Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators

Article *in* IEEE Transactions on Information Theory · April 1999

DOI: 10.1109/18.749011 · Source: IEEE Xplore

2 authors, including:

Xiaohong Chen
Yale University
95 PUBLICATIONS **4,735** CITATIONS

97-11

UNIVERSITY OF CALIFORNIA, SAN DIEGO

DEPARTMENT OF ECONOMICS

IMPROVED RATES AND ASYMPTOTIC NORMALITY FOR NONPARAMETRIC NEURAL NETWORK ESTIMATORS

BY

XIAOHONG CHEN

AND

HALBERT WHITE

DISCUSSION PAPER 97-11
MAY 1997

Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators[*]

by

Xiaohong Chen
Department of Economics
University of Chicago

and

Halbert White
Department of Economics and
Institute for Neural Computation
University of California, San Diego

April 1997

[*] We thank Andrew Barron and Yuly Makovoz for helpful comments. White's participation was supported by NSF Grant SBR-9511253.

Abstract

Barron (1993) obtained a deterministic approximation rate (in L_2-norm) of $r^{-1/2}$ for a class of single hidden layer feedforward artificial neural networks (ANN) with r hidden units and sigmoid activation functions when the target function satisfies certain smoothness conditions. Hornik, Stinchcombe, White, and Auer (HSWA, 1994) extended Barron's result to a class of ANNs with possibly non-sigmoid activation approximating the target function and its derivatives simultaneously. Recently Makovoz (1996) obtained an improved degree of approximation rate $r^{-(1+1/d)/2}$ for Barron's ANNs with sigmoid activation function where d is the dimension of the domain of the target function.

When applying Barron's ANNs with sigmoid activation functions to nonparametrically estimate a regression function (the target), Barron (1994) obtained a root mean square convergence rate of $O_P([n/\log n]^{-1/4})$ for a minimum complexity regression estimator with i.i.d. observations, where n is the sample size (number of training examples). Unfortunately, this rate is not fast enough to establish root-n asymptotic normality for plug-in estimates of functionals of the regression function, according to a recent result obtained by Chen and Shen (1996).

In this paper, we first obtain an improved approximation rate (in Sobolev norm) of $r^{-1/2-\alpha/(d+1)}$, $0<\alpha\leq 1$, for HSWA's ANNs with possibly non-sigmoid activation functions, where α is related to the choice of activation function. We then obtain a root mean square convergence rate of $O_P([n/\log(n)]^{-(1+2\alpha/(d+1))/[4(1+\alpha/(d+1))]})=o_P(n^{-1/4})$ for general nonparametric ANN sieve extremum estimators, by letting the number of hidden units r_n increase with the sample size n on the order of $(r_n)^{2(1+\alpha/(d+1))}\log(r_n)=O(n)$. Our rates are valid for i.i.d. as well as for uniform mixing and absolutely regular (β–mixing) stationary time series data. Among other things, this rate provides theoretical justification for the popularity of ANN models in fitting multivariate financial data, since many nonlinear financial time series are plausibly modeled as β-mixing processes. In addition, the rate is fast enough to deliver root-n asymptotic normality for plug-in estimates of smooth functionals using general ANN sieve estimators. As interesting applications to nonlinear time series, we establish rates for ANN sieve estimators of three different target functions: a multivariate conditional mean function, a joint density, and a conditional density. We also obtain root-n asymptotic normality results for semiparametric models and average derivative statistics.

1. Introduction

Artificial neural networks (ANNs) are useful classes of flexible approximators that are attracting increasing attention in engineering, as well as in economics and finance, medicine, and other application areas. Critical to both the theoretical understanding of the strengths and weaknesses of these models and to their practical application is the development of techniques that can be used to conduct statistical inference about the phenomenon modeled by the artificial neural network. At present, valid inferential techniques are available for the case in which the phenomenon of interest can be represented exactly by an artificial neural network of finite complexity, or, when the phenomenon of interest cannot be so represented, one conducts inference about the optimal finite complexity network approximation rather than about the phenomenon itself (White, 1989; Stinchcombe and White, 1997). So far unavailable are methods that will permit statistical inference directly about the phenomenon of interest when the neural network does not necessarily provide an exact representation with finite complexity. As this is the generic situation, the absence of such methods represents a significant gap in our understanding and a significant deficiency in the procedures available to those interested in applying neural network models. The purpose of this paper is to provide the missing theory and tools for inference for the generic case.

To provide an arbitrarily accurate approximation to an unknown target function of interest (e.g. a regression function) using a sample of data, it is necessary to fit the network model in such a way that its allowed complexity increases appropriately with the sample size n (White, 1990). The resulting function estimate can then be viewed as a nonparametric sieve estimator (Grenander, 1981) analogous to the more familiar kernel and series estimators of nonparametric statistics. Inference about a phenomenon of interest is often conducted using these standard nonparametric estimators by directing attention to an appropriate functional of interest of the unknown target function (e.g., the derivative averaged over a suitable region) and estimating the functional of interest by replacing the unknown target function by its nonparametric estimate. When the functional is sufficiently smooth, it is well known that such "plug-in" estimators based on kernel or series estimators have an asymptotic normal distribution under the parametric root-n standardization (Goldstein and Messer, 1992), permitting the desired inference to be conducted using standard methods.

Recently, Shen (1997) established that plug-in estimators for smooth functionals using sieve maximum-likelihood estimators with i.i.d. data can also have a root-n asymptotic normal distribution, provided that the sieve maximum-likelihood estimator converges in root mean squared error to the unknown target function at the rate of $o_P(n^{-1/4})$. Chen and Shen (1996) extend this result to plug-in estimators for smooth functionals using sieve extremum estimators (including ANN sieve estimators) with time series dependent data. Unfortunately, the best known convergence rate for ANNs to estimate a regression function is presently $O_P([n/\log n]^{-1/4})$, obtained by Barron (1994) for sigmoid activation functions with i.i.d. data, by Modha and Masry (1996a) for sigmoid activation functions with m-dependent data, and by Chen and Shen (1996) for possibly nonsigmoid activation functions with stationary uniform mixing as well as β-mixing dependent data. To accomplish our present goal, we first sharpen the convergence rate in root mean square error of ANN sieve estimators with possibly nonsigmoid activation functions to the desired rate $o_P(n^{-1/4})$.

