Neural Networks, Vol. 3, pp. 535–549, 1990
Printed in the USA.

0893-6080/90 $3.00 + .00

ORIGINAL CONTRIBUTION

Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings

HALBERT WHITE

University of California, San Diego

(*Received* 27 *February* 1989; *revised and accepted* 31 *January* 1990)

Abstract—*It has been recently shown (e.g., Hornik, Stinchcombe & White, 1989, 1990) that sufficiently complex multilayer feedforward networks are capable of representing arbitrarily accurate approximations to arbitrary mappings. We show here that these approximations are learnable by proving the consistency of a class of connectionist nonparametric regression estimators for arbitrary (square integrable) regression functions. The consistency property ensures that as network "experience" accumulates (as indexed by the size of the training set), the probability of network approximation error exceeding any specified level tends to zero. A key feature of the demonstration of consistency is the proper control of the growth of network complexity as a function of network experience. We give specific growth rates for network complexity compatible with consistency. We also consider automatic and semi-automatic data-driven methods for determining network complexity in applications, based on minimization of a cross-validated average squared error measure of network performance. We recommend cross-validated average squared error as a generally applicable criterion for comparing relative performance of differing network architectures and configurations.*

Keywords—Learning, Feedforward networks, Nonparametric, Regression, Convergence.

1. INTRODUCTION

Recently, Hornik, Stinchcombe and White (1989, 1990) (HSWa,b) have demonstrated that sufficiently complex multilayer feedforward networks (e.g., Rumelhart, Hinton, & Williams, 1986) are capable of arbitrarily accurate approximations to arbitrary mappings (i.e., measurable or continuous functions). (See also Carroll and Dickinson, 1989; Cybenko, 1989; Funahashi, 1989; Hecht-Nielsen, 1989; and Stinchcombe & White, 1989a). An unresolved issue is that of "learnability", that is, whether there exist methods allowing the network weights corresponding to these approximations to be learned from empirical observation of such mappings. It is the purpose of this paper to prove that these approximations are indeed learnable by establishing the statistical property of consistency for a certain class of learning methods for such networks. The consistency property ensures that as network experience accumulates (as indexed by the size of the training set n) the probability of network approximation error exceeding any specified level tends to zero. This establishes the nonparametric regression capability of multilayer feedforward network models.

A key feature of our analysis is the proper control of network complexity as a function of network experience n. Arbitrarily accurate approximation to an arbitrary function will generally require an arbitrarily complex network. In practice, only networks of finite complexity can be implemented. Too complex a network (relative to n) learns too much—it memorizes the training set and therefore generally performs poorly at generalization tasks. Too simple a network learns too little—it fails to extract the desired relationships from the training set and thus also performs poorly at generalization tasks. Our theory provides specific growth rates for network complexity that asymptotically (i.e., as $n \to \infty$) avoid the dangers of both overfitting and underfitting the training set. Moreover, these rates can be combined with data-driven methods for determining network complexity in a manner that guarantees their consistency also.

Acknowledgments: This research was supported by a grant from the Guggenheim Foundation and by NSF grant SES-8806990. The author is grateful for helpful conversations with Maxwell Stinchcombe, Whitney Newey, and James Powell, and for helpful suggestions by the editor and referees. The author is responsible for any errors.

Requests for reprints should be sent to Halbert White, Department of Economics, D-008, University of California, San Diego, La Jolla, CA 92093.

The learning methods treated here are extremely computationally demanding. Thus, they lay no claim to biological or cognitive plausibility. They do, however, provide some appreciation for the computational effort required to train multilayer feedforward networks to approximate arbitrary mappings arbitrarily well. Our methods are feasible in modern (especially parallel) computing environments and are thus of some value to researchers interested in practical applications. In particular, we recommend general application of cross-validated measures of network performance.

To keep this paper to a manageable size, we limit its scope strictly to the issue of consistency (learnability). This by no means exhausts the relevant issues; it leaves many relevant questions unaddressed, and indeed makes possible the formulation of further important questions. We leave these to other work, and merely identify them as they arise.

2. HEURISTICS

Consider a sequence $\{Z_t\} = \{Z_t, t = 1, 2, \ldots\}$ of identically distributed random column vectors. Suppose we are interested in the relationship between some elements of Y_t of Z_t and the remaining elements X_t. We write $Z_t = (Y_t', X_t')'$, where a prime superscript denotes vector transposition. For example, in classification or pattern recognition problems Y_t is a binary or multinomial variable designating class membership and X_t is a set of variables influencing the classification. In forecasting problems, Y_t is the set of variables (digital or analog) that we wish to forecast on the basis of variables X_t, which may itself contain past values of Y_t. In image enhancement problems, Y_t encodes a target image, and X_t encodes a degraded version of the image. In pattern completion problems, Y_t is the missing part of a pattern, and X_t is the supplied part of the pattern (we suppose the same part is always to be supplied for this application).

Regardless of whether a deterministic or stochastic relationship exists between Y_t and X_t, a natural object of interest in such situations is the conditional expectation of Y_t given X_t, written $E(Y_t|X_t)$. (See White, 1989a.) This can be represented as a regression function,

$$\theta_o(X_t) = E(Y_t|X_t).$$

For example, when Y_t can take on only the values 0 or 1, $\theta_o(x)$ gives the probability that $Y_t = 1$ given that $X_t = x$. When Y_t can assume a continuum of values, $\theta_o(x)$ gives the expected value for Y_t given that $X_t = x$. We may also write

$$Y_t = \theta_o(X_t) + \varepsilon_t,$$

where $\varepsilon_t \equiv Y_t - E(Y_t|X_t)$ is a random error with conditional expectation zero given X_t. When the relationship between Y_t and X_t is deterministic, ε_t is always zero; otherwise, ε_t is nonzero with positive probability. In the situation treated here, the regression function θ_o is assumed to be entirely unknown. Our problem is to learn (estimate, approximate) the mapping θ_o from a realization of the sequence $\{Z_t\}$.

In practice, we observe a realization of only a finite part of the sequence $\{Z_t\}$, a "training set" or "sample" of size n (i.e., a realization of $Z^n \equiv (Z_1', \ldots, Z_n')'$). Because θ_o is an element of a space of functions (say Θ), we have essentially no hope of learning θ_o in any complete sense from a sample of fixed finite size. Nevertheless, it is possible to approximate θ_o to some degree of accuracy using a sample of size n, and to construct increasingly accurate approximations with increasing n. In what follows we will refer to such a procedure interchangeably as learning, estimation, or approximation. A learning rule is correspondingly defined in the weakest possible sense as a sequence of mappings, say $\{\hat{\theta}_n\} = \{\hat{\theta}_n, n = 1, 2, \ldots\}$, from the underlying probability space generating the observed phenomenon of interest to the space Θ in which θ_o, the object of interest describing this phenomenon, lies. Because $\hat{\theta}_n$ is a mapping from a probability space, it is stochastic.

A minimal property for any learning rule is that of *consistency*. A stochastic sequence $\{\hat{\theta}_n\}$ is consistent for θ_o if the probability that $\hat{\theta}_n$ exceeds any specified level of approximation error relative to θ_o tends to zero as the sample size n tends to infinity. Procedures that are not consistent will always make errors in classification, recognition, forecasting, enhancement, or pattern completion (forms of generalization) that are eventually avoided by a consistent procedure. The only errors ultimately made by a consistent procedure are the inherently unavoidable errors (ε_t) arising from any fundamental randomness or fuzziness in the true relation between X_t and Y_t. Consequently, our focus is on using connectionist networks to obtain learning rules $\{\hat{\theta}_n\}$ consistent for an arbitrary regression function θ_o, and we therefore identify the issue of the "learnability" of a particular class of mappings with the existence of a consistent learning rule for that class. This distinguishes our focus of attention from such standard concept-learnability issues as treated for example by Haussler (1989).

We apply the method of sieves, (Grenander, 1981; Geman & Hwang 1982; White & Wooldridge, 1990), a general nonparametric statistical procedure in which an object of interest θ_o lying in a general (i.e., not necessarily finite-dimensional) space Θ is approximated using a sequence of parametric models

in which the dimensionality of the parameter space grows along with the sample size. To succeed, the approximating parametric models must be capable of arbitrarily accurate approximation to elements of Θ as the underlying parameter space grows. For this reason, Fourier series (e.g., Gallant & Nychka, 1987) and spline functions (e.g., Wahba, 1975; Wahba & Wold, 1975; Cox, 1984) are commonly used in this context. Among others, HSWa,b establish that multilayer feedforward networks also have universal approximation properties. Without this, attempts at nonparametric estimation using these network models would be doomed from the outset.

For concreteness and simplicity, we consider only single output single hidden layer feedforward networks. Our approach generalizes straightforwardly to the multi-output multi-hidden layer case. Specifically, we write the output of a q hidden unit feedforward network given input x as

$$f^q(x, \delta^q) \equiv \beta_0 + \sum_{j=1}^{q} \beta_j \psi(\bar{x}'\gamma_j),$$

where $\delta^q = (\beta^{q\prime}, \gamma^{q\prime})'$ is the $p \times 1$ ($p \equiv q(r + 2) + 1$) vector of network weights (parameters) (let $\beta^q = (\beta_0, \beta_1, \ldots, \beta_q)'$, $\gamma^q \equiv (\gamma_1', \ldots, \gamma_q')'$, $\gamma_j \equiv (\gamma_{j0}, \ldots, \gamma_{jr})'$ for $j = 1, \ldots, q$), ψ is the (given) hidden unit activation function, and $\bar{x} \equiv (1, x')'$. For simplicity and without loss of generality, we put no squashing function at the output unit. Figure 1 depicts the corresponding network architecture.

We construct a sequence of approximations to θ_o by letting network complexity q grow with n at an appropriate rate, and for given n (hence given q) selecting weights $\hat{\delta}_n$ so that $\hat{\theta}_n \equiv f^{q_n}(\cdot, \hat{\delta}_n)$ provides an approximation to the unknown regression function θ_o that is the best possible in an appropriate sense, given the sample information.

To formulate precisely a solution to the problem of finding $\hat{\theta}_n$, we define

$$T(\psi, q, \Delta) \equiv \Big\{ \theta \in \Theta : \theta(\cdot) = f^q(\cdot, \delta^q), \sum_{j=0}^{q} |\beta_j| \le \Delta, \sum_{j=1}^{q} \sum_{i=0}^{r} |\gamma_{ji}| \le q\Delta \Big\},$$

the set of output functions of all single hidden layer feedforward networks with q hidden units having activation functions ψ, and with weights satisfying a particular restriction on their sum norm, indexed by Δ. We construct a sequence of "sieves" $\{\Theta_n(\psi)\}$ by specifying sequences $\{q_n\}$ and $\{\Delta_n\}$ and setting $\Theta_n(\psi) = T(\psi, q_n, \Delta_n)$, $n = 1, 2, \ldots$. The sieve $\Theta_n(\psi)$ becomes finer (less escapes) as $q_n \to \infty$ and $\Delta_n \to \infty$. For given sequences $\{q_n\}$ and $\{\Delta_n\}$, the "connectionist sieve estimator" $\hat{\theta}_n$ is defined as a solution to the least squares problem (appropriate for learning $E(Y_t|X_t)$)

$$\min_{\theta \in \Theta_n(\psi)} n^{-1} \sum_{t=1}^{n} [Y_t - \theta(X_t)]^2, \qquad n = 1, 2, \ldots. \tag{2.1}$$

Associated with $\hat{\theta}_n$ is an estimator $\hat{\delta}_n$ of dimension $p_n \times 1$ ($p_n \equiv q_n(r + 2) + 1$) such that $\hat{\theta}_n(\cdot) = f^{q_n}(\cdot, \hat{\delta}_n)$. Sections 3 and 4 are devoted to specifying precise conditions on $\{Z_t\}$, ψ, $\{q_n\}$, and $\{\Delta_n\}$ that ensure the consistency of $\hat{\theta}_n$ for θ_o.

Before describing these conditions, we must emphasize that obtaining at least a near global solution to (2.1) is central to this approach. Consequently, optimization methods that deliver only local optima, such as the method of back-propagation or iterative Newton methods will be inadequate. Instead, global optimization methods (as described by Baba, 1989; or White, 1989a, Section 4.a) are required. This is one source of the advertised computational burden of this approach. For simplicity, we suppose for now that a global solution to (2.1) is available. At the end of this section, we describe the modifications to our set-up that permit treatment of cases in which only approximately optimal solutions are available.

For simplicity throughout we suppose that $\{Z_t\}$ is a bounded stochastic process, without loss of generality taking values in the hypercube $\mathbb{I}^v \equiv \times_{i=1}^{v} [0,1]$. We also assume that θ_o is a square integrable function on $\mathbb{I}^r$, $r \equiv v - 1$. The activation function ψ is assumed to be bounded and satisfy a Lipschitz condition. The Lipschitz condition holds whenever ψ is continuously differentiable with bounded derivative, as for the logistic and hyperbolic tangent squashers commonly used in applications. Non-sigmoid activation functions are also permitted.

