

# Minimum Complexity Regression Estimation with Weakly Dependent Observations

Dharmendra S. Modha, *Member, IEEE*, and Elias Masry, *Fellow, IEEE*

**Abstract**— The minimum complexity regression estimation framework, due to Barron, is a general data-driven methodology for estimating a regression function from a given list of parametric models using independent and identically distributed (i.i.d.) observations. We extend Barron's regression estimation framework to  $m$ -dependent observations and to strongly mixing observations. In particular, we propose abstract minimum complexity regression estimators for dependent observations, which may be adapted to a particular list of parametric models, and establish upper bounds on the statistical risks of the proposed estimators in terms of certain deterministic indices of resolvability. Assuming that the regression function satisfies a certain Fourier-transform-type representation, we examine minimum complexity regression estimators adapted to a list of parametric models based on neural networks and, by using the upper bounds for the abstract estimators, we establish rates of convergence for the statistical risks of these estimators. Also, as a key tool, we extend the classical Bernstein inequality from i.i.d. random variables to  $m$ -dependent processes and to strongly mixing processes.

**Index Terms**— Minimum complexity regression estimation, mixing processes, neural networks, rates of convergence, Bernstein inequality.

## I. INTRODUCTION

LET  $\{X_i, Y_i\}_{i=-\infty}^{\infty}$  be a bivariate stationary random process, such that  $X_1$  takes values in  $\mathbb{R}^d$  and  $Y_1$  takes values in  $\mathbb{R}$ . Define the regression function, namely, the conditional mean of  $Y_1$  given  $X_1$ , by

$$f^*(x) = E[Y_1 | X_1 = x], \quad x \in \mathbb{R}^d.$$

In general, the regression function  $f^*$  can only be assumed to satisfy weak smoothness conditions. In other words,  $f^*$ , in general, is not a member of a finite-dimensional parametric family of functions. Thus any model depending only on some finite set of parameters will be generically inadequate to approximate  $f^*$ . In contrast, in this paper, we consider a list of parametric models of increasing dimensionality which approximate  $f^*$  more and more accurately as their dimension  $n$  increases.

Manuscript received August 30, 1994; revised March 11, 1996. This work was supported by the Office of Naval Research under Grant N00014-90-J-1175. The material in this paper was presented in part at the 1994 IEEE-IMS Workshop on Information theory and Statistics, Alexandria, VA, October 27–29, 1994.

The authors are with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0407 USA.  
Publisher Item Identifier S 0018-9448(96)06891-5.

Given  $N$  observations

$$\{X_i, Y_i\}_{i=1}^N$$

drawn from

$$\{X_i, Y_i\}_{i=-\infty}^{\infty}$$

and given a suitable list of parametric models of increasing dimensionality, we are interested in estimating the regression function  $f^*$ , in a data-driven fashion, so as to achieve the smallest statistical risk.

Statistical risk in estimating  $f^*$  using a parametric model has two components: approximation error (“bias”) and estimation error (“variance”). Generally speaking, a model with a larger dimension has a smaller bias but a larger variance, while a model with a smaller dimension has a smaller variance but a larger bias. Consequently, to minimize the statistical risk in estimating  $f^*$  from a list of parametric models of increasing dimensionality, a tradeoff between the bias and the variance must be found. The tradeoff can be achieved by judiciously selecting the dimension of the model used to estimate  $f^*$ . Minimum complexity regression estimation framework (also called complexity regularization) is a data-driven methodology for selecting the model dimension so as to achieve such a tradeoff among (possibly) nonlinearly parametrized models, see Barron [3], Barron and Cover [6], and Rissanen [24]. It is closely related to Vapnik's method of structural risk minimization [29]. For related work see Farago and Lugosi [13], Lugosi and Zeger [18], [19], and McCaffrey and Gallant [21].

In this paper, we extend the minimum complexity regression estimation framework from independent and identically distributed (i.i.d.) observations to more general cases of  $m$ -dependent [14] observations and strongly mixing [27] observations. Previously, White and Wooldridge [31] and White [30] considered cross-validated regression estimators for strongly mixing processes and established convergence, without rates, of their estimators. In contrast, we consider minimum complexity regression estimators and obtain rates of convergence.

In Section III, we propose abstract minimum complexity regression estimators, which may be adapted to a particular list of parametric models. The proposed estimators are obtained by minimizing a certain complexity regularized empirical loss (see (23) and (24)). We then establish, in Theorem 3.1, upper bounds on the statistical risks of the proposed estimators in terms of certain deterministic indices of resolvability—which are, in turn, relatively easier to upper-bound for a particular

list of parametric models of interest. The proof of Theorem 3.1 uses ideas from Barron [3] and McCaffrey and Gallant [21]; the proof also relies on certain Bernstein-type inequalities for dependent observations which are derived in Section IV. The main difference between the results of Section III and those of [3] for i.i.d. observations is that for dependent observations, the “effective number of observations” are found to be less than the number of observations  $N$ ; in other words, we find different tradeoffs between the bias and the variance. Also, unlike Barron, we do not restrict the parameter space of the models to be countable.

In Section II, we apply the abstract ideas of Section III (namely, Theorem 3.1) to neural networks. Specifically, assuming that the observations  $\{X_i, Y_i\}_{i=1}^N$  are either  $m$ -dependent or strongly mixing, that  $X_1$  and  $Y_1$  are bounded, and that  $f^*$  admits a certain Fourier-transform-type representation, we examine minimum complexity regression estimators based on a list of parametric models constructed from neural networks (see (6)–(9)). Furthermore, in Theorem 2.1, we establish rates of convergence, independent of the dimension  $d$ , for the statistical risks of these estimators based on neural networks. Theorem 2.1 extends previous results of Barron [5], for minimum complexity regression estimators based on neural networks, from i.i.d. observations to dependent observations.

In Section IV, we extend the classical Bernstein inequality [9], [28] to  $m$ -dependent processes and to strongly mixing processes; these extensions are used in the proof of Theorem 3.1. Previously, Bosq [7] established a Bernstein-type inequality for uniformly mixing processes, a class of processes smaller than strongly mixing processes. Also, Carbon [8, Proposition 1] and White and Wooldridge [31, Theorem 3.3] established exponential inequalities for strongly mixing processes. However, inequalities of Carbon and White and Wooldridge are of a different form than the classical Bernstein inequality in that they contain a lesser power of  $\zeta_2$  as compared to our Theorem 4.3. Consequently, their inequalities lead to a weaker upper bound on the model variance, and hence do not permit as good a tradeoff between the bias and the variance (or, equivalently, as good a rate of convergence) as that obtained here. The inequalities of Section IV (and the related Hoeffding inequality for strongly mixing processes in Modha [22]) should also be of independent interest. For example: i) they may be useful in establishing a rate of convergence for the uniform strong law of large numbers for strongly mixing processes (see Pollard [23] and Vapnik [29] for the i.i.d. case); ii) they may furnish the exponential bounds (on the tail probabilities) required to invoke a certain chaining argument while establishing functional central limit theorems for strongly mixing processes (see Andrews and Pollard [1]) and, in a related setting, they may help avoid the detour to independent blocks in Arcones and Yu [2], Doukhan, Massart, and Rio [12], and Yu [33]; and finally, iii) they may help avoid the detour to Bradley’s strong approximation theorem in certain estimation-theoretic proofs, for example, see Masry [20]. A more detailed analysis is needed to ascertain whether using our inequalities, in the above cited contexts, leads to more refined results and/or to simpler proofs. Furthermore, our inequalities require an exponential decay for the strong

mixing coefficient whereas an algebraic decay was sufficient in [1], [2], [12], [20], and [33].

In the Appendix, we gather some simple but useful results.

## II. REGRESSION ESTIMATION USING NEURAL NETWORKS

### A. Two Notions of Dependence

Let  $\{Z_i \equiv (X_i, Y_i)\}_{i=-\infty}^{\infty}$  be a stationary random process on a probability space  $(\Omega, \mathcal{F}, P)$ . For  $-\infty < i < \infty$ , let  $\mathcal{F}_i^{\infty}$  and  $\mathcal{F}_{-\infty}^i$  denote the  $\sigma$ -algebras of events generated by the random variables  $\{Z_j, j \geq i\}$  and  $\{Z_j, j \leq i\}$ , respectively.