We achieve this sharpening by improving on Barron's (1993) path-breaking degree of approximation results for ANNs. Barron established a root mean square approximation rate of $r^{-1/2}$ for a class of single hidden layer feedforward networks with r hidden units having sigmoid activation functions, provided that the target function is suitably smooth, in the sense that its Fourier transform has a bounded first moment of the magnitude distribution. Interestingly, this rate depends neither on the dimension (say d) of the domain of the target function, nor on the number of derivatives (a measure of smoothness) of the target function. Subsequently, Hornik, Stinchcombe, White, and Auer (1994) (HSWA) obtained the same rate $r^{-1/2}$ for a single layer feedforward networks with possibly nonsigmoid activation approximating the target function and its derivatives simultaneously, provided that the target function satisfies certain smoothness conditions related to Barron's. Recently, for Barron's feedforward networks with sigmoid activation functions, Makovoz (1996) obtained an improved root mean square approximation rate $r^{-(1+1/d)/2}$. In this paper, we first extend Makovoz's result to HSWA's class of feedforward networks with possibly nonsigmoid activation approximating the target function and its derivatives simultaneously. In particular, we obtain a new approximation rate (in Sobolev norm) of $r^{-1/2-\alpha/(d+1)}$, $0<\alpha\leq 1$, for HSWA's ANNs with possibly non-sigmoid activation functions, where α is related to the choice of activation

function. We then obtain a root mean square convergence rate of $O_P([n/\log(n)]^{-(1+2\alpha/(d+1))/[4(1+\alpha/(d+1))]}) = o_P(n^{-1/4})$ for general nonparametric ANN sieve extremum estimators, by letting the number of hidden units r_n increase with the sample size n on the order of $(r_n)^{2(1+\alpha/(d+1))}\log(r_n) = O(n)$. This rate is shown to hold for i.i.d. as well as for uniform (ϕ-) mixing and absolutely regular (β-mixing) data generating processes. The results of Chen and Shen (1996) then immediately deliver the desired root-n asymptotic normality for plug-in estimators of smooth functionals using general ANN sieve estimators.

The plan of the paper is as follows: Section 2 gives the new improved approximation rate for ANNs with possibly non-sigmoid activation functions. Section 3 gives the convergence rate for ANN sieve extremum estimates for dependent data. As interesting applications to nonlinear time series, we give new improved convergence rates for ANN sieve estimators of a multivariate conditional mean function, a joint density, and a conditional density. Section 4 gives root-n asymptotic normality for plug-in ANN sieve estimators of smooth functionals. We discuss applications to estimating semiparametric models and average derivative statistics of a regression function. Section 5 concludes. Technical proofs are given in the Mathematical Appendix.

2. Improved Degree of Approximation Results

Barron (1993) and HSWA (1994) establish their approximation rates by applying a fundamental approximation property of convex combinations of families of functions in a Hilbert space, i.e., Maurey's theorem in Pisier (1981), which can be stated loosely as follows: when the target function (say f) belongs to the closure of the convex hull of a symmetric norm-bounded subset (say A) of a Hilbert space, the degree of approximation in the Hilbert norm metric $\inf_{g_r \in A_r} \|f - g_r\|$ will be $const. \times r^{-1/2}$, where A_r consists of all functions of the form $g_r = \sum_{i=1}^{r} b_i \phi_i$, $\phi_i \in A$, $b_i \in R$, $\sum_{i=1}^{r} |b_i| \leq 1$. Recently Makovoz (1996) obtained $\inf_{g_r \in A_r} \|f - g_r\| \leq const. \times \varepsilon_r(A)\, r^{-1/2}$, where $\varepsilon_r(A)$ is the infimum of $\varepsilon > 0$ such that A can be covered by at most r sets of diameter less than or equal to ε. This leads to a refinement of Maurey's theorem when A is a (relatively) compact set since then $\varepsilon_r(A) \to 0$ as $r \to \infty$.

When applying this refinement to Barron's ANNs with sigmoid activation function, Makovoz (1996) obtained an approximation rate (in L_2-norm) of $r^{-1/2-1/(2d)}$ by establishing that $\varepsilon_r(A) = O(r^{-1/(2d)})$ for Barron's ANN with sigmoid activation function ψ (a bounded measurable function on R satisfying $\psi(y) \to 1$ as $y \to \infty$, $\psi(y) \to 0$ as $y \to -\infty$), and

$$A = \{ \psi_{a,\theta} : \psi_{a,\theta}(x) = \psi(a^T x + \theta),\, a \in R^d,\, \sum_{i=1}^{d} |a_i| = c_1,\, \theta \in R,\ |\theta| \le c_2 \}.$$

Here and in what follows the notation ^{T} denotes vector transposition.

For a variety of applications in engineering, robotics, economics, finance, physics, and other fields, it is of interest to identify those ANN classes that can well approximate the target function and its derivatives simultaneously. We achieve this by sharpening HSWA's approximation rate (in a weighted Sobolev norm), using Makovoz's refinement of Maurey's theorem.

Throughout, we focus on the ANN class studied in HSWA. To the extent possible, our notation in this section will coincide with theirs. First, we let the target function $f: R^d \to R$ (e.g. a regression function) have a Fourier representation such that $f \in F_d^{m+1}$, where

$$F_d^{m+1} = \{ f: R^d \to R : f(x) = \int \exp(ia^T x) d\sigma_f(a),\, \|\sigma_f\|_{m+1} \equiv \int l(a)^{m+1}\, d\,|\sigma_f|(a) < \infty \}, \qquad (2.1)$$

where σ_f is a complex measure on R^d, $|\sigma_f|$ denotes the total variation of σ_f, $l(a) \equiv \max[\,|a|, 1\,]$, and $|a| \equiv (a^T a)^{1/2}$. Notice that this is an analog of the smoothness conditions on target functions imposed by Barron (1993, 1994): $f(x) = \int_{R^d} \exp(ia^T x) \tilde{f}(a)\, da$, with $\int_{R^d} |a|\, |\tilde{f}(a)|\, da < \infty$, where $\tilde{f}$ is the Fourier transform of f and $|a| = \sum_{i=1}^{d} |a_i|$.

