|^p < \infty$ $(1 \le p < \infty)$, then there always exists a (not necessarily unique) measurable function m^* that minimizes the L_p-error $(\mathbf{E}|m^*(X) - Y|^p)^{1/p}$. Take, e.g.,

$$m^*(x) = \inf\{z : \mathbf{E}(|z - Y|^p | X = x) \le \mathbf{E}(|t - Y|^p | X = x), \forall t\}.$$

Denote the error of the L_p-optimal predictor by J_p^*, that is

$$J_p^* = \inf_m (\mathbf{E}|m(X) - Y|^p)^{1/p} = (\mathbf{E}|m^*(X) - Y|^p)^{1/p}$$

where the expectation is taken with respect to the joint distribution ν of (X,Y). Assume that we do not know anything about the distribution of the pair (X,Y), but a collection of independent, identically distributed (i.i.d.) copies

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

of (X,Y) is available, where D_n is independent of (X,Y). Our aim is to estimate good predictors from the data, that is, to construct a function $m_n(x) = m_n(x, D_n)$ such that its L_p-error is close to the optimum J_p^*. Denote its error by the random variable

$$J_p(m_n) = \left(\mathbf{E}\left(|m_n(X) - Y|^p \middle| D_n\right)\right)^{1/p}.$$

Clearly, the estimated predictor m_n is good, if its error $J_p(m_n)$ is close to the optimum J_p^*. A desirable property of an estimate is that its error converges to the optimum as the sample size n grows. This concept is formulated in the following definition.

Definition 1: We call a sequence of estimators $\{m_n\}$ *consistent* for a given distribution of (X,Y), if

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

$\{m_n\}$ is *universally consistent* if it is consistent for *any* distribution of (X,Y) satisfying $\mathbf{E}|Y|^p < \infty$.

Consistency may be defined in terms of other modes of convergence, too. The reason we adopt (the strong notion of) almost sure convergence is because it provides information about the behavior of the estimate for the given realization of the training data.

The main results of the paper are estimators that are universally consistent. These estimators are based on empirical risk minimization, which is described in Section II. Sections VI and VII give our two main applications, where universal consistency of neural network estimates and generalized linear estimates are demonstrated. Sections III and IV contain some general tools for studying estimates based on empirical risk minimization, while Section V gives lemmas that are necessary for the neural network results in Section VII.

The following examples illustrate why this notion of universal consistency is important.

Remark 1 (Curve Fitting): If Y is a function of X, that is, $Y = h(X)$ for some measurable h, then clearly $J_p^* = 0$, and the problem of minimizing $J_p(m_n) - J_p^*$ reduces to approximating the unknown h in L_p

$$J_p(m_n) - J_p^* = \left(\mathbf{E}\left(|m_n(X) - h(X)|^p \middle| D_n\right)\right)^{1/p} = \left(\int |m_n(x) - h(x)|^p \mu(dx)\right)^{1/p}$$

where the available data are observations of $h(x)$ at random points $X_1, \cdots, X_n$. If the unknown function h is an indicator of a set, then the problem reduces to the basic question of the theory of *concept learning*, where the estimator m_n is typically an indicator function (see, e.g., Valiant [67], Blumer *et al.* [11]).

Manuscript received March 30, 1993; revised April 11, 1994. The material in this paper was presented at the IEEE International Symposium on Information Theory, Trondheim, Norway, June 1994. This research was supported in part by the National Science Foundation under Grant NCR-92-96231.

G. Lugosi is with the Department of Mathematics and Computer Science, Faculty of Electrical Engineering, Technical University of Budapest, Budapest, Hungary.

K. Zeger is with the Coordinated Science Laboratory, Department of Electrical and Computer Engineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA.

IEEE Log Number 9410410.

Remark 2 (L_2-Error, Regression Estimation, and Pattern Recognition): If $p = 2$, then $J_2^* = \sqrt{\mathbf{E}(Y - m^*(X))^2}$, where $m^*(x) = \mathbf{E}(Y|X = x)$ is the regression function, and $J_2(m_n) - J_2^* \to 0$ if and only if

$$\mathbf{E}((m_n(X) - Y)^2|D_n) - \mathbf{E}(m^*(X) - Y)^2 = \mathbf{E}((m_n(X) - m^*(X))^2|D_n) \to 0$$

which is the usual notion of L_2-consistency for regression function estimates. Several *local averaging* type regression function estimates are known to be universally consistent, such as *k-nearest neighbor* estimates (Stone [65], Devroye, Györfi, Krzyżak, and Lugosi [22]), *kernel estimates* (Devroye and Wagner [25], Spiegelman and Sacks [64], Devroye and Krzyżak [23]), and *partitioning estimates* (Breiman, Friedman, Olshen, and Stone [12], Devroye and Györfi [21], Györfi [42]). Estimating the regression function is closely related to pattern recognition. In the 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, whose *error probability* $\mathbf{P}\{g_n(X) \neq Y|D_n\}$ is to be minimized. The function that minimizes the error probability is given by

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

and is called the *Bayes decision*. Its error probability $\mathbf{P}\{g^*(X) \neq Y\}$ is the *Bayes-risk*. As observed by Van Ryzyn [69], Wolverton and Wagner [77], Glick [34], Györfi [40], and Devroye and Wagner [24], good estimators of $m^*(x)$ provide classifiers with small error probability. Namely, if a classifier g_n is defined as

$$g_n(x) = \begin{cases} -1, & \text{if } m_n(x) \leq 0 \\ 1, & \text{otherwise} \end{cases}$$

then

$$\mathbf{P}\{g_n(X) \neq Y|D_n\} - \mathbf{P}\{g^*(X) \neq Y\} \leq \left(\mathbf{E}\left((m_n(X) - m^*(X))^2\middle|D_n\right)\right)^{1/2}$$

that is, if $J_2(m_n) - J_2^* \to 0$, then the error probability of the obtained classifier approaches the Bayes-risk.

Remark 3 (L_1-Error, Conditional Median Estimation, and Pattern Recognition): Next we discuss the case $p = 1$. It is well known that if the median of the conditional distribution of Y given $X = x$ exists, then it is equal to $m^*(x)$, the function that minimizes the L_1-error $\mathbf{E}|m(X) - Y|$. Consistency (in probability) of local averaging-type conditional quantile estimates were established by Stone [65], while neural network estimation of conditional quantiles was studied by White [76]. Again, a connection to pattern recognition can be established as follows. Assume that Y can take only two values: $Y \in \{-1, 1\}$. Then it is easy to see that a function that minimizes the L_1 error is also binary valued, and can be written as

$$m^*(x) = \begin{cases} -1, & \text{if } \mathbf{P}\{Y = -1|X = x\} \geq 1/2 \\ 1, & \text{otherwise} \end{cases}$$

which is just the Bayes-classifier: $m^* = g^*$. Now, if we have a (not necessarily binary valued) estimator m_n such that the difference between the L_1-errors $J_1(m_n) - J_1^*$ is small, then it is natural to define a decision rule as

$$g_n(x) = \begin{cases} -1, & \text{if } m_n(x) < 0 \\ 1, & \text{otherwise.} \end{cases}$$

The following lemma asserts that L_1-consistency of m_n implies consistency of g_n:

Lemma 1:

$$\mathbf{P}\{g_n(X) \neq Y|D_n\} - \mathbf{P}\{g^*(X) \neq Y\} \leq J_1(m_n) - J_1^*.$$

Proof: If $g_n(x) = g^*(x)$ then clearly

$$\mathbf{P}\{g_n(X) \neq Y|X = x, D_n\} - \mathbf{P}\{g^*(X) \neq Y|X = x\} = 0$$

so it suffices to consider the case when $g_n(x) \neq g^*(x)$. By straightforward calculation we get

$$\mathbf{P}\{g_n(X) \neq Y|X = x, D_n\} - \mathbf{P}\{g^*(X) \neq Y|X = x\} = |1 - 2\mathbf{P}\{Y = 1|X = x\}|$$

and

$$\mathbf{E}\Big(|m_n(X) - Y|\Big|X = x, D_n\Big) - \mathbf{E}\Big(|g^*(X) - Y|\Big|X = x\Big) \geq |1 - 2\mathbf{P}\{Y = 1|X = x\}| \cdot (1 + \min(|m_n(x)|, 1)).$$

Therefore, for every x

$$\mathbf{P}\{g_n(X) \neq Y|X = x, D_n\} - \mathbf{P}\{g^*(X) \neq Y|X = x\} \leq \mathbf{E}\Big(|m_n(X) - Y|\Big|X = x, D_n\Big) - \mathbf{E}\Big(|g^*(X) - Y|\Big|X = x\Big).$$

Integrating both sides with respect to μ completes the proof. ■

II. Empirical Risk Minimization

Our method of constructing an estimator m_n is to choose it as a function from a class of functions $\mathcal{F}$ that minimizes the *empirical error*

$$J_{p,n}(f) = \left(\frac{1}{n}\sum_{j=1}^{n} |f(X_j) - Y_j|^p\right)^{1/p}$$

that is

$$J_{p,n}(m_n) \leq J_{p,n}(f), \quad \text{for } f \in \mathcal{F}.$$

Remark: Here we assumed the existence of minimizing functions, though not necessarily their uniqueness. It is easy to see that in the special cases studied in Sections VI and VII the minima indeed exist. In cases where the minima do not exist, the same analysis can be carried out with functions whose error is arbitrarily close to the infimum, but for the sake of simplicity we stay with the assumption of existence throughout the paper. Also, similar analysis may be carried out for estimators that approximately minimize the empirical error, i.e., when $J_{p,n}(m_n)$ is sufficiently close to the optimum.

