Proceedings of the 28th Conference
on Decision and Control
Tampa, Florida • December 1989

STATISTICAL PROPERTIES OF ARTIFICIAL NEURAL NETWORKS

Andrew R. Barron [1]

Department of Electrical and Computer Engineering
and Department of Statistics
University of Illinois
Champaign, Illinois 61820

ABSTRACT

Convergence properties of empirically estimated neural networks are examined. In this theory, an appropriate size feedforward network is automatically determined from the data. The networks we study include two and three layer networks with an increasing number of simple sigmoidal nodes, multiple layer polynomial networks, and networks with certain fixed structures but an increasing complexity in each unit. Each of these classes of networks is dense in the space of continuous functions on compact subsets of d-dimensional Euclidean space, with respect to the topology of uniform convergence. In this talk we show how, with the use of an appropriate complexity regularization criterion, the statistical risk of network estimators converges to zero as the sample size increases. Bounds on the rate of convergence are given in terms of an index of the approximation capability of the class of networks.

I. INTRODUCTION

A variety of learning network methods for empirical classification and function fitting have a common framework as given in [1]. These methods include traditional neural network models (e.g., as represented in [2], [3]), polynomial learning networks [4], some nonparametric statistical techniques such as projection pursuit [5,6,7], and new generalizations of these techniques as described in [1]. In this presentation we show that the statistical risk of sequences of network estimators converges to zero, as the size of the training sample increases. Given a class of network structures, a complexity regularization criterion is used to estimate the appropriate size and, in some cases, to estimate the connectivity as well as the coefficients of the network. Bounds on the statistical rate of convergence are given in terms of an index of the approximation capability of the given class of networks.

Although the statistical convergence theory can be applied to any collection of functions to be optimized in accordance with the criterion, we focus here on functions represented as feedforward networks. Networks express functions $f(x_1,...,x_d)$ of several variables as a composition of basic functions (also called units, network nodes, or sometimes artificial neurons). The units of the network are parameterized as a nonlinear transformation of a linear combination of many variables, i.e., $g(z,\theta) = h(\sum\theta_j z_j+\theta_0)$, or as a linear combination of nonlinear transformations of a few variables, i.e., $g(z,\theta) = \sum\theta_j\phi_j(z)$. Important special cases include sigmoidal units, where h is a fixed bounded continuous nondecreasing function , and polynomial units, where the $\phi_j(z)$ are basis functions for the polynomials of a given degree. In addition, networks often include one or more units which simply take a linear combination of the specified inputs. Here z refers to the vector of variables which are input to a given node. The network diagram is a tree which specifies the composition, i.e., for each node it identifies which preceding units and/or which original input variables are input to the node. The layers of the network are the sets of functions which occupy the same depth in the tree, and the width is the number of nodes in a layer.

In Section II some basic approximation properties of networks are stated. The criterion used for network estimation is developed in Section III, followed by the main result on the statistical convergence of estimated networks in Section IV. Implications for empirical classification are in Section V.

II. NETWORK APPROXIMATION

Some basic approximation properties of networks are known and summarized in the following.

Network Approximation Theorem: *The following classes of networks are dense in the space of continuous functions on compact subsets of d-dimensional Euclidean space, in the sense that for any such function f^* there exists a sequence of network functions f_n that converges uniformly to f^*.*

(1) The class of two-layer networks with an unrestricted number of sigmoidal units on the first layer and a linear unit on the second layer. These networks implement functions of the form $\sum_j\beta_j h(\sum_{k=1}^d\alpha_{jk}x_k+\alpha_{j,0})$ where h is any fixed sigmoidal function.

(2) The class of three-layer networks with linear units on the first layer, polynomial units in one variable on the second layer, and a linear unit on the third layer. These networks implement functions of the form $\sum_j\beta_j g_j(\sum_{k=1}^d\alpha_{jk}x_k)$ where the g_j are polynomials in one variable.