We first consider deterministic methods for network complexity growth. We prove that the appropriate choice for Δ_n is such that $\Delta_n \to \infty$ as $n \to \infty$ and $\Delta_n = o(n^{1/4})$, that is, $n^{-1/4}\Delta_n \to 0$ as $n \to \infty$. The proper choice for $\{q_n\}$ depends on $\{\Delta_n\}$ and on the dependence properties of $\{Z_t\}$. When $\{Z_t\}$ is an independent identically distributed (i.i.d.) sequence, Theorem 3.5 below establishes that $\hat{\theta}_n$ is consistent for θ_o, provided that $q_n \to \infty$ as $n \to \infty$ and $q_n \Delta_n^4 \log q_n \Delta_n = o(n)$. When $\{Z_t\}$ is a stationary mixing process, Theorem 3.5 establishes that $\hat{\theta}_n$ is consistent for θ_o, provided that $q_n \to \infty$ as $n \to \infty$ and $q_n \Delta_n^2 \log q_n \Delta_n = o(n^{1/2})$. Mixing processes are a class of time

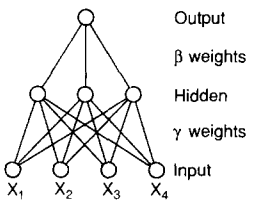

FIGURE. 1. Single hidden layer feedforward network.

series processes that can exhibit considerable short-run dependence, but display a form of asymptotic independence, in that events involving elements of $\{Z_t\}$ separated by increasingly greater time intervals are increasingly closer to independence. The allowable growth rate for network complexity is slower in the case of mixing processes than for the case of independent processes because new information arrives at a slower rate. (See White, 1984, for a further discussion of mixing processes.)

The allowed growth rates for network complexity are not immediately obvious from the conditions stated. To gain some insight we consider the allowed rates in more detail. First, we see a clear trade-off between the allowed rate of increase of q_n and that of Δ_n: the slower of the rate of increase of Δ_n, the greater the permissible rate for q_n. Because the approximation improves as network complexity increases, it is desirable to permit Δ_n to grow at only a modest rate in order to accommodate rapid growth in q_n.

Accordingly, suppose $\Delta_n = O(\log n)$ (i.e., $\Delta_n \leq c \log n$ for some $0 < c < \infty$). This satisfies $\Delta_n = o(n^{1/4})$, because $n^{-1/4}\Delta_n \leq cn^{-1/4}\log n \to 0$ as $n \to \infty$. For the independent case, we now require $q_n \Delta_n^4 \log q_n \Delta_n = o(n)$ (i.e., $n^{-1} q_n \Delta_n^4 \log q_n \Delta_n \to 0$). With $\Delta_n = O(\log n)$, it suffices that $n^{-1} q_n (\log n)^4 \log(q_n \log n) \to 0$. We put $q_n = O(n^\alpha)$, $\alpha > 0$, and seek admissible choices for α: it now suffices that $n^{-1} n^\alpha (\log n)^4 (\alpha \log n + \log\log n)$. Because $\log n > \log\log n$, we need $n^{\alpha-1}(\log n)^5 \to 0$. Because $n^{-\varepsilon}(\log n)^5 \to 0$ for any $\varepsilon > 0$, we can take $\alpha = 1 - \varepsilon$ for any small $\varepsilon > 0$. Thus, $q_n = O(n^{1-\varepsilon})$ and $\Delta_n = O(\log n)$ satisfy the required conditions for independent observations. Similarly, for mixing processes $q_n = O(n^{(1-\varepsilon)/2})$ and $\Delta_n = O(\log n)$ suffice.

These allowed growth rates for network complexity q_n are fairly rapid, as Table 1 illustrates. The entries are scaled so that unit complexity is achieved at $n = 100$. Thus, if five hidden units are appropriate in the independent case when $n = 100$, 477 hidden units can be supported by a sample of 10,000 and 45,600 hidden units can be supported by a sample of one million. With mixing processes, allowable growth rates are slower. If five hidden units are appropriate in the mixing case when $n = 100$, then 49 hidden units can be supported by a sample of 10,000. A sample of one million supports 477 hidden units in this case. Even though the growth rates in the mixing case appear dramatically slower than for the independent case, they are still rather generous.

These relatively rapid growth rates for network complexity are achieved by taking Δ_n to grow at the modest rate of $\log n$. It is helpful to explore the implications of this choice. Recall that we require $\sum_{j=0}^{q_n} |\beta_j| \leq \Delta_n$ and $\sum_{j=1}^{q_n} \sum_{i=0}^{r} |\gamma_{ji}| \leq q_n \Delta_n$. For the latter condition, it suffices that $\sum_{i=0}^{r} |\gamma_{ji}| \leq \Delta_n$. Because r is

TABLE 1
Relative Network Complexity for Consistency as a Function of Sample Size, $\epsilon = .01$

n	10^2	10^4	10^6	10^8	10^{10}
$n^{1-\epsilon}$	1	95.50	9.12×10^3	8.70×10^5	8.32×10^7
$n^{(1-\epsilon)/2}$	1	9.77	95.50	9.33×10^2	9.12×10^3

fixed and $\Delta_n \leq c \log n$ will increase with n, the restrictions imposed on γ are quite mild. (In fact, there is even room for letting the number of inputs r increase with n, so that greater amounts of information can be exploited as $n \to \infty$ in attempting to hit the target; we leave this to other work for the sake of simplicity.) Because of the mildness of the restrictions on γ, we do not consider them further.

The force of our condition is borne by β. We require that $\sum_{j=0}^{q_n} |\beta_j| \leq c \log n$ for some $0 < c < \infty$. With q_n increasing at a rate only slightly slower than n, this inequality can be satisfied only if an increasing proportion of the β_j's are close to zero. An elementary calculation yields

$$\sum_{j=0}^{q_n} (j+1)^{-1} = \sum_{j=1}^{q_n+1} j^{-1} \leq 1 + \int_1^{q_n+1} j^{-1}\, dj.$$

Now

$$\int_1^{q_n+1} j^{-1}\, dj = \log j \Big]_1^{q_n+1} = \log(q_n + 1) \leq \log q_n + \frac{1}{q_n}$$

so that

$$\sum_{j=0}^{q_n} (j+1)^{-1} \leq \log q_n + 1 + \frac{1}{q_n}.$$

When $q_n = O(n^{1-\varepsilon})$, we obtain from this inequality that $\sum_{j=0}^{q_n} (j+1)^{-1} = O(\log n)$. Hence, for $\sum_{j=0}^{q_n} |\beta_j| = O(\log n)$ it suffices that $|\beta_j| = O(j^{-1})$. When j is large, $|\beta_j|$ will be decreasing at the rate j^{-1} under this condition; this is a rather rapid approach to zero.

If less stringent restrictions are desired for $\sum_{j=1}^{q_n} |\beta_j|$, they can be achieved by trading off restrictions on $\sum_{j=1}^{q_n}$ for restrictions on q_n. For example, if restrictions analogous to those encountered for γ in the previous example are desired, we could set $\Delta_n \propto q_n \log n$. This choice is compatible with $q_n = O(n^{1/5-\varepsilon})$ for the independent case or $q_n = O(n^{1/6-\varepsilon})$ for the mixing case, as some algebra will verify. These are much slower growth rates than in the previous example.

Despite these results for growth rates, they provide no practical method for choosing network complexity for a sample of given size n. One appealing method for making this choice is the method of cross-validation (Stone, 1974). The rationale for this method is straightforward. First, consider a naive approach. Suppose $\{\Delta_n\}$ is given, and put $\Theta_n(\psi, q) = T(\psi, q,$

Δ_n). For given n, network performance with q hidden units can be (naively) measured by the smallest attainable average squared error over the training set.

$$\min_{\theta \in \Theta_n(\psi, q)} n^{-1} \sum_{t=1}^{n} [Y_t - \theta(X_t)]^2.$$

Unfortunately, this measure is over-optimistic: it is downward biased (as a measure of minimum risk, $\min_{\theta \in \Theta_n(\psi, q)} E([Y_t - \theta(X_t)]^2)$) and is increasingly downward biased as q increases. The bias arises because the observation for $(Y_t', X_t')'$ is used in arriving at $\hat{\theta}_n^q$ (say), the solution to the above minimization problem. This makes the contribution to squared error of the tth observation ($[Y_t - \hat{\theta}_n^q(X_t)]^2$) too small. Indeed, if network complexity were selected by choosing q to make average squared error as small as possible, average squared error could be driven to zero by choosing $q = n$, a clearly inappropriate choice.

The method of cross-validation effectively avoids these difficulties. The "delete one" cross-validated average squared error averages an estimate of squared error for each observation (say observation t) that uses an estimator for θ_o (say $\hat{\theta}_{n(t)}^q$) that *ignores* the information contributed by that (tth) observation. (For example, a "brute force" choice for $\hat{\theta}_{n(t)}^q$ can be obtained as the solution to the problem $\min_{\theta \in \Theta_n(\psi, q)} n^{-1} \sum_{\tau \neq t} [Y_\tau - \theta(X_\tau)]^2$.) The cross-validated performance measure is formally defined as

$$C_n(q) \equiv n^{-1} \sum_{t=1}^{n} [Y_t - \hat{\theta}_{n(t)}^q(X_t)]^2.$$

Because $\hat{\theta}_{n(t)}^q$ ignores information from the tth observation, $[Y_t - \hat{\theta}_{n(t)}^q(X_t)]^2$ is a measure of out-of-sample performance and is thus less biased than $[Y_t - \hat{\theta}_n^q(X_t)]^2$, an in-sample measure. The cross-validated average squared error inherits this relative lack of bias by construction and consequently provides a measure of network performance superior to average squared error. Because of its relative unbiasedness, we recommend its general use.

A completely automatic method for determining network complexity appropriate for any specific application is given by choosing the number of hidden units $\hat{q}_n$ to be the smallest solution to the problem

$$\min_{q \in \mathcal{N}_n} C_n(q),$$

where $\mathcal{N}_n$ is some appropriate choice set, a subset of $\mathbb{N} \equiv \{1, 2, \ldots\}$. For the moment, we can suppose that $\mathcal{N}_n = \{1, 2, \ldots, n\}$. The choice $\hat{q}_n$ will be called the "cross-validated complexity" of the network. The "cross-validated connectionist sieve estimator" is then the solution $\tilde{\theta}_n$ to the problem

$$\min_{\theta \in \Theta_n(\psi, \hat{q}_n)} n^{-1} \sum_{t=1}^{n} [Y_t - \theta(X_t)]^2.$$

The method of cross-validation has proven widely successful in other statistical applications; we may reasonably expect it or some variant (e.g., White, 1989b) to be similarly successful here.

To ensure the consistency of $\tilde{\theta}_n$ for θ_o, we cannot choose $\{\Delta_n\}$ and $\{N_n\}$ arbitrarily. In both the independent and mixing cases, it is again appropriate to take $\Delta_n = o(n^{1/4})$. As a crude but expedient approach we ensure consistency here by taking $N_n = \{\underline{q}_n, \ldots, \bar{q}_n\}$, where $\underline{q}_n \to \infty$ and $\bar{q}_n$ satisfies the rate conditions for q_n discussed earlier. These conditions control $\hat{q}_n$ in just the right way to ensure that $\Theta_n(\psi, \hat{q}_n)$ eventually fills Θ appropriately. It may be possible to remove the controls on $\underline{q}_n$ without affecting consistency, but we leave this to other work.

In applications it is convenient to pick $\underline{q}_n = \lambda \bar{q}_n$ for some $\lambda < 1$. All of the discussion regarding $\{q_n\}$ and $\{\Delta_n\}$ for the deterministically determined sieves $\{\Theta_n(\psi)\}$ then applies directly to $\{\bar{q}_n\}$ and $\{\Delta_n\}$. Applications entail optimization of $C_n(q)$ over $N_n = [\lambda \bar{q}_n, \bar{q}_n] \cap \mathbb{N}$. An advantage of choosing N_n in this way is that it can greatly reduce the range of values that must be considered for q. Even so, the range of N_n may still be large. Exhaustive search over N_n is thus likely to be expensive; more sophisticated univariate optimization methods are available (e.g., Scales, 1982).

A disadvantage of these procedures is that they are not fully automatic, as we require values for auxiliary control parameters Δ_n, $\bar{q}_n$ and either $\underline{q}_n$ or λ. For this reason, we call these consistent procedures "semi-automatic." Nevertheless, it is typically straightforward to pick intuitively plausible values for these control parameters for a given problem with given n. Given these, values for all other n can be readily computed (e.g., using the numbers of Table 1 or numbers similarly constructed).