As HSWA stated, any $f \in F_d^{m+1}$ and its derivatives can be represented as

$$D^\alpha f(x) = \int_{R^d} \int_R a^\alpha D^{|\alpha|} \psi(a^T x + \theta)\, d\nu(a, \theta),$$

for all multi-indices $0 \le |\alpha| \le m$ ($|\alpha| \equiv \sum_{i=1}^{d} \alpha_i$), for some $\psi \in B_1^m$ and some $\nu \in M_d^m$, where M_d^m is the space of all signed measures ν on $R^d \times R$ for which

$$\|\nu\|_{M_d^m} \equiv \int_{a \in R^d} \int_{\theta \in R} l(a)^m\, d\,|\nu|(a, \theta) < \infty,$$

and B_d^m is a weighted Sobolev space of all functions on R^d that have continuous and uniformly bounded

(partial) derivatives up through order m. For all $h \in B_d^m$, define the norm

$$\|h\|_{B_d^m} \equiv \max_{0 \le |\alpha| \le m} \sup_{x \in R^d} |D^\alpha h(x)| < \infty .$$

Suppose that the network input at time t, X_t, has distribution μ (the same for all t) with compact support S in R^d, $d \in \mathbb{N}$, having non-empty interior. Without confusion, we also denote $B_d^m = (B_d^m, \|\cdot\|_{m,\mu})$, the Hilbert space completion under the inner product induced norm (with $\rho_{m,\mu}$ the associated metric) :

$$\|h\|_{m,\mu} \equiv [\sum_{|\alpha|=0}^{m} (\|D^\alpha h\|_{L_2(\mu)})^2]^{1/2}, \quad \|D^\alpha h\|_{L_2(\mu)}^2 \equiv \int_{R^d} |D^\alpha h(x)|^2 d\mu(x) .$$

We approximate any target function $f \in F_d^m$ using the ANN class

$$G_d^m(\psi, B, r) \equiv \{ g : g(x) = \sum_{j=1}^{r} \beta_j \, l(a_j)^{-m} \psi(a_j^T x + \theta_j), \; a_j \in R^d, \; \theta_j, \beta_j \in R, \; \sum_{j=1}^{r} |\beta_j| \le B \}, \quad (2.2)$$

where $\psi \in B_1^m$.

For $\psi \in B_1^m$ and $\nu \in M_d^m$, define the function $T_\psi \nu(\cdot)$ on R^d as

$$T_\psi \nu(x) \equiv \int_{a \in R^d} \int_{\theta \in R} \psi(a^T x + \theta) \, d\nu(a, \theta),$$

and define

$$A_\psi \equiv \{ \psi_{a,\theta} : \psi_{a,\theta}(x) = l(a)^{-m} \psi(a^T x + \theta), \; a \in R^d, \theta \in R \} .$$

The following result is a refinement of HSWA's theorem 2.1, and is a simple consequence of Makovoz's theorem 1 and HSWA's lemma A.1.

Lemma 2.1: *For $\psi \in B_1^m$ and $\nu \in M_d^m$,* let $B \ge \|\nu\|_{M_d^m}$. *Then* :

$$\rho_{m,\mu}[T_\psi \nu, G_d^m(\psi, B, r)] \le const. \times B \, r^{-1/2} \varepsilon_r(A_\psi),$$

where $\varepsilon_r(A_\psi)$ is the infimum of the $\varepsilon > 0$ such that A_ψ can be covered by at most r sets of diameter less than or equal to ε.

Lemma 2.1 will deliver an improved approximation rate if one can choose ψ such that A_ψ is a relatively compact set, as then $\varepsilon_r(A_\psi) \to 0$ as $r \to \infty$. For this we impose a Hölder condition.

Assumption H: *There exists an $\alpha \in (0,1]$ associated with $\psi \in B_1^m$ such that for all x in the compact*

support S,

$$\| \psi_{a,\theta} - \psi_{a_1,\theta_1} \|_{B_1^m} \leq const. \times [\| a - a_1 \| + \| \theta - \theta_1 \|]^{\alpha}.$$

The following result is a refinement of HSWA's corollary 2.4.

Theorem 2.2: *Let* $f \in F_d^{m+1}$. *Suppose that* $\psi \in B_1^m$ *has integrable derivatives up to order m and satisfies Assumption H. Suppose that* ψ *and* μ *are compactly supported. Let* $B \geq const. \times \| \sigma_f \|_{m+1}$. *Then*: $\rho_{m,\mu}[f, G_d^m(\psi, B, r)] \leq B\, r^{-1/2-\alpha/d^*}$, *where* $d^* = d$ *if* ψ *is homogeneous*; $d^* = d+1$ *otherwise*.

In the following, we relax the condition that ψ has absolutely integrable derivatives up to order *m*. To facilitate our subsequent statistical applications, we also allow *B* and r to depend on sample size *n*, and we denote the resulting ANN sieve as

$$\begin{aligned} G_n &\equiv G_d^m(\psi, B_n, 2^k r_n) \\ &\equiv \{ g : g(x) = \sum_{j=1}^{2^k r_n} \beta_j\, l(a_j)^{-m}\, \psi(a_j^T x + \theta_j),\ a_j \in R^d,\ \theta_j, \beta_j \in R,\ \sum_{j=1}^{2^k r_n} |\beta_j| \leq B_n \}, \end{aligned} \tag{2.3}$$

where $\psi \in B_1^m$ is *k-finite* for some $k \geq m$, i.e., ψ is a nonzero element in B_1^m and $0 < \int_R |D^k \psi(z)|\, dz < \infty$ for some $k \geq m$. HSWA give many examples of *k-finite* activation functions ψ, e.g., the Heaviside, logistic, tanh, cosine squasher, and other sigmoid functions for $k = 1$.

Corollary 2.3: *Suppose that* μ *has compact support and that* $\psi \in B_1^m$ *and is k-finite for some* $k \geq m$ *and satisfies Assumption H. For any* $f \in F_d^{m+1}$, *there exists* $\pi_n f \equiv g_n \in G_n$ *such that the following hold*:

(a) *If* $D^k \psi$ *has compact support, let* $B_n \geq C_1 \| \sigma_f \|_{m+1}$. *Then:* $\| \pi_n f - f \|_{m,\mu} \leq c \| \sigma_f \|_{m+1} (r_n)^{-1/2-\alpha/d^*}$;

(b) *If* $\int_R |\exp(\lambda |t|) D^j \psi(t)|\, dt < \infty$ *for some* $\lambda > 0$ *and all* $0 \leq j \leq m$, *let* $B_n \geq C_2 \| \sigma_f \|_{m+1} \log(r_n)$. *Then*: $\| \pi_n f - f \|_{m,\mu} \leq c \| \sigma_f \|_{m+1} (r_n)^{-1/2-\alpha/d^*} \log(r_n)$ *for some finite* $c > 0$.

3. Improved Convergence Rates for ANN Sieve Extremum Estimators

We suppose that the data in our sample are generated by a stationary stochastic process.

Assumption 3.1: (a) $\{ Z_t \}$ *is a stationary sequence of random* $p \times 1$ *vectors, each having joint distribu-*

tion P_o.