*Definition 2.1:* For  $m \geq 0$ ,  $\{Z_i\}_{i=-\infty}^{\infty}$  is called  $m$ -dependent [14], if  $\mathcal{F}_{-\infty}^0$  and  $\mathcal{F}_{m+1}^{\infty}$  are independent.

Set  $N^{(m)} = \lfloor N/(m+1) \rfloor$ , where  $N$  denotes the number of observations.  $N^{(m)}$  arises from the Bernstein inequality for  $m$ -dependent processes (Theorem 4.2) and is called the “effective number of observations” for  $m$ -dependent processes.

*Definition 2.2:*  $\{Z_i\}_{i=-\infty}^{\infty}$  is called *strongly mixing* [27], if

$$\sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_j^{\infty}} |P[AB] - P[A]P[B]| = \alpha(j) \rightarrow 0 \quad \text{as } j \rightarrow \infty.$$

$\alpha(j)$  is called the strong mixing coefficient.

*Assumption 2.1 Exponentially Strongly Mixing:* Assume that the strong mixing coefficient satisfies

$$\alpha(j) \leq \bar{\alpha} \exp(-cj^{\beta}), \quad j \geq 1$$

for some  $\bar{\alpha} > 0$ ,  $\beta > 0$ , and  $c > 0$ , where the constants  $\beta$  and  $c$  are assumed to be known.

Assumption 2.1 is satisfied by a large class of processes, for example, certain linear processes (which includes certain ARMA processes) satisfy the assumption with  $\beta = 1$  [32], and also certain aperiodic, Harris-recurrent Markov processes (which includes certain bilinear processes, nonlinear ARX processes, and ARCH processes [11]) satisfy the assumption [10, Theorem 1]. As a trivial example, i.i.d. random variables satisfy the assumption with  $\beta = \infty$ .

Set

$$N^{(\alpha)} = \left\lfloor N \left[ \left\{ \frac{8N}{c} \right\}^{1/(\beta+1)} \right]^{-1} \right\rfloor \quad (1)$$

where  $N$  denotes the number of observations and  $\lfloor u \rfloor$  ( $\lceil u \rceil$ ) denotes the greatest (least) integer less (greater) than or equal to  $u$ .  $N^{(\alpha)}$  arises from the Bernstein inequality for strongly mixing processes (Theorem 4.3) and is called the “effective number of observations” for strongly mixing processes.

$N^{(m)}$  and  $N^{(\alpha)}$  play the same role in our analysis as that played by the number of observations  $N$  in the i.i.d. case.

### B. A Class of Target Regression Functions

*Assumption 2.2 Compactness:* Assume that  $Y_1$  takes values in a known fixed interval  $[a, a+b]$ , for some  $a \in \mathbb{R}$  and for some  $b > 0$ .

Assumption 2.2 is introduced here with the hindsight that the minimum complexity regression estimation framework developed in Section III requires it; in particular, it is necessary

to enable us to use the exponential inequalities derived in Theorems 4.2 and 4.3. Assumption 2.2 implies that the regression function  $f^*$  also takes values in the interval  $[a, a + b]$ .

For  $w = (w_1, \dots, w_d)$  and  $x = (x_1, \dots, x_d)$  in  $\mathbb{R}^d$ , let

$$w \cdot x = \sum_{i=1}^d w_i x_i$$

denote the usual inner product on  $\mathbb{R}^d$  and let

$$\|w\|_1 = \sum_{i=1}^d |w_i|$$

denote a norm on  $\mathbb{R}^d$ . The class of regression functions of interest is characterized by the following assumption.

*Assumption 2.3 Barron [4]:* Assume that

- $X_1$  takes values in  $\mathcal{B}_X = [-1, 1]^d$ , and that
- there exists a complex-valued function  $\tilde{f}$  on  $\mathbb{R}^d$  such that for  $x \in \mathcal{B}_X$ , we have

$$f^*(x) = f^*(0) + \int_{\mathbb{R}^d} (e^{iw \cdot x} - 1) \tilde{f}(w) dw$$

and that

$$\int_{\mathbb{R}^d} \|w\|_1 |\tilde{f}(w)| dw \leq C' < \infty$$

for some known  $C' > 0$ . Set  $C = \max\{1, C'\}$ .

Part b) of Assumption 2.3 implies that the regression function  $f^*$  has an inverse Fourier-transform-type representation on the set  $\mathcal{B}_X$ . Specifically,  $f^*$  has an extension  $f^{*e}$  outside the compact set  $\mathcal{B}_X$ , such that the extended function  $f^{*e}$  possesses a uniformly continuous gradient whose Fourier transform is absolutely integrable [4]. Assumption 2.3 characterizes a class of functions for which neural networks can provide rates of approximation independent of the dimension  $d$ .

### C. Neural Networks

In this subsection, we use various results of Barron [5] to construct a list of parametric models based on neural networks, which is specifically designed to well approximate the class of functions characterized by Assumption 2.3.

We assume that  $\phi: \mathbb{R} \rightarrow [0, 1]$  is a Lipschitz continuous sigmoidal function such that its tails approach the tails of the unit step function at least polynomially fast.

*Assumption 2.4 [5]:* Assume that

- $\phi(u) \rightarrow 1$  as  $u \rightarrow \infty$  and  $\phi(u) \rightarrow 0$  as  $u \rightarrow -\infty$ .
- $|\phi(u)| \leq 1$  and  $|\phi(u) - \phi(v)| \leq D'_1 |u - v|$  for all  $u, v \in \mathbb{R}$  and for some  $D'_1 > 0$ . Set  $D_1 = \max\{1, D'_1\}$ .
- $|\phi(u) - 1_{\{u > 0\}}| \leq D'_2 / |u|^{D_3}$  for  $u \in \mathbb{R}$ ,  $u \neq 0$ , and for some  $D_3 > 0$  and  $D'_2 > 0$ . Set  $D_2 = \max\{1, D'_2\}$ .

Fix  $n \geq 1$ . We now proceed to define a neural networks with  $n$  “hidden units.” Let

$$\gamma_n = n(d + 2) + 1 \quad (2)$$

represent the number of real-valued parameters parameterizing such a network. For  $0 \leq i \leq n$ , let  $c_i \in \mathbb{R}$ ; for

$1 \leq i \leq n$ , let  $a_i \in \mathbb{R}^d$  and let  $b_i \in \mathbb{R}$ . We define a  $\gamma_n$ -dimensional parameter vector

$$\nu = (a_1, a_2, \dots, a_n; b_1, b_2, \dots, b_n; c_0, c_1, \dots, c_n).$$

Now, define a neural networks with  $n$  hidden units  $f_{(n, \nu)}: \mathbb{R}^d \rightarrow \mathbb{R}$  parameterized by  $\nu$  as

$$f_{(n, \nu)}(x) = \text{clip} \left( c_0 + \sum_{i=1}^n c_i \phi(a_i \cdot x + b_i) \right), \quad x \in \mathbb{R}^d \quad (3)$$

where

$$\text{clip}(t) = a 1_{\{t < a\}} + t 1_{\{a \leq t \leq a+b\}} + (a+b) 1_{\{a+b < t\}}.$$

In (3), “clip” serves to restrict the range of  $f_{(n, \nu)}$  to  $[a, a + b]$  without disturbing the capacity of  $f_{(n, \nu)}$  to approximate  $f^*$ . Let  $\tau_0 = \max\{1, (b + 3)/(2e) - 1\}$  and let the rate at which the hidden unit weights, namely,  $a_i$  and  $b_i$ , are allowed to grow as a function of  $n$  be

$$\tau_n = \tau_0 2^{(2D_3+1)/D_3} D_2^{1/D_3} n^{(D_3+1)/(2D_3)} \quad (4)$$

where  $D_1$ ,  $D_2$ , and  $D_3$  are as in Assumption 2.4. We now define a compact subset of  $\mathbb{R}^{\gamma_n}$ , namely,

$$S_n = \left\{ \nu: c_0 \in [a, a + b], \sum_{i=1}^n |c_i| \leq C, \right. \\ \left. \max_{1 \leq i \leq n} \|a_i\|_1 \leq \tau_n, \max_{1 \leq i \leq n} |b_i| \leq \tau_n \right\}. \quad (5)$$

Assumptions 2.3 and 2.4 and the choices (3)–(5) are made, with hindsight, to fulfill the hypothesis of [5, Corollary 1].