In practice, full optimization of network performance over the sample can be difficult or impossible to attain. Instead, estimated weights will be more or less optimal depending on the intensity of computational effort. Results analogous to the consistency results just described continue to hold when approximate rather than exact optimization is carried out, although the strength of the results varies with the assumed degree of computational effort, as one should expect. Our analysis is conducted by letting the data determine an appropriate (e.g., approximately optimal) degree of network complexity, say $\ddot{q}_n$. As before, we require $\underline{q}_n \leq \ddot{q}_n \leq \bar{q}_n$. For this choice of complexity, again let the fully optimal solution to the least squares problem be denoted $\tilde{\theta}_n$, so that $\tilde{\theta}_n$ now solves

$$\min_{\theta \in \Theta_n(\psi, \ddot{q}_n)} n^{-1} \sum_{t=1}^{n} [Y_t - \theta(X_t)]^2.$$

Instead of obtaining $\tilde{\theta}_n$, we suppose that we obtain an approximate solution $\ddot{\theta}_n$ to this problem, characterized by the closeness of network performance at $\ddot{\theta}_n$ to optimal network performance. Specifically, given a "tolerance level" ζ_n (some positive number), we assume that $\ddot{\theta}_n$ is chosen so that

$$\left| n^{-1} \sum_{t=1}^{n} [Y_t - \ddot{\theta}_n(X_t)]^2 - n^{-1} \sum_{t=1}^{n} [Y_t - \tilde{\theta}_n(X_t)]^2 \right| \leq \zeta_n.$$

The smaller is ζ_n, the closer is $\ddot{\theta}_n$ to being an exact solution to the network learning problem, and, in general, the more the computational effort required to obtain $\ddot{\theta}_n$. When ζ_n is large, we can afford to be more sloppy in finding $\ddot{\theta}_n$.

To obtain our results, we suppose that a sequence $\{\zeta_n\}$ of tolerances is specified, together with a limit ζ_o to which the tolerances tend as $n \to \infty$. The limiting tolerance ζ_o may be zero, in which case we demand an exact global solution in the limit. In this case, we can again establish the consistency of $\ddot{\theta}_n$ for θ_o. Alternatively, the limiting tolerance may be nonzero; for example, we could require only a fixed tolerance $\zeta_n = \zeta_o > 0$ for all n. In this case, consistency of $\ddot{\theta}_n$ for θ_o cannot be guaranteed (so that learning is incomplete), but network performance in the limit is adversely affect to a precise extent governed by ζ_o. The conditions on $\{Z_t\}$, ψ, $\{q_n\}$, $\{\bar{q}_n\}$, and $\{\Delta_n\}$ under which these results hold are identical to those previously described.

To summarize, multilayer feedforward networks can learn arbitrarily accurate approximations to unknown functions in the sense that they admit a learning rule possessing the consistency property. To ensure consistency, network complexity must be properly controlled; this control can be data-driven. Exact optimization for each n is not necessary; approximate optimization is permitted. However, consistency is guaranteed only when exact optimization occurs in the limit.

3. CONSISTENCY OF CONNECTIONIST SIEVE ESTIMATORS

We now state formal results establishing the consistency of connectionist sieve estimators. We first treat deterministic choice of connectionist sieves $\{\Theta_n\}$, and then treat the cross-validated connectionist sieve estimators, assuming exact optimization. We next give results assuming approximate optimization. Our results follow from more general statements given in the next section.

The first assumption describes the data generating process.

ASSUMPTION A.1. *Let $(\Omega, \mathbf{F}, P)$ be a complete probability space. The observed data are the realization of a stochastic process $\{Z_t: \Omega \to \mathbb{I}^\nu, \nu \in \mathbb{N}, t = 1, 2, \ldots\}$ on $(\Omega, \mathbf{F}, P)$, and P is such that either*

(i) *$\{Z_t\}$ is* i.i.d.; *or*

(ii) *$\{Z_t\}$ is a stationary mixing process with either $\phi(k) = \phi_o \rho_o^k$ or $\alpha(k) = \alpha_o \rho_o^k$, $k \geq 1$, for some constants $\phi_o, \alpha_o > 0$, $0 < \rho_o < 1$.* (Formal definitions of ϕ and α are given in the next section.)

For convenience, we assume that Z_t takes values in the unit cube. When $\{Z_t\}$ is a bounded sequence, this can always be ensured by appropriate scaling and shifting. This assumption ensures the existence of $E(Y_t)$ and hence of $E(Y_t|X_t)$, where as before we partition Z_t as $Z_t = (Y_t, X_t')'$, with Y_t a scalar random variable for simplicity. Let $\mathbf{B}^r = \mathbf{B}(\mathbb{I}^r)$ denote the Borel σ-field generated by the open sets of $\mathbb{I}^r$. An element A of $\mathbf{B}^r$ is a subset of $\mathbb{I}^r$ for which the probability $P[X_t \in A]$ is defined. Define the measure μ as $\mu(A) = P[X_t \in A]$ for each A in $\mathbf{B}^r$; μ measures the relative frequency with which patterns X_t belong to the set A. Let $L^2(\mathbb{I}^r, \mu)$ denote the collection of all measurable functions $\theta: \mathbb{I}^r \to \mathbb{R}$ such that $\|\theta\|_2 = [\int_{\mathbb{I}^r} |\theta(x)|^2 \mu(dx)]^{1/2} < \infty$ (the square integrable functions), and define the metric $\rho_2(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2$ for $\theta_1, \theta_2 \in L^2(\mathbb{I}^r, \mu)$. The metric ρ_2 measures the distance between two functions θ_1 and θ_2 belonging to $L^2(\mathbb{I}^r, \mu)$ in terms of weighted root mean squared error, where the weighting is with respect to μ. Two functions θ_1 and θ_2 differing on a set of μ-measure zero will have $\rho_2(\theta_1, \theta_2) = 0$. We shall not distinguish between such functions; technically, functions differing only on a set of μ-measure zero are taken as forming an equivalence class. Uniqueness of functions of $L^2(\mathbb{I}^r, \mu)$ is with respect to these equivalence classes.

The space $(L^2(\mathbb{I}^r, \mu), \rho_2)$ is a complete separable metric space. We take the object of interest to be the function $\theta_o: \mathbb{I}^r \to \mathbb{R}$ such that $E(Y_t|X_t) = \theta_o(X_t)$, and we impose the following structure on θ_o.

ASSUMPTION A.2. *With $(\Theta, \rho) = (L^2(\mathbb{I}^r, \mu), \rho_2)$, θ_o is the unique element of Θ such that $E(Y_t|X_t) = \theta_o(X_t)$.* □

This structure is convenient and plausible in many applications.

Next we formally define the connectionist sieve used to approximate elements of Θ.

DEFINITION 3.1. *Let $\psi: \mathbb{R} \to \mathbb{R}$ be a given bounded function. For any $q \in \mathbb{N}$ and $\Delta \in \mathbb{R}^+$, define a subset $T(\psi, q, \Delta)$ of Θ as*

$$T(\psi, q, \Delta) = \left\{ \theta \in \Theta: \theta(x) = \beta_0 + \sum_{j=1}^{q} \beta_j \psi(\bar{x}'\gamma_j), \right.$$

$$\bar{x} = (1, x')', \text{ for all } x \text{ in } \mathbb{I}^r,$$

$$\left. \sum_{j=0}^{q} |\beta_j| \leq \Delta, \sum_{j=1}^{q} \sum_{i=0}^{r} |\gamma_{ji}| \leq q\Delta \right\}.$$

For given ψ and sequences $\{q_n\}$, $\{\Delta_n\}$, define the sequence of (single hidden layer) connectionist sieves $\{\Theta_n(\psi)\}$ as

$$\Theta_n(\psi) \equiv T(\psi, q_n, \Delta_n), \qquad n = 1, 2, \ldots. \qquad \square$$

This definition implies that when $\theta \in \Theta_n(\psi)$, we have $\theta(x) = f^{q_n}(x, \delta^{q_n})$ in the notation of section 2. We consider only single hidden layer connectionist sieves for notational simplicity. The results of section 4 also apply to feedforward networks with an arbitrary number of hidden layers; the results of this section can be extended to such cases.

To obtain the consistency result, we must place specific conditions on ψ, and q_n and Δ_n. A necessary condition for consistency is that $\cup_{n=1}^{\infty}\Theta_n(\psi)$ is ρ-dense in Θ (i.e., $\cup_{n=1}^{\infty}\Theta_n(\psi)$ contains an element θ as close as we like to any element θ^* of Θ, with distance measured by ρ). The universal approximation results for single hidden layer feedforward networks established by HSW a,b make it possible to specify conditions on ψ ensuring the required denseness property. To state these conditions, we first define some terms. We say that ψ is a squashing function if $\psi: \mathbb{R} \to I$, $\psi(a) \to 0$ as $a \to -\infty$, $\psi(a) \to 1$ as $a \to \infty$ and ψ is monotonic. (In probability theoretic terms, ψ is a cumulative distribution function.) We say that ψ is l-finite if $\psi: \mathbb{R} \to \mathbb{R}$ is continuously differentiable of order l ($0 \le l < \infty$) and $0 < \int |D^l\psi(x)|\, dx < \infty$, where $D^l\psi$ denotes the lth derivative of ψ. ($D^0\psi = \psi$; ψ is assumed continuous when $l = 0$.) If ψ is a differentiable squashing function, then ψ is 1-finite, but not 0-finite. A continuous nondifferentiable squashing function is not l-finite for any l.

Conditions on ψ, q_n and Δ_n ensuring the required denseness are given by the next result.

LEMMA 3.2. *Let $\{q_n \in \mathbb{N}\}$ and $\{\Delta_n \in \mathbb{R}^+\}$ be increasing sequences tending to infinity with n. If ψ is a squashing function or if ψ is l-finite for any integer l, $0 \le l < \infty$, then $\cup_{n=1}^{\infty}\Theta_n(\psi)$ is ρ-dense ($\rho = \rho_2$) in $\Theta = L^2(I^r, \mu)$.* $\square$

Lemma 3.2 ensures that by proper choice of ψ, q_n, and Δ_n we can make $\Theta_n(\psi)$ sufficiently "big," avoiding underfitting. To avoid overfitting, $\Theta_n(\psi)$ must be prevented from becoming too big too fast. This will require restrictions on q_n and Δ_n. The derivation of these restrictions is rather technical, and is postponed to the next section. A further restriction on ψ is also helpful. We say that ψ satsifies a Lipschitz condition if $|\psi(a_1) - \psi(a_2)| \le L\,|a_1 - a_2|$ for all a_1, $a_2 \in \mathbb{R}$ and some $L \in \mathbb{R}^+$. For convenience, let $\mathbf{L}$ denote the set of all activation functions $\psi: \mathbb{R} \to \mathbb{R}$ such that ψ is bounded, satisfies a Lipschitz condition, and is either a squashing function or is l-finite, $0 \le l < \infty$. The following condition ensures that $\Theta_n(\psi)$ does not permit overfitting in the limit.

ASSUMPTION A.3. *$\Theta_n(\psi) = T(\psi, q_n, \Delta_n)$ where $\psi \in \mathbf{L}$, and $\{q_n\}$ and $\{\Delta_n\}$ are such that q_n and Δ_n are increasing with n, $q_n \to \infty$ and $\Delta_n \to \infty$ as $n \to \infty$, $\Delta_n = o(n^{1/4})$ and either* (i) $q_n\Delta_n^4 \log q_n\Delta_n = o(n)$ *or* (ii) $q_n\Delta_n^2 \log q_n\Delta_n = o(n^{1/2})$. $\square$

We can now state the desired consistency result.

THEOREM 3.3. *Given Assumptions* A.1(i), A.2 and A.3(i) *or* A.1(ii), A.2, *and* A.3(ii), *there exists a measurable connectionist sieve estimator $\hat{\theta}_n: \Omega \to \Theta$ such that*

$$n^{-1}\sum_{t=1}^{n}[Y_t - \hat{\theta}_n(X_t)]^2 = \min_{\theta \in \Theta_n(\psi)} n^{-1}\sum_{t=1}^{n}[Y_t - \theta(X_t)]^2, \qquad n = 1, 2, \ldots.$$

Further, $\rho(\hat{\theta}_n, \theta_o) \xrightarrow{p} 0$ (i.e., for all $\varepsilon > 0$ $P[\omega \in \Omega: \rho(\hat{\theta}_n(\omega), \theta_o) > \varepsilon] \to 0$ as $n \to \infty$. $\square$

This theorem delivers the fundamental result that connectionist network models are capable of consistently estimating (learning) an arbitrary conditional expectation function θ_o, an element of the class of functions square integrable on I^r. The result gives precise specifications for the growth in network complexity sufficient to achieve consistency, although as we saw in section 2 these specifications still permit a wide latitude for choice of network complexity. The consistency result holds for a wide variety of activation functions such as the logistic or hyperbolic tangent squashers, choices common in practice.

Consistency of cross-validation-based estimates for θ_o obtains under an appropriate modification of Assumption A.3.