Clearly, we need a large class of functions in order to be able to get small errors for any distribution of (X, Y). On the other hand, if the class is too large (e.g., the class of all measurable functions, or the class of all continuous functions with bounded support), it may *overfit* the data, that is, the empirical error of a function in the class may be small, while

its true error is large. Asymptotic properties of this method of minimizing the empirical risk were studied by several authors such as Vapnik [70] and Haussler [44]. Empirical risk minimization has also become known in the statistics literature as "minimum contrast estimation," e.g., by Nemirovskii [53], Nemirovskii *et al.* [52], Van de Geer [68], and Birgé and Massart [10]. They typically consider picking the empirical optimum from a collection of *fixed* functions, and study how far it is from the true optimum *in the class*. The situation is similar in the theory of learning, where classes of functions, for which empirical minimization picks a function with small error, are called *learnable*, and are usually characterized by having finite *VC-dimension* (see, e.g., Blumer, Ehrenfeucht, Haussler, and Warmuth [11]). These classes, however, are usually too "small" to approximate arbitrary functions, and therefore fail to provide universally consistent estimators. To resolve this conflict, one can adopt different strategies. One of the more interesting techniques is called *complexity regularization* (Barron and Cover [9], Barron [5], [6], and [8]), where one adds a term to the empirical error that penalizes functions with high "complexity." They also apply their results to the special cases discussed in this paper. The method of *structural risk minimization*, developed by Vapnik and Chervonenkis [72] and closely related to complexity regularization, offers an automatic way of selecting correct-sized classes. The approach we investigate here in depth is different. We let the class of candidate functions change (i.e., grow) as the sample-size n grows. This principle is sometimes called the "method of sieves," introduced by Grenander [39]. Its consistency and rates of convergence have been exhaustively studied primarily for nonparametric maximum-likelihood density estimation and least squares regression function estimation by Geman and Hwang [33], Gallant [32], Shen and Wong [62], and Wong and Shen [78]. This is the approach discussed by Devroye [20] for pattern recognition in general, and by White [75], as well as Faragó and Lugosi [30] for neural networks. We should also mention here that apart from arithmetic means, other estimators of the error can also be minimized over functions in the class, and these estimators may perform better. The work of Buescher and Kumar [13], [14] formulates a more general theory.

In all of our applications, the approximating function classes are finite-dimensional, i.e., they can be smoothly parametrized by finitely many parameters. This seems to be necessary, as our goal is to obtain estimators that are consistent for all distributions.

Formally, let $\{\mathcal{F}_n\}$ be a sequence of classes of functions, and define m_n as a function in $\mathcal{F}_n$ that minimizes the empirical error

$$J_{p,n}(m_n) \le J_{p,n}(f), \quad \text{for } f \in \mathcal{F}_n.$$

For analyzing how close the error of the estimator $J_p(m_n)$ is to the optimum J_p^*, we will use the following decomposition:

$$J_p(m_n) - J_p^* = \left(J_p(m_n) - \inf_{f\in\mathcal{F}_n} J_p(f)\right) + \left(\inf_{f\in\mathcal{F}_n} J_p(f) - J_p^*\right).$$

The first term on the right-hand side tells us about the "learnability" of $\mathcal{F}_n$, that is, how well the empirical minimization performs over this class. We will refer to this term as the *estimation error*. The second term, which we call the *approximation error*, describes how rich the class $\mathcal{F}_n$ is, that is, how well the best function in the class performs. Here the main problem is to balance the tradeoff between the approximation potential and the estimability of the class, that is, to determine how fast the class should grow to get universally consistent estimators, if possible at all. The main tools for analyzing such estimators are approximation properties of the classes (i.e., denseness theorems), and exponential distribution-free probability inequalities for the uniform estimability of the error over the class. These exponential inequalities are necessary, since we need distribution-free rate-of-convergence results for the estimation error in order to be able to choose the size of the class without knowing the distribution. If Y is bounded (as, for example, in the pattern recognition problem), then this can be handled in a relatively straightforward way using uniform large deviation inequalities originated mainly by Vapnik and Chervonenkis [71], [73] (also see Vapnik [70], Devroye [20], White [75], or Haussler [44]).

In this paper we introduce techniques to extend results to unbounded variables. After developing general principles and techniques, we exploit them to obtain universal consistency for estimators based on linear combinations of fixed basis functions. Namely, if $\psi_1, \psi_2, \cdots$ is a sequence of real-valued functions on $\mathbb{R}^d$, then the estimator m_n takes the form

$$m_n(x) = \sum_{i=1}^{k} a_i \psi_i(x)$$

where the coefficients $a_1, \cdots, a_k$ are determined from the data D_n.

Our other main examples are estimators m_n realized by *neural networks of one hidden layer*, that is, by functions of the form

$$f_{\theta_k}(x) = \sum_{i=1}^{k} c_i \sigma(a_i^T x + b_i) + c_0$$

where the *sigmoid* $\sigma : \mathbb{R} \to [0, 1]$ is a monotone nondecreasing function converging to 0 as $x \to -\infty$, and to 1 as $x \to \infty$. k is the number of hidden *neurons*, and

$$\theta_k = \{a_1, \cdots, a_k, b_1, \cdots, b_k, c_0, \cdots, c_k\}$$

(where $a_1, \cdots, a_k \in \mathbb{R}^d; b_1, \cdots, b_k, c_0, \cdots, c_k \in \mathbb{R}$) is the set of *parameters (or weights)* that specify the network. Our aim is to adjust the weights of the network as functions of the data D_n such that the function realized by the obtained network is a good—desirably consistent—estimator of m^*.

III. Approximation Error

Here we deal with the convergence of the approximation error

$$\inf_{f\in\mathcal{F}_n} J_p(f) - J_p^*.$$

Clearly, if f' minimizes $(\mathbf{E}|f(X)-m^*(X)|^p)^{1/p}$ over f in $\mathcal{F}_n$, then

$$\inf_{f\in\mathcal{F}_n} J_p(f) - J_p^* \le J_p(f') - J_p^* \le (\mathbf{E}|f'(X)-m^*(X)|^p)^{1/p}$$

by the triangle inequality, therefore, the approximation error goes to zero if

$$\lim_{n\to\infty} \inf_{f\in\mathcal{F}_n} (\mathbf{E}|f(X)-m^*(X)|^p)^{1/p} = 0.$$

Since our aim is to establish *universal* consistency, we require this convergence for *every* measure μ and $m^* \in L_p(\mu)$. If, for example, the classes are nested, that is, $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ for every $n > 0$, then this is equivalent to requiring that the set

$$\bigcup_{n=1}^{\infty} \mathcal{F}_n$$

be *dense* in $L_p(\mu)$ for all μ.

IV. Estimation Error

This section is devoted to investigating the almost sure convergence of the estimation error

$$J_p(m_n) - \inf_{f\in\mathcal{F}_n} J_p(f).$$

The usual way to investigate the above quantity is to exploit the inequality (Devroye [20], Haussler [44])

$$J_p(m_n) - \inf_{f\in\mathcal{F}_n} J_p(f) \le 2\sup_{f\in\mathcal{F}_n} \left| (\mathbf{E}|f(X)-Y|^p)^{1/p} - \left(\frac{1}{n}\sum_{j=1}^n |f(X_j)-Y_j|^p\right)^{1/p} \right|$$

by using uniform laws of large numbers to estimate the right-hand side. In our case, we need nonasymptotic uniform large-deviation inequalities, since the class the supremum is taken over changes with the sample size n. These types of inequalities are available (Vapnik and Chervonenkis [71], [73], Pollard [56]) if the random variable $f(X)-Y$ is uniformly bounded for $f \in \mathcal{F}_n$ with probability one, that is, if for each n there exists a constant $B_n \in (0,\infty)$ such that $\mathbf{P}\{|f(X)-Y| \le B_n\} = 1$. If Y is bounded by a number $B > 0$ with probability one, then this condition is satisfied if the class of functions $\mathcal{F}_n$ is uniformly bounded by some $B_n' < \infty$, that is, for every $f \in \mathcal{F}_n$, and $x \in \mathbb{R}^d$ we have $|f(x)| \le B_n'$. Then $B_n = 2\max\{B, B_n'\}$ is an almost sure bound for $|f(X)-Y|$. Note that in order to get the desired denseness property that is required for the convergence of the approximation error, B_n' has to approach infinity as n grows. The situation is more problematic if Y (and therefore, possibly $m^*(X)$) is not bounded. In this case it is much harder to obtain exponentional probability inequalities for the above supremum. Fortunately, however, in the case of empirical minimization the situation is much nicer. This is asserted by the theorem below. A similar result for estimators based on local averaging was given by Györfi [42]. We briefly comment here on other approaches taken in similar situations. Vapnik [70] developed one-sided inequalities so that nonuniformly bounded function classes may be handled. In a similar setup Shen and Wong [62] used adaptive truncation and large deviation inequalities based on L_2 bracketing metric entropy.