(3) The class of four-layer networks with polynomial units in one variable on the first and third layers, linear units on the second and fourth layers, and the structure $\sum_{j=1}^{2d+1}\beta_j g_j(\sum_{k=1}^d\alpha_{jk}g_{jk}(x_k))$, which has a fixed number of units.

(4) The class of polynomial networks, unrestricted in the number of layers, where the degree and the number of inputs to each unit may be constrained to be as small as 2. Constraints may also be imposed in the width of the layers, provided the inputs to a layer are permitted to emanate from any preceding layer or from original input variables. These networks implement arbitrary polynomials in d variables.

[1]Work supported in part by an Office of Naval Research grant N00014-89-J-1811 and by a NSF Postdoctoral Research Fellowship.

CH2642-7/89/0000-0280$1.00 © 1989 IEEE

Remarks: Case (1) is proved in Cybenko [8], where a more general form for the functions h is permitted. In the proof for case (3) we use a result of Kolmogorov as described in Lorentz [9] on the exact representation of functions using compositions, together with the Weierstrass Theorem. The proof for cases (2) and (4) involves a more direct consideration of polynomials.

For each of the above classes of networks, it is seen that there are countably many possible structures or families of networks $F = \{f(\cdot \mid \theta) : \theta \in \mathbf{R}^k\}$, each of which depends continuously on the parameters θ, and these families can be enumerated in a systematic way $F_1, F_2, \ldots$ with increasing dimensions $k_1, k_2, \ldots$.

The Theorem illustrate a range of options in the tradeoff between the number and the complexity of units. In particular, cases (1) and (4) use arbitrarily many units of low complexity, while case (3) uses a fixed number of units of unrestricted complexity.

In cases (2) and (3), spline basis functions for an arbitrarily fine grid of knots may be used in place of the one-dimensional polynomials in the Theorem. Splines have certain advantage over polynomials near the boundaries of their domains. However, splines have the disadvantage that for very smooth functions the rates of approximation saturate at a slower convergence rate than with polynomial approximations.

A consequence of the Theorem is that each of the given classes of functions is also dense in the space of square-integrable functions on compact subsets of $\mathbf{R}^d$ in the sense that for any such f^* there exists a sequence of network functions f_n such that $\lim \int (f_n - f^*)^2 = 0$. This follows by using the fact that functions in L^2 are arbitrarily well approximated by bounded continuous functions. A similar conclusion also holds for other loss functions which are weaker than the L^∞ norm.

Larger classes of networks than considered in the Theorem can be obtained by removing restrictions on the number of layers or the type of units. Of course, such larger classes of networks are also dense. In some cases, enlarging the class of candidate networks may improve the rates of approximation.

The approximation rate in L^2, using polynomials (or sufficiently high order splines) which are written as sums of products of the one dimensional basis functions, is known to be of order $(1/m)^r$ as $m \to \infty$, uniformly over all functions f^* for which the rth order partial derivatives have L^2 norm bounded by a constant, where m is the degree of the polynomial approximation in each coordinate (see, e.g., [10], [11]). The total number of terms in this polynomial is $(m+1)^d$. Polynomial networks can be arranged such the same approximation rates are obtained using a network of order m^d parameters. However, in high dimensions the exponential growth of the number of parameters with the dimension d precludes the use of approximations based on the traditional expansions. The attractiveness of the network methods is that, in practice, function in high dimensions are often closely approximated by compositions of lower dimensional smooth functions. The aim then is to search for an appropriate composition within a given class of networks.

III. PERFORMANCE CRITERIA

Before giving our convergence result for statistically estimated networks, we address the issues of the choice of the loss function and the estimation criterion.

Let $(X_i, Y_i)_{i=1}^n$ be independent observations drawn from the unknown joint distribution of random variables X, Y, where the support of X is in $\mathbf{R}^d$. Let $\hat{f}_n$ denote a network function estimated from this data. Given a distortion or loss function $d(Y, f(X))$, the network $\hat{f}_n$ is typically chosen by attempting to minimize the empirical average loss $(1/n)\sum_{i=1}^n d(Y_i, f(X_i))$, or by attempting to minimize the empirical loss with a penalty added for the complexity of the network functions.