ASSUMPTION B.3. *Let $\underline{\Theta}_n(\psi) = T(\psi, \underline{q}_n, \Delta_n)$ and $\overline{\Theta}_n = T(\psi, \overline{q}_n, \Delta_n)$ where $\psi \in \mathbf{L}$, and suppose that $\{\underline{q}_n\}$, $\{\overline{q}_n\}$ and $\{\Delta_n\}$ are such that $\underline{q}_n < \overline{q}_n$, $\{\underline{q}_n\}$, $\{\overline{q}_n\}$ and $\{\Delta_n\}$ are increasing, $\underline{q}_n \to \infty$ and $\Delta_n \to \infty$ as $n \to \infty$, $\Delta_n = o(n^{1/4})$ and either* (i) $\overline{q}_n\Delta_n^4 \log \overline{q}_n\Delta_n = o(n)$ *or* (ii) $\overline{q}_n\Delta_n^2 \log \overline{q}_n\Delta_n = o(n^{1/2})$. $\square$

The consistency result for the cross-validated sieve estimator is a straightforward consequence of Theorems 4.2 and 4.1 of the next section, provided that $\hat{\Theta}_n(\psi) \equiv \Theta_n(\psi, \hat{q}_n) = T(\psi, q_n, \Delta_n)$ has an analytic graph. This demonstration is straightforward, and is relegated to the Appendix. We have the following result.

THEOREM 3.4. *Given Assumptions* A.1(i), A.2 *and* B.3(i) *or* A.1(ii), A.2 *and* B.3(ii), *there exists a measurable cross-validated sieve estimator $\tilde{\theta}_n: \Omega \to \Theta$ such that*

$$n^{-1}\sum_{t=1}^{n}[Y_t - \tilde{\theta}_n(X_t)]^2 = \min_{\theta \in \hat{\Theta}_n(\psi)} n^{-1}\sum_{t=1}^{n}[Y_t - \theta(X_t)]^2, \qquad n = 1, 2, \ldots.$$

where $\hat{\Theta}_n(\psi) = \Theta_n(\psi, \hat{q}_n)$ *and* $\hat{q}_n$ *is cross-validated network complexity given in section* 2, *based on* $\hat{\theta}^{q}_{m(t)}$ *measurable-*$\mathbf{F}/\mathbf{B}(\Theta)$, $t = 1, \ldots, n$, $q = \underline{q}_n, \ldots, \bar{q}_n$, $n = 1, 2, \ldots$.

Further, $\rho(\ddot{\theta}_n, \theta_o) \xrightarrow{p} 0$. □

The data-driven cross-validation procedure for determination of network complexity thus delivers a consistent estimator for θ_o (i.e., a learning procedure capable of eventually approximating an arbitrary conditional expectation to any desired level of accuracy with probability arbitrarily close to one).

To obtain results for the case of approximate optimization, we impose the following condition.

ASSUMPTION B.4. *Let* $\ddot{q}_n: \Omega \to \{\underline{q}_n, \ldots, \bar{q}_n\}$ *be a measurable mapping*, $n = 1, 2, \ldots$, *and with* $\bar{\theta}_n$: $\Omega \to \bar{\Theta}_n$ *denoting a measurable solution to the problem* $\min_{\theta \in \Theta_n(\psi, \ddot{q}_n)} n^{-1} \sum_{t=1}^{n} [Y_t - \theta(X_t)]^2$, *let* $\ddot{\theta}_n$: $\Omega \to \bar{\Theta}_n$ *be a measurable mapping such that* $\ddot{\theta}_n(\omega) \in \Theta_n(\psi, \ddot{q}_n(\omega))$ *for each* ω *in* Ω *and*

$$|n^{-1} \sum_{t=1}^{n} [Y_t - \ddot{\theta}_n(X_t)]^2 - n^{-1} \sum_{t=1}^{n} [Y_t - \bar{\theta}_n(X_t)]^2| \leq \zeta_n,$$

$n = 1, 2, \ldots$, *where* $\{\zeta_n\}$ *is a sequence of positive real numbers such that* $\zeta_n \to \zeta_o \geq 0$ *as* $n \to \infty$. □

We define the set $\Theta^*(\zeta) \equiv \{\theta \in \Theta: |E([Y_t - \theta(X_t)]^2) - E([Y_t - \theta_o(X_t)]^2)| \leq \zeta\} = \{\theta \in \Theta: \rho(\theta, \theta_o) \leq \zeta^{1/2}\}$. The desired result is

THEOREM 3.5. *Given Assumptions* A.1(i), A.2, B.3(i), *and* B.4 *or* A.1(ii), A.2, B.3(ii), *and* B.4, *for any* $\xi > 0$ *we have* $P[\ddot{\theta}_n \in \Theta^*(\zeta_o + \xi)] \to 1$ *as* $n \to \infty$. *If in addition* $\zeta_o = 0$, *then* $\rho(\ddot{\theta}_n, \theta_o) \xrightarrow{p} 0$. □

The final conclusion delivers the consistency of $\ddot{\theta}_n$ for θ_o when optimization becomes exact in the limit ($\zeta_o = 0$). The first conclusion provides a precise statement of the effects of approximate optimization. In the limit, the approximate estimator belongs to a neighborhood of the true mapping θ_o with probability approaching unity. The smaller is ζ_n, the smaller is the neighborhood.

4. THEORETICAL FOUNDATIONS

This section contains theoretical results underlying those of the previous section. A variety of other network learning results can be obtained from the results given here. Our results follow as a corollary to Theorem 2.1 of White and Wooldridge (1990) (WW). For simplicity, the result here is stated so as to be easily applied to stationary stochastic processes. Its validity is not limited to this case, however. Where possible, notation follows that of WW. Other notation and definitions are as in Stinchcombe and White (1989b) (SW). We write $\mathbf{B}(\cdot)$ to denote the Borel σ-field generated by the open sets of the argument set, $\mathrm{gr}(\cdot)$ denotes the graph of the indicated correspondence, and $\mathbf{A}(\cdot)$ is the collection of analytic sets of the indicated σ-field.

THEOREM 4.1. (a) *Let* $(\Omega, \mathbf{F}, P)$ *be a complete probability space and let* (Θ, ρ) *be a metric space. For* $n = 1, 2, \ldots$, *let* Θ_n *be a complete separable Borel subset of* Θ *and let* $\hat{\Theta}_n$: $\Omega \to \Theta$ *be a correspondence with* $\mathrm{gr}\,\hat{\Theta}_n \in \mathbf{A}(\mathbf{F} \otimes \mathbf{B}(\Theta_n))$ *such that for each* ω *in* Ω $\hat{\Theta}_n(\omega) \subset \Theta_n$, *and the set* $\hat{\Theta}_n(\omega)$ *is nonempty and compact. Let* Q_n: $\Omega \times \Theta \to \bar{\mathbb{R}}$ *be* $\mathbf{F} \otimes \mathbf{B}(\Theta)$*-measurable, and suppose that* $Q_n(\omega, \cdot)$ *is lower semicontinuous on* Θ_n *for each* ω *in* Ω, $n = 1, 2, \ldots$.

Then for each $n = 1, 2, \ldots$ *there exists a function* $\hat{\theta}_n$: $\Omega \to \Theta_n$ *measurable-*$\mathbf{F}/\mathbf{B}(\Theta_n)$ *(hence-*$\mathbf{F}/\mathbf{B}(\Theta)$*) such that* $Q_n(\omega, \hat{\theta}_n(\omega)) = \min_{\theta \in \hat{\Theta}_n(\omega)} Q_n(\omega, \theta)$ *for all* ω *in* Ω.

(b) *In addition, suppose* $\{\Theta_n\}$ *and* $\{\bar{\Theta}_n\}$ *are increasing sequences of compact subsets of* Θ *such that* $\cup_{n=1}^{\infty} \Theta_n$ *is dense in* Θ *and* $\Theta_n \subseteq \hat{\Theta}_n(\omega) \subseteq \bar{\Theta}_n$ *for all* ω *in* Ω, $n = 1, 2, \ldots$. *Suppose there exists a function* $\bar{Q}$: $\Theta \to \bar{\mathbb{R}}$ *such that for all* $\epsilon > 0$

$$P[\omega: \sup_{\theta \in \bar{\Theta}_n} |Q_n(\omega, \theta) - \bar{Q}(\theta)| > \epsilon] \to 0 \quad \text{as } n \to \infty, \tag{4.1}$$

and for $\theta_o \in \Theta$

$$\inf_{\theta \in \eta^c(\theta_o, \varepsilon)} \bar{Q}(\theta) - \bar{Q}(\theta_o) > 0, \tag{4.2}$$

where $\eta^c(\theta_o, \varepsilon) \equiv \{\theta \in \Theta: \rho(\theta, \theta_o) \geq \varepsilon\}$, *and* $\bar{Q}$ *is continuous at* θ_o. *Then* $\rho(\hat{\theta}_n, \theta_o) \xrightarrow{p} 0$. □

In our application, $(\Omega, \mathbf{F}, P)$ is the space on which the stochastic process $\{Z_t\}$ is defined. The properties of P determine whether $\{Z_t\}$ is an independent or a mixing sequence. The space Θ contains the object of interest (the unknown regression function θ_o), and ρ is a metric that measures distance in this space (weighted mean squared error). Q_n is the criterion function (squared error) optimized to arrive at an estimator $\hat{\theta}_n$. The set $\hat{\Theta}_n(\omega)$ over which optimization is carried out may depend on the data through ω; this permits treatment of cross-validation procedures. In some applications it may be natural to have Q_n defined only on the graph of $\hat{\Theta}_n(\omega)$ or on $\Omega \times \Theta_n$ instead of on all of $\Omega \times \Theta$ as we assume. Lemma 2.1 of SW establishes that defining Q_n on $\Omega \times \Theta$ results in no loss of generality, as there generally exists an appropriate measurable extension to $\Omega \times \Theta$ of Q_n originally defined on $\mathrm{gr}\,\hat{\Theta}_n$ or $\Omega \times \Theta_n$.

Part (a) establishes the existence of a measurable estimator $\hat{\theta}_n$. Without measurability we cannot make probability statements about $\hat{\theta}_n$, such as statements

about consistency. Part (b) establishes consistency of $\hat{\theta}_n$ for θ_o. The object θ_o is distinguished by its role as minimizer of $\bar{Q}$ (condition (4.2)), the limit to which Q_n converges uniformly (condition (4.1)). This uniform convergence can be verified in particular stochastic contexts; our next result permits this for i.i.d. and stationary mixing processes. The other notable assumption of part (b) is the existence of nonstochastic sets Θ_n and $\bar{\Theta}_n$ bounding $\hat{\Theta}_n(\omega)$. The behavior of Θ_n ensures that $\hat{\Theta}_n(\omega)$ becomes sufficiently dense in $\bar{\Theta}$, so that an element of $\hat{\Theta}_n(\omega)$ can well approximate an element of Θ. The behavior of $\bar{\Theta}_n$ ensures that $\hat{\Theta}_n(\omega)$ does not increase too fast and that Q_n converges to $\bar{Q}$ uniformly in an appropriate sense. The constants $\underline{q}_n$ and $\bar{q}_n$ of the previous section determine Θ_n and $\bar{\Theta}_n$ in our application.

We now give a result permitting verification of condition (4.1), related to Corollary 2 of Haussler (1989). Instead of using the concept of V–C dimension (Vapnik & Chervonenkis, 1971) we use the concept of metric entropy (Kolmogorov & Tihomirov, 1961; see also Lorentz, 1966). The metric entropy of a closed set K, denoted $H(\varepsilon)$, is the logarithm of the number of sets of radius ε (with respect to a specified metric, not necessarily the same as that in Theorem 4.1) required to cover K. We also make use of exponential inequalities available for both independent and dependent stochastic processes. To specify precisely the allowed dependence, we utilize mixing measures of stochastic dependence, in particular, uniform ($\phi-$) and strong ($\alpha-$) mixing. These are defined as

$$\phi(k) = \sup_t \sup_{A \in \mathbf{F}_1^t, B \in \mathbf{F}_{t+k}^\infty, P(A) \neq 0} |P(B|A) - P(B)|$$

$$\alpha(k) = \sup_t \sup_{A \in \mathbf{F}_1^t, B \in \mathbf{F}_{t+k}^\infty} |P(A \cap B) - P(A)P(B)|$$

where $\mathbf{F}_1^t \equiv \sigma(Z_1, \ldots, Z_t)$ is the σ-field generated by $\{Z_1, \ldots, Z_t\}$, and $\mathbf{F}_t^\infty \equiv \sigma(Z_t, \ldots)$ is the σ-field generated by $\{Z_t, Z_{t+1}, \ldots\}$. For a discussion of $\phi(k)$ and $\alpha(k)$ and the properties of mixing processes $\{Z_t\}$ (i.e., processes for which $\phi(k) \to 0$ or $\alpha(k) \to 0$ as $k \to \infty$), we refer to White (1984). For our next result, the metric ρ need not be identical to that of Theorem 4.1.