Theorem 1: If

$$\sup_{f\in\mathcal{F}_n} \left| (\mathbf{E}|f(X)-Y|^p)^{1/p} - \left(\frac{1}{n}\sum_{j=1}^n |f(X_j)-Y_j|^p\right)^{1/p} \right| \to 0$$

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

$$J_p(m_n) - \inf_{f\in\mathcal{F}_n} J_p(f) \to 0$$

almost surely for *every* distribution of (X,Y) such that $\mathbf{E}|Y|^p < \infty$.

Proof: Let $L > 0$ be an arbitrary fixed number and introduce the following "truncated" random variables:

$$Y_L = \begin{cases} Y, & \text{if } |Y| \le L \\ L\,\mathrm{sgn}(Y), & \text{otherwise} \end{cases}$$

and

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

for $j = 1, \cdots, n$, where $\mathrm{sgn}(x) = 2I_{\{x\ge 0\}} - 1$. 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 $J_p(f)$ over $\mathcal{F}_n$.

Observe that the triangle inequality implies

$$\begin{aligned} J_p(m_n) &= \left(\mathbf{E}\left(|m_n(X)-Y|^p \middle| D_n\right)\right)^{1/p} \\ &\le \left(\mathbf{E}\left(|m_n(X)-Y_L|^p \middle| D_n\right)\right)^{1/p} + (\mathbf{E}|Y_L-Y|^p)^{1/p} \end{aligned}$$

and similarly

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

Combining the two inequalities above we obtain

$$\begin{aligned} J_p(m_n) - \inf_{f\in\mathcal{F}_n} J_p(f) &= \left(\mathbf{E}\left(|m_n(X) - Y|^p \middle| D_n\right)\right)^{1/p} \\ &\quad - \inf_{f\in\mathcal{F}_n} (\mathbf{E}|Y - f(X)|^p)^{1/p} \\ &\le \left(\mathbf{E}\left(|m_n(X) - Y_L|^p \middle| D_n\right)\right)^{1/p} \\ &\quad - \inf_{f\in\mathcal{F}_n} (\mathbf{E}|Y_L - f(X)|^p)^{1/p} \\ &\quad + 2(\mathbf{E}|Y_L - Y|^p)^{1/p}. \end{aligned} \tag{1}$$

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

$$\left(\mathbf{E}\left(|m_n(X)-Y_L|^p\middle|D_n\right)\right)^{1/p}-\inf_{f\in\mathcal{F}_n}\left(\mathbf{E}|Y_L-f(X)|^p\right)^{1/p}$$

$$=\left(\mathbf{E}\left(|m_n(X)-Y_L|^p\middle|D_n\right)\right)^{1/p}-\left(\frac{1}{n}\sum_{j=1}^{n}|m_n(X_j)-Y_{j,L}|^p\right)^{1/p}$$

$$+\left(\frac{1}{n}\sum_{j=1}^{n}|m_n(X_j)-Y_{j,L}|^p\right)^{1/p}-\inf_{f\in\mathcal{F}_n}\left(\mathbf{E}|Y_L-f(X)|^p\right)^{1/p}$$

$$\le\sup_{f\in\mathcal{F}_n}\left|\left(\mathbf{E}|f(X)-Y_L|^p\right)^{1/p}-\left(\frac{1}{n}\sum_{j=1}^{n}|f(X_j)-Y_{j,L}|^p\right)^{1/p}\right|$$

$$+\left(\frac{1}{n}\sum_{j=1}^{n}|Y_j-Y_{j,L}|^p\right)^{1/p}+\left(\frac{1}{n}\sum_{j=1}^{n}|m_n(X_j)-Y_j|^p\right)^{1/p}$$

$$-\inf_{f\in\mathcal{F}_n}\left(\mathbf{E}|Y_L-f(X)|^p\right)^{1/p}\tag{2}$$

$$\le\sup_{f\in\mathcal{F}_n}\left|\left(\mathbf{E}|f(X)-Y_L|^p\right)^{1/p}-\left(\frac{1}{n}\sum_{j=1}^{n}|f(X_j)-Y_{j,L}|^p\right)^{1/p}\right|$$

$$+\left(\frac{1}{n}\sum_{j=1}^{n}|Y_j-Y_{j,L}|^p\right)^{1/p}+\left(\frac{1}{n}\sum_{j=1}^{n}|\hat{m}_n(X_j)-Y_j|^p\right)^{1/p}$$

$$-\inf_{f\in\mathcal{F}_n}\left(\mathbf{E}|Y_L-f(X)|^p\right)^{1/p}\tag{3}$$

$$\le\sup_{f\in\mathcal{F}_n}\left|\left(\mathbf{E}|f(X)-Y_L|^p\right)^{1/p}-\left(\frac{1}{n}\sum_{j=1}^{n}|f(X_j)-Y_{j,L}|^p\right)^{1/p}\right|$$

$$+2\left(\frac{1}{n}\sum_{j=1}^{n}|Y_j-Y_{j,L}|^p\right)^{1/p}+\left(\frac{1}{n}\sum_{j=1}^{n}|\hat{m}_n(X_j)-Y_{j,L}|^p\right)^{1/p}$$

$$-\inf_{f\in\mathcal{F}_n}\left(\mathbf{E}|Y_L-f(X)|^p\right)^{1/p}\tag{4}$$

$$\le 2\sup_{f\in\mathcal{F}_n}\left|\left(\mathbf{E}|f(X)-Y_L|^p\right)^{1/p}-\left(\frac{1}{n}\sum_{j=1}^{n}|f(X_j)-Y_{j,L}|^p\right)^{1/p}\right|$$

$$+2\left(\frac{1}{n}\sum_{j=1}^{n}|Y_j-Y_{j,L}|^p\right)^{1/p}\tag{5}$$

where (2) and (4) follow from the triangle inequality, while (3) exploits the defining optimality property of m_n. Combining this with (1), and using the strong law of large numbers we get

$$\limsup_{n\to\infty}\left(J_p(m_n)-\inf_{f\in\mathcal{F}_n}J_p(f)\right)$$

$$\le 2\cdot\limsup_{n\to\infty}\left(\sup_{f\in\mathcal{F}_n}\left|\left(\mathbf{E}|f(X)-Y|^p\right)^{1/p}-\left(\frac{1}{n}\sum_{j=1}^{n}|f(X_j)-Y_j|^p\right)^{1/p}\right|\right)$$

$$+4\left(\mathbf{E}|Y_L-Y|^p\right)^{1/p}\quad\text{a.s.}$$

The first term of the right-hand side is zero almost surely by the condition of the theorem, while the second term can be made arbitrarily small by appropriate choice of L; therefore, the proof is completed. ∎

The main message of the above theorem is that we can always assume that Y is bounded (though we cannot assume that the bound is known), in which case using uniformly bounded classes of functions we will be able to derive the desired exponential inequalities.

If $|f(X)-Y|^p\le B_n$ for $f\in\mathcal{F}_n$, then inequalities of the following type can be derived:

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}_n}\left|\mathbf{E}|f(X)-Y|^p-\frac{1}{n}\sum_{j=1}^{n}|f(X_j)-Y_j|^p\right|>\epsilon\right\}\le C(n,\epsilon)e^{-n\epsilon^2c_0/B_n^2}$$

where $C(n,\epsilon)$ is the complexity of the class $\mathcal{F}_n$ expressed in terms of either the *Vapnik–Chervonenkis shatter coefficient* or the *covering number* of the class. We investigate some of these inequalities in the next section.