Loss Functions

For a given loss or distortion function $d(Y, f(X))$, we let f^* be the function which minimizes the expected distortion $Ed(Y, f(X))$ over all measurable functions on $\mathbf{R}^d$. For the loss functions we investigate, such a optimum function $f^*(x)$ exists and is related explicitly to the conditional distribution of Y given $X = x$. The difference between the expected distortion at a choice f and at the optimal f^* is denoted

$$r(f, f^*) = Ed(Y, f(X)) - Ed(Y, f^*(X)).$$

For network estimators $\hat{f}_n$, the expected value of $r(\hat{f}_n, f^*)$ is the statistical risk that we desire to bound. Note that $r(\hat{f}_n, f^*)$ quantifies the ability of the estimate to generalize, on the average, to new data from the distribution of X, Y.

The most common choice for the loss function $d(Y, f(X))$ is the squared error $(Y - f(X))^2$, although other choices can also be handled in the theory. In the case of a dichotomous random variable Y, assumed to take values in $\{-1, +1\}$ as is traditional for neural networks, other reasonable loss functions include the zero-one loss function $1_{\{Y \neq \operatorname{sgn}(f(X))\}}$, and the logistic loss function $-Yf(X) + \log(e^{f(X)} + e^{-f(X)})$. A general class of loss functions are those which take the form $d(Y, f(X)) = -\log p(Y \mid f(X))$ where $p(y \mid f(x))$ models the conditional density of Y given X. In particular, the squared error loss corresponds to a Gaussian conditional density and the logistic loss corresponds to the Bernoulli model written in exponential form $p(y \mid x) = e^{yf(x)}/(e^{f(x)} + e^{-f(x)})$, $y = \pm 1$, where $f(x)$ models one-half the log-odds ratio in favor of class $+1$ versus class -1. The logistic loss function is the same, except for a linear rescaling, as the conditional log-likelihood function used in traditional logistic regression.

For the squared error loss, $r(f, f^*)$ reduces to $E((f(X) - f^*(X))^2)$ and $f^*(x) = E[Y \mid X = x]$. In particular, in the dichotomous case, $f^*(x)$ is the difference between the conditional probabilities of class $+1$ and class -1. For the zero-one loss, $r(f, f^*)$ is the difference between the probability of error based on f and the Bayes optimal probability of error. For loss functions based on a family of conditional densities, f^* is the choice which makes the conditional density $p(y \mid f^*(x))$ closest to the true conditional density in the relative entropy sense. If the family of conditional densities is correctly specified, i.e., if it includes the true conditional density, then $r(f, f^*)$ is the average relative entropy distance between $p(\cdot \mid f^*(x))$ and $p(\cdot \mid f(x))$.

There is not agreement as to the best loss function for networks even in the dichotomous case. The zero-one loss, the squared error loss, and the loss function based on the Bernoulli model may all seem reasonable. However, there are clear computational difficulties in the zero-one case due to the lack of differentiability of the empirical loss. For the squared error loss with dichotomous Y, since $f^*(x)$ is necessarily in the range $[-1,1]$, it is suggested that the output of the network be clipped so as not to take values outside this range. This is sometimes accomplished by directly incorporating a sigmoidal transformation in the final node of the network (although such a transformation of the final node is not used in the network structures in the four cases of the approximation theorem above). However, the presence of such a sigmoidal transformation or clipping leads to the unfortunate property that the objective function is not convex, and often not unimodal, as a function of the parameters of the final unit of the network (indeed this lack of unimodality of the objective function persists even for a networks with a single sigmoidal unit and the squared error loss function). A natural recommendation for the dichotomous case is to use the logistic loss function, for which the range of $f(x)$ may be unrestricted on the real line, and to apply the logistic loss function to the linear combination in the final unit of the network, without first transforming this linear combination. This choice of loss function makes the objective function convex in the parameters of the final unit. (Subsequent to the training step, the logistic transformation of the output $e^{f(X)}/(e^{f(X)}+e^{-f(X)})$ may then be used to estimate the conditional probability of class 1 given X, or it may taken as a useful transformation to be fed to additional layers in the evolution of the network.) The point is that since multiple-layer networks have an ever-present multimodality problem, it is best not to compound the difficulty by a problematic choice of the loss function.