We are interested in an aspect of the data generating process P_o indexed by a value $\theta_o = T(P_o)$ belonging to a parameter space Θ, infinite dimensional in general. Θ corresponds to F_d^{m+1} and θ_o corresponds to f in the notation of the previous section (The notation of this section corresponds to that of the sieve estimation literature.). For example, with $Z_t = (Y_t, X_t^T)^T$ we may be interested in $\theta_o(X_t) = E(Y_t \mid X_t)$, the conditional expectation or "regression" of Y_t given X_t. We suppose in particular that θ_o is defined as the solution of an extremal problem, as follows.

Assumption 3.2: *$\theta_o \in \Theta \subseteq F_d^{m+1}$ is such that*

$$E[L_n(\theta_o)] \geq E[L_n(\theta)]$$

for all $\theta \in \Theta$, where E denotes expectation under P_o and

$$L_n(\theta) \equiv n^{-1} \sum_{t=1}^{n} l(Z_t, \theta),$$

where $l : R^p \times \Theta \to \bar{R}$ is measurable - $B(R^p \times \Theta)$ with $E(l(Z_t, \theta)) < \infty$ for all $\theta \in \Theta$.

We interpret θ_o as maximizing the expected average quasi-log-likelihood L_n. For regression, set $l(Z_t, \theta) = -[Y_t - \theta(X_t)]^2/2$. Note that the assumed stationarity of Z_t ensures that the solution to the extremal problem does not depend on n.

Our goal is to conduct inference about a smooth functional of θ_o, say $\gamma(\theta_o)$. For convenience, we take $\gamma(\theta_o)$ to be real-valued. There is no further difficulty in treating the case in which $\gamma(\theta_o)$ is a finite dimensional real vector. For example, if $\theta_o(X_t) = E(Y_t \mid X_t)$, we may be interested in conducting inference about the expected derivative of θ_o, $\gamma(\theta_o) = E(D\theta_o(X_t))$.

To estimate θ_o, we consider approximate sieve extremum estimators, which approximately maximize L_n over a subset Θ_n of Θ:

Assumption 3.3: *There exists a measurable function $\hat{\theta}_n$ such that*

$$L_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta_n} L_n(\theta) - O(\varepsilon_n^2) \text{ with } \varepsilon_n \to 0 \text{ as } n \to \infty$$

where $\Theta_n \equiv G_n$, with

$$G_n \equiv \{ g : g(x) = \sum_{j=1}^{2^k r_n} \beta_j \, l(a_j)^{-m} \psi(a_j^T x + a_{0,j}) \}, \tag{2.3'}$$

where

$$\max_{1 \le j \le 2^k r_n} \sum_{i=0}^{d} |a_{i,j}| \le c_n, \quad \sum_{j=1}^{2^k r_n} |\beta_j| \le B_n .$$

The sequence $\{ \Theta_n \}$ is called a sieve (Grenander, 1981). By choosing $\Theta_n \equiv G_n$ as defined in (2.3') we ensure that $\{ \Theta_n \}$ is a sequence of nondecreasing sets whose union is dense in Θ (see White, 1990). Because our sieve is constructed from ANN output functions, we call $\hat{\theta}_n$ an ANN sieve extremum estimator. Thus, the network is "trained" (estimated) by maximizing the "training objective function" L_n over the weights allowed by G_n. The "plug-in" ANN sieve estimator of $\gamma(\theta_o)$ is simply $\gamma(\hat{\theta}_n)$.

We obtain our desired convergence rates by applying theorem 1 of Chen and Shen (1996). For this it suffices to extend Assumption 3.1 and to impose some further technical requirements.

Assumption 3.1:(b) $\{ Z_t \}$ *has compactly supported elements, and is either:*

(i) i.i.d. or m-dependent; or

(ii) uniform mixing with $\phi(j) \le \phi_0 \, j^{-\xi}$ *for some* $\phi_0 > 0$, $\xi > 2$*; or*

(iii) β*-mixing with* $\beta(j) \le \beta_0 \, j^{-\xi}$ *for some* $\beta_0 > 0$, $\xi > 2$.

These conditions restrict the the memory of the $\{ Z_t \}$ process in a mild way; White (1990) imposes similar conditions. Additional regularity conditions imposed on the quasi-log-likelihood l are as follows:

Assumption 3.4:

(a) For all small $\varepsilon > 0$

$$\sup_{\{ \theta \in \Theta_n : \| \theta_o - \theta \| \le \varepsilon \}} var \, (l(Z_t, \theta) - l(Z_t, \theta_o)) \le C_1 \varepsilon^2,$$

where $\| \cdot \|$ *is a semi-norm associated with any metric on* Θ *equivalent to* $[E(L_n(\theta_o)) - E(L_n(\theta))]^{1/2}$ *in the neighborhood of* θ_o.

(b) *For any* $\delta > 0$ *there exists a constant* $s \in (0, 2)$ *such that*

$$\sup_{\{\theta \in \Theta_n : \|\theta_o - \theta\| \le \varepsilon\}} | l(Z_t, \theta) - l(Z_t, \theta_o) | \le \delta^s U(Z_t),$$

with $E | U(Z_t) |^{\kappa} \le C_3$ *for some* $\kappa \ge 2$ *and* $s\kappa \ge 2$.

We can now state a sharpened rate result for the convergence of the NN sieve extremum estimator to the function of interest, θ_o.

Theorem 3.1: *Suppose Assumptions H and 3.1-3.4 hold with (i)* $\kappa \ge 2, s\kappa \ge 2$*; (ii)* $\kappa > 2$, $s\kappa \ge 2$*; or (iii)* $\kappa > 2$, $s\kappa > 2$ *holding, depending on which of Assumptions 3.1(b) (i), (ii) or (iii) hold. We also choose* c_n *= const. in (2.3').*

(a) *If* $D^k \psi$ *has compact support,* $B_n \ge C_1 \|\sigma_{\theta_o}\|_{m+1}$ *, and* r_n *is such that*

$(r_n)^{2(1+\alpha/d^*)} \log r_n = O(n)$ *, then* $\|\hat{\theta}_n - \theta_o\| = O_P([n/\log n]^{-(1+(2\alpha/d^*))/[4(1+(\alpha/d^*))]})$.

(b) *If* $\int_R | \exp(\lambda | t |) D^j \psi(t) | dt < \infty$ *for some* $\lambda > 0$ *and all* $0 \le j \le m$,

$B_n \ge C_2 \|\sigma_{\theta_o}\|_{m+1} \log(r_n)$ *, and* r_n *is such that* $(r_n)^{2(1+\alpha/d^*)} / \log r_n = O(n)$, *then*

$\|\hat{\theta}_n - \theta_o\| = O_P([n \log n]^{-(1+(2\alpha/d^*))/[4(1+(\alpha/d^*))]} \log n)$.