V. Shatter Coefficients and Covering Numbers

In this section we present some lemmas that will be used to obtain consistency for neural network and generalized linear estimates. As Theorem 1 clearly demonstrates, in order to prove consistency, it suffices to study uniform deviations of averages from their expectations. Let $\mathcal{F}$ be a class of real-valued functions defined on $\mathbb{R}^d$, and let $Z_1,\cdots,Z_n$ be i.i.d., $\mathbb{R}^d$ valued random variables. For our purposes it suffices to assume that functions in $\mathcal{F}$ are nonnegative and uniformly bounded, that is, there is a positive number B such that $0\le f(x)\le B$ for all $x\in\mathbb{R}^d$ and for all $f\in\mathcal{F}$. By Hoeffding's inequality,

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}f(Z_i)-\mathbf{E}f(Z_1)\right|>\epsilon\right\}\le 2e^{-2n\epsilon^2/B^2}$$

for any $f\in\mathcal{F}$. However, we need information about

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(Z_i)-\mathbf{E}f(Z_1)\right|>\epsilon\right\}.$$

Vapnik and Chervonenkis [73] were the first to obtain bounds for the probability above. Our basic tool is an inequality involving the notion of covering numbers defined as follows. Let A be a bounded subset of $\mathbb{R}^d$. For every $\epsilon>0$ the

L_1-covering number, denoted by $N(\epsilon, A)$, is defined as the cardinality of the smallest finite set in $\mathbb{R}^d$ such that for every $z \in A$ there is a point $t \in \mathbb{R}^d$ in this finite set such that $(1/d)\,\|z - t\|_1 \le \epsilon$, where $\|\cdot\|_1$ denotes the l_1 norm in $\mathbb{R}^d$. We will mainly be interested in covering numbers of special sets. Let $z^{(n)} = (z_1, \cdots, z_n)$ be a vector of n fixed points in $\mathbb{R}^d$, and define the following set:

$$\mathcal{F}(z^{(n)}) = \{(f(z_1), \cdots, f(z_n)); f \in \mathcal{F}\} \subset \mathbb{R}^n$$

that is, $\mathcal{F}(z^{(n)})$ is the space of functions in $\mathcal{F}$ restricted to $z_1, \cdots, z_n$. The L_1 covering number of $\mathcal{F}(z^{(n)})$ is $N(\epsilon, \mathcal{F}(z^{(n)}))$. If $Z^{(n)} = (Z_1, \cdots, Z_n)$ is a sequence of i.i.d. random variables, then $N(\epsilon, \mathcal{F}(Z^{(n)}))$ is a random variable. As the next inequality shows, this random variable plays a central role in the theory of uniform large deviations.

Lemma 2 (Pollard, [56]): For any $\epsilon > 0$

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbf{E}f(Z_1)\right| > \epsilon\right\} \le 4\mathbf{E}\Big(N(\epsilon/16, \mathcal{F}(Z^{(n)}))\Big)e^{-n\epsilon^2/128B^2}.$$

Next we recall the concept of the *shatter coefficient* of a class of sets. Let $\mathcal{A}$ be a collection of measurable sets in $\mathbb{R}^d$. For $z_1, \cdots, z_n \in \mathbb{R}^d$, let $N_{\mathcal{A}}(z_1, \cdots, z_n)$ be the number of different sets in

$$\{\{z_1, \cdots, z_n\} \cap A; A \in \mathcal{A}\}$$

and define the shatter coefficient as

$$s(\mathcal{A}, n) = \max_{z_1,\cdots,z_n\in\mathbb{R}^d} N_{\mathcal{A}}(z_1, \cdots, z_n).$$

The shatter coefficient measures, in a sense, the richness of the class $\mathcal{A}$. Clearly, $s(\mathcal{A}, n) \le 2^n$. If $N_{\mathcal{A}}(z_1, \cdots, z_n) = 2^n$ for some $(z_1, \cdots, z_n)$, then we say that $\mathcal{A}$ *shatters* the set $\{z_1, \cdots, z_n\}$. If $s(\mathcal{A}, n) < 2^n$, then there exist n points, such that for some subset of it there is no set in $\mathcal{A}$ that contains exactly that subset of the n points. In other words, $\mathcal{A}$ does not shatter those n points. The largest integer $k \ge 1$ satisfying $s(\mathcal{A}, k) = 2^k$ is denoted by $V_{\mathcal{A}}$, and is called the *Vapnik–Chervonenkis dimension* (or VC dimension) of the class $\mathcal{A}$. If $s(\mathcal{A}, n) = 2^n$ for all n, then by definition, $V_{\mathcal{A}} = \infty$.

First we list a few interesting properties of shatter coefficients $s(\mathcal{A}, n)$ and the VC-dimension $V_{\mathcal{A}}$ of a class of sets $\mathcal{A}$. The following lemma, usually attributed to Sauer [60], describes the relationship between the VC-dimension and shatter coefficients of a class of sets.

Lemma 3 (Sauer [60]): If a class of sets $\mathcal{A}$ has VC-dimension $V_{\mathcal{A}}$, then for every $n \ge V_{\mathcal{A}}$

$$s(\mathcal{A}, n) \le \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}.$$

Lemma 3 has some very surprising implications. Probably the most important is the following corollary.

Corollary 1: If $2 < V_{\mathcal{A}} < \infty$, then for every $n \ge 1$

$$s(\mathcal{A}, n) \le n^{V_{\mathcal{A}}}.$$

This means that for any fixed class $\mathcal{A}$, the shatter coefficients $s(\mathcal{A}, n)$ are either equal to 2^n for every n, or they are bounded by a polynomial in n.

The following is a general, and very useful result. (For the proof see also Pollard [56].)

Lemma 4 (Cover [16]): Let $\mathcal{G}$ be an r-dimensional vector space of real functions on $\mathbb{R}^d$. The class of sets

$$\mathcal{A} = \{\{x : g(x) \ge 0\}; g \in \mathcal{G}\}$$

has VC-dimension $V_{\mathcal{A}} = r$.

Next we discuss properties of covering numbers, and their connection to shatter coefficients of certain classes of sets. The next result is a straightforward extension of inequalities found in Nobel [54], Nolan and Pollard [55], and Pollard [57, p. 22].

Lemma 5: Let $\mathcal{F}_1, \cdots, \mathcal{F}_k$ be classes of real functions on $\mathbb{R}^d$. For arbitrary, fixed points $z^{(n)} = (z_1, \cdots, z_n) \in \mathbb{R}^{dn}$ define the sets $\mathcal{F}_1(z^{(n)}), \cdots, \mathcal{F}_k(z^{(n)})$ in $\mathbb{R}^n$ by

$$\mathcal{F}_j(z^{(n)}) = \{(f(z_1), \cdots, f(z_n)); f \in \mathcal{F}_j\}, \quad j = 1, \cdots, k.$$

Also, let

$$\mathcal{F}(z^{(n)}) = \{(f(z_1), \cdots, f(z_n)); f \in \mathcal{F}\}$$

for the class of functions

$$\mathcal{F} = \{f_1 + \cdots + f_k; f_j \in \mathcal{F}_j, \ j = 1, \cdots, k\}.$$

Then for every $\epsilon > 0$ and $z^{(n)}$

$$N(\epsilon, \mathcal{F}(z^{(n)})) \le \prod_{j=1}^{k} N(\epsilon/k, \mathcal{F}_j(z^{(n)})).$$

Lemma 6 (Pollard [57, p. 23]): Let $\mathcal{F}$ and $\mathcal{G}$ be classes of bounded real functions on $\mathbb{R}^d$, where $|f(x)| \le B_1$ and $|g(x)| \le B_2$ for every $x \in \mathbb{R}^d$, $f \in \mathcal{F}$, and $g \in \mathcal{G}$. For arbitrary, fixed points $z^{(n)} = (z_1, \cdots, z_n) \in \mathbb{R}^{dn}$ define the sets $\mathcal{F}(z^{(n)})$ and $\mathcal{G}(z^{(n)})$ in $\mathbb{R}^n$ as in Lemma 5. Let

$$\mathcal{H}(z^{(n)}) = \{(h(z_1), \cdots, h(z_n)); h \in \mathcal{H}\}$$

for the class of functions

$$\mathcal{H} = \{fg; f \in \mathcal{F}, g \in \mathcal{G}\}.$$

Then for every $\epsilon > 0$ and $z^{(n)}$

$$N\Big(\epsilon, \mathcal{H}(z^{(n)})\Big) \le N\Big(\epsilon/(2B_2), \mathcal{F}(z^{(n)})\Big) \cdot N\Big(\epsilon/(2B_1), \mathcal{G}(z^{(n)})\Big).$$

Now, we recall the notion of packing numbers. Let $\mathcal{F}$ be a class of real-valued functions on $\mathbb{R}^d$, and μ an arbitrary probability distribution on $\mathbb{R}^d$. Let $g_1, \cdots, g_m$ be a finite collection of functions from $\mathcal{F}$ with the property that for any two of them

$$\int_{\mathbb{R}^d} |g_i(x) - g_j(x)|\mu(dx) \ge \epsilon.$$

The largest m for which such a collection exists is called the *packing number* of $\mathcal{F}$ (corresponding to μ), and it is

denoted by $M(\epsilon, \mathcal{F})$. But if μ places $1/n$ probability on each of $z_1, \cdots, z_n$, then $M(\epsilon, \mathcal{F}) = M(\epsilon, \mathcal{F}(z^{(n)}))$, and it is easy to see (e.g., Kolmogorov and Tikhomirov [48]) that

$$M(2\epsilon, \mathcal{F}(z^{(n)})) \le N(\epsilon, \mathcal{F}(z^{(n)})) \le M(\epsilon, \mathcal{F}(z^{(n)})).$$

An important feature of a class of functions $\mathcal{F}$ is the VC-dimension $V_{\mathcal{F}^+}$ of the following class of subsets of $\mathbb{R}^d \times \mathbb{R}$:

$$\mathcal{F}^+ = \{\{(x,t) : t \le f(x)\}; f \in \mathcal{F}\}.$$

This importance is made clear by the following lemma, which is Haussler's [44] result, based on earlier ideas by Dudley [26] and Pollard [56]. It connects the packing number of $\mathcal{F}$ with the VC-dimension of the class of sets $\mathcal{F}^+$.