Concerning the optimization of criteria based on the empirical loss function, the convexity of the loss as a function of the parameters in a given unit has clear advantages for adaptive synthesis strategies which build up a network one unit at a time. This convexity for single units permits the successful use of Gauss-Newton or other derivative-based search routines in an adaptive synthesis program. Adaptive synthesis may be regarded as an approach to initializing the structure and parameters of a network for subsequent fine-tuning of all the parameters of the network. Examples of practical adaptive synthesis algorithms for networks include ASPN for polynomial networks (see [1]), related GMDH algorithms [4], projection pursuit [5,6,7], and some extensions of these methods introduced in [1]. Combinations of global random and local derivative-based search strategies have also been successful in some practical applications, see [4, Chapter 2]. In practice, it appears that adaptive synthesis and guided random search strategies achieve near optimum levels of performance. However, a gap exists between the theory and practice of network estimation, because of the difficulty in guaranteeing global optimization of the multimodal performance surfaces intrinsic to multiple-layer nonlinear networks. In the theory we develop below, we do not bridge this gap. Instead, we examine statistical convergence properties assuming that a criterion is globally optimized over a sequence of candidate families of networks.

Complexity Regularization

Since the size and structure of a network are to be estimated as well as the parameters, a key concern is to use a criterion which will avoid the problem of statistical overfit to the observed data, which would be the inevitable result of unconstrained minimization of the empirical loss. A common but not fully satisfactory way to sidestep the overfit problem is to restrict the optimization to a family of network functions of given complexity. The difficulty with imposing such constraints is that it is generally not possible to determine a priori the appropriate size network. Another approach is to use one of several related criteria (cross validation, Mallows C_p, Akaike's AIC, or the predicted squared error [4, Chapter 4]) which have been shown to have certain asymptotic optimality properties for selection problems with nested linear models, see [12],[13]. Experimentally, there have been numerous practical successes in selecting network structures by the use of the predicted squared error criterion (see [4, Chapter 2]). However, it is not yet clear whether the theory for nested linear models will carry over to the general network estimation context, even under smoothness assumptions. The approach which we examine here is an extension of the minimum description length criterion [14,15], for which we are able to obtain convergence theory in the network estimation context. A complexity-based penalty is added to the empirical loss, which more severly penalizes overly complex networks than does the AIC or predicted squared error criteria. Here we motivate the criterion in the context of bounds on the statistical risk of network estimators.

There are two contributions to the risk $E(r(\hat{f}_n,f^*))$ of a network estimator: namely, the approximation error $r(f,f^*)$ achieved by network functions f in the given class as an approximation to the desired function f^*, and the estimation error, which is due to the discrepancy between the empirical average and the theoretical average of the loss function for the estimated network. By techniques in [16], [17], or [18], this discrepancy between empirical and theoretical averages can be shown to be uniformly bounded by $O(\sqrt{C_n/n})$, in probability, for families of networks of complexity bounded by C_n. (Essentially, C_n is taken there as the logarithm of the number of functions required to approximate functions in the class to within a prescribed accuracy.) By generalizations of these techniques and by the techniques we recently used in [15], it is shown that the estimation error can be bounded in probability by a multiple of $\sqrt{C_n(f)/n}$ for arbitrary bounded loss functions and by $r(f,f^*) + C_n(f)/n$ for the squared error loss and for the likelihood based loss functions, uniformly for all candidate networks f, where instead of requiring a uniform complexity bound, we permit unbounded complexities $C_n(f)$ that may depend on the candidate networks. (These "complexities" $C_n(f)$ are arbitrary numbers satisfying a summability requirement, as given in equation (5) below, in accordance with an information-theoretic interpretation.) Thus we are led to complexity regularization criteria and to corresponding indices of approximation. Depending on whether bounds of order $\sqrt{C_n(f)/n}$ or $C_n(f)/n$ arise in controlling the estimation error, we add the appropriate complexity penalty to the empirical loss to define the criterion for network estimation, for in this way it seen that the minimizer of the empirical criterion has a performance essentially as good as that achievable by the theoretical analog of the criterion.