Previously Barron (1994) and Modha and Masry (1996a) have applied the minimum complexity method to estimate a regression mean target function via ANNs with sigmoid activation functions. They established a root mean square convergence rate of $O_p([n/\log n]^{-1/4})$ for i.i.d. data and *m*-dependent data respectively. Modha and Masry (1996b) also apply the minimum complexity method to estimate a multivariate density via ANNs with sigmoid activation function, and they obtained the same convergence rate (in expected Hellinger distance) for i.i.d. data. In the following two examples, by applying our Theorem 3.1, we obtain faster convergence rates for i.i.d. data as well as for stationary β-mixing data.

Example 3.2: *Nonparametric multivariate regression via sigmoid activation functions.* Suppose $\{Y_t, X_t\}$ is generated according to

$$Y_t = \theta_o(X_t) + e_t, \quad E[e_t | X_t] = 0, \quad Var[e_t | X_t] = \sigma^2(X_t) < \infty,$$

where $\{Y_t, X_t\}$ is a stationary sequence which is either *uniform mixing* satisfying Assumption 3.1(b)(ii)

or β*–mixing* satisfying Assumption 3.1(b)(iii). Let X_t have the (unknown) distribution μ with compact support S in R^d, and suppose that $\theta_o \in F_d^{m+1}$. Suppose $E[\sigma^4(X_t)]<\infty$ and $E[|e_t|^{2+\zeta}]<\infty$ for some $\zeta > m^{-1}$, with $0<m<\infty$.

Let $l(\theta, Y_t, X_t) = -(Y_t - \theta(X_t))^2/2$. Let $\|\theta - \theta_o\|^2 = E[\theta(X_t) - \theta_o(X_t)]^2$. We estimate θ_o using the ANN sieve G_n with sigmoid activation function such as the Heavyside or logistic.

Case 1: Heavyside (Unit step) activation function: By Makovoz(1996), Assumption H is satisfied with $B_1^m = L_2, \alpha = 1/2,$ and $d^* = d$. It is easy to show that all the assumptions for Theorem 3.1 (a) are satisfied, and we obtain the convergence rate $O_p([n/\log n]^{-(1+(1/d))/[4(1+(1/2d))]})$.

Case 2: Logistic activation function: We can approximate any sigmoid activation function, e.g., the logistic, via the unit step function and obtain the same rate as that for Case 1. Alternatively, as we can verify directly, Assumption H is satisfied for the logistic with $B_1^m = L_2, \alpha = 1$, and $d^* = d+1$. We then get the rate $O_p([n/\log n]^{-(1+2/(d+1))/[4(1+1/(d+1))]})$. Note that these two methods will give the same rate when $d = 1$, but the latter affords an improvement when $d > 1$.

Example 3.3: *Joint density, Conditional density.* Suppose that $\{Y_t, X_t\}$ is a stationary process which is either *uniform mixing* or β*-mixing*. Suppose $Y_t \in R^1$, and $X_t \in R^{d-1}$, $d \geq 2$. We allow X_t to include past Y_t's and other stationary random variables W_t. Let $Z_t \equiv (Y_t, X_t^T)^T$ have (joint) density $f^*: R^d \to R^1$. We are interested in estimating the joint density f^* as well as the conditional density $f^*_{Y|X}$. Suppose for simplicity that the support of Y_t is [0,1], and the support of X_t is $[0,1]^{d-1}$.

(i) Joint density: Since one can always express the true joint density $f^*(\cdot) \equiv \exp(\theta_o(\cdot))/\int_{[0,1]^d} \exp(\theta_o(z))\,dz$ for some measurable function $\theta_o: R^d \to R^1$, estimating f^* is equivalent to estimating θ_o. Suppose that $\theta_o \in F_d^{m+1}$; then we can choose

$$l(\theta, Z_t) = \log(f(Z_t)) = \theta(Z_t) - \log \int_{[0,1]^d} \exp(\theta(z))\,dz .$$

We will use the expected Hellinger distance between two densities f and f^* as defining a norm $\|\theta - \theta_o\| \equiv \rho(\theta, \theta_o)$, where

$$\rho(\theta, \theta_o)^2 = E \int_{[0,1]^d} [(f(z))^{1/2} - (f^*(z))^{1/2}]^2\,dz$$

It is easy to see that the metric ρ is equivalent to the Kullback-Leibler distance $E[\log(f^*(Z_t)) - \log(f(Z_t))] = E[l(\theta_o, Z_t) - l(\theta, Z_t)]$.

(ii) Conditional density: Again we write the true conditional density $f^*_{Y|X}(x,y) \equiv \exp(\theta_o(x,y)) / \int_{[0,1]} \exp(\theta_o(x,y))\,dy$ for some measurable function $\theta_o : R^d \to R^1$. Suppose that $\theta_o \in F_d{}^{m+1}$; then we can choose

$$l(\theta, Y_t, X_t) = \log(f_{Y|X}(X_t, Y_t)) = \theta(X_t, Y_t) - \log \int_{[0,1]} \exp(\theta(X_t, y))\,dy .$$

Again we use the expected Hellinger distance between two conditional densities $f_{Y|X}$ and $f^*_{Y|X}$ to define $\|\theta - \theta_o\|$. Again it is easy to see that this metric is equivalent to the Kullback-Leibler distance $E[l(\theta_o, Y_t, X_t) - l(\theta, Y_t, X_t)]$.

We estimate θ_o using the ANN sieve G_n with either logistic activation function, or any *k*-finite activation function where $D^k\psi$ has compact support for some $k \geq m$, and satisfying Assumption H with $\alpha = 1$ and $d^* = d + 1$. Using Theorem 3.1 (a), we obtain the convergence rate $O_p([n/\log n]^{-(1+2/(d+1))/[4(1+1/(d+1))]})$.

Notice that one can use other parameterization of the density too. For example, we can assume that $\theta_o = (f^*)^{1/2}$, and assume that $\theta_o \in F_d^{m+1}$. For simplicity, we assume that $f^* \geq const. > 0$ over the support. We use $l(\theta, Z_t) = \log(\theta^2(Z_t))$, and the distance $\|\theta - \theta_o\|^2 = \int_{[0,1]^d} [\theta(z) - \theta_o(z)]^2\,dz$. We can again estimate θ_o using the ANN sieve G_n with additional constraint: $\theta \geq 0$, $\int_{[0,1]^d} \theta^2(z)\,dz = 1$. Then we can obtain the same convergence rate.