### D. Minimum Complexity Regression Estimators and Rates of Convergence

We now construct the advertised minimum complexity regression estimators and establish upper bounds on their statistical risks. This subsection uses the abstract minimum complexity regression estimators to be presented in Section III-B (see Example 3.1).

Let  $\kappa(N)$  be as in Theorem 2.1. For each fixed  $1 \leq n \leq \kappa(N)$ , compute the least squares estimator with  $n$  hidden units as

$$\hat{\nu}(n, N) = \arg \min_{\nu \in S_n} \left\{ \frac{1}{N} \sum_{i=1}^N [Y_i - f_{(n, \nu)}(X_i)]^2 \right\} \quad (6)$$

where  $f_{(n, \nu)}$  is as in (3). Write

$$\hat{f}_{n, N} = f_{(n, \hat{\nu}(n, N))}. \quad (7)$$

Given the sequence of least squares estimators, namely,  $\{\hat{f}_{n, N}\}_{1 \leq n \leq \kappa(N)}$ , how should one formulate an estimator for  $f^*$  that achieves the smallest statistical risk? In particular, for a very small  $n$ ,  $\hat{f}_{n, N}$  has a small variance but a large bias; on the other hand, for a very large  $n$ ,  $\hat{f}_{n, N}$  has a small bias but a large variance. A tradeoff between the bias and the variance can be achieved by selecting the number of hidden

units (that is, the model dimension) in a data-driven fashion. Specifically, *compute*

$$\hat{n} = \arg \min_{1 \leq n \leq \kappa(N)} \left\{ \frac{1}{N} \sum_{i=1}^N [Y_i - \hat{f}_{n,N}(X_i)]^2 + \lambda \frac{L_n(\varepsilon(N)) + 2 \ln(n+1)}{N} \right\} \quad (8)$$

where

$$L_n(\varepsilon) = [n(d+2) + 1] \ln(8\tau_n e/\varepsilon)$$

is as in (18), and  $\lambda$ ,  $\bar{N}$ ,  $\kappa(N)$ , and  $\varepsilon(N)$  are as in Theorem 2.1. The first term on the right-hand side of (8) is known as the “empirical loss” of  $\hat{f}_{n,N}$ , while the second term is known as the “complexity” of  $\hat{f}_{n,N}$ . Equation (8) allows  $\hat{n}$  to take a larger value, only if the resulting increase in the complexity is offset by a matching decrease in the empirical loss.

Finally, define the *minimum complexity regression estimator* for  $f^*$  as

$$\hat{f}_N = \hat{f}_{\hat{n},N}. \quad (9)$$

In other words,  $\hat{f}_N$  is the element corresponding to  $\hat{n}$  in the sequence  $\{\hat{f}_{n,N}\}_{1 \leq n \leq \kappa(N)}$ . We now establish rates of convergence for the statistical risk (measured by mean integrated squared error) of  $\hat{f}_N$ .

**Theorem 2.1 Rates of Convergence:** Suppose that Assumptions 2.2, 2.3, and 2.4 hold.

**Condition (m):** Suppose that  $\{X_i, Y_i\}_{i=-\infty}^{\infty}$  is  $m$ -dependent. Then, let  $\bar{N} = N^{(m)}$ , let  $\kappa(N) \geq \lceil \sqrt{N^{(m)}} \rceil$ , and for some  $D_4 \geq 1/2$  let  $\varepsilon(N) = (N^{(m)})^{-D_4}$ .

**Condition ( $\alpha$ ):** Suppose that  $\{X_i, Y_i\}_{i=-\infty}^{\infty}$  is strongly mixing and that Assumption 2.1 holds. Then, let  $\bar{N} = N^{(\alpha)}$ , let  $\kappa(N) \geq \lceil \sqrt{N^{(\alpha)}} \rceil$ , and for some  $D_4 \geq 1/2$  let  $\varepsilon(N) = (N^{(\alpha)})^{-D_4}$ .

Then, either under Condition (m) or under Condition ( $\alpha$ ), and for  $\lambda > 5b^2/3$  and  $\bar{N} \geq 2$

$$E \int_{\mathbb{R}^d} [\hat{f}_N(x) - f^*(x)]^2 dP_X(x) \leq \tilde{K} \sqrt{\frac{\log \bar{N}}{N}} \quad (10)$$

where  $P_X$  denotes the marginal distribution of  $X_1$  and the constant  $\tilde{K}$  can be read from (13) and (14).

**Proof:** For any two measurable functions  $g_1, g_2: \mathbb{R}^d \rightarrow \mathbb{R}$ , define the integrated squared error between them as

$$r(g_1, g_2) = \int_{\mathbb{R}^d} [g_1(x) - g_2(x)]^2 dP_X(x). \quad (11)$$

To establish Theorem 2.1, we proceed in two steps as follows.

1) Define the *index of resolvability* corresponding to  $\hat{f}_N$  as

$$R_N(f^*) = \min_{1 \leq n \leq \kappa(N)} \left\{ \min_{\nu \in S_n} [r(f(n, \nu), f^*)] + \lambda \frac{L_n(\varepsilon(N)) + 2 \ln(n+1)}{N} \right\}. \quad (12)$$

We first establish, in Lemma 2.1, an upper bound on the overall statistical risk  $E[r(\hat{f}_N, f^*)]$  of the estimator  $\hat{f}_N$  in terms of the index of resolvability  $R_N(f^*)$ , by invoking Theorem 3.1. The proof of Theorem 3.1 can be found, in an abstract setting, in Section III-C; it uses techniques from Barron [3] and McCaffrey and Gallant [21] and also uses the Bernstein inequalities for dependent processes derived in Section IV.

2) We next establish, in Lemma 2.2, an upper bound on the index of resolvability  $R_N(f^*)$  using ideas in [5].

**Lemma 2.1. A Bound on Statistical Risk:** Suppose that Assumption 2.2 holds. Let  $\lambda > 5b^2/3$  and  $\bar{N} \geq 2$ . Then, either under Condition (m) or under Condition ( $\alpha$ ) of Theorem 2.1,

$$E[r(\hat{f}_N, f^*)] \leq \frac{1+\eta}{1-\eta} R_N(f^*) + \frac{6b(4D_1C)\varepsilon(N)}{1-\eta} + \frac{4\tilde{\alpha}\lambda}{(1-\eta)\bar{N}} \quad (13)$$

where  $\eta$  and  $\tilde{\alpha}$  are as in Theorem 3.1.

**Proof:** It follows from Example 3.1 that all the hypotheses of Theorem 3.1 hold, and hence the lemma follows by setting  $\delta = (4D_1C)\varepsilon(N)$ .  $\square$

**Lemma 2.2 A Bound on the Index of Resolvability:** Suppose that Assumptions 2.3 and 2.4 hold. Then, either under Condition (m) or under Condition ( $\alpha$ ) of Theorem 2.1,

$$R_N(f^*) \leq K_9 \sqrt{\frac{\log \bar{N}}{N}}$$

where the constant  $K_9$  is as in (14).