Definition: Given a collection Γ_n of network functions, numbers $C_n(f)$, $f \in \Gamma_n$, satisfying the summability condition (5), and a positive constant λ, the method of *complexity regularization* chooses the network estimate to minimize one of the following two criteria

$$\frac{1}{n}\sum_{i=1}^{n} d(Y_i, f(X_i)) + \lambda\left(\frac{1}{n}C_n(f)\right)^{1/2} \tag{1}$$

or

$$\frac{1}{n}\sum_{i=1}^{n} d(Y_i, f(X_i)) + \lambda\frac{1}{n}C_n(f). \tag{2}$$

The theoretical analog of these criteria leads to the following indices of approximation, which quantify the tradeoff between the complexity and accuracy of approximations to f^*,

$$R_n^{(1)}(f^*) = \min_{f\in\Gamma_n}\left(r(f, f^*) + \lambda\frac{1}{n}C_n(f)\right)^{1/2} \tag{3}$$

and

$$R_n^{(2)}(f^*) = \min_{f\in\Gamma_n}\left(r(f, f^*) + \lambda\frac{1}{n}C_n(f)\right). \tag{4}$$

The latter quantity is the index of resolvability introduced in an information-theoretic context in [15].

The requirement that is imposed on the numbers $C_n(f)$ is the following summability condition:

$$\sum_{f\in\Gamma_n} e^{-C_n(f)} \le c_0, \tag{5}$$

for some finite constant c_0. There is an information-theoretic interpretation of a related summability condition, with $c_0 = 1$: this is the Kraft-McMillan inequality which is necessary and sufficient for the existence of uniquely decodable binary codes, with codelengths $C_n(f)\log_2 e$, for $f \in \Gamma_n$.

The countable collection Γ_n is typically chosen such that the networks in every family in the class of interest are accurately represented. Although the theorems below apply with any choice for Γ_n, a particularly reasonable choice, given a sequence of network structures $F_1, F_2, \ldots$, is to take Γ_n to consist of the networks in each F_j for which the parameter values are in the regular grid of points spaced at width $1/\sqrt{n}$ in each coordinate. Corresponding to this choice we can arrange to satisfy the summability condition by taking $C_n(f)$ to be

$$C_n(f) = \frac{k}{2}\log n + c(f), \tag{6}$$

where k is the number of parameters in the network structure F_j that contains f, and $c(f)$ is independent of n. For network functions which depend smoothly on the parameters, these choices for Γ_n and $C_n(f)$ achieve roughly the best tradeoff between complexity and accuracy of approximations in each given family, as can be shown using the techniques in [14], or [15,Sec.6]. A more refined analysis as in [15], suggests advantages of certain variable-width spacings of the parameter values for functions in Γ_n; nevertheless, the best asymptotic density of points per unit volume remains of order $n^{k/2}$, which leads to $C_n(f) = (k/2)\log n + O(1)$, and all such choices will lead to the same rates of convergence for the indices of approximation. Here we will be content to use equal spacings.