Lemma 7 (Haussler, [44]):

$$M(\epsilon, \mathcal{F}) \le 2\left(\frac{2eB}{\epsilon}\log\frac{2eB}{\epsilon}\right)^{V_{\mathcal{F}^+}}.$$

The quantity $V_{\mathcal{F}^+}$ is sometimes called the *pseudo-dimension* of $\mathcal{F}$ (see Haussler [44]). It is immediate from Lemma 4 that if $\mathcal{F}$ is a linear space of functions of dimension r, then its pseudo-dimension is r. The following lemma is another property of the pseudo-dimension that will be useful later. It is proved, for example, in Haussler's paper [44]:

Lemma 8 (Nolan and Pollard [55], and Dudley [27]): Let $g : [0, B] \to \mathbb{R}$ be a fixed nondecreasing function, and define the class $\mathcal{G} = \{g \circ f; f \in \mathcal{F}\}$. Then

$$V_{\mathcal{G}^+} \le V_{\mathcal{F}^+}.$$

VI. Series Methods

Our first application of the principles of the previous sections is to the family of linear estimators. Here the estimated function is a linear combination of a certain number of fixed basis functions $\psi_1, \psi_2, \ldots, \psi_{k_n}$. The coefficients are picked to minimize the empirical error. In order to achieve consistency, the number of functions k_n in the linear combination has to grow, as the sample size n grows, but not too rapidly. At the same time, for every n, the possible range of the coefficients in the linear combination has to be restricted as $\sum_{i=1}^{k_n} |a_i| \le \beta_n$, where, again, to obtain consistency, β_n has to grow, but not too rapidly. These estimators are closely related to the Fourier series estimates of a density. These density estimates were studied in the works of Cencov [15], Schwartz [61], Kronmal and Tarter [49], Tarter and Kronmal [66], Specht [63], Greblicki [35], and Greblicki and Pawlak [36], [37], [38]. Series based regression function estimation was investigated by, e.g., Cox [17] and Härdle [43]. The estimate for curve fitting and pattern recognition is also related to the so-called "potential function method" (see Aizerman, Braverman and Rozonoer [1], [2], [3]). Our consistency theorem is formulated as follows:

Theorem 2: Let $p \in [1, \infty)$. Let $\psi_1, \psi_2, \cdots$ be a uniformly bounded sequence of functions such that the set of all finite linear combinations of the ψ_j's

$$\bigcup_{k=1}^{\infty}\left\{\sum_{j=1}^{k} a_j\psi_j(x) : \ a_1, a_2, \cdots, a_k \in \mathbb{R}\right\}$$

is dense in $L_p(\mu)$ for any probability measure μ. Let the coefficients $a_1^*, \cdots, a_{k_n}^*$ minimize the empirical error

$$\frac{1}{n}\sum_{i=1}^{n}\left|\sum_{j=1}^{k_n} a_j\psi_j(X_i) - Y_i\right|^p$$

under the constraint

$$\sum_{i=1}^{k_n}|a_i| \le \beta_n$$

for every $j = 1, \cdots, k_n$, and denote the empirically optimal estimator m_n as

$$m_n(x) = \sum_{j=1}^{k_n} a_j^*\psi_j(x).$$

Then if k_n and β_n satisfy

$$k_n \to \infty, \quad \beta_n \to \infty, \quad \text{and} \quad \frac{k_n\beta_n^{2p}\log(\beta_n)}{n} \to 0$$

then

$$J_p(m_n) - J_p^* \to 0$$

in probability, for all distributions of (X, Y) with $\mathbf{E}|Y|^p < \infty$. If we additionally assume that $\beta_n^{2p} = o(n^{1-\delta})$ for some $\delta > 0$, then $J_p(m_n) - J_p^* \to 0$ almost surely, that is, the estimate m_n is universally consistent.

Proof: We can assume without loss of generality that $|\psi_j(x)| \le 1$ for every $x \in \mathbb{R}^d$ and every j. We apply the usual decomposition into estimation and approximation errors

$$J_p(m_n) - J_p^* = \left(J_p(m_n) - \inf_{f\in\mathcal{F}_n} J_p(f)\right) + \left(\inf_{f\in\mathcal{F}_n} J_p(f) - J_p^*\right).$$

By the denseness assumption and the conditions $\beta_n \to \infty$ and $k_n \to \infty$, the class of functions

$$\bigcup_{n=1}^{\infty}\mathcal{F}_n$$

is dense in $L_p(\mu)$ for any μ by the argument in Section III, where

$$\mathcal{F}_n = \left\{\sum_{j=1}^{k_n} a_j\psi_j; \ \sum_{j=1}^{k_n}|a_j| \le \beta_n\right\}.$$

To show that the estimation error

$$J_p(m_n) - \inf_{f\in\mathcal{F}_n} J_p(f)$$

converges to zero almost surely, we use Theorem 1. By Theorem 1, it is enough to show that if $|Y| \le L$ a.s. for some $L < \infty$, then

$$\sup_{\sum_{j=1}^{k_n}|a_j|\le\beta_n}\left|\left(\frac{1}{n}\sum_{i=1}^{n}\left|\sum_{j=1}^{k_n} a_j\psi_j(X_i) - Y_i\right|^p\right)^{1/p} - \left(\mathbf{E}\left|\sum_{j=1}^{k_n} a_j\psi_j(X) - Y\right|^p\right)^{1/p}\right| \to 0$$

almost surely. This convergence certainly holds if

$$\sup_{\sum_{j=1}^{k_n}|a_j|\le\beta_n}\left|\frac{1}{n}\sum_{i=1}^{n}\left|\sum_{j=1}^{k_n}a_j\psi_j(X_i)-Y_i\right|^p - \mathbf{E}\left|\sum_{j=1}^{k_n}a_j\psi_j(X)-Y\right|^p\right| = \sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}h(X_i,Y_i)-\mathbf{E}h(X,Y)\right|\to 0$$

where the class of functions $\mathcal{H}_n$ is defined by

$$\mathcal{H}_n=\left\{h(x,y)=\left|\sum_{j=1}^{k_n}a_j\psi_j(x)-y\right|^p;\sum_{j=1}^{k_n}|a_j|\le\beta_n\right\}.$$

We can use the inequality $|a+b|^p\le|2\max\{|a|,|b|\}|^p$ and the fact that $|Y|$ is bounded by some L with probability one to see that

$$\begin{aligned}h(X,Y)&=\left|\sum_{j=1}^{k_n}a_j\psi_j(X)-Y\right|^p\\&\le\left|\sum_{j=1}^{k_n}|a_j|+L\right|^p\quad\text{a.s.}\\&\le\left(2\max\left\{\sum_{j=1}^{k_n}|a_j|,L\right\}\right)^p\\&\le 2^p\cdot\max\{\beta_n^p,L^p\}.\end{aligned}$$