**Proof:**

$$\begin{aligned} R_N(f^*) &= \min_{1 \leq n \leq \kappa(N)} \left\{ \min_{\nu \in S_n} [r(f(n, \nu), f^*)] + \lambda \frac{L_n(\varepsilon(N)) + 2 \ln(n+1)}{N} \right\} \\ &\stackrel{(a)}{\leq} \min_{1 \leq n \leq \lceil \sqrt{\bar{N}} \rceil} \left\{ \frac{4C^2}{n} + \lambda \frac{n(d+2) + 1}{N} \right. \\ &\quad \left. \cdot \ln \frac{4\tau_n e}{\varepsilon(N)} + \lambda \frac{2 \ln(n+1)}{N} \right\} \\ &\stackrel{(b)}{\leq} \min_{1 \leq n \leq \lceil \sqrt{\bar{N}} \rceil} \left\{ \frac{4C^2}{n} + \lambda \frac{n(d+3)}{N} \right. \\ &\quad \left. \cdot \ln(K_1 n^{K_2} (\bar{N})^{D_4}) + \lambda \frac{2 \ln(n+1)}{N} \right\} \\ &\stackrel{(c)}{\leq} \min_{1 \leq n \leq \lceil \sqrt{\bar{N}} \rceil} \left\{ \frac{K_3}{n} + \frac{nK_4}{N} \right. \\ &\quad \left. \cdot \ln(K_1 (\bar{N})^{2K_5}) + \frac{\lambda \ln 4 (\bar{N})^2}{N} \right\} \\ &\stackrel{(d)}{\leq} \min_{1 \leq n \leq \lceil \sqrt{\bar{N}} \rceil} \left\{ \frac{K_3}{n} + \frac{nK_6}{N} \ln(K_7 (\bar{N})^2) \right\} \end{aligned}$$

$$R_N(f^*) \stackrel{(f)}{\leq} K_9 \sqrt{\frac{\ln \bar{N}}{\bar{N}}} \tag{14}$$

where (a) follows from Barron [5, Corollary 1], (18), and since  $\kappa(N) \geq \lfloor \sqrt{N} \rfloor$ ; (b) follows from (4) by setting  $K_1 = 8\epsilon\tau_0 2^{(2D_3+1)/D_3} D_2^{1/D_3}$  and by setting  $K_2 = (D_3 + 1)/(2D_3)$  and also since  $\epsilon(N) = (\bar{N})^{-D_4}$ ; (c) follows since  $n \leq \lfloor \sqrt{\bar{N}} \rfloor \leq \bar{N}$  and by setting  $K_3 = 4C^2$ ,  $K_4 = \lambda(d + 3)$ ,  $K_5 = (K_2 + D_4)/2$ ; (d) follows by setting  $K_6 = 2 \max \{K_4 K_5, \lambda\}$  and  $K_7 = \max \{(K_1)^{1/K_5}, 4\}$ ; (e) follows by setting  $K_8 = 4K_6 \ln K_7$ ; (f) follows by selecting

$$n = \left\lfloor \sqrt{\bar{N}/\ln \bar{N}} \right\rfloor$$

and by setting  $K_9 = (K_3 + 2K_8)$ . Also, since  $\bar{N} \geq 2$ , we have

$$1 \leq \left\lfloor \sqrt{\bar{N}/\ln \bar{N}} \right\rfloor \leq \left\lfloor \sqrt{\bar{N}} \right\rfloor. \quad \square$$

Theorem 2.1 now follows from Lemmas 2.1 and 2.2, if  $D_4 \geq 1/2$ . □

*E. Discussion*

*Remark 2.1:* The minimum complexity regression estimators for dependent observations in Section II-D differ from the corresponding estimator of Barron [5] for i.i.d. observations, in that, in our case, the effective number of observations  $\bar{N}$  and not the actual number of observations  $N$  appears in the second term on the right-hand side in (8). Correspondingly, the rate of convergence obtained in Theorem 2.1 for dependent observations is  $O((\ln \bar{N}/\bar{N})^{1/2})$ , whereas the rate of convergence obtained in [5] for i.i.d. observations was  $O((\ln N/N)^{1/2})$ . Consequently, for  $m$ -dependent observations, since  $\bar{N} = N^{(m)} = \lfloor N/(m + 1) \rfloor$ , the rate obtained in our case is identical to that obtained by Barron. However, for strongly mixing observations, since  $\bar{N} = N^{(\alpha)} \sim N^{\beta/(\beta+1)}$ , the rate obtained in our case is slower than that obtained by him. Technically, the decrease in the rate in the strongly mixing case is due to the corresponding decrease in the rate of decay in the upper bound in the Bernstein inequality for strongly mixing processes (compare Theorems 4.1 and 4.3). In a similar context, while analyzing their regression estimators for strongly mixing processes, White and Wooldridge [31] found that models with smaller dimensions are required, to achieve weak consistency, in the mixing case as compared with the independent case.

*Remark 2.2:* Notice that if we set  $m = 0$  (under Condition (m)) or  $\beta = \infty$  (under Condition (α)) in Theorem 2.1, then we recover the i.i.d. result of Barron [5] as a special case. However, observe that in (6) we compute the least squares estimator by minimization over the entire parameter space, whereas Barron computed his estimator by minimization over a certain finite grid of parameters.

*Remark 2.3:* For strongly mixing observations, we now compare the rate of convergence obtained in Theorem 2.1 to that achieved by the classical nonparametric kernel estimator.

Suppose that Assumptions 2.1, 2.2, 2.3, and 2.4 hold; then we have from Theorem 2.1 (under Condition (α)) and from (1) that

$$E \int_{\mathbb{R}^d} [\hat{f}_N(x) - f^*(x)]^2 dP_X(x) = O\left(\sqrt{\frac{\log N}{N^{\beta/(\beta+1)}}}\right). \tag{15}$$

Noticeably, the exponent of  $N$  in the rate of convergence does not depend on the dimension  $d$ . While formulating the minimum complexity regression estimator  $\hat{f}_N$  we only assumed (see Assumption 2.3) that

$$\int_{\mathbb{R}^d} \|w\|_1 |\tilde{f}(w)| dw \leq C < \infty.$$

It may be possible to achieve a faster rate of convergence for  $\hat{f}_N$  under Assumption 2.3 with

$$\int_{\mathbb{R}^d} \|w\|_1^s |\tilde{f}(w)| dw < \infty, \quad s > 1$$

but no method of proof is currently available.

Now, on the other hand, suppose that the regression function  $f^*$  has continuous and bounded partial derivatives of total order  $s$  and suppose that the strong mixing coefficient decays algebraically, that is,  $\alpha(j) = o(1/j^2)$ ,  $j \geq 1$ . Let  $\tilde{f}_N$  denote a nonrecursive kernel estimator [25], [26] which uses a kernel of order  $s$ . Then, it is known that with an optimal choice of the corresponding bandwidth parameter, we have for each  $x \in \mathbb{R}^d$

$$E[\tilde{f}_N(x) - f^*(x)]^2 \sim \frac{1}{N^{2s/(2s+d)}}. \tag{16}$$

The exponent of  $N$  in the rate of convergence depends on  $d$ , and hence  $\tilde{f}_N$  delivers progressively poorer performance as  $d$  increases, that is, suffers from the curse of dimensionality.

Notice that the estimators  $\hat{f}_N$  and  $\tilde{f}_N$  are formulated under different assumptions and use different measures of performance, respectively, mean integrated squared error and mean squared error;<sup>1</sup> thus direct comparison of (15) with (16) is not possible. However, since Assumption 2.3 implies that the regression function  $f^*$  has bounded and continuous partial derivatives of total order 1, if we set  $s = 1$  for  $\tilde{f}_N$ , then roughly  $\hat{f}_N$  outperforms  $\tilde{f}_N$  if  $d > 2(1+2/\beta)$ . Finally, observe that we require that the strong mixing coefficient decays exponentially fast, a fairly stringent condition, when compared to the algebraic decay permitted by the kernel estimator.

*Remark 2.4:* For the sake of simplicity, while formulating the estimator  $\hat{f}_N$  we required that the constant  $C$  is known. Using ideas in [5, eqs. (31) and (32)], it is easy to extend our estimators to cover the case when  $C$  is unknown.

<sup>1</sup>To the best of our knowledge, although rates of convergence in the mean integrated squared error sense are available for kernel density estimators, no such rates are available for kernel regression estimators—even for i.i.d. observations.

*Remark 2.5:* While formulating the estimator  $\hat{f}_N$  we required that the set of parameters  $S_n$  be compact. By introducing a prior density (satisfying certain regularity conditions) on the set of parameters and by proceeding essentially as in [5, p. 129], it is possible to eliminate the compactness assumption and still obtain the same rate of convergence for the resulting estimator as that obtained in Theorem 2.1. However, we do not pursue such an extension here, since i) the resulting estimator must search over a wider domain; ii) the estimator once again involves the level of discretization, namely  $\varepsilon(N)$ , in its computation; and iii) a more elaborate abstract estimation framework is required to accommodate a prior.