THEOREM 4.2. *Let $(\Omega, \mathbf{F}, P)$ be a complete probability space, let (Θ, ρ) be a metric space, and let $\{\Theta_n\}$ be a (nonstochastic) increasing sequence of separable subsets of Θ. Let $H_n(\varepsilon)$ be the metric entropy of Θ_n, and put $G_n(\varepsilon) \equiv \exp H_n(\varepsilon)$.*

Let $\{s_n: \mathbb{I}^v \times \Theta_n \to \bar{\mathbb{R}}, n = 1, 2, \ldots\}$ and $\{m_n: \mathbb{I}^v \times \Theta_n \to \bar{\mathbb{R}}^+, n = 1, 2, \ldots\}$ be sequences of functions such that $s_n(\cdot, \theta)$ and $m_n(\cdot, \theta)$ are continuous on $\mathbb{I}^v$ for each θ in Θ_n. Suppose there exists a sequence $\{d_n: \Theta_n \to \mathbb{R}^+\}$ and a constant $\lambda > 0$ such that for each $z \in \mathbb{I}^v$ and θ^o in Θ_n

$$|s_n(z, \theta) - s_n(z, \theta^o)| < m_n(z, \theta^o)\rho(\theta, \theta^o)^\lambda$$

for all θ in $\eta_n(\theta^o) \equiv \{\theta \in \Theta_n: \rho(\theta, \theta^o) < d_n(\theta^o)\}$. Put $\bar{s}_n \geq \sup_{z \in \mathbb{I}^v} \sup_{\theta \in \Theta_n} |s_n(z, \theta)|$ and $\bar{m}_n \geq \sup_{z \in \mathbb{I}^v} \sup_{\theta \in \Theta_n} m_n(z, \theta)$. Put $\underline{d}_n \equiv \inf_{\theta \in \Theta_n} d_n(\theta)^\lambda$, and suppose $\bar{m}_n \geq \underline{d}_n^{-1}$.

Let $\{Z_t: \Omega \to \mathbb{I}^v\}$ be a stochastic process on $(\Omega, \mathbf{F}, P)$. (i) *If $\{Z_t\}$ is an i.i.d. sequence then for any $\varepsilon > 0$ and for all n sufficiently large*

$$P\left[\sup_{\theta \in \Theta_n} |n^{-1} \sum_{t=1}^n [s_n(Z_t, \theta) - E(s_n(Z_t, \theta))]| > \varepsilon\right] \leq 2G_n([\varepsilon/6\bar{m}_n]^{1/\lambda})[\exp(-6n/7) + \exp(-\varepsilon^2 n/\bar{s}_n^2[18 + 4\varepsilon])].$$

If in addition $\{\Theta_n\}$ is such that $n^{-1}\bar{s}_n^2 \to 0$ as $n \to \infty$ and for all $\varepsilon > 0$

$$(\bar{s}_n^2/n)\, H_n([\varepsilon/6\bar{m}_n]^{1/\lambda}) \to 0 \qquad \text{as } n \to \infty \tag{4.3}$$

then for any $\varepsilon > 0$

$$P\left[\sup_{\theta \in \Theta_n} |n^{-1} \sum_{t=1}^n [s_n(Z_t, \theta) - E(s_n(Z_t, \theta))]| > \varepsilon\right] \to 0 \qquad \text{as } n \to \infty.$$

(ii) *If $\{Z_t\}$ is a stationary mixing process with either $\phi(k) = \phi_o \rho_o^k$ or $\alpha(k) = \alpha_o \rho_o^k$, $\phi_o, \alpha_o > 0$, $0 < \rho_o < 1$, $k \geq 1$, then there exist constants $0 < c_1, c_2 < \infty$ not depending on n such that for any $\varepsilon > 0$ and all n sufficiently large*

$$P\left[\sup_{\theta \in \Theta_n} |n^{-1} \sum_{t=1}^n [s_n(Z_t, \theta) - E(s_n(Z_t, \theta))]| > \varepsilon\right] \leq c_1 G_n([\varepsilon/6\bar{m}_n]^{1/\lambda})[\exp(-c_2 n^{1/2}) + \exp(-c_2 \varepsilon n^{1/2}/6\bar{s}_n)].$$

If in addition $\{\Theta_n\}$ is such that $n^{-1}\bar{s}_n^2 \to 0$ as $n \to \infty$ and for all $\varepsilon > 0$

$$(\bar{s}_n^2/n) H_n([\varepsilon/6\bar{m}_n]^{1/\lambda}) \to 0 \quad \text{as } n \to \infty \tag{4.4}$$

then for any $\varepsilon > 0$

$$P\left[\sup_{\theta \in \Theta_n} \left| n^{-1} \sum_{t=1}^n [s_n(Z_t, \theta) - E(s_n(Z_t, \theta))] \right| > \varepsilon\right] \to 0 \qquad \text{as } n \to \infty.$$

Except for the metric entropy conditions (4.3) or (4.4), the conditions of this result are straightforward to verify in applications. The metric entropy $H_n(\varepsilon)$ is merely a little tedious to determine in most applications; our next result provides bounds for our application. Given $H_n(\varepsilon)$, choosing Θ_n appropriately is fairly direct, allowing verification of condition (4.1) with $Q_n(\cdot, \theta) = n^{-1} \sum_{t=1}^n s_n(Z_t, \theta)$.

A (loose) bound for the metric entropy of single hidden layer feedforward networks is given by the following result.

LEMMA 4.3. *Let $\psi \in \mathbf{L}$, $q \in \mathbb{N}$ and $\Delta \in \mathbb{R}^+$ be given. Let ρ_∞ denote the uniform metric, $\rho_\infty(\theta_1, \theta_2) \equiv$*

$\sup_{x \in I^r} |\theta_1(x) - \theta_2(x)|$, $\theta_1, \theta_2 \in T(\psi, q, \Delta)$, *and let* $H_\varepsilon(\psi, q, \Delta)$ *denote the metric entropy of* $T(\psi, q, \Delta)$ *with respect to* ρ_∞. *Then for all* $\varepsilon > 0$ *sufficiently small,*

$$H_\varepsilon(\psi, q, \Delta) \leq p \log 8/\varepsilon + p \log [\Delta + rL\Delta^r] + p \log p$$

where $p \equiv q(r + 2) + 1$. □

To apply Theorem 4.2 to connectionist regression, we must specify Q_n, s_n and m_n; from these $\bar{s}_n$ and $\bar{m}_n$ needed for (4.3) and (4.4) follow. The method of least squares is implemented by taking

$$Q_n(\cdot, \theta) = n^{-1} \sum_{t=1}^{n} [Y_t - \theta(X_t)]^2,$$

so that $s_n(z, \theta) = [y - \theta(x)]^2$. Because $(a^2 - b^2) = (a + b)(a - b)$, we have

$$\begin{aligned} |s_n(z, \theta) - s_n(z, \theta^*)| &\leq |2y - (\theta(x) + \theta^*(x))| \cdot |\theta(x) - \theta^*(x)| \\ &\leq \Big(\sup_{\theta \in \Theta_n(\omega)} 2|y - \theta(x)|\Big) \rho_s(\theta, \theta^*), \end{aligned}$$

so we take $m_n(z, \theta) = \sup_{\theta \in \Theta_n(\psi)} 2|y - \theta(x)|$, $\lambda = 1$ and $d_n(\theta) = 1$. Without loss of generality, we may take $\sup_{x \in I^r} |\theta(x)| \geq 1$ (this supremum is finite because the Lipschitz condition of ψ ensures the continuity of θ), so that

$$\begin{aligned} \sup_{z \in I^{r+1}} m_n(z, \theta) &\leq \sup_{z \in I^{r+1}} \sup_{\theta \in \Theta_n(\psi)} 2|y| + 2|\theta(x)| \\ &\leq \sup_{x \in I^r} \sup_{\theta \in \Theta_n(\psi)} 4|\theta(x)|, \end{aligned}$$

as $|y| \leq 1$. As ψ is assumed bounded, we may take the bound to be unity, without loss of generality. We then put $\bar{m}_n = 4\Delta_n \geq \sup_{x \in I^r} \sup_{\theta \in \Theta_n(\psi)} 4|\theta(x)|$. Similarly,

$$[y - \theta(x)]^2 \leq y^2 + 2|y| \cdot |\theta(x)| + \theta(x)^2 \leq \sup_{x \in I^r} 4\theta(x)^2,$$

and we take $\bar{s}_n = 4\Delta_n^2 \geq \sup_{x \in I^r} \sup_{\theta \in \Theta_n(\psi)} 4\theta(x)^2$.

Provided that the metric entropy conditions (4.3) and (4.4) hold, we obtain from Theorem 4.2 that (4.1) holds (i.e., $Q_n(\cdot, \theta)$ converges uniformly to $\bar{Q}(\theta) = E([Y_t - \theta(X_t)]^2)$). Using Lemma 4.3, we obtain conditions on $\{q_n\}$ and $\{\Delta_n\}$ ensuring the validity of the metric entropy conditions (4.3) and (4.4).

LEMMA 4.4. *Let* $\{q_n \in \mathbb{N}\}$ *and* $\{\Delta_n \in \mathbb{N}^+\}$ *be sequences such that* $\{\Delta_n\}$ *is increasing and* $\Delta_n \to \infty$ *as* $n \to \infty$. *Put* $\bar{s}_n = 4\Delta_n^2$, $\bar{m}_n = 4\Delta_n$, $\lambda = 1$, *and for given* $\psi \in \mathbf{L}$ *and* $\varepsilon > 0$ *put* $H_n(\varepsilon) = H_\varepsilon(\psi, q_n, \Delta_n)$.

(i) *If* $q_n \Delta_n^4 \log q_n \Delta_n = o(n)$, *then for any* $\varepsilon > 0$

$$(\bar{s}_n^2/n) H_n([\varepsilon/6\bar{m}_n]^{1/r}) \to 0 \quad \text{as } n \to \infty;$$

(ii) *If* $q_n \Delta_n^2 \log q_n \Delta_n = o(n^{1/2})$, *then for any* $\varepsilon > 0$

$$(\bar{s}_n/n^{1/2}) H_n([\varepsilon/6\bar{m}_n]^{1/r}) \to 0 \quad \text{as } n \to \infty. \quad \square$$

Here we see the growth rates for network complexity appearing in Assumptions A.3 and B.3. The growth rate for Δ_n, $\Delta_n = o(n^{1/4})$, follows from the requirement in Theorem 4.2 that $n^{-1} \bar{s}_n^2 \to 0$, with $\bar{s}_n$ set to $4\Delta_n^2$.

The result for approximate optimization follows as a corollary to Theorem 4.1.

COROLLARY 4.5. *Let the conditions and definitions of Theorem 4.1 hold. For* $\zeta \geq 0$, *define* $\Theta_n(\omega, \zeta) = \{\theta \in \Theta_n(\omega): |Q_n(\omega, \theta) - Q_n(\omega, \ddot{\theta}_n(\omega))| \leq \zeta\}$ *and* $\Theta^*(\zeta) = \{\theta \in \Theta: |\bar{Q}(\theta) - \bar{Q}(\theta_0)| \leq \zeta\}$. *Let* $\{\zeta_n \in \mathbb{R}^+\}$ *satisfy* $\zeta_n \to \zeta_0 \geq 0$ *as* $n \to \infty$, *and let* $\{\ddot{\theta}_n: \Omega \to \bar{\Theta}_n\}$ *be a sequence of measurable functions such that* $\ddot{\theta}_n(\omega) \in \Theta_n(\omega, \zeta_n)$, $n = 1, 2, \ldots$. *Then for all* $\zeta > 0$, $P[\ddot{\theta}_n \in \Theta^*(\zeta_0 + \zeta)] \to 1$ *as* $n \to \infty$. *If in addition* $\zeta_0 = 0$, *then* $\rho(\ddot{\theta}_n, \theta_0) \xrightarrow{p} 0$. □

5. SUMMARY AND DIRECTIONS FOR FURTHER RESEARCH

Our results resolve the learnability issue for multilayer feedforward networks by establishing the consistency of connectionist sieve estimators for an unknown regression function. These networks thus possess nonparametric regression capabilities. Network complexity plays a fundamental role in our analysis. We consider both deterministic and data-driven methods for controlling the growth of network complexity as a function of network experience. The data-driven methods are based on a cross-validation measure of network performance, cross-validated average squared error, that we advocate for general use in evaluating network performance. This is not the only appropriate or useful measure (e.g., see Barron, 1989), but it offers a considerable improvement over naive methods. We also consider the implications of approximate rather than exact optimization of network performance.

Consistency is a minimal property for a learning rule; the analysis here is inherently limited by our focus on this property. There are numerous important areas for further investigation. Especially desirable is further analysis regarding the rate of convergence of connectionist sieve estimators. Among other things, this will help make possible more informed application of the procedures described here. The results and approach of Severini and Wong (1987) may prove helpful in this investigation. Such results will permit further evaluation of the trade-offs between Δ_n and q_n required to ensure convergence at the best possible rate, and will facilitate comparisons with other nonparametric regression methods, such as kernel (Bierens, 1987) and spline (Cox, 1984) methods. Rate of convergence results may also be useful in establishing asymptotic optimality proper-

ties (as in Li, 1987) for cross-validation procedures. Another area for further research is greater automation of consistent and optimal procedures. Developing and implementing well-behaved methods for computing $\hat{\theta}^q_{n(t)}$ is especially important for practical application of the cross-validation approach.

Straightforward extensions of our analysis yield nonparametric estimators of conditional functionals other than the conditional expectations studied here, such as conditional variance or conditional quantiles. These can be estimated with proper choice of Q_n, as discussed in WW. In fact, it is possible to use multilayer feedforward networks to obtain nonparametric estimates of joint or conditional density, using the approach described by Geman and Hwang (1982, section 6). Such multilayer feedforward networks would be essentially self-organizing, and could be used in pattern completion problems in which arbitrary parts of the pattern are supplied and the arbitrary remainder has to be completed.