Thus we conclude that if $\beta_n^p\ge L^p$, then $0\le h(X,Y)\le 2^p\beta_n^p$ almost surely. Therefore, Lemma 2 asserts that if n is large enough such that $\beta_n\ge L$ is satisfied, then

$$\mathbf{P}\left\{\sup_{h\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^{n}h(X_i,Y_i)-\mathbf{E}h(X,Y)\right|>\epsilon\right\}\le 4\mathbf{E}N\left(\frac{\epsilon}{16},\mathcal{H}_n(Z^{(n)})\right)e^{-n\epsilon^2/(128\cdot 2^{2p}\beta_n^{2p})}$$

where $Z^{(n)}=((X_1,Y_1),\cdots,(X_n,Y_n))$. Next we estimate $N\big((\epsilon/16),\mathcal{H}_n(z^{(n)})\big)$ for any fixed $z^{(n)}$. First consider two functions $h_1(x,y)=|f_1(x)-y|^p$ and $h_2(x,y)=|f_2(x)-y|^p$. Then for any probability measure ν on $\mathbb{R}^d\times[-L,L]$, using the inequality

$$||a|^p-|b|^p|\le p|a-b|\cdot|\max\{a,b\}|^{p-1}$$

we get

$$\begin{aligned}\int|h_1(x,y)-h_2(x,y)|\nu(dx,dy)&=\int||f_1(x)-y|^p-|f_2(x)-y|^p|\nu(dx,dy)\\&\le p(2\beta_n)^{p-1}\int|f_1(x)-f_2(x)|\mu(dx)\end{aligned}$$

where μ is the marginal distribution of ν on $\mathbb{R}^d$. Therefore, for any $z^{(n)}=(x_1,y_1),\cdots,(x_n,y_n)$ and ϵ,

$$N(\epsilon,\mathcal{H}_n(z^{(n)}))\le N\left(\frac{\epsilon}{p(2\beta_n)^{p-1}},\mathcal{F}_n(x^{(n)})\right)$$

where $\mathcal{F}_n$ is the class of functions

$$\mathcal{F}_n=\left\{\sum_{j=1}^{k_n}a_j\psi_j;\sum_{j=1}^{k_n}|a_j|\le\beta_n\right\}.$$

Therefore, it is enough to estimate the covering number corresponding to $\mathcal{F}_n$. But $\mathcal{F}_n$ is a subset of a linear space of functions, and therefore, its pseudo-dimension satisfies $V_{\mathcal{F}^+}\le k_n$ (Lemma 4). So by Lemma 7

$$\begin{aligned}N\left(\frac{\epsilon}{p(2\beta_n)^{p-1}},\mathcal{F}(x^{(n)})\right)&\le 2\left(\frac{e2^{p+1}\beta_n^p}{\epsilon/(p(2\beta_n)^{p-1})}\log\frac{2^{p+1}e\beta_n^p}{\epsilon/(p(2\beta_n)^{p-1})}\right)^{k_n}\\&\le 2\left(\frac{ep2^{2p}\beta_n^{2p-1}}{\epsilon}\right)^{2k_n}.\end{aligned}$$

Therefore, summarizing, we have

$$\begin{aligned}&\mathbf{P}\left\{\sup_{\sum_{j=1}^{k_n}|a_j|\le\beta_n}\left|\frac{1}{n}\sum_{i=1}^{n}\left|\sum_{j=1}^{k_n}a_j\psi_j(X_i)-Y_i\right|^p-\mathbf{E}\left|\sum_{j=1}^{k_n}a_j\psi_j(X)-Y\right|^p\right|>\epsilon\right\}\\&\le 8\left(\frac{ep2^{2p}\beta_n^{2p-1}}{\epsilon/16}\right)^{2k_n}\exp\left[-n\epsilon^2/(128\cdot 2^{2p}\beta_n^{2p})\right]\\&=8\exp\left(2k_n\log\frac{ep2^{2p}\beta_n^{2p-1}}{\epsilon/16}-\frac{n\epsilon^2}{128\cdot 2^{2p}\beta_n^{2p}}\right)\end{aligned}$$

which goes to zero if $(1/n)k_n\beta_n^{2p}\log(\beta_n)\to 0$. It is easy to see that if, in addition, for some $\delta>0$, $\beta_n^{2p}/n^{1-\delta}\to 0$, then strong universal consistency follows by applying the Borel–Cantelli lemma for the last probability. ■

VII. Neural Networks

In our second example we show that it is possible to obtain universally consistent estimators using neural networks. For a limited class of distributions (i.e., for distributions, where both X and Y are of bounded support) White [75] proved L_2-consistency in probability for certain estimators. Almost sure consistency for the same class of distributions can be obtained by using Haussler's [44] results. For a smaller class of distributions, Mielniczuk and Tyrcha [51] obtained L_2-consistency for arbitrary sigmoids. Universal consistency for the pattern recognition problem was shown by Faragó and Lugosi [30] for threshold function networks. Barron [6], [8] used the complexity regularization principle to prove consistency and a rate of convergence for curve fitting by neural networks.

A *neural network of one hidden layer with k hidden neurons* is a real-valued function on $\mathbb{R}^d$ of the form

$$f_{\theta_k}(x)=\sum_{i=1}^{k}c_i\sigma(a_i^Tx+b_i)+c_0$$

where the *sigmoid* $\sigma:\mathbb{R}\to[0,1]$ is a monotone non-decreasing function converging to 0 as $x\to-\infty$ and 1

as $x \to \infty$. $\theta_k = \{a_1, \cdots, a_k, b_1, \cdots, b_k, c_0, \cdots, c_k\}$ is the set of parameters that specify the network $(a_1, \cdots, a_k \in \mathbb{R}^d; b_1, \cdots, b_k, c_0, \cdots, c_k \in \mathbb{R})$. We choose the parameters that minimize the empirical error. However, in order to obtain consistency, again, as in the previous section, we have to restrict the range of some parameters. Here we have to impose some restriction on the c_i's. This is in contrast to results by White [75], [76] where the range of the a_i's and b_i's had to be restricted, too. The next consistency theorem states that with a certain regulation of the parameters c_i and k_n, empirical risk minimization provides universally consistent neural network estimates. We emphasize that we do not have to impose any additional condition on the sigmoid function.

Theorem 3: Define a sequence of classes of neural networks $\mathcal{F}_1, \mathcal{F}_2, \cdots$ as

$$\mathcal{F}_n = \left\{ \sum_{i=1}^{k_n} c_i \sigma(a_i^T x + b_i) + c_0; \right. \\ \left. a_i \in \mathbb{R}^d,\ b_i \in \mathbb{R},\ \sum_{i=1}^{k_n} |c_i| \le \beta_n \right\}$$

and let m_n be a function that minimizes the empirical L_p-error in $\mathcal{F}_n$, i.e.

$$\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^p \le \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^p, \quad \text{if } f \in \mathcal{F}_n.$$

Then if k_n and β_n satisfy

$$k_n \to \infty, \quad \beta_n \to \infty, \quad \text{and} \quad \frac{k_n \beta_n^{2p} \log(k_n \beta_n)}{n} \to 0$$

then

$$J_p(m_n) - J_p^* \to 0$$

in probability, for all distributions of (X, Y) with $\mathbf{E}|Y|^p < \infty$. If, in addition, there exists a $\delta > 0$ such that $\beta_n^{2p}/n^{1-\delta} \to 0$, then $J_p(m_n) - J_p^* \to 0$ almost surely, that is, the estimate m_n is universally consistent.

In order to be able to handle the approximation error, we need a denseness theorem for feedforward neural networks. In 1989, Cybenko [18], Hornik, Stinchcombe, and White [47], and Funahashi [31] proved independently, that feedforward neural networks with one hidden layer are dense with respect to the supremum norm on compact sets in the set of continuous functions. In other words, every continuous function on $\mathbb{R}^d$ can be approximated arbitrarily closely, uniformly over any compact set by functions realized by neural networks. For a survey of such denseness results we refer the reader to Barron [4] and Hornik [46]. Here, as seen in Section III, we need denseness in $L_p(\mu)$ for any probability measure μ.

Lemma 9 (Hornik [45]): For every probability measure μ on $\mathbb{R}^d$, every measurable function $f : \mathbb{R}^d \to \mathbb{R}$ with $\int |f(x)|^p \mu(dx) < \infty$, and every $\epsilon > 0$, there exists a neural network $h(x)$ such that

$$\left(\int |f(x) - h(x)|^p \mu(dx) \right)^{1/p} < \epsilon.$$

Proof of Theorem 3: We can proceed similarly as in the proof of Theorem 2; it is only the estimation of covering numbers that requires additional consideration. It follows from the argument in Section III that the approximation error, $\inf_{f \in \mathcal{F}_n} J_p(f) - J_p^*$, converges to zero as $k_n, \beta_n \to \infty$, if the union of the $\mathcal{F}_n$'s is dense in $L_p(\mu)$ for every μ (Lemma 9).