*Remark 2.6:* Recently, Hornik *et al.* [17] and Yukich, Stinchcombe, and White [34] have established generalized approximation bounds analogous to [5, Corollary 1] in Sobolev norms and in sup norm, respectively. It may be possible to use their results (for example, [17, Theorems 2.1 and 2.3] and [34, Theorems 2.1 and 2.2]) to substantially relax Assumption 2.4 and to relax Part b) of Assumption 2.3; however, since, unlike them, we restrict our parameter space  $S_n$  to be compact, we are unable to employ their results here. Even though it is possible to dispose off the compactness assumption in our framework (see Remark 2.5), a result analogous to [4, Theorem 3] is still required before we may exploit the generalized approximation bounds of Hornik *et al.* and of Yukich, Stinchcombe, and White.

### III. MINIMUM COMPLEXITY REGRESSION ESTIMATION FRAMEWORK

The principal result of this section, namely Theorem 3.1, is derived in an abstract setting and under minimal structure on the underlying space of parameters. This not only simplifies the proof of the theorem, but also widens the scope of the result. In particular, Theorem 3.1 is not limited to neural networks (as used in Lemma 2.1 of Section II), but may also apply to trigonometric series, polynomials, and wavelets.

Throughout this section, fix the number of observations  $N \geq 1$  and let  $B_X \subseteq \mathbb{R}^d$  denote the support of  $X_1$ .

#### A. Abstract Parameter Spaces and Abstract Complexities

For each integer  $n \geq 1$ , let  $\gamma_n$  denote a model dimension, for example, see (2), and let  $S_n$  denote a compact subset of  $\mathbb{R}^{\gamma_n}$ . The set  $S_n$  will serve as a collection of parameters associated with the model dimension  $\gamma_n$ , for example, see (5). For every  $\nu \in S_n$ , let  $f_{(n,\nu)}$  denote a real-valued function on  $B_X$  parameterized by  $(n, \nu)$ , for example, see (3). The following condition is required to invoke the exponential inequalities in Theorems 4.2 and 4.3.

*Assumption 3.1:* For each integer  $n \geq 1$  and for every  $\nu \in S_n$ , assume that  $f_{(n,\nu)}$  takes values in  $[a, a + b]$ .

To make possible the union bound argument required in Lemma 3.2, we now introduce a certain finite subset of  $S_n$ . Let  $\rho_n$  denote a metric on  $\mathbb{R}^{\gamma_n}$ . For  $\varepsilon \in (0, 1]$ , let  $T_n(\varepsilon)$  denote an  $(\varepsilon, \rho_n)$ -net of the set  $S_n$ ; in other words, for every  $\nu_1 \in S_n$  there exists a  $\nu_2 \in T_n(\varepsilon)$  such that  $\rho_n(\nu_1, \nu_2) \leq \varepsilon$ . Assume that  $T_n(\varepsilon) \subset S_n$ . Actual construction of  $T_n(\varepsilon)$  is not required here, it suffices that it exists and that an upper bound

on its cardinality is known. Let  $L_n(\varepsilon)$  be such that

$$\ln \#(T_n(\varepsilon)) \leq L_n(\varepsilon) \quad (17)$$

where  $\#$  denotes the cardinality operator. In other words,  $L_n(\varepsilon)$  is an upper bound on the natural log of the  $\varepsilon$ -metric entropy of the set  $S_n$  with respect to the metric  $\rho_n$ . In practice, the upper bound  $L_n(\varepsilon)$  should be as tight as possible to obtain the best possible estimators.

*Example 3.1:* Let notations be as in Section II. Let  $\rho_n$  denote a metric on  $\mathbb{R}^{\gamma_n}$  defined as in Barron [5, eq. (19)]. It follows from [5, Lemma 2] by using (4) that for every  $0 < \varepsilon \leq 1$  and for every  $C \geq 1$ , there exists a  $(\varepsilon, \rho_n)$ -net of  $S_n$ , namely  $T_n(\varepsilon)$ , such that

$$\ln \#(T_n(\varepsilon)) \leq [n(d+2) + 1] \ln \frac{4\tau_n e}{\varepsilon} \equiv L_n(\varepsilon) \quad (18)$$

where we use the precision  $\varepsilon/2$  on the right-hand side to ensure that  $T_n(\varepsilon) \subset S_n$ .

We now introduce the following assumption to furnish the continuity argument required in Lemma 3.2.

*Assumption 3.2:* For every  $n \geq 1$ , there exists a strictly increasing function (in  $\varepsilon$ )  $\varpi_n(\cdot): (0, 1] \rightarrow (0, \infty)$  such that for all  $\varepsilon \in (0, 1]$  and for all  $\nu_1 \in S_n$  and  $\nu_2 \in T_n(\varepsilon)$  with  $\rho_n(\nu_1, \nu_2) \leq \varepsilon$ , we have

$$\sup_{x \in B_X} |f_{(n,\nu_1)}(x) - f_{(n,\nu_2)}(x)| \leq \varpi_n(\varepsilon).$$

Assumption 3.2 implies that the function  $\varpi_n$  is invertible; let  $\varpi_n^{-1}$  denote the inverse. Observe that the inverse  $\varpi_n^{-1}(\delta)$  is defined for all  $0 < \delta \leq \varpi_n(1) < \infty$  and takes values in the range  $(0, 1]$ . Assumption 3.2 is equivalent to saying that the class of parametric functions  $\{f_{(n,\nu)}: \nu \in S_n\}$  can be covered in the supremum norm (over  $B_X$ ) by the finite class of functions  $\{f_{(n,\nu)}: \nu \in T_n(\varpi_n^{-1}(\delta))\}$ .

*Example 3.1 Continued:* It follows from [5, Lemma 1], by invoking Assumption 2.2 and Part b) of Assumption 2.4, that Assumption 3.2 holds with  $\varpi_n(\varepsilon) = 4D_1C\varepsilon$ . For all  $0 < \delta \leq \varpi_n(1) = 4D_1C$ , the inverse of  $\varpi_n$  can be written as

$$\varpi_n^{-1}(\delta) = \frac{\delta}{4D_1C}. \quad (19)$$

Let  $\Theta_\kappa$  denote a collection of parameters of different dimensions, with the index  $n$  less than or equal to  $\kappa$ . Each of the parameters comes packaged with the index of its dimension; formally, we write

$$\Theta_\kappa = \bigcup_{n=1}^{\kappa} \{(n, \nu): \nu \in S_n\}. \quad (20)$$

It follows from (20) that every  $\theta \in \Theta_\kappa$  must be of the form  $\theta = (n, \nu)$  for some  $1 \leq n \leq \kappa$  and for some  $\nu \in S_n$ ; then, define

$$f_\theta = f_{(n,\nu)} \quad (21)$$

and for every  $0 < \delta \leq \varpi_n(1)$  define the "description complexity" of the parameter  $\theta$  as

$$L(\theta, \delta) = 2 \ln(n+1) + L_n(\varpi_n^{-1}(\delta)) \quad (22)$$

where  $\varpi_n$  is as in Assumption 3.2 and  $L_n(\varpi_n^{-1}(\delta))$  is obtained from (17) by substituting  $\varepsilon = \varpi_n^{-1}(\delta)$ . In words, for each fixed  $1 \leq n \leq \kappa$  and for each fixed  $0 < \delta \leq \varpi_n(1)$ , we assign a constant complexity, namely  $L_n(\varpi_n^{-1}(\delta))$ , to each parameter  $\nu \in S_n$ . Consequently, the right-hand side of (22) does not depend on  $\nu$ , it only depends on  $n$ .

### B. Abstract Estimators and Indices of Resolvability

For any natural number  $\kappa$ , for any real number  $\delta$ , where  $0 < \delta \leq \min_{1 \leq n \leq \kappa} \varpi_n(1)$ , and for any real number  $\lambda$ , write

$$\hat{\theta}_N = \arg \min_{\theta \in \Theta_\kappa} \left\{ \frac{1}{N} \sum_{i=1}^N [Y_i - f_\theta(X_i)]^2 + \lambda \frac{L(\theta, \delta)}{N} \right\} \quad (23)$$

where  $\Theta_\kappa$  is as in (20),  $f_\theta$  is as in (21),  $L(\theta, \delta)$  is as in (22), and  $\bar{N}$  is as in Theorem 3.1. Define the *minimum complexity regression estimator* as

$$\hat{f}_N = f_{\hat{\theta}_N} \quad (24)$$

and define the *index of resolvability* corresponding to  $\hat{f}_N$  as

$$R_N(f^*) = \min_{\theta \in \Theta_\kappa} \left\{ r(f_\theta, f^*) + \lambda \frac{L(\theta, \delta)}{N} \right\} \quad (25)$$

where  $r(\cdot, \cdot)$  is as in (11).