Finally, we mention that it is often useful to have available statistical confidence regions for linear or nonlinear functionals of the nonparametric estimator. For example, we might want a 95% confidence interval for the conditional expectation of Y_t given $X_t = x$, a specified value. This interval may be obtainable using methods similar to those of Andrews (1988).

MATHEMATICAL APPENDIX

Unless otherwise noted, all definitions and notations are as given in the text.

Proof of Lemma 3.2. Let $\Sigma^r(\psi) \equiv \{g: \mathbb{R}^r \to \mathbb{R}: g(x) = \beta_0 + \sum_{j=1}^q \beta_j \psi(x'\gamma_j), x \in \mathbb{I}^r, \beta_j \in \mathbb{R}, \gamma_j \in \mathbb{R}^{r+1}$ $j = 1, \ldots, q, q \in \mathbb{N}\}$. It follows immediately from Theorem 2.4 of Hornik, Stinchcombe, and White (1989) or Corollary 3.6 of Hornik, Stinchcombe, and White (1990) and Theorem 3.14 of Rudin (1974) that $\Sigma^r(\psi)$ is ρ_2-dense in $L^2(\mathbb{I}^r, \mu)$. Let $g: \mathbb{I}^r \to \mathbb{R}$ be an arbitrary element of $\Sigma^r(\psi)$, so that for some $q \in \mathbb{N}$, $\beta_j \in \mathbb{R}, j = 0, \ldots, q, \gamma_j \in \mathbb{R}^{r+1}, j = 1, \ldots, q$, we have $g(x) = \beta_0 + \sum_{j=1}^q \beta_j \psi(x'\gamma_j)$. Because $q_n \to \infty$ and $\Delta_n \to \infty$, we can always pick n sufficiently large that $\sum_{j=0}^q |\beta_j| \leq \Delta_n$, $\sum_{j=1}^q \sum_{i=0}^r |\gamma_{ji}| \leq q_n \Delta_n$, and $q \leq q_n$. Thus, for n sufficiently large g belongs to $\Theta_n(\psi) = T(\psi, q_n, \Delta_n)$ and therefore to $\cup_{n=1}^\infty \Theta_n(\psi)$. Because g is arbitrary, $\Sigma^r(\psi) \subset \cup_{n=1}^\infty \Theta_n(\psi)$. It follows from the ρ_2-denseness of $\Sigma^r(\psi)$ in $L^2(\mathbb{I}^r, \mu)$ that $\cup_{n=1}^\infty \Theta_n(\psi)$ is ρ_2-dense in $L^2(\mathbb{I}^r, \mu)$. □

Proof of Theorem 3.3. We apply Theorem 4.1. By Assumption A.1, $(\Omega, \mathbf{F}, P)$ is a complete probability space; $(\Theta, \rho) = (L^2(\mathbb{I}^r, \mu), \rho_2)$ is a complete separable metric space (e.g., Kolmogorov and Fomin, 1970, Theorem 37.5, Problem 37.4). Consequently, we may take $\Theta_n = \Theta, n = 1, 2, \ldots, \hat{\Theta}_n(\cdot) = \Theta_n(\psi) \equiv T(\psi, q_n, \Delta_n)$ is nonempty and compact (for all n sufficiently large) given Assumption A.3, and $\hat{\Theta}_n$ is an analytic correspondence because it is nonstochastic. $Q_n(\omega, \theta) \equiv n^{-1} \sum_{t=1}^n [Y_t(\omega) - \theta(X_t(\omega))]^2$ is $\mathbf{F} \times \mathbf{B}(\Theta)$ – measurable as a consequence of Lemma 2.2 of Stinchcombe and White (1989b) (SW) because for every ω in Ω $Q_n(\omega, \cdot)$ is $(\rho_2 -)$ continuous and for every θ in the separable metric space Θ $Q_n(\cdot, \theta)$ is measurable. Continuity of $Q_n(\omega, \cdot)$ implies lower semi-continuity. The existence of a measurable connectionist sieve estimator $\hat{\theta}_n: \Omega \to \Theta$ now follows from Theorem 4.1(a).

By Assumption A.3, $\{\underline{\Theta}_n = \Theta_n = T(\psi, q_n, \Delta_n)\}$ is an increasing sequence of compact subsets of Θ. By Lemma 3.2 $\cup_{n=1}^\infty \underline{\Theta}_n$ is dense in Θ. Arguments sketched in the text of section 4 and Lemmas 4.3 and 4.4 apply to establish the existence of $\overline{Q}: \Theta \to \mathbb{R}$, $\overline{Q}(\theta) = E([Y_t - \theta(X_t)]^2)$ such that $P[\omega: \sup_{\theta \in \Theta_n} |Q_n(\omega, \theta) - \overline{Q}(\theta)| > \varepsilon] \to 0$ as $n \to \infty$, as a consequence of Theorem 4.2, given Assumptions A.1(i) and A.3(i) or A.1(ii) and A.3(ii). Now $\overline{Q}(\theta) - \overline{Q}(\theta_o) = \rho(\theta, \theta_o)^2$, so $\inf_{\theta \in \Pi^c(\theta_o, \varepsilon)} \overline{Q}(\theta) - \overline{Q}(\theta_o) = \inf_{\theta \in \Pi^c(\theta_o, \varepsilon)} \rho(\theta, \theta_o)^2 \geq \varepsilon^2 > 0$ and $\overline{Q}(\theta) = \rho(\theta, \theta_o)^2 + \overline{Q}(\theta_o)$ is continuous at θ_o. That $\rho(\hat{\theta}_n, \theta_o) \xrightarrow{p} 0$ now follows from Theorem 4.1(b). □

The proof of Theorem 3.4 makes use of the following lemmas.

LEMMA A.1. *Let* $(\Theta, \rho) = (L^2(\mathbb{I}^r, \mu), \rho_2)$ *and let* $T(\psi, q, \Delta)$ *be as in Definition* 3.1 *with* $\psi \in \mathbf{L}$. *Let* $(\Omega, \mathbf{F})$ *be a measurable space and let* $\hat{q}: \Omega \to N$ *be measurable-*$\mathbf{F}/\mathbf{B}(N)$, *where* $N \subset \mathbb{N}$ *and* $\mathbb{N}$ *is endowed with the discrete topology. Then for every* ψ *and* Δ, $T(\psi, \hat{q}, \Delta): \Omega \rightrightarrows \Theta$ *has* gr $T(\psi, \hat{q}, \Delta) \in \mathbf{F} \otimes \mathbf{B}(\Theta)$. □

Proof. For each ψ and Δ, $T(\psi, q, \Delta)$ is a compact set, as it is the continuous image of the compact set $B \times \Gamma$, $B = \{\beta: \sum_{j=0}^q |\beta_j| \leq \Delta\}$, $\gamma = \{\gamma: \sum_{j=1}^q \sum_{i=0}^r |\gamma_{ji}| \leq q\,\Delta\}$. Then gr $T(\psi, \hat{q}, \Delta) = \cup_{n \in N}(\hat{q}^{-1}(n) \times T(\psi, n, \Delta)) \in \mathbf{F} \otimes \mathbf{B}(\Theta)$, because $\hat{q}$ is measurable ($\hat{q}^{-1}(n) \in \mathbf{F}$ for all $n \in N$) and $T(\psi, n, \Delta)$ is compact for all $n \in N$ (so $T(\psi, n, \Delta) \in \mathbf{B}(\Theta)$). □

LEMMA A.2. *Suppose Assumption A.1 holds, and put* $N_n = [\underline{q}_n, \overline{q}_n] \cap \mathbb{N}$ *for* $\underline{q}_n, \overline{q}_n \in \mathbb{N}$, $\underline{q}_n \leq \overline{q}_n$, $n = 1, 2, \ldots$. *For each* $n = 1, 2, \ldots, t = 1, \ldots, n$ *and* $q \in N_n$, *let* $\hat{\theta}^q_{n(t)}$ *be measurable-*$\mathbf{F}/\mathbf{B}(\Theta)$. *Then for each* n *there exists* $\hat{q}_n: \Omega \to N_n$ *measurable-*$\mathbf{F}/\mathbf{B}(N_n)$ *such that*

$$\hat{q}_n \in N_n \equiv \operatorname*{argmin}_{q \in N_n} n^{-1} \sum_{t=1}^n [Y_t - \hat{\theta}^q_{n(t)}(X_t)]^2,$$

and for all ω *in* Ω *and* $\tilde{q}_n \in N_n$ *we have* $\hat{q}_n(\omega) \leq \tilde{q}_n(\omega)$. □

Proof. We apply Corollary 2.4 of SW. Condition (i) of SW holds given Assumption A.1, and condition (ii) holds as $(N_n, \mathbf{B}(N_n))$ is a complete separable metric space under the discrete topology. For each q in N_n, $Q_n(\cdot, q) \equiv n^{-1} \sum_{t=1}^n [Y_t - \hat{\theta}^q_{n(t)}(X_t)]^2$ is measurable-$\mathbf{F}$ given the measurability of the elements of $\{\hat{\theta}^q_{n(t)}\}$ and $\{Z_t\}$. Because N_n is a finite set Q_n: $\Omega \times N_n \to \mathbb{R}$ is measurable-$\mathbf{F} \otimes \mathbf{B}(N_n)$, so that condition (iii) holds. Condition (iv.b) holds as N_n does not depend on ω, so that N_n is an analytic set. Because N_n is a finite set, a minimum exists. Thus, by Corollary 2.4 of SW, $\hat{q}_n \in N_n^*$ and $\hat{q}_n$ is measurable-$\mathbf{F}/\mathbf{B}(N_n)$. Further, the arbitrary selection of Corollary 2.4 can be taken to be such that for all ω in Ω and $\hat{q}_n \in N_n^*$ we have $\hat{q}_n(\omega) \leq \bar{q}_n(\omega)$. □

Proof of Theorem 3.4. The proof is analogous to that of Theorem 3.3, except for some details. Lemmas A.1–A.2 immediately imply that gr $\hat{\Theta}_n(\psi) \in \mathbf{F} \otimes \mathbf{B}(\Theta) \subset \mathbf{A}(\mathbf{F} \otimes \mathbf{B}(\Theta))$. The denseness of $\cup_{n=1}^\infty \underline{\Theta}_n$ follows from Lemma 3.2 under Assumption B.3. By construction, $\underline{\Theta}_n \subseteq \hat{\Theta}_n(\psi) \subseteq \bar{\Theta}_n$, as $\underline{q}_n \leq \hat{q}_n \leq \bar{q}_n$. Condition (4.1) holds by Theorem 4.2 for $\{\bar{\Theta}_n\}$, using the same argument as in Theorem 3.3. As the rest of the argument is unchanged, the conditions of Theorem 4.1 hold and the result follows. □

Proof of Theorem 3.5. The result follows immediately from Theorem 4.5, as the conditions of Theorem 4.1 hold as established in the proof of Theorem 3.4. □

Proof of Theorem 4.1. (a): The result follows immediately from Corollary 2.4 of SW, as we directly impose their conditions (i)–(iii) and (iv.b) on $(\Omega, \mathbf{F})$ and (Θ_n, ρ). (Completeness and separability of Θ_n ensure that it is Souslin, condition (ii).) We also impose their conditions (v.a) ($\hat{\Theta}_n(\omega)$ nonempty and compact) and (vi.b) ($Q_n(\omega, \cdot)$ lower semi-continuous on Θ_n for each ω), which ensure that $\{\omega$: $Q_n(\omega, \cdot)$ achieves its infimum in $\hat{\Theta}_n(\omega)\} = \Omega$.

(b) We first verify the conditions of Theorem 2.1 of White and Wooldridge (1990) (WW). The conditions imposed in (a) suffice for all of their conditions through their (2.1) and (2.2), as $\hat{\theta}_n(\omega)$ minimizes $Q_n(\omega, \theta)$ in $\hat{\Theta}_n(\omega)$. It remains to verify their (2.3) and (2.4), that is, there exists a nonstochastic sequence $\{\theta_n^* \in \Theta\}$ such that

$$P^*[\omega: \theta_n^* \in \hat{\Theta}_n(\omega)] \to 1 \quad \text{as } n \to \infty \qquad \text{(A.1)}$$

and for all $\varepsilon > 0$

$$P^*\left[\omega: \inf_{\theta \in \eta_n^c(\theta_n^*, \varepsilon, \omega)} Q_n(\omega, \theta) - Q_n(\omega, \theta_n^*) > 0\right] \to 1 \quad \text{as } n \to \infty, \qquad \text{(A.2)}$$

where $\eta_n^c(\theta_n^*, \varepsilon, \omega) \equiv \{\theta \in \hat{\Theta}_n(\omega)$: $\rho(\theta, \theta_n^*) \geq \varepsilon\}$ and P^* is the outer measure associated with $(\Omega, \mathbf{F}, P)$.