To handle the estimation error, we use Theorem 1 again, which implies that we can assume $|Y| \le L$ almost surely, for some L, and then we have to show that

$$\sup_{f \in \mathcal{F}_n} \left| (\mathbf{E}|f(X) - Y|^p)^{1/p} - \left(\frac{1}{n}\sum_{j=1}^{n} |f(X_j) - Y_j|^p \right)^{1/p} \right| \to 0.$$

Proceeding exactly as in the proof of Theorem 2 we obtain

$$\mathbf{P}\left\{ \sup_{f \in \mathcal{F}_n} \left| \mathbf{E}|f(X) - Y|^p - \frac{1}{n}\sum_{j=1}^{n} |f(X_j) - Y_j|^p \right| > \epsilon \right\} \\ \le 4\mathbf{E} N\left(\frac{\epsilon}{16p(2\beta_n)^{p-1}}, \mathcal{F}_n(X^{(n)}) \right) e^{-n\epsilon^2/(128 \cdot 2^{2p} \beta_n^{2p})}$$

if $k_n \beta_n \ge L$, so we have to upper-bound the covering number $N(\epsilon, \mathcal{F}_n(x^{(n)}))$. This can be done by applying the series of Lemmas from Section V. Define the following three collections of functions:

$$\mathcal{G}_1 = \{a^T x + b;\ a \in \mathbb{R}^d, b \in \mathbb{R}\}$$
$$\mathcal{G}_2 = \{\sigma(a^T x + b);\ a \in \mathbb{R}^d, b \in \mathbb{R}\}$$
$$\mathcal{G}_3 = \{c\sigma(a^T x + b);\ a \in \mathbb{R}^d, b \in \mathbb{R}, c \in [-\beta_n, \beta_n]\}.$$

By Lemma 4, $V_{\mathcal{G}_1^+} = d + 1$. This implies by Lemma 8 that $V_{\mathcal{G}_2^+} \le d + 1$, so by Lemma 7, for any $x^{(n)}$

$$N(\epsilon, \mathcal{G}_2(x^{(n)})) \le 2\left(\frac{2e}{\epsilon}\right)^{2(d+1)}.$$

Now, Lemma 6 allows us to estimate covering numbers of $\mathcal{G}_3$

$$N(\epsilon, \mathcal{G}_3) \le \frac{4}{\epsilon} N(\epsilon/(2\beta_n), \mathcal{G}_2) \le \left(\frac{4e\beta_n}{\epsilon}\right)^{2d+3}$$

if $\beta_n > 2/e$. Finally, we can apply Lemma 5 to obtain

$$N(\epsilon, \mathcal{F}_n) \le \frac{2\beta_n(k_n+1)}{\epsilon} N(\epsilon/(k_n+1), \mathcal{G}_3)^{k_n} \\ \le \left(\frac{4e(k_n+1)\beta_n}{\epsilon}\right)^{k_n(2d+3)+1}.$$

Thus substituting this bound into the probability inequality above we get for n large enough

$$\mathbf{P}\left\{ \sup_{f \in \mathcal{F}_n} \left| \mathbf{E}|f(X) - Y|^p - \frac{1}{n}\sum_{j=1}^{n} |f(X_j) - Y_j|^p \right| > \epsilon \right\} \\ \le 4\left(\frac{64pe(k_n+1)2^{p-1}\beta_n^p}{\epsilon}\right)^{k_n(2d+3)+1} e^{-n\epsilon^2/(128 \cdot 2^{2p}\beta_n^{2p})}$$

which goes to zero if

$$\frac{k_n \beta_n^{2p} \log(k_n \beta_n)}{n} \to 0$$

and almost sure convergence is guaranteed by the Borel–Cantelli lemma if the additional condition on $k_n \beta_n$ holds. ■

VIII. Concluding Remarks

In this paper we developed some general tools for proving universal consistency in a very strong sense for estimators based on empirical risk minimization. We demonstrated the usefulness of the tools for two basic examples, namely, we established consistency of generalized linear estimators and neural network estimators. Finally, some remarks are in order.

Remark 4 (Pattern Recognition): By the discussion in Remarks 2 and 3 we see that L_1 and L_2 consistency imply consistency in error probability for classification functions. This means, using Theorems 2 and 3, that minimizing the empirical L_1 or L_2 errors lead to consistent generalized linear, or neural network classifiers. However, in order to obtain good classifiers, it may seem more natural to pick a classification rule that minimizes the empirical error probability, that is, the mean number of errors

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

from a class of classifiers $\mathcal{G}_n$. Indeed, it is easy to show examples, where if the class from which we pick is fixed, then minimizing an L_p-error yields much worse classifiers then minimizing the empirical error probability, even though, consistency can be obtained by appropriately increasing the class of functions. The method of minimizing the empirical error probability was extensively studied by Devroye [20], including series methods. For neural network classifiers Faragó and Lugosi [30] proved its consistency. Note that if the class of functions $\mathcal{F}_n$ contains binary-valued functions only, as in Vapnik's book [70], then the two methods are equivalent, but our methods of proving consistency do not work in that case.

Remark 5 (Algorithms): An important reason why minimizing the L_2-error is much more popular in practice than minimizing the empirical error probability for classification, is that usually it is algorithmically much simpler. For example, for series methods *stochastic approximation* algorithms are available. If the dimension of the generalized linear classifier k_n is fixed, then stochastic approximation asymptotically provides the minimizing coefficients. For more information about these methods we refer to Robbins and Monro [58], Aizerman, Braverman, and Rozonoer [1]–[3], Fabian [28], Györfi [41], as well as Ljung, Pflug, and Walk [50]. Györfi [40] introduced an algorithm for minimizing the L_1-error. Similarly, for training neural networks, attempting to minimize the squared error is the most widely used principle, mainly using the back-propagation method (see Rumelhart, Hinton, and Williams [59], White [74], Fabian [29]).

Remark 6 (Rates of Convergence): We have considered the problem of distribution-free almost sure convergence of estimators, but not how fast the error of these estimators converges. Devroye [19] proved that there is no universal rate of convergence in pattern recognition, that is, there is no classifier whose error probability converges to the Bayes-risk at a certain rate for every possible distribution. By the inequalities in Remark 2 and Lemma 1, Devroye's theorem applies for L_1 and L_2 consistent estimators, too. Therefore, without imposing additional assumptions on the joint distribution of (X, Y), there is no hope to obtain upper bounds for the rate of convergence. It is relatively straightforward to obtain upper bounds for the rate of convergence for the estimation error from our analysis, if one assumes some tail conditions for the distribution of Y. Analysis of the rate of convergence of the approximation error is usually more involved. One typically has to take a closer look at the approximation properties of $\mathcal{F}_n$ for the class of functions in which m^* can lie, under the assumptions imposed on the distribution. For series methods these types of results can be found among results of classical approximation theory, while more recently, some remarkable approximation properties of neural networks have been explored by Barron [7]. To obtain upper bounds for the overall error one has to choose the parameters of $\mathcal{F}_n$ to balance the tradeoff between the approximation and estimation errors.

Acknowledgment

The authors wish to thank two anonymous referees and the associate editor, A. Barron, for useful comments and for bringing relevant references to our attention.

References

[1] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "The method of potential functions for the problem of restoring the characteristic of a function converter from randomly observed points," *Automat. Remote Contr.*, vol. 25, pp. 1546–1556, 1964.

[2] ——, "The probability problem of pattern recognition learning and the method of potential functions," *Automat. Remote Contr.*, vol. 25, pp. 1307–1323, 1964.

[3] ——, "Theoretical foundations of the potential function method in pattern recognition learning," *Automat. Remote Contr.*, vol. 25, pp. 917–936, 1964.

[4] A. R. Barron, "Statistical properties of artificial neural networks," in *Proc. 28th Conf. on Decision and Control* (Tampa, FL, 1989.

[5] ——, "Approximation and estimation errors for artificial neural networks," in *Computational Learning Theory: Proc. 4th Annual Workshop.* Morgan Kaufman, 1991.

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

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

[8] ——, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, pp. 115–133, 1994.

[9] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1034–1054, 1991.

[10] L. Birgé and P. Massart, "Rates of convergence for minimum contrast estimators," *Prob. Theory Related Fields*, vol. 97, pp. 113–150, 1993.

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

[12] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees.* Belmont, CA: Wadsworth Int., 1984.

[13] K. L. Buescher and P. R. Kumar, "Learning by canonical smooth estimation, Part I: Simultanious estimation," submitted to *IEEE Trans. Automat. Contr.*, 1994.

[14] ——, "Learning by canonical smooth estimation, Part II: Learning and choice of model complexity," submitted to *IEEE Trans. Automat. Contr.*, 1994.

[15] N. N. Cencov, "Evaluation of an unknown distribution density from observations," *Sov. Math.–Dokl.*, vol. 3, pp. 1559–1562, 1962.

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

[17] D. D. Cox, "Approximation of least squares regression on nested subspaces," *Annals Statist.*, vol. 16, pp. 713–732, 1988.