*Example 3.1 Continued:* It follows from (20)–(22), and by setting  $\delta = (4D_1C)\varepsilon(N)$  that for neural networks the abstract estimator  $\hat{f}_N$  and the abstract index of resolvability  $R_N(f^*)$  may be written as in (9) and (12), respectively.

*Theorem 3.1:* Suppose Assumptions 2.2, 3.1, and 3.2 hold.

*Condition (m):* Suppose that  $\{X_i, Y_i\}_{i=-\infty}^\infty$  is  $m$ -dependent, then let  $\bar{N} = N^{(m)}$  and let  $\tilde{\alpha} = 1$ .

*Condition ( $\alpha$ ):* Suppose that  $\{X_i, Y_i\}_{i=-\infty}^\infty$  is strongly mixing and that Assumption 2.1 holds, then let  $\bar{N} = N^{(\alpha)}$  and let  $\tilde{\alpha} = (1 + 4e^{-2\bar{\alpha}})$ .

Then, either under Condition (m) or under Condition ( $\alpha$ ), and for  $\lambda > 5b^2/3$ ,  $\bar{N} \geq 2$ , for all natural numbers  $\kappa$ , and for all  $0 < \delta \leq \min_{1 \leq n \leq \kappa} \varpi_n(1)$

$$E[r(\hat{f}_N, f^*)] \leq \frac{1 + \eta}{1 - \eta} R_N(f^*) + \frac{6b\delta}{1 - \eta} + \frac{4\tilde{\alpha}\lambda}{(1 - \eta)\bar{N}} \quad (26)$$

where  $\eta = b^2/(\lambda - 2b^2/3)$ .

The proof can be found in Section III-C.

Theorem 3.1 has the same structure as the corresponding result in Barron [3] except for the additional term  $(6b\delta)/(1 - \eta)$ —which arises since we do not restrict the parameter space  $\Theta_\kappa$  to be countable.

The smaller the parameter  $\delta$ , generally speaking, the larger the complexity  $L(\cdot, \delta)$ , and hence larger the corresponding index of resolvability. Thus to obtain tighter bounds on the index, we should select  $\delta$  to be as large as possible. Specifically, although the choice  $\delta = O(1/\bar{N})$  is always available, for particular cases of interest a larger  $\delta$  may be viable, for example, the choice  $\delta \leq (4D_1C)/\sqrt{\bar{N}}$  works for neural networks.

The index of resolvability was first introduced by Barron and Cover [6] in the context of density estimation and universal data compression for i.i.d. observations, and later used by Barron [3] in the context of regression estimation also for i.i.d. observations. The significance of the index stems from Theorem 3.1, where we establish that the statistical risk of the minimum complexity regression estimator is bounded from above by the index. Thus the consistency of the estimator follows, if the index goes to zero as  $N \rightarrow \infty$ . Moreover, if the index tends to zero at a certain rate, then we can also conclude the same for the statistical risk of the estimator. The rate at which the index converges to zero depends on the tradeoff between the complexity of the functions in  $\{f_\theta: \theta \in \Theta_\infty\}$  and the accuracy of their approximation to  $f^*$ . Finally, the index of resolvability is a deterministic quantity, and hence is relatively easy to upper-bound in particular cases of interest, for example, see Lemma 2.2.

### C. Proof of Theorem 3.1

The proof relies heavily on the Bernstein inequality for strongly mixing processes established in Theorem 4.3 of Section IV and uses techniques of Barron [3] and McCaffrey and Gallant [21]. For simplicity, we assume throughout that Condition ( $\alpha$ ) holds; the result under Condition (m) follows similarly by using Theorem 4.2.

For any measurable function  $g_1: \mathbb{R}^d \rightarrow \mathbb{R}$ , define

$$\hat{r}_N(g_1, f^*) = \frac{1}{N} \sum_{i=1}^N [Y_i - g_1(X_i)]^2 - \frac{1}{N} \sum_{i=1}^N [Y_i - f^*(X_i)]^2. \quad (27)$$

*Lemma 3.1:* Suppose that Assumptions 2.2, 3.1, and 3.2 hold. Then, for all  $0 < \delta \leq \min_{1 \leq n \leq \kappa} \varpi_n(1)$ , for all  $\theta \in \Theta_\kappa(\delta)$ , for all  $\bar{N} \geq 2$ , for all  $\lambda > 5b^2/3$ , and for all  $\tilde{\delta} > 0$

$$P \left\{ (1 - \eta)r(f_\theta, f^*) \geq \hat{r}_N(f_\theta, f^*) + \lambda \frac{(L(\theta, \delta) + \ln 1/\tilde{\delta})}{\bar{N}} \right\} \leq \tilde{\alpha}\tilde{\delta}e^{-L(\theta, \delta)}.$$

*Proof:* For  $i = 1, 2, \dots, N$ , we write

$$U_i = -\{(Y_i - f_\theta(X_i))^2 - (Y_i - f^*(X_i))^2\} + r(f_\theta, f^*) \quad (28)$$

and observe that  $\{U_i\}_{i=1}^N$  are identically distributed,  $E[U_1] = 0$ ,  $|U_1| \leq 2b^2$ , and  $E|U_1|^2 \leq 2b^2r(f_\theta, f^*)$  [3]. It follows from (27) and (28) that

$$\frac{1}{N} \sum_{i=1}^N U_i = -\hat{r}_N(f_\theta, f^*) + r(f_\theta, f^*). \quad (29)$$

Since Condition ( $\alpha$ ) of Theorem 3.1 holds, we can now apply the Craig–Bernstein inequality in Theorem 4.3 to (29) with  $Z_i = \{X_i, Y_i\}$ ,  $d_1 = 2b^2$ ,  $3\zeta_1 = 1/\lambda$ , and  $\tau = L(\theta, \delta) + \ln 1/\tilde{\delta}$  to conclude that for  $\tilde{\delta} > 0$  and for  $\bar{N} = N^{(\alpha)} \geq 2$

we have

$$P \left\{ \begin{aligned} &r(f_\theta, f^*) - \hat{r}_N(f_\theta, f^*) \\ &\geq \lambda \frac{(L(\theta, \delta) + \ln 1/\delta)}{\bar{N}} + \frac{E|U_1|^2}{2\lambda \left(1 - \frac{2b^2}{3\lambda}\right)} \\ &\leq \tilde{\alpha}\tilde{\delta}e^{-L(\theta, \delta)}. \end{aligned} \right\}$$

Since  $E|U_1|^2 \leq 2b^2r(f_\theta, f^*)$ , the lemma now follows from Lemma A2, where we let  $\tilde{\alpha} = (1 + 4e^{-2\tilde{\alpha}})$  and  $\eta = b^2/(\lambda - 2b^2/3)$ . Also, observe that if  $\lambda > 5b^2/3$ , then  $\eta < 1$ .  $\square$

*Lemma 3.2:* Suppose that Assumptions 2.2, 3.1, and 3.2 hold. Then, for all  $0 < \delta \leq \min_{1 \leq n \leq \kappa} \varpi_n(1)$ , for all  $\bar{N} \geq 2$ , for all  $\tilde{\delta} > 0$ , and for all  $\lambda > 5b^2/3$

$$P \left\{ \begin{aligned} &(1 - \eta)r(\hat{f}_N, f^*) \geq \hat{r}_N(\hat{f}_N, f^*) \\ &+ \lambda \frac{(L(\hat{\theta}_N, \delta) + \ln 1/\tilde{\delta})}{\bar{N}} + 6b\tilde{\delta} \end{aligned} \right\} \leq \tilde{\alpha}\tilde{\delta}. \quad (30)$$

*Proof:* Since  $\hat{\theta}_N$  takes values in  $\Theta_\kappa$ , a possibly uncountable set, we first establish a result analogous to (30), namely (34), for a certain "projection" of  $\hat{\theta}_N$ , namely  $\check{\theta}_N$ , on a finite set. We then establish (30) from (34) using a certain continuity argument due to McCaffrey and Gallant [21].