It follows by argument identical to the proof of Proposition 2.4 of WW that there exists a nonstochastic sequence $\{\theta_n^* \in \underline{\Theta}_n\}$ such that: θ_n^* is a solution to $\inf_{\theta \in \underline{\Theta}_n} \bar{Q}(\theta)$; for all $\varepsilon > 0$

$$\liminf_n \left\{ \inf_{\theta \in \eta^c(\theta_n^*, \varepsilon)} \bar{Q}(\theta) - \bar{Q}(\theta_n^*) \right\} > 0,$$

where $\eta^c(\theta_n^*, \varepsilon) \equiv \{\theta \in \Theta$: $\rho(\theta, \theta_n^*) \geq \varepsilon\}$; and $\rho(\theta_n^*, \theta_o) \to 0$ as $n \to \infty$. Because $\underline{\Theta}_n \subseteq \hat{\Theta}_n(\omega)$, $\{\theta_n^*\}$ satisfies (A.1).

Because $\eta^c(\theta_n^*, \varepsilon)$ contains $\bar{\eta}_n^c(\theta_n^*, \varepsilon) \equiv \{\theta \in \bar{\Theta}_n$: $\rho(\theta, \theta_n^*) \geq \varepsilon\}$, it follows that

$$\liminf_n \left\{ \inf_{\theta \in \bar{\eta}_n^c(\theta_n^*, \varepsilon)} \bar{Q}(\theta) - \bar{Q}(\theta_n^*) \right\} > 0.$$

By assumption $P[\omega$: $\sup_{\theta \in \bar{\Theta}_n} |Q_n(\omega, \theta) - \bar{Q}(\theta)| > \varepsilon] \to 0$ as $n \to \infty$ for any $\varepsilon > 0$; because $\underline{\Theta}_n \subseteq \bar{\Theta}_n$, we also have $\theta \in \bar{\Theta}_n$. It now follows immediately from the argument of Corollary 2.3 of WW that

$$P^*\left[\omega: \inf_{\theta \in \bar{\eta}_n^c(\theta_n^*)} Q_n(\omega, \theta) - Q_n(\omega, \theta_n^*) > 0\right] \to 1 \quad \text{as } n \to \infty.$$

Because $\hat{\Theta}_n(\omega) \subseteq \bar{\Theta}_n$, $\bar{\eta}_n^c(\theta_n^*, \varepsilon)$ contains $\eta_n^c(\theta_n^*, \varepsilon, \omega)$ so that

$$\left\{\omega: \inf_{\theta \in \bar{\eta}_n^c(\theta_n^*, \varepsilon)} Q_n(\omega, \theta) - Q_n(\omega, \theta_n^*) > 0\right\} \subseteq \left\{\omega: \inf_{\theta \in \eta_n^c(\theta_n^*, \varepsilon, \omega)} Q_n(\omega, \theta) - Q_n(\omega, \theta_n^*) > 0\right\}.$$

It follows immediately from the monotonicity of P^* that (A.2) holds, so that $\rho(\hat{\theta}_n, \theta_n^*) \xrightarrow{P^*} 0$ by Theorem 2.1 of WW.

Because $\hat{\theta}_n$ is measurable (by (a) above), $\rho(\hat{\theta}_n, \theta_n^*) \xrightarrow{P} 0$. Because $\rho(\theta_n^*, \theta_o) \to 0$, the triangle inequality yields $\rho(\hat{\theta}_n, \theta_o) \xrightarrow{P} 0$. □

Proof of Theorem 4.2. We verify the conditions of Theorem 2.5 of WW. The conditions on $(\Omega, \mathbf{F}, P)$, (Θ, ρ) and $\{\Theta_n\}$ are imposed directly. The functions $s_n(Z_t, \cdot)$ and $m_n(Z_t, \cdot)$ correspond to s_{nt} and m_{nt} of WW. The required measurability and Lipschitz conditions for these hold by assumption. We take M_n of WW to be $M_n = n\bar{m}_n$.

Under conditions (i), ($\{Z_t\}$ i.i.d.) it follows from the Bernstein inequality (e.g., Proposition 3.2 of WW) that

$$P\left[\left|\sum_{t=1}^n s_n(Z_t, \theta) - E(s_n(Z_t, \theta))\right| > \Delta\right] \leq 2 \exp[-\Delta^2/(2\sigma_n^2 + 4\bar{s}_n\Delta/3)],$$

where $\sigma_n^2 \equiv \text{var}(\sum_{t=1}^n s_n(Z_t, \theta))$, and we use the fact that $|s_n(Z_t, \theta) - E(s_n(Z_t, \theta))| \leq 2\bar{s}_n$. Further, $\sigma_n^2 \leq n\bar{s}_n^2$, and we may take $\bar{s}_n \geq 1$ (implying $\bar{s}_n \leq \bar{s}_n^2$)

without loss of generality, so that

$$P\left[\left|\sum_{t=1}^{n} s_n(Z_t, \theta) - E(s_n(Z_t, \theta))\right| > \Delta\right] \leq 2\exp[-\Delta^2/\bar{s}_n^2(2n + 4\Delta/3)].$$

Note that this bound is independent of θ, and define $\Gamma_n^s(\Delta) \equiv 2\exp[-\Delta^2/\bar{s}_n^2(2n + 4\Delta/3)]$. Application of the Bernstein inequality with m_n in place of s_n leads to an analogous inequality,

$$P\left[\left|\sum_{t=1}^{n} m_n(Z_t, \theta) - E(m_n(Z_t, \theta))\right| > \Delta\right] \leq 2\exp[-\Delta^2/(2n\bar{m}_n^2 + 4\bar{m}_n\Delta/3)].$$

This bound is also independent of θ; we define $\Gamma_n^m(\Delta) \equiv 2\exp[-\Delta^2/(2n\bar{m}_n^2 + 4\bar{m}_n\Delta/3)]$.

Theorem 2.5 of WW applies to yield the desired inequality, provided that $n = O(M_n \underline{d}_n)$. Because $M_n = n\bar{m}_n$ and $\bar{m}_n \geq \underline{d}_n^{-1}$, we see immediately that this holds. Consequently, by Theorem 2.5 of WW it follows that for all $\varepsilon > 0$ and all n sufficiently large

$$P\left[\sup_{\theta\in\Theta_n}\left|n^{-1}\sum_{t=1}^{n}[s_n(Z_t, \theta) - E(s_n(Z_t, \theta))]\right| > \varepsilon\right] \leq G_n([\varepsilon/6\bar{m}_n]^{1/2})[\Gamma_n^m(2M_n) + \Gamma_n^s(\varepsilon n/3)].$$

Now $\Gamma_n^m(2M_n) = 2\exp[-6n/7]$ and $\Gamma_n^s(\varepsilon n/3) = 2\exp[-\varepsilon^2 n/\bar{s}_n^2(18 + 4\varepsilon)]$. Substituting these expressions into the expression above gives the first result of part (i).

To obtain the second result (uniform convergence to zero), it suffices that

$$G_n([\varepsilon/6\bar{m}_n]^{1/2})\exp[-6n/7] \to 0 \quad \text{as } n \to \infty$$

and

$$G_n([\varepsilon/6\bar{m}_n]^{1/2})\exp[-\varepsilon^2 n/\bar{s}_n^2(18 + 4\varepsilon)] \to 0 \quad \text{as } n \to \infty$$

for any $\varepsilon > 0$. Because the force of the condition occurs when ε is small and because we take $\bar{s}_n \geq 1$, the second condition suffices for the first. For this it suffices that

$$\varepsilon^2 n/\bar{s}_n^2(18 + 4\varepsilon) - H_n([\varepsilon/6\bar{m}_n]^{1/2}) \to \infty \quad \text{as } n \to \infty$$

for any $\varepsilon > 0$. Now

$$\varepsilon^2 n/\bar{s}_n^2(18 + 4\varepsilon) - H_n([\varepsilon/6\bar{m}_n]^{1/2}) = (n/\bar{s}_n^2)[\varepsilon^2/(18 + 4\varepsilon) - (\bar{s}_n^2/n)H_n([\varepsilon/6\bar{m}_n]^{1/2})] \geq (n/\bar{s}_n^2)[\varepsilon^2/2(18 + 4\varepsilon)],$$

where inequality holds for all n sufficiently large because $(\bar{s}_n^2/n)H_n([\varepsilon/6\bar{m}_n]^{1/2}) \to 0$. Because $n^{-1}\bar{s}_n^2 \to 0$ we have $n/\bar{s}_n^2 \to \infty$, and the required divergence is verified.

The argument for the mixing case (ii) is analogous, using a mixing analog of the Bernstein inequality (Bosq, 1975, Theorem 3.3 of WW). The condition $(\bar{s}_n/n^{1/2})H_n([\varepsilon/6\bar{m}_n]^{1/2}) \to 0$ can be similarly verified to suffice for the uniform convergence result. □

Proof of Lemma 4.3. We construct an ε-net for $T(\psi, q, \Delta)$. (Recall that $T_\varepsilon = \{t_1, \ldots, t_h\}$ is an ε-net for T if each θ in T there is at least one element t_k of T_ε such that $\rho_\varepsilon(\theta, t_k) \leq \varepsilon$.) Then $H_\varepsilon(\psi, q, \Delta) \leq \log h$ (e.g., Lorentz, 1966, chap. 10).

Let $\eta > 0$ be given, and let $B_\eta \equiv \{b_k \in B, k = 1, \ldots, l\}$, $C_\eta \equiv \{c_k \in \Gamma, k = 1, \ldots, m\}$ be (l_1 − norm) η − nets for $B = \{\beta: \|\beta\| \leq \Delta\} \subset \mathbb{R}^{q+1}$ and $\Gamma = \{\gamma: \|\gamma\| \leq q\Delta\} \subset \mathbb{R}^{q(r+1)}$. (For notational convenience $\|\cdot\|$ here denotes the l_1-norm in spaces $\mathbb{R}^v$ of whatever dimension.) Let $D_\eta = B_\eta \times C_\eta$, and put $\tilde{T}_\eta = \{\theta \in T(\psi, q, \Delta): \delta \in D_\eta\}$ in an obvious use of notation. We choose η so that $\tilde{T}_\eta$ is an ε-net.

Let $\theta \in T(\psi, q, \Delta)$ be arbitrary, with corresponding parameters $\delta = (\beta', \gamma')'$. Then there exists $d = (b', c')'$ in D_η such that $\|\beta - b\| \leq \eta$ and $\|\gamma - c\| \leq \eta$. Let t be the element of $\tilde{T}_\eta$ corresponding to d. The triangle inequality gives

$$|\theta(x) - t(x)| = \left|\beta_0 + \sum_{i=1}^{q}\beta_i\psi(\tilde{x}'\gamma_i) - b_0 - \sum_{i=1}^{q} b_i\psi(\tilde{x}'c_i)\right| \leq \left|\beta_0 - b_0 + \sum_{i=1}^{q}(\beta_i - b_i)\psi(\tilde{x}'\gamma_i)\right| + \left|\sum_{i=1}^{q} b_i(\psi(\tilde{x}'\gamma_i) - \psi(\tilde{x}'c_i))\right|.$$

For the first term, the fact that ψ is bounded (set the bound to unity) and the triangle inequality give

$$\left|\beta_0 - b_0 + \sum_{i=1}^{q}(\beta_i - b_i)\psi(\tilde{x}'\gamma_i)\right| \leq \sum_{i=0}^{q}|\beta_i - b_i| = \|\beta - b\| \leq \eta.$$

For the second term, the triangle inequality gives

$$\left|\sum_{i=1}^{q} b_i(\psi(\tilde{x}'\gamma_i) - \psi(\tilde{x}'c_i))\right| \leq \sum_{i=1}^{q}|b_i|\cdot|\psi(\tilde{x}'\gamma_i) - \psi(\tilde{x}'c_i)| \leq \left(\sum_{i=1}^{q}|b_i|\right)\left(\sum_{i=1}^{q}|\psi(\tilde{x}\gamma_i) - \psi(\tilde{x}'c_i)|\right) \leq \Delta\sum_{j=1}^{q}|\psi(\tilde{x}'\gamma_i) - \psi(\tilde{x}'c_i)|.$$

The Lipschitz condition imposed on ψ and the triangle inequality give

$$|\psi(\tilde{x}'\gamma_i) - \psi(\tilde{x}'c_i)| \leq L|\tilde{x}'\gamma_i - \tilde{x}'c_i| = L\left|\sum_{i=0}^{r}\tilde{x}_i(\gamma_{ij} - c_{ij})\right| \leq L\sum_{i=0}^{r}|x_i|\cdot|\gamma_{ij} - c_{ij}| \leq L\left(\sum_{i=0}^{r}|x_i|\right)\left(\sum_{i=0}^{r}|\gamma_{ij} - c_{ij}|\right) \leq rL\sum_{i=0}^{r}|\gamma_{ij} - c_{ij}|.$$

Consequently, $\Sigma_{j=1}^{q} |\psi(\tilde{x}'\gamma_j) - \psi(\tilde{x}'c_j)| \leq rL \Sigma_{j=1}^{q} \Sigma_{i=1}^{r} |\gamma_{ij} - c_{ij}| \leq rL\eta$, so that

$$\left| \sum_{j=1}^{q} b_j(\psi(\tilde{x}'\gamma_j) - \psi(\tilde{x}'c_j)) \right| \leq rL\ \Delta\eta.$$

It follows that

$$\rho_x(\theta, t) \equiv \sup_{x \in I^r} |\theta(x) - t(x)| \leq \eta(1 + rL\ \Delta) = \varepsilon,$$

provided that we choose $\eta = \varepsilon/(1 + rL\ \Delta)$. Because θ is arbitrary and $t \in \tilde{T}_\eta$, $T_\varepsilon \equiv \tilde{T}_{\varepsilon/(1+rL\Delta)}$ is an ε-net for $T(\psi, q, \Delta)$.