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

[19] L. Devroye, "Any discrimination rule can have an arbitrarily bad probability of error for finite sample size," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-4, pp. 154–157, 1982.
[20] L. Devroye, "Automatic pattern recognition: A study of the probability of error," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, pp. 530–543, 1988.
[21] L. Devroye and L. Györfi, "Distribution-free exponential bound on the L_1 error of partitioning estimates of a regression function," in F. Konecny, J. Mogyoródi, and W. Wertz, Eds., *Proc. 4th Pannonian Symp. on Mathematical Statistics*. Budapest, Hungary: Akadémiai Kiadó, 1983, pp. 67–76.
[22] L. Devroye, L. Györfi, A. Krzyżak, and G. Lugosi, "On the strong universal consistency of nearest neighbor regression function estimates," to appear in the*Annals Stat.*, Sept. 1994.
[23] L. Devroye and A. Krzyżak, "An equivalence theorem for L_1 convergence of the kernel regression estimate," *J. Statist. Planning and Inference*, vol. 23, pp. 71–82, 1989.
[24] L. Devroye and T. J. Wagner, "Nonparametric discrimination and density estimation," Tech. Rep. 183, Electron. Res. Cen., Univ. of Texas, 1976.
[25] ——, "Distribution-free consistency results in nonparametric discrimination and regression function estimation," *Annals Statist.*, vol. 8, pp. 231–239, 1980.
[26] R. M. Dudley, "Central limit theorems for empirical measures," *Annals Probab.*, vol. 6, pp. 899–929, 1978.
[27] ——, "Universal Donsker classes and metric entropy," *Annals Probab.*, vol. 15, pp. 1306–1326, 1987.
[28] V. Fabian, "Stochastic approximation," in J. S. Rustagi, Ed., *Optimizing Methods in Statistics*. New York, London: Academic Press, 1971, pp. 439–470.
[29] ——, "On neural network models and stochastic approximation," preprint, 1992.
[30] A. Faragó and G. Lugosi," "Strong universal consistency of neural network classifiers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1146–1151, 1993.
[31] K. Funahashi,"On the approximate realization of continuous mappings by neural networks," *Neural Net.*, vol. 2, pp. 183–192, 1989.
[32] A. R. Gallant, *Nonlinear Statistical Models*. New York: Wiley, 1987.
[33] S. Geman and C. R. Hwang, "Nonparametric maximum likelihood estimation by the method of sieves," *Annals Statist.*, vol. 10, pp. 401–414, 1982.
[34] N. Glick, "Sample-based multinomial classification," *Biometrics*, vol. 29, pp. 241–256, 1973.
[35] W. Greblicki, "Asymptotic efficiency of classifying procedures using the Hermite series estimate of multivariate probability densities," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 364–366, 1981.
[36] W. Greblicki and M. Pawlak, "Classification using the Fourier series estimate of multivariate density functions," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-11, pp. 726–730, 1981.
[37] ——, "A classification procedure using the multiple Fourier series," *Inform. Sci.*, vol. 26, pp. 115–126, 1982.
[38] ——, "Almost sure convergence of classification procedures using Hermite series density estimates," *Pattern Recogn. Lett.*, vol. 2, pp. 13–17, 1983.
[39] U. Grenander, *Abstract Inference*. New York: Wiley, 1981.
[40] L. Györfi, "An upper bound of error probabilities for multihypothesis testing and its application in adaptive pattern recognition," *Probl. Contr. and Inform. Theory*, vol. 5, pp. 449–457, 1975.
[41] ——, "Adaptive linear procedures under general conditions," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 262–267, 1984.
[42] ——, "Universal consistencies of a regression estimate for unbounded regression functions," in G. Roussas, Ed., *Nonparametric Functional Estimation and Related Topics* (NATO ASI Series). Dordrecht, The Netherlands: Kluwer, 1991, pp. 329–338.
[43] W. Härdle, *Applied Nonparametric Regression*. Cambridge, UK: Cambridge Univ. Press, 1990.
[44] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Inform. and Comput.*, vol. 100, pp. 78–150, 1992.
[45] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Net.*, vol. 4, pp. 251–257, 1991.
[46] ——, "Some new results on neural network approximation," *Neural Net.*, vol. 6, pp. 1069–1072, 1993.
[47] K. Hornik, M. Stinchcombe, and H. White, "Multi-layer feedforward networks are universal approximators," *Neural Net.*, vol. 2, pp. 359–366, 1989.
[48] A. N. Kolmogorov and V. M. Tikhomirov, "ϵ-entropy and ϵ-capacity of sets in function spaces," *Transl. Amer. Math. Soc.*, vol. 17, pp. 277–364, 1961.
[49] R. A. Kronmal and M. E. Tarter, "The estimation of probability densities and cumulatives by Fourier series methods," *J. Amer. Statist. Assoc.*, vol. 63, pp. 925–952, 1968.
[50] L. Ljung, G. Pflug, and H. Walk, *Stochastic Approximation and Optimization of Random Systems*. Basel, Boston, Berlin: Birkhäuser, 1992.
[51] J. Mielniczuk and J. Tyrcha, "Consistency of multilayer perceptron regression estimators," *Neural Net.*, to appear, 1993.
[52] A. S. Nemirovskiy, B. T. Polyak, and A. B. Tsybako, "Rate of convergence of nonparametric estimators of maximum-likelihood type," *Probl. Inform. Transmission*, vol. 21, pp. 258–272, 1985.
[53] A. S. Nemirovski, "Nonparametric estimation of smooth regression functions," *Eng. Cybern.*, vol. 23, no. 6, pp. 1–11, 1985.
[54] A. B. Nobel, "On uniform laws of averages," Ph.D. dissertation, Dep. Statist., Stanford Univ., Stanford, CA, 1992.
[55] D. Nolan and D. Pollard, "U-processes: Rates of convergence," *Annals Statist.*, vol. 15, pp. 780–799, 1987.
[56] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984.
[57] ——, *Empirical Processes: Theory and Applications* (NSF-CBMS Regional Conference Series in Probability and Statistics, Hayward, CA, Alexandria, VA, 1990).
[58] H. Robbins and S. Monro, "A stochastic approximation method," *Annals Math. Stat.*, vol. 22, pp. 400–407, 1951.
[59] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in D. E. Rumelhart, J. L. McCelland, and the PDP Research Group, Eds., *Parallel Distributed Processing, Vol. I*. Cambridge, MA: M.I.T. Press, 1986.
[60] N. Sauer, "On the density of families of sets," *J. Combinatorial Theory Ser. A*, vol. 13, pp. 145–147, 1972.
[61] S. C. Schwartz, "Estimation of probability density by an orthogonal series," *Annals Math. Stat.*, vol. 38, pp. 1261–1265, 1967.
[62] X. Shen and W. H. Wong, "Convergence rate of sieve estimates," to appear in the *Annals Statist.*, vol. 22. pp. 580–615, June 1994.
[63] D. F. Specht, "Series estimation of a probability density function," *Technometrics*, vol. 13, pp. 409–424, 1971.
[64] C. Spiegelman and J. Sacks, "Consistent window estimation in nonparametric regression," *Annals Stat.*, vol. 8, pp. 240–246, 1980.
[65] C. J. Stone, "Consistent nonparametric regression," *Annals Stat.*, vol. 8, pp. 1348–1360, 1977.
[66] M. E. Tarter and R. A. Kronmal, "On multivariate density estimates based on orthogonal expansions," *Annals Math. Stat.*, vol. 41, pp. 718–722, 1970.
[67] L. G. Valiant, "A theory of the learnable," *Commun. ACM*, vol. 27, pp. 1134–1142, 1984.
[68] S. Van de Geer, "Estimating a regression function," *Annals Stat.*, vol. 18, pp. 907–924, 1990.
[69] J. Van Ryzin, "Bayes risk consistency of classification procedures using density estimation," *Sankhya* , ser. A, vol. 28, pp. 161–170, 1966.
[70] V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data*. New York: Springer-Verlag, 1982.
[71] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. and Applic.*, vol. 16, pp. 264–280, 1971.
[72] ——, *Theory of Pattern Recognition*. Moscow, USSR: Nauka, 1974 (in Russian); German translation: *Theorie der Zeichenerkennung*. Berlin, Germany: Akademie-Verlag, 1979.
[73] ——, "Necessary and sufficient conditions for the uniform convergence of means to their expectations," *Theory Prob. and Appl.*, vol. 26, pp. 821–832, 1981.
[74] H. White, "Some asymptotic results for learning in single hidden-layer feedforward network models," *J. Amer. Statist. Assoc.*, vol. 84, pp. 1003–1013, 1989.
[75] ——, "Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings," *Neural Net.*, vol. 3, pp. 535–549, 1990.
[76] ——, "Nonparametric estimation of conditional quantiles using neural networks," in *Proc. 23rd Symp. of the Interface: Computing Science and Statistics*, 1991.
[77] C. T. Wolverton and T. J. Wagner, "Asymptotically optimal discriminant functions for pattern classification," *IEEE Trans. Syst., Sci., Cybern.*, vol. 15, pp. 258–265, 1969.
[78] W. H. Wong and X. Shen, "Probability inequalities for likelihood ratios and convergence rates of sieve MLE's, Tech. Rep. 346, Dep. Stat., University of Chicago, Chicago, IL, 1992.