Let  $S_n, T_n, \rho_n$ , and  $\varpi_n$  be as in Section III-A. For  $0 < \delta \leq \min_{1 \leq n \leq \kappa} \varpi_n(1)$ , define

$$\Xi_\kappa(\delta) = \bigcup_{n=1}^{\kappa} \{(n, \nu) : \nu \in T_n(\varpi_n^{-1}(\delta))\}. \quad (31)$$

Since the set  $T_n(\varpi_n^{-1}(\delta))$  is finite by assumption and since  $\kappa$  is a natural number, it follows that the set  $\Xi_\kappa(\delta)$  is finite. For every  $\varepsilon \in (0, 1]$  and for every  $\nu \in S_n$ , let  $\pi_n(\nu, \varepsilon)$  denote the element of the finite set  $T_n(\varepsilon)$  that is closest to  $\nu$  in the metric  $\rho_n$ ; formally, let  $\pi_n(\nu, \varepsilon)$  denote the lexicographically smallest element of the set

$$\{\nu_1 \in T_n(\varepsilon) : \rho_n(\nu, \nu_1) \leq \rho_n(\nu, \nu_2) \text{ for all } \nu_2 \in T_n(\varepsilon)\}.$$

Using (20), we may write  $\hat{\theta}_N$  in (23) as  $(\hat{n}, \hat{\nu})$ , where  $\hat{\nu} \in S_{\hat{n}}$ ; then define

$$\check{\theta}_N \equiv \check{\theta}_N(\delta) = (\hat{n}, \pi_{\hat{n}}(\hat{\nu}, \varpi_{\hat{n}}^{-1}(\delta))). \quad (32)$$

In words,  $\check{\theta}_N$  is a projection of  $\hat{\theta}_N \in \Theta_\kappa$  onto the finite set  $\Xi_\kappa$ . Write  $\check{f}_N = f_{\check{\theta}_N}$ . Since  $\hat{\theta}_N$  and  $\check{\theta}_N$  both have the same dimension  $\hat{n}$ , it follows from (22) that

$$L(\check{\theta}_N, \delta) = L(\hat{\theta}_N, \delta). \quad (33)$$

Write

$$A(f_\theta) = (1 - \eta)r(f_\theta, f^*) - \lambda \frac{(L(\theta, \delta) + \ln 1/\tilde{\delta})}{\bar{N}}.$$

Since the random variable  $\check{\theta}_N \in \Xi_\kappa(\delta)$ , we can write

$$\begin{aligned} &P\{A(\check{f}_N) \geq \hat{r}_N(\check{f}_N, f^*)\} \\ &= P \left\{ \bigcup_{\theta \in \Xi_\kappa(\delta)} (\{A(\check{f}_N) \geq \hat{r}_N(\check{f}_N, f^*)\} \cap \{\check{\theta}_N = \theta\}) \right\} \\ &\stackrel{(a)}{=} \sum_{\theta \in \Xi_\kappa(\delta)} P\{\{A(f_\theta) \geq \hat{r}_N(f_\theta, f^*)\} \cap \{\check{\theta}_N = \theta\}\} \\ &\leq \sum_{\theta \in \Xi_\kappa(\delta)} P\{A(f_\theta) \geq \hat{r}_N(f_\theta, f^*)\} \\ &\stackrel{(b)}{\leq} \tilde{\alpha}\tilde{\delta} \sum_{\theta \in \Xi_\kappa(\delta)} e^{-L(\theta, \delta)} \\ &\stackrel{(c)}{\leq} \tilde{\alpha}\tilde{\delta} \end{aligned}$$

where (a) follows since  $\check{f}_N = f_\theta$  on the set  $\{\check{\theta}_N = \theta\}$  and from the union of disjoint events bound; (b) follows from Lemma 3.1; and (c) follows from (22). In other words, we have shown that

$$P \left\{ \begin{aligned} &(1 - \eta)r(\check{f}_N, f^*) \geq \hat{r}_N(\check{f}_N, f^*) \\ &+ \lambda \frac{(L(\hat{\theta}_N, \delta) + \ln 1/\tilde{\delta})}{\bar{N}} \end{aligned} \right\} \leq \tilde{\alpha}\tilde{\delta} \quad (34)$$

where we have used (33).

Since  $\{Y_i\}_{i=1}^N$  are bounded (Assumption 2.2),  $\hat{f}_N$  and  $\check{f}_N$  are bounded (Assumption 3.1), and  $\hat{f}_N$  and  $\check{f}_N$  are close in the supremum norm (Assumption 3.2), we can show (after some algebraic manipulations similar to [21, Lemma 2]) that

$$\begin{aligned} &P\{(1 - \eta)r(\hat{f}_N, f^*) - \hat{r}_N(\hat{f}_N, f^*) \\ &\geq (1 - \eta)r(\check{f}_N, f^*) - \hat{r}_N(\check{f}_N, f^*) + 6b\tilde{\delta}\} = 0. \end{aligned} \quad (35)$$

The lemma now follows by combining (34) and (35) using Lemma A5.  $\square$

Define

$$\theta_N^* = \arg \min_{\theta \in \Theta_\kappa} \left\{ r(f_\theta, f^*) + \lambda \frac{L(\theta, \delta)}{\bar{N}} \right\} \quad (36)$$

where  $f_\theta$  is as in (21),  $r(\cdot, \cdot)$  is as in (11),  $L(\theta, \delta)$  is as in (22), and  $\lambda$  and  $\bar{N}$  are as in Theorem 3.1. Write  $f_N^* = f_{\theta_N^*}$ .

*Lemma 3.3:* Suppose all hypotheses of Lemma 3.2 hold. Then

$$P \left\{ \begin{aligned} &(1 - \eta)r(\hat{f}_N, f^*) \geq \hat{r}_N(\hat{f}_N, f^*) \\ &+ \lambda \frac{(L(\theta_N^*, \delta) + \ln 1/\tilde{\delta})}{\bar{N}} + 6b\tilde{\delta} \end{aligned} \right\} \leq \tilde{\alpha}\tilde{\delta}.$$

*Proof:* Note that  $\hat{\theta}_N$  minimizes (23), and that  $\hat{f}_N = f_{\hat{\theta}_N}$  and  $f_N^* = f_{\theta_N^*}$ . Thus it follows that

$$\begin{aligned} & \left\{ \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{f}_N(X_i))^2 + \lambda \frac{L(\hat{\theta}_N, \delta)}{N} \right\} \\ & \leq \left\{ \frac{1}{N} \sum_{i=1}^N (Y_i - f_N^*(X_i))^2 + \lambda \frac{L(\theta_N^*, \delta)}{N} \right\} \end{aligned}$$

and hence the lemma follows from (27), Lemma 3.2, and Lemma A2.  $\square$

*Lemma 3.4:* Suppose that Assumptions 2.2 and 3.1 hold. Then, for all  $N \geq 2$  and for all  $\tilde{\delta} > 0$

$$P \left\{ \hat{r}_N(f_N^*, f^*) \geq (1 + \eta)r(f_N^*, f^*) + \lambda \frac{\ln 1/\tilde{\delta}}{N} \right\} \leq \tilde{\alpha}\tilde{\delta}.$$

*Proof:* The lemma follows by applying the Craig-Bernstein inequality in Theorem 4.3 to the sum

$$\begin{aligned} & \frac{1}{N} \sum_{i=1}^N \{(Y_i - f_N^*(X_i))^2 - (Y_i - f^*(X_i))^2\} - r(f_N^*, f^*) \\ & = \hat{r}_N(f_N^*, f^*) - r(f_N^*, f^*) \end{aligned}$$

with  $Z_i = \{X_i, Y_i\}$ ,  $d_1 = 2b^2$ ,  $3\zeta_1 = 1/\lambda$ , and  $\tau = \ln 1/\tilde{\delta}$  and by simplifying as in Lemma 3.1.  $\square$