Let # be the cardinality operator. Because $\#\tilde{T}_\eta = \#D_\eta$ and $D_\eta = B_\eta \times C_\eta$, we have $\#\tilde{T}_\eta = (\#B_\eta)(\#C_\eta)$. It follows from Theorems IX, X, and V of Kolmogorov and Tihomirov (1961) and elementary geometrical arguments that for all ε (hence η) sufficiently small $\#B_\eta \leq 2(2\Delta/\eta)^{q+1}$ and $\#C_\eta \leq 2(2q\Delta/\eta)^{q(r+1)}$. Therefore with $\eta = \varepsilon/(1 + rL\ \Delta)$, $p = q(r + 2) + 1$, we have

$$\begin{aligned} \log \#T_\varepsilon &\leq \log 4 + [q(r + 2) + 1]\log 2\Delta/\eta \\ &\quad + q(r + 1)\log q \\ &\leq \log 4 + p \log 2/\varepsilon + p \log(\Delta + rL\ \Delta^2) + p \log p. \end{aligned}$$

Because $H_\varepsilon(\psi, q, \Delta) = \log \#T_\varepsilon$ and $\log 4 < p \log 4$ we have

$$H_\varepsilon(\psi, q, \Delta) \leq p \log 8/\varepsilon + p \log(\Delta + rL\ \Delta^2) + p \log p.$$

□

Proof of Lemma 4.4. (i) With $\bar{m}_n = 4\Delta_n$ and $\lambda = 1$, for any $\varepsilon > 0$

$$\begin{aligned} H_n([\varepsilon/6\bar{m}_n]^{1/\lambda}) &= H_n(\varepsilon/24\Delta_n) \\ &\leq p_n \log 192\ \Delta_n/\varepsilon + p_n \log [\Delta_n + rL\ \Delta_n^2] + p_n \log p_n \end{aligned}$$

from Lemma 4.3. Because $\Delta_n \to \infty$, for any $\varepsilon > 0$ there exists N_ε sufficiently large that $\Delta_n \geq 192/\varepsilon$ for all $n > N_\varepsilon$. Also, for all n sufficiently large $\Delta_n^2 \geq \Delta_n$ and $\Delta_n^2 > (1 + rL)$. Without loss of generality suppose this is true for all $n > N_\varepsilon$. Consequently,

$$\begin{aligned} &p_n \log 192\ \Delta_n/\varepsilon + p_n \log[\Delta_n + rL\ \Delta_n^2] \\ &\quad \leq p_n \log \Delta_n^2 + p_n \log \Delta_n^2(1 + rL) \\ &\quad = p_n \log \Delta_n^2 + p_n \log \Delta_n^2 + p_n \log(1 + rL) \\ &\quad \leq 6\ p_n \log \Delta_n, \end{aligned}$$

so that $H_n([\varepsilon/6\bar{m}_n]^{1/\lambda}) \leq 6p_n \log p_n\ \Delta_n$. With $\bar{s}_n = 4\ \Delta_n^2$

$$(\bar{s}_n^2/n)H_n([\varepsilon/6\bar{m}_n]^{1/\lambda}) \leq 96n^{-1}p_n\ \Delta_n^4 \log p_n\ \Delta_n.$$

Because $q_n\ \Delta_n^4 \log q_n\ \Delta_n = o(n)$, we have $p_n\ \Delta_n^4 \log p_n\ \Delta_n = o(n)$, so that $(\bar{s}_n^2/n)H_n([\varepsilon/6\bar{m}_n]^{1/\lambda}) \to 0$ as required.

(ii) Similarly,

$$(\bar{s}_n/n^{1/2})H_n([\varepsilon/6\bar{m}_n]^{1/\lambda}) \leq 24n^{-1/2}p_n\ \Delta_n^2 \log p_n\ \Delta_n.$$

Because $q_n\ \Delta_n^2 \log q_n\ \Delta_n = o(n^{1/2})$, the result follows. □

Proof of Corollary 4.5. Suppose that for some $\xi > 0$ $P[\ddot{\theta}_n \in \Theta^*(\zeta_o + \xi)] \not\to 1$. Then $P[F] \equiv P[\omega; \ddot{\theta}_n(\omega) \notin \Theta^*(\zeta_o + \xi)\ i.o.] > 0$. Pick $\omega \in F$ and a subsequence $\{n'\}$ of $\{n\}$ such that $\ddot{\theta}_{n'}(\omega) \notin \Theta^*(\zeta_o + \xi)$. Because $\sup_{\theta \in \Theta_n} |Q_n(\cdot, \theta) - \bar{Q}(\theta)| \to 0$ and $\rho(\ddot{\theta}_n, \theta_o) \xrightarrow{P} 0$, we can without loss of generality pick $\omega \in F$ and a further subsequence $\{n''\}$ such that $\sup_{\theta \in \Theta_n} |Q_n(\omega, \theta) - \bar{Q}(\theta)| \to 0$ and $\rho(\ddot{\theta}_n(\omega), \theta_o) \to 0$ (e.g., Theorem 2.4.4 of Lukacs, 1975). By the triangle inequality

$$\begin{aligned} |\bar{Q}(\ddot{\theta}_n(\omega)) - \bar{Q}(\theta_o)| &\leq |\bar{Q}(\ddot{\theta}_n(\omega)) - Q_n(\omega, \ddot{\theta}_n(\omega))| \\ &\quad + |Q_n(\omega, \ddot{\theta}_n(\omega)) - Q_n(\omega, \tilde{\theta}_n(\omega))| \\ &\quad + |Q_n(\omega, \tilde{\theta}_n(\omega)) - \bar{Q}(\tilde{\theta}_n(\omega))| + |\bar{Q}(\tilde{\theta}_n(\omega)) - \bar{Q}(\theta_o)|. \end{aligned}$$

By uniform convergence, the fact that $\rho(\tilde{\theta}_{n''}(\omega), \theta_o) \to 0$, and the continuity of $\bar{Q}$ at θ_o, it follows that given any $\varepsilon > 0$ we have for all n'' sufficiently large that

$$\begin{aligned} &|\bar{Q}(\ddot{\theta}_{n''}(\omega)) - \bar{Q}(\theta_o)| \\ &\qquad \leq |Q_{n''}(\omega, \ddot{\theta}_{n''}(\omega)) - Q_{n''}(\omega, \tilde{\theta}_{n''}(\omega))| + 3\varepsilon \end{aligned}$$

so that for all n'' sufficiently large

$$\begin{aligned} &|Q_{n''}(\omega, \ddot{\theta}_{n''}(\omega)) - Q_{n''}(\omega, \tilde{\theta}_{n''}(\omega))| \\ &\quad \geq |\bar{Q}(\ddot{\theta}_{n''}(\omega)) - \bar{Q}(\theta_o)| - 3\varepsilon \\ &\quad > \zeta_o + \xi - 3\varepsilon, \end{aligned}$$

because $\ddot{\theta}_{n''}(\omega) \notin \Theta^*(\zeta_o + \xi)$. But by construction of $\tilde{\theta}_{n''}(\omega)$, we have $|Q_{n''}(\omega, \ddot{\theta}_{n''}(\omega)) - Q_{n''}(\omega, \tilde{\theta}_{n''}(\omega))| \leq \zeta_{n''} \leq \zeta_o + \xi - 3\varepsilon$ for $\varepsilon > 0$ sufficiently small and all n'' sufficiently large, a contradiction. Hence, for all $\xi > 0$, $P[\ddot{\theta}_n \in \Theta^*(\zeta_o + \xi)] \to 1$.

When $\zeta_o = 0$, we have that $P[\ddot{\theta}_n \in \Theta^*(\xi)] \to 1$ as $n \to \infty$ for all $\xi > 0$. It follows from (4.2) that for all $\varepsilon > 0$ $P[\ddot{\theta}_n \in \eta(\theta_o, \varepsilon)] \to 1$, so that $\rho(\ddot{\theta}_n, \theta_o) \xrightarrow{P} 0$. □

REFERENCES

Andrews, D.W.K. (1988). *Asymptotic normality of series estimators for various nonparametric and semi-parametric estimators* (Cowles Foundation Discussion Paper 874). Yale University.

Baba, N. (1989). A new approach for finding the global minimum of error function of neural networks. *Neural Networks*, **2**, 367–374.

Barron, A. (1989). Statistical properties of artificial neural networks. In *Proceedings of the 28th IEEE Conference on Decision and Control*, Tampa, Florida (pp. 1:280–285). New York: IEEE Press.

Bierens, H. (1987). Kernel estimators of regression functions. In T. Bewley (Ed.), *Advances in econometrics—Fifth World Congress* (Vol. I, pp. 99–144). New York: Cambridge University Press.

Bosq, D. (1975). Inégalité de Bernstein pour les processus stationnaires et mélangeants. Applications. *C.R. Acad. Sc. Paris, Série A*, **281**, 1095–1098.

Carroll, S.M., & Dickinson, B.W. (1989). Construction of neural nets using the radon transform. In *Proceedings of the International Joint Conference on Neural Networks*, Washington, D.C. (pp. 1:607–611). New York: IEEE Press.

Cox, D.D. (1984). Multivariate smoothing splines. *SIAM Journal of Numerical Analysis*, **21**, 789–813.

Cybenko, G. (1989). Approximation by superpositions of a sigmoid function. *Mathematics of Control, Signals and Systems*, **2**, 303–314.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, **2**, 183–192.

Gallant, A.R., Nychka, D.W. (1987). Seminonparametric maximum likelihood estimation. *Econometrica*, **55**, 363–390.

Geman, S., & Hwang, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, **10**, 401–414.

Grenander, U. (1981). *Abstract inference*. New York: Wiley.

Haussler, D. (1989). *Generalizing the PAC model for neural net and other learning applications*, (UC Santa Cruz Computer Research Laboratory Technical Report UCSC-CRL-89-30).

Hecht-Nielsen, R. (1989). Theory of the backpropagation neural network. In *Proceedings of the International Joint Conference on Neural Networks*, Washington, D.C., (pp. I:593–606). New York: IEEE Press.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359–366.

Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, **3**, this issue.

Kolmogorov, A.N., & Fomin, S.V. (1970). *Introductory real analysis*. New York: Dover.

Kolmogorov, A.N., & Tihomirov, V.M. (1961). ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations 2*, **17**, 277–364.

Li, K-C. (1987). Asymptotic optimality for C_p, C_l, cross-validation and generalized cross validation: Discrete index set. *The Annals of Statistics*, **15**, 958–975.

Lorentz, G.G. (1966). *Approximation of functions*. New York: Holt, Rinehart and Winston.

Lukacs, E. (1975). *Stochastic convergence*. New York: Academic Press.

Rudin, W. (1974). *Real and complex analysis*. New York: McGraw-Hill.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Scales, L.E. (1982). *Introduction to non-linear optimization*. New York: Springer-Verlag.

Severini, T.A., & Wong, W.H. (1987). *Convergence rates of maximum likelihood and related estimates in general parameter spaces*. (University of Chicago Department of Statistics Working Paper).

Stinchcombe, M., & White, H. (1989a). Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. In *Proceedings of the International Joint Conference on Neural Networks*, Washington, D.C. (pp. I:613–617). New York: IEEE Press.

Stinchcombe, M., & White, H. (1989b). *Some measurability results for extrema of random functions over random sets*. (UCSD Department of Economics Discussion Paper 89-18).

Stone, M. (1974). Cross-validitory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, **36**, 111–133.

Vapnik, V.N., & Chervonenkis, A. Ya. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, **16**, 264–280.

Wahba, G. (1975). Smoothing noisy data with spline functions. *Numerical Mathematics*, **24**, 383–393.

Wahba, G., & Wold, S. (1975). A completely automatic French curve: Fitting spline functions by cross validation. *Communications in Statistics*, **4**, 1–17.

White, H. (1984). *Asymptotic theory for Econometricians*. Orlando: Academic Press.

White, H. (1989a). Learning in artificial neural networks: A statistical perspective. *Neural Computation* 1, 425–464.

White, H. (1989b). *A practical cross-validation method for multilayer feedforward networks* (Internal HNC Report #89-03).

White, H., & Wooldridge, J. (1990). Some results on sieve estimation with dependent observations. In W. Barnett, J. Powell, and G. Tauchen (Eds.), *Nonparametric and semi-parametric methods in econometrics and statistics*. New York: Cambridge University Press, to appear.