4. Asymptotic Normality for Plug-in Estimators via ANN Sieves

Let $\gamma : \Theta \to R$ be some known functional, and let $\gamma_o = \gamma(\theta_o)$ be the object of interest. Let $\hat{\theta}_n$ be the ANN sieve estimator; then $\gamma(\hat{\theta}_n)$ is the ''plug-in'' ANN sieve estimator of γ_o.

Suppose that for all $\theta \in \Theta$ and all z, $l(\theta, z) - l(\theta_o, z)$ can be approximated using the linear functional $l'_{\theta_o}[\theta - \theta_o, z]$, which is defined as $\lim_{\tau \to 0} [l(\xi(\theta_o, \tau), z) - l(\theta_o, z)]/\tau$, where $\xi(\theta_o, \tau) \in \Theta$ is a path in τ connecting θ_o and θ such that $\xi(\theta_o, 0) = \theta_o$ and $\xi(\theta_o, 1) = \theta$. Define

$$R[\theta-\theta_o, z] \equiv l(\theta, z) - l(\theta_o, z) - l'_{\theta_o}[\theta-\theta_o, z].$$

Assumption 4.1: *For any* $\theta \in \Theta$,

$$|\gamma(\theta)-\gamma(\theta_o)-\gamma'_{\theta_o}[\theta-\theta_o]| \leq O(\|\theta-\theta_o\|^2), \quad \text{as} \quad \|\theta-\theta_o\| \to 0, \tag{4.1}$$

where $\gamma'_{\theta_o}[\theta-\theta_o]$ *is linear in* $\theta-\theta_o$, *and* $\|\gamma'_{\theta_o}\| < \infty$, *with*

$$\|\gamma'_{\theta_o}\| \equiv \sup_{\theta \in \Theta: \|\theta-\theta_o\|>0} (|\gamma'_{\theta_o}[\theta-\theta_o]| / \|\theta-\theta_o\|).$$

By the Riesz representation theorem, there exists $v^* \in \bar{V}$ such that for any $\theta \in \Theta$, $\gamma'_{\theta_o}[\theta-\theta_o] = <\theta-\theta_o, v^*>$, where $\bar{V}$ is the completion of $\Theta - \{\theta_o\}$ under the norm $\|\gamma'_{\theta_o}\|$ and $<,>$ is the associated inner product. Let $u^* = v^*$ or $-v^*$, and define $\xi^*(\theta, \varepsilon_n) = (1-\varepsilon_n)\theta + \varepsilon_n(u^* + \theta_o)$ with $\varepsilon_n = o(n^{-1/2})$. Define

$$g(\theta, Z_t) \equiv R[\theta-\theta_o, Z_t] - R[\pi_n \xi^*(\theta, \varepsilon_n) - \theta_o, Z_t]. \tag{4.2}$$

Assumption 4.2: $g(\theta, Z_t)$ *satisfies Assumption 3.4 for all* $\theta \in \{\theta \in \Theta_n : \|\theta-\theta_o\| \leq \delta_n\}$ *where* $\|\hat{\theta}_n - \theta_o\| = o_p(\delta_n)$.

Assumption 4.3: $\sigma_*^2 \equiv \lim_{n\to\infty} n^{-1} Var(\sum_{t=1}^{n} l'_{\theta_o}[v^*, Z_t])$ *is positive and finite.*

The following result is a simple consequence of our Theorem 3.1 and CS's theorem 2.

Theorem 4.1: *Suppose that Assumptions H and 3.1, 4.1 - 4.3 hold. Let* $\hat{\theta}_n$ *be the ANN sieve estimator from Theorem 3.1 (a). Then*: $n^{1/2}(\gamma(\hat{\theta}_n) - \gamma(\theta_o)) => N(0, \sigma_*^2)$.

Example 4.2: *Semi-parametric model.* Suppose time-series data $\{Y_t\}$ are generated according to

$$Y_t = X_{1,t}^T b_o + \eta_o(X_{2,t}) + e_t, \quad E[e_t | X_t] = 0, \quad E[e_t^2 | X_t] = \sigma^2(X_t),$$

where $X_t = (X_{1,t}, X_{2,t})$ and $X_{i,t} = (Y_{t-1}, ..., Y_{t-p_i}, U_t, ..., U_{t-q_i+1})^T$ for $i = 1, 2$. U_t does not contain current and past Y_t's and is stationary β-mixing in R^d with compact support. Let $d_i \equiv p_i + d\, q_i$. The parameters of interest are $b_o \in R^{d_1}$ and $\eta_o \in B_{d_2}^m$ for $m \geq 1$. Under mild conditons, we have that $\{Y_t\}$ is a stationary β–mixing process satisfying Assumption 3.1(b)(iii). Suppose $E[\sigma^4(X_t)] < \infty$ and

$E[|e_t|^{2+\zeta}]<\infty$ for some $\zeta > m^{-1}$.

Let $\theta_o = (b_o, \eta_o)$, and for convenience write $\theta(X_t) = X_{1,t}^T b + \eta(X_{2,t})$ and

$$\|\theta - \theta_o\|^2 = E[X_{1,t}^T (b - b_o) + (\eta(X_{2,t}) - \eta_o(X_{2,t}))]^2 .$$

Let $l(\theta, Y_t, X_t) = -0.5(Y_t - (X_{1,t}^T b + \eta(X_{2,t})))^2$, let $\Theta = A \times D$, where A is the closure of an open bounded set in R^{d_1}, and put $D = F_{d_2}^{m+1}$. Let $\Theta_n = A \times G_n$, and suppose that the activation function ψ is k–finite for $k \geq m$, and $D^k \psi$ has compact support. By Theorem 3.1 (a), we have $\|\hat{\eta}_n - \eta_o\| = o_p(n^{-1/4})$.