There is freedom in the choice of $c(f)$ subject to the summability requirement. One method for assigning $c(f)$ is to use prior probabilities $P(j) > 0$, $\sum_j P(j) = 1$, assigned to the network structures F_j, $j = 1,2,\ldots$, and prior distributions W_j, with continuous density functions $w_j(\theta)$, assigned to the parameter spaces for each network structure. Then we may set $c(f) = -\log P(j) - \log W_j(A)/\delta^k$, for $f = f(\cdot\,|\,\theta) \in F_j$ and $\theta \in A \in \Pi$. Here Π is a partition of $\mathbf{R}^k$ into cubes of constant width $\delta > 0$, each of which contains $(\sqrt{n}\,\delta)^k$ of the points in Γ_n for every n. In practice, it is convenient to use the density function $w_j(\theta)$ in place of $W_j(A)/\delta^k$ for small δ (the summability requirement will continue to hold if for some $\delta > 0$, there exists an integrable function $\overline{w}_j(\theta)$ dominating $w_j(\theta')$ for $|\theta' - \theta| < \delta$). This leads to the choice

$$C_n(f) = \frac{k}{2}\log n - \log w_j(\theta) - \log P(j). \tag{7}$$

A computationally convenient choice is to take $w_j(\theta)$ to be a multivariate Gaussian density which makes the coordinates of θ independent with mean zero and fixed variance σ_0^2. Indeed, when attempting to minimize the empirical loss with $C_n(f)$ added as a penalty, the Gaussian choice for the prior density prevents singularity for each iteration in Gauss-Newton search algorithms. Nevertheless, the most important term in the complexity based criteria is the $(k/2)\log n$ because of its primary role in controlling the growth of the size of the network.

IV. CONVERGENCE RESULTS

Now we present the main results on statistical convergence properties of network estimators. The network structure and parameters are estimated using a complexity regularization criterion. Thus let

$$\hat{f}_n^{(1)} = \arg\min_{f\in\Gamma_n}\left(\frac{1}{n}\sum_{i=1}^{n} d(Y_i, f(X_i)) + \lambda\left(\frac{1}{n}C_n(f)\right)^{1/2}\right) \tag{8}$$

and

$$\hat{f}_n^{(2)} = \arg\min_{f\in\Gamma_n}\left(\frac{1}{n}\sum_{i=1}^{n} d(Y_i, f(X_i)) + \lambda\frac{1}{n}C_n(f)\right). \tag{9}$$

The first estimator is used for loss functions such as the zero-one loss for which the target rate of convergence of the risk would be close to $1/\sqrt{n}$. The second estimator is used for squared error and log-likelihood based loss functions for which close to $1/n$ would be the target rate.

Let $R_n^{(1)}(f^*)$ and $R_n^{(2)}(f^*)$ be the indices of approximation as defined in equations (3) and (4) above. It is generally the case that these indices converge to zero, whenever the class of network functions is dense at f^*. Indeed, suppose for each network family F_j, when restricted to networks f in the family, that $r(f, f^*)$ depends continuously on the parameters and that for each $\varepsilon > 0$, the sequence of subsets of parameter values for candidate networks in Γ_n that satisfy $C_n(f) \le n\varepsilon$ is dense in the usual Euclidean sense (e.g., as is the case for $C_n(f)$ defined as in (6)). Then $\lim R_n^{(1)}(f^*) = 0$ and $\lim R_n^{(2)}(f^*) = 0$, whenever $\lim_{j\to\infty} \inf_{f\in F_j} r(f, f^*) = 0$. In this way, network approximation theorems such as in Section II, above, may be used to show that R_n tends to zero, but leave open the question of characterizing the rate of approximation. For evaluation

of rates in related contexts of density estimation see [15]. We remark that for $C_n(f)$ as in (6) and for smooth networks, it is typically the case that $R_n^{(1)}(f^*) \le O(\sqrt{k_n (\log n)/n})$ where $k_n = k_n(f^*)$ is the dimension of the network that achieves the minimum in the definition of $R_n^{(1)}(f^*)$. Similarly $R_n^{(2)}(f^*) \le O(k_n (\log n)/n)$ where now k_n is the dimension of the optimum network in the definition of $R_n^{(2)}(f^*)$.