We would like to obtain root-*n* normality for $\hat{b}_n - b_o$. Let $W_t \equiv X_{1,t} - E(X_{1,t} \mid X_{2,t})$. Suppose that $\Sigma \equiv E[W_t W_t^T]$ is a positive definite $d_1 \times d_1$ matrix. Let $\lambda = (\lambda_1, ..., \lambda_{d_1})^T$ be an arbitrary fixed unit vector in R^{d_1}. Consider $\gamma(\theta_o) = \lambda^T b_o$. It is easy to see that $\gamma'_{\theta_o}[\theta - \theta_o] = \lambda^T (b - b_o)$, and CS's $\omega = \infty$. After some calculation, we have $\|v^*\|^2 = \lambda^T \Sigma^{-1} \lambda$, and $v^* = (\Sigma^{-1}\lambda, -(\Sigma^{-1}\lambda)^T E[X_{1,t} \mid X_{2,t}])$. Since

$$l'_{\theta_o}[\theta - \theta_o] = [X_{1,t}^T (b - b_o) + (\eta - \eta_o)] e_t$$

we have

$$l'_{\theta_o}[v^*] = [(\Sigma^{-1}\lambda)^T X_{1,t} - (\Sigma^{-1}\lambda)^T E[X_{1,t} \mid X_{2,t}]] e_t = (\Sigma^{-1}\lambda)^T W_t e_t$$

By Theorem 4.1, we get

$$n^{1/2} \lambda^T (\hat{b}_n - b_o) \Rightarrow N(0, \sigma_*^2), \quad \sigma_*^2 \equiv \lim_{n\to\infty} n^{-1} E[\sum_{t=1}^{n} (\Sigma^{-1}\lambda)^T W_t e_t]^2 .$$

Hence we get $n^{1/2}(\hat{b}_n - b_o) \Rightarrow N(0, \Omega)$ with

$$\Omega = \Sigma^{-1} (E[e_1^2 W_1 W_1^T] + \sum_{j=2}^{\infty} E[e_1 e_j W_1 W_j^T] + \sum_{j=2}^{\infty} E[e_j e_1 W_j W_1^T]) \Sigma^{-1} .$$

In addition, if $E[e_t^2 \mid X_t] = \sigma^2$, independent of X_t, then $\Omega = \sigma^2 \Sigma^{-1}$, which reaches the semiparametric asymptotic efficiency bound.

Example 4.3: *Functionals of a regression function.* In Example 3.2, suppose we want to estimate a certain smooth functional of the conditional mean function θ_o, for example, $\gamma(\theta_o) = \int_{R^d} [(D_i \theta_o(x))^2] d\mu(x)$, where $D_i \theta$ is the partial derivative of θ with respect to the *i*-th component of *x*, $i = 1, ..., d$. If μ is

known, then the plug-in estimator is simply $\gamma(\hat{\theta}_n) = \int_{R^d} [(D_i \hat{\theta}_n(x))^2] \, d\mu(x)$. If μ is unknown, let μ_n denote the empirical cdf of X_t; then a plug-in estimator is

$$\gamma(\hat{\theta}_n, \mu_n) = n^{-1} \sum_{t=1}^{n} (D_i \hat{\theta}_n(X_t))^2 \equiv \int_{R^d} [(D_i \hat{\theta}_n(x))^2] \, d\mu_n(x).$$

After simple calculations, and by Theorem 4.1, we have:

$n^{1/2} (\int (D_i \hat{\theta}_n)^2 \, d\mu - \int (D_i \theta_o)^2 \, d\mu) \Rightarrow N(0, \sigma^2_{*,1})$ where

$$\sigma^2_{*,1} = 4E[D_i \theta_o(X_1) e_1] + 8 \sum_{j=2}^{\infty} Cov(D_i \theta_o(X_1) e_1, D_i \theta_o(X_j) e_j).$$

By the delta method, we have

$$n^{1/2} (\int (D_i \hat{\theta}_n)^2 \, d\mu_n - \int (D_i \theta_o)^2 \, d\mu) =$$
$$n^{1/2} (\int (D_i \hat{\theta}_n)^2 \, d\mu - \int (D_i \theta_o)^2 \, d\mu) + \int (D_i \theta_o)^2 \, n^{1/2} d[\mu_n - \mu] + o_p(1).$$

By a Hilbert-valued central limit theorem for the empirical process $\mu_n - \mu$ with stationary uniform mixing or $\beta-$ mixing, see e.g., Dehling (1983), and by the continuous mapping theorem, we obtain:

$\int (D_i \theta_o)^2 \, n^{1/2} d[\mu_n - \mu] \Rightarrow N(0, \sigma^2_{*,2})$, where

$$\sigma^2_{*,2} = Var([D_i \theta_o]^2(X_1)) + 2 \sum_{j=2}^{\infty} Cov([D_i \theta_o]^2(X_1), [D_i \theta_o]^2(X_j)).$$

Hence, $n^{1/2} (\int (D_i \hat{\theta}_n)^2 \, d\mu_n - \int (D_i \theta_o)^2 \, d\mu) \Rightarrow N(0, \sigma^2_*)$ with $\sigma^2_* = \sigma^2_{*,1} + \sigma^2_{*,2}$.

5. Conclusion

We have used recent results of Chen and Shen (1996) and sharpened degree of approximation results for ANNs to obtain improved rates for the root mean square convergence of ANN sieve extremum estimators. Our rates are sufficiently fast as to ensure the root-*n* asymptotic normality for ANN-based plug-in estimates of smooth functionals of interest. This provides for the first time theory and tools relevant for conducting statistical inference about phenomena modeled using ANNs in the generic case in which an ANN of finite complexity cannot provide an exact representation of the phenomenon of interest.

6. Appendix

Proof of Lemma 2.1: By Lemma A.1 of HSWA, $T_\psi \mu$ is in the $\rho_{m,\mu}$ closure of the convex hull of $G_d^m(B)$. The Assertion now follows from theorem 1 in Makovoz (1996). □.

Lemma A: *Suppose that* $\psi \in B_1^m$ *satisfies Assumption H, and that both* ψ *and* μ *are compactly supported. Then*: $\varepsilon_r(A_\psi) = O(r^{-\alpha/d^*})$.

Proof: Given the conditions on ψ and μ, we can choose

$$A_\psi = \{\psi_{a,\theta} : \sum_{i=1}^{d} |a_i| \le c_1, |\theta| \le c_2\} \quad \text{if } \psi \text{ is not homogeneous ;}$$

and

$$A_\psi = \{\psi_{a,\theta} : \sum_{i=1}^{d} |a_i| = c_1, |\theta| \le c_2\} \quad \text{if } \psi \text{ is homogeneous .}$$

By Assumption H, we obtain an $O(\varepsilon^\alpha)$-net for A_ψ in $\rho_{m,\mu}$-norm if we can find an ε-net for the set:

$$P_{nh} \equiv \{(a,\theta) \in R^{d+1} : |a| \le c_1, |\theta| \le c_2\} \quad \text{if } \psi \text{ is not homogeneous ;}$$

or

$$P_h \equiv \{(a,\theta) \in R^{d+1} : |a| = c_1, |\theta| \le c_2\} \quad \text{if } \psi \text{ is homogeneous .}$$

Clearly one needs $O((1/\varepsilon)^{d+1})$ elements to build an ε-net for P_{nh} and $O((1/\varepsilon)^d)$ elements for P_h. Thus an ε-net for A_ψ has cardinality $O(\varepsilon^{-d^*/\alpha})$. This gives $\varepsilon_r(A_\psi) = O(r^{-\alpha/d^*})$. □

Proof of Theorem 2.2: The proof is the same as that for HSWA's corollary 2.4 except using our Lemma 2.1 and Lemma A instead of HSWA's theorem 2.1 □.