Our main result, which we now give, shows that the statistical rate of convergence of network estimators is bounded by the rate of approximation.

The result requires in some cases that $d(Y,f(X))$ be almost surely bounded. This is forced by a constraint on the support of Y and by clipping the functions $f(X)$, or by explicit choice of a bounded loss function. For loss functions based on the log-likelihood, with a correctly specified family of conditional densities, no boundedness of the loss is required.

Network Convergence Theorem: *Assume that the indices of approximation $R_n^{(1)}(f^*)$ and $R_n^{(2)}(f^*)$ tend to zero as $n \to \infty$. If the range of $d(Y,f(X))$ for every f in Γ_n is in a fixed interval of length b, and if $\lambda > b/\sqrt{2}$ in the definition of the complexity regularized estimator $\hat{f}_n^{(1)}$, then the statistical risk of the network estimator converges to zero at rate bounded by $R_n^{(1)}(f^*)$, i.e.,*

$$E(r(\hat{f}_n^{(1)},f^*)) \le O(R_n^{(1)}(f^*)). \tag{10}$$

For the squared error loss function, if the support of Y and of each function $f(X)$ is in a known interval of length b, then with $\lambda \ge 4b^2$ in the definition of the estimator $\hat{f}_n^{(2)}$, the mean squared error converges to zero at rate bounded by $R_n^{(2)}(f^*)$, i.e.,

$$E((\hat{f}_n^{(2)}(X) - f^*(X))^2) \le O(R_n^{(2)}(f^*)). \tag{11}$$

If the loss function is $d(Y,f(X)) = -\log p(Y \mid f(X))$ where the true conditional density is $p(Y \mid f^*(X))$, then for all $\lambda > 1$ in the definition of the estimator $\hat{f}_n^{(2)}$, the expected squared Hellinger distance between the conditional densities converges at rate bounded by the index of resolvability $R_n^{(2)}(f^*)$, i.e.,

$$E(d_H^2(\hat{f}_n^{(2)}(X),f^*(X))) \le O(R_n^{(2)}(f^*)). \tag{12}$$

The squared Hellinger distance is $d_H^2(f(x),f^*(x)) = \int ((p(y \mid f(x)))^{1/2} - (p(y \mid f^*(x))^{1/2})^2 \lambda(dy)$ (where $\lambda(dy)$ is the dominating measure, typically Lebesque's measure in the continuous case and counting measure in the discrete case). Since the L^1 distance, which takes the form $\int |p(y \mid f(x)) - p(y \mid f^*(x))| \lambda(dy)$, is known to not be greater than twice the Hellinger distance, a consequence of (12) is that the expected square of the L^1 distance also converges at rate bounded by $R_n^{(2)}(f^*)$.

For the Gaussian error case, the Hellinger distance can be evaluated and lower bounded as in [15,p.33]. It is seen that for any $c > 0$, the risk $E(\min((\hat{f}_n(X) - f^*(X))^2, c))$ converges to zero at rate bounded by $R_n^{(2)}(f^*)$.

For the first two conclusions of the theorem, the proof uses a bound on the probability of the event that $r(\hat{f}_n,f^*) > t$ in terms of a sum of probabilities of related events for each $f \in \Gamma_n$, to which inequalities of Hoeffding and Bennett can be applied. The bounds on the probabilities are then integrated for $t > 0$ to obtain the indicated bounds on the risk. The third conclusion is based on results we recently gave in [15]. The proof there also uses a union of events bound and inequalities of Chernoff.

V. CLASSIFICATION

Now we discuss the implications of the theory for network based classifiers. Here we restrict attention to dichotomous $Y \in \{-1,1\}$. The sign of the estimate $\hat{f}_n(x)$ is used as the decision rule. Let $P_e^{(f)} = P(E_f)$ denote the probability of the error event $E_f = \{Y \ne \text{sgn}(f(X))\}$, for which the minimum value is the Bayes optimal probability of error $P_e^{f^*}$ achieved by discriminant functions such as the log odds ratio which are positive if and only if $P(Y=1 \mid X=x) > P(Y=-1 \mid X=x)$. We desire to show that with the network classifiers, the probability of error converges to the Bayes optimum probability of error. Since the difference $P_e^{(f_n)} - P_e^{(f^*)}$ is positive, it converges to zero in probability at a given rate, if the expected value is shown to converge at that rate. Thus we examine $\bar{P}_e^{(f_n)} = E(P_e^{(f_n)})$.

Network Classification Theorem: *Under the conditions of the Network Convergence Theorem, if $\hat{f}_n^{(1)}$ is estimated using the zero-one loss and $\lambda > 1/\sqrt{2}$, then $\bar{P}_e^{(f_n)}$ converges to the Bayes optimal probability of error at rate bounded by $R_n^{(1)}(f^*)$. If $\hat{f}_n^{(2)}$ is estimated using the squared error loss with the network functions clipped to take values in the interval $[-1,1]$ and we take $\lambda \ge 16$, or if it is estimated using the logistic loss function and we take $\lambda > 1$, then $\bar{P}_e^{(f_n)}$ converges to the Bayes optimal probability of error at rate bounded by $\sqrt{R_n^{(2)}(f^*)}$.*

The proof shows that $P_e^{(f)} - P_e^{(f^*)}$ is not greater than $(E(f(X) - f^*(X))^2)^{1/2}$ in the squared error case, it is not greater than the expected L^1 distance between $p(y \mid f(x))$ and $p(y \mid f^*(x))$ in the logistic case, and it is equal to $r(f,f^*)$ in the case of the zero-one loss. The conclusion on classification rates then follows by application of the Network Convergence Theorem.

REFERENCES

[1] A. R. Barron and R. L. Barron, "Statistical learning networks: a unifying view," *Computing Science and Statistics: Proc. 20th Symp. Interface.*, Ed Wegman, editor, Amer. Statist. Assoc., Washington, DC., pp.192-203, 1988.

[2] J. A. Anderson and E. Rosenfeld, *Neurocomputing: Foundations of Research*, MIT press, 1988.

[3] D. E. Rumelhart, J. L. McClelland, et. al. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, 1986.

[4] S. J. Farlow, *Self Organizing Methods in Modeling: GMDH Type Algorithms*, New York: Marcel Dekker, 1984.

[5] J. H. Friedman and W. Stuetzle, "Projection pursuit classification," Department of Statistics Tech. Rep., Stanford University, Stanford, California, 1980.

[6] J. H. Friedman and W. Stuetzle, "Projection pursuit regression," *J. Amer. Statist. Assoc.* vol.76, pp.817-823, 1981.

[7] P. J. Huber, "Projection pursuit (with discussion)," *Ann. Statist.*, vol.13, pp.435-525, June 1985.

[8] G. Cybenko, "Approximations by Superpositions of Sigmoidal Functions," To appear in *Mathematics of Control, Signals & Systems*, 1989.

[9] G. G. Lorentz, *Approximation of Functions*, New York: Holt, Rinehart, and Winston, 1966.

[10] C. Canuto and A. Quarteroni, "Approximation results for orthogonal polynomials in Sobolev spaces," *Math. Comp.* vol.38, pp.67-86, 1982.

[11] D. D. Cox, "Approximation of least squares regression on nested subspaces," *Ann. Statist.*, vol.18, pp.713-732, 1988.

[12] R. Shibata, "An optimal selection of regression variables," *Biometrika* vol.68, pp.45-54, 1981.

[13] R. L. Eubank, *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker, 1988.

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

[15] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," Submitted to *IEEE Trans. Information Theory*, January 1989.

[16] L. Devroye, "Automatic pattern recognition: a study of the probability of error," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol.10, pp.530-543, 1988.

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

[18] D. Haussler, "Generalizing the PAC model: sample size bounds from metric dimension-based uniform convergence results" *Conf. Computational Learning Theory*, Preprint, 1989.