Proof of Corollary 2.3: The proof is the same as that for HSWA's corollaries 2.4, 2.5 and 3.1, except using our Lemma 2.1 and Lemma A instead of HSWA's theorem 2.1 □.

Proof of Theorem 3.1: We prove this by applying Chen and Shen's (1996) (CS) theorem 1, and we use some of CS's notation in the following: Assumption 3.1 implies CS's Condition A.1; Assumption 3.2 implies CS's Conditions A.2 and A.4. After some calculation, we have the following bound for the L_2-

metric entropy with bracketing: $H(w, G_n) \leq 2^k r_n B_n (d+1) \log(2^k r_n B_n (d+1)/w)$, which allows us to obtain δ_n from CS's Condition A.3:

$$\delta^{-2} \int_{\delta^2}^{\delta} [H(w, G_n)]^{1/2} dw \leq const. \times n^{1/2}.$$

Now CS's theorem 1 gives the convergence rate $\|\hat{\theta}_n - \theta_o\| = O_p(\max[\delta_n, \|\pi_n \theta_o - \theta_o\|])$.

For Case (b), we have $B_n = const.$, and

$$\|\pi_n \theta_o - \theta_o\| \leq \|\pi_n \theta_o - \theta_o\|_{m,\mu} \leq const. \times r_n^{-1/2-\alpha/d^*}.$$

From CS's Condition A.3, $\delta_n = const. \times [r_n \log(r_n)]^{1/2} n^{-1/2}$. The final convergence rate is obtained by setting $\delta_n = \|\pi_n \theta_o - \theta_o\|$, which yields $r_n^{2(1+\alpha/d^*)} \log(r_n) = O(n)$, and $\|\hat{\theta}_n - \theta_o\| = O_p([n/\log n]^{-(1+(2\alpha/d^*))/[4(1+(\alpha/d^*))]})$.

For Case (a), we have $B_n = const. \times \log(r_n)$, and

$$\|\pi_n \theta_o - \theta_o\| \leq \|\pi_n \theta_o - \theta_o\|_{m,\mu} \leq const. \times r_n^{-1/2-\alpha/d^*} \log(r_n).$$

From CS's Condition A.3, we get $\delta_n = [r_n \log(r_n)]^{1/2} \log(B_n) n^{-1/2}$. Because

$$\max([r_n \log(r_n)]^{1/2} \log\log(r_n) n^{-1/2}, r_n^{-1/2-\alpha/d^*} \log(r_n)) = o(1),$$

we can restrict our attention to the set $S = \{\theta \in \Theta_n : B_n \leq const.\}$. Clearly, $\hat{\theta}_n \in S$ by CS's theorem 1.

Now we can repeat the argument in (a) using the local metric entropy $H(w,S) \leq const. \times r_n \log(const. \times r_n / w)$ in CS's Condition A.3, and we obtain a smaller $\delta_n = const. \times [r_n \log(r_n)]^{1/2} n^{-1/2}$, as in (a). Now we get the desired rate by setting $r_n^{2(1+\alpha/d^*)} / \log r_n = const. \times n$, and $\|\hat{\theta}_n - \theta_o\| = O_p([n \log n]^{-(1+(2\alpha/d^*))/[4(1+(\alpha/d^*))]} \log n)$. □.

Proof of Theorem 4.1: We prove this by applying CS's theorem 2. Assumptions 3.1 - 3.4 and 4.2 imply that CS's Assumptions B.1, B.3 and B.4 are satisfied. CS's Assumption B.2 is also satisfied by our definition of norm $\|\cdot\|$. Assumtions 3.1 - 3.4, 4.2, and 4.3 imply CS's condition B.5:

$$n^{-1/2} \sum_{t=1}^{n} (l'_{\theta_o}[v^*, Z_t] - E\, l'_{\theta_o}[v^*, Z_t]) \Rightarrow N(0, \sigma_*^2).$$

Assumption 4.1 together with Theorem 3.1 (a) imply that $\|\hat{\theta}_n - \theta_o\|^{\omega} = o_p(n^{-1/4})$ with $\omega = 2$. Thus all the

conditions in CS's theorem 2 are satisfied and the result follows. □

References

1. Barron, A. (1993) Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Information Theory*, 39, 930-45.

2. Barron, A. (1994) Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14, 115-133.

3. Chen, X. and X. Shen (1996) Asymptotic properties of sieve extremum estimates for weakly dependent data with applications. Manuscript, University of Chicago.

4. Dehling, H. (1983) Limit theorems for sums of weakly dependent Banach space valued random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 63, 393-432.

5. Goldstein, L., and K. Messer (1992) Optimal plug-in estimators for nonparametric functional estimation. *Annals of Statistics*, 20, 1306-1328.

6. Grenander, U. (1981) *Abstract Inference*. Wiley, New York.

7. Hornik, K., M. Stinchcombe, H. White, and P. Auer (1994) Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. *Neural Computation*, 6, 1262-75.

8. Makovoz, Y. (1996) Random approximants and neural networks. *Journal of Approximation Theory*, 85, 98-109.

9. Modha, D., and E. Masry (1996a) Minimum complexity regression estimation with weakly dependent observations. *IEEE Trans. Information Theory*, 42, 2133-2145.

10. Modha, D., and E. Masry (1996b) Rate of convergence in density estimation using neural networks. *Neural Computation*, 8, 1107-1122.

11. Pisier, G. (1981) Remarques sur un resultat non publie de B. Maurey. *Seminaire D'analyse Fonctionelle* 1980-1981, Ecole Polytechnique, Palaiseau.

12. Shen, X. (1997) On the efficiency of methods of sieves and penalization. Forthcoming, *Ann. of Statistics*

13. Stinchcombe, M., and H. White (1997) Consistent specification testing with nuisance parameters present only under the alternative. Forthcoming, *Econometric Theory*.

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

15. White, H. (1990) Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3, 535-549.