Combining Lemmas 3.3 and 3.4 using Lemma A5 and ignoring the term  $-\eta\lambda L(\theta_N^*, \delta)/N$  using Lemma A2, we have that

$$\begin{aligned} P \left\{ (1 - \eta)r(\hat{f}_N, f^*) \geq (1 + \eta) \left( r(f_N^*, f^*) + \lambda \frac{L(\theta_N^*, \delta)}{N} \right) \right. \\ \left. + \lambda \frac{2 \ln 1/\tilde{\delta}}{N} + 6b\delta \right\} \leq 2\tilde{\alpha}\tilde{\delta}. \end{aligned}$$

It now follows from (25) and (36) and by setting

$$\tilde{\delta} = \exp[-(Nt/2\lambda)] \text{ for } t > 0$$

that

$$P\{W \geq t\} \leq 2\tilde{\alpha} \exp \left\{ -\frac{Nt}{2\lambda} \right\} \quad (37)$$

where

$$W = (1 - \eta)r(\hat{f}_N, f^*) - (1 + \eta)r(f_N^*, f^*) - 6b\delta.$$

It is easy to see that  $|W| < \infty$  and hence  $E|W| < \infty$ . Equation (26) now follows from (37) and Lemma A6. The proof of Theorem 3.1 is now complete.  $\square$

#### IV. BERNSTEIN INEQUALITIES FOR DEPENDENT PROCESSES

In this section, we extend the classical Bernstein inequality for i.i.d. random variables to  $m$ -dependent processes and to strongly mixing processes. The extended inequalities are used, in this paper, in the proof of Theorem 3.1 of Section III; they may also be of independent interest.

*Theorem 4.1 i.i.d. Random Variables:* Let an integer  $N \geq 1$  be given. Let  $\{U_i\}_{i=1}^N$  be i.i.d. random variables on the probability space  $(\Omega, \mathcal{F}, P)$  such that  $|U_1| \leq d_1$  a.s. and  $E[U_1] = 0$ . Then, the following probability inequalities hold.

a) *Craig-Bernstein Inequality* [9]: For all  $\tau \in \mathbb{R}$  and for all  $0 < \zeta_1 < 1/d_1$

$$P \left\{ \frac{1}{N} \sum_{i=1}^N U_i \geq \frac{\tau}{3\zeta_1 N} + \frac{3\zeta_1 E|U_1|^2}{2(1 - \zeta_1 d_1)} \right\} \leq e^{-\tau}.$$

b) *Bernstein Inequality* [28, p. 855]: For all  $\zeta_2 > 0$

$$P \left\{ \frac{1}{N} \sum_{i=1}^N U_i \geq \zeta_2 \right\} \leq \exp \left[ -\frac{\zeta_2^2 N}{2 \left( E|U_1|^2 + \frac{\zeta_2 d_1}{3} \right)} \right].$$

*Theorem 4.2  $m$ -Dependent Processes:* For a given integer  $m \geq 0$ , let  $\{Z_i\}_{i=-\infty}^{\infty}$  be a stationary  $m$ -dependent [14] process on the probability space  $(\Omega, \mathcal{F}, P)$ . Let an integer  $N \geq 1$  be given. Let  $\psi: \mathbb{R} \rightarrow \mathbb{R}$  be some measurable function. For each integer  $-\infty < i < \infty$ , let  $U_i = \psi(Z_i)$ . Assume that  $|U_1| \leq d_1$  a.s. and that  $E[U_1] = 0$ . Set  $N^{(m)} = \lfloor N/(m+1) \rfloor$ . Then, for all  $N \geq (m+1)$ , the following probability inequalities hold.

a) *Craig-Bernstein Inequality:* For all  $\tau \in \mathbb{R}$  and for all  $0 < \zeta_1 < 1/d_1$

$$P \left\{ \frac{1}{N} \sum_{i=1}^N U_i \geq \frac{\tau}{3\zeta_1 N^{(m)}} + \frac{3\zeta_1 E|U_1|^2}{2(1 - \zeta_1 d_1)} \right\} \leq e^{-\tau}.$$

b) *Bernstein Inequality:* For all  $\zeta_2 > 0$

$$P \left\{ \frac{1}{N} \sum_{i=1}^N U_i \geq \zeta_2 \right\} \leq \exp \left[ -\frac{\zeta_2^2 N^{(m)}}{2 \left( E|U_1|^2 + \frac{\zeta_2 d_1}{3} \right)} \right].$$

*Proof:* The proof is similar to the proof of Theorem 4.3 (see below) and is omitted.  $\square$

*Remark 4.1:* Observe that by setting  $m = 0$  in Theorem 4.2 we recover Theorem 4.1.

*Theorem 4.3 Strongly Mixing Processes:* Let  $\{Z_i\}_{i=-\infty}^{\infty}$  be a stationary strongly mixing [27] process on the probability space  $(\Omega, \mathcal{F}, P)$  with the mixing coefficient satisfying Assumption 2.1, that is,

$$\alpha(j) \leq \bar{\alpha} \exp(-cj^\beta), \quad j \geq 1, \bar{\alpha} > 0, \beta > 0, c > 0. \quad (38)$$

Let an integer  $N \geq 1$  be given. For each integer  $-\infty < i < \infty$ , let  $U_i = \psi(Z_i)$ , where  $\psi$  is some real-valued Borel measurable function. Assume that  $|U_1| \leq d_1$  a.s. and that  $E[U_1] = 0$ . Set

$$N^{(\alpha)} = \left\lfloor N \left[ \left\{ \frac{8N}{c} \right\}^{1/(\beta+1)} \right]^{-1} \right\rfloor. \quad (39)$$

Then, the following probability inequalities hold.

- a) *Craig-Bernstein Inequality*: For all  $N^{(\alpha)} \geq 2$ , for all  $\tau \in \mathbb{R}$ , and for all  $0 < \zeta_1 < 1/d_1$

$$P \left\{ \frac{1}{N} \sum_{i=1}^N U_i \geq \frac{\tau}{3\zeta_1 N^{(\alpha)}} + \frac{3\zeta_1 E|U_1|^2}{2(1-\zeta_1 d_1)} \right\} \leq (1 + 4e^{-2\bar{\alpha}}) e^{-\tau}. \quad (40)$$

- b) *Bernstein Inequality*: For all  $N^{(\alpha)} \geq 2$  and for all  $\zeta_2 > 0$

$$P \left\{ \frac{1}{N} \sum_{i=1}^N U_i \geq \zeta_2 \right\} \leq (1 + 4e^{-2\bar{\alpha}}) \exp \left[ -\frac{\zeta_2^2 N^{(\alpha)}}{2(E|U_1|^2 + \frac{\zeta_2 d_1}{3})} \right]. \quad (41)$$

*Remark 4.2*: Observe that by setting  $\beta = \infty$  in Theorem 4.3 and by ignoring the multiplicative constant  $(1 + 4e^{-2\bar{\alpha}})$  we recover Theorem 4.1.

*Proof*: Write

$$V_N = \sum_{i=1}^N U_i.$$

We now proceed by the method of blocks, and partition the set  $\{1, 2, \dots, N\}$  into  $k_N$  blocks. Each block will contain approximately  $l_N = \lfloor N/k_N \rfloor$  terms. Let  $h_N = (N - k_N l_N) < k_N$  denote the remainder when we divide  $N$  by  $k_N$ . For simplicity of notation, we will write  $k = k_N$ ,  $l = l_N$ , and  $h = h_N$ .

We now construct  $k$  blocks as follows. Define  $\bar{l}_j$ , the number of terms in the  $j$ th block, as

$$\bar{l}_j = \begin{cases} l+1, & \text{if } j = 1, 2, \dots, h \\ l, & \text{if } j = h+1, h+2, \dots, k. \end{cases}$$

In other words, the first  $h$  blocks each contain  $l+1$  terms, while the last  $(k-h)$  blocks each contain  $l$  terms. Then

$$\sum_{j=1}^k \bar{l}_j = \sum_{j=1}^h \bar{l}_j + \sum_{j=h+1}^k \bar{l}_j = h(l+1) + (k-h)l = N. \quad (42)$$

For  $j = 1, 2, \dots, k$ , we define the  $j$ th block as

$$V_{j,N} = U_j + U_{j+k} + \dots + U_{j+(\bar{l}_j-1)k} = \sum_{i=1}^{\bar{l}_j} U_{j+(i-1)k}$$

such that

$$V_N = \sum_{j=1}^k V_{j,N}.$$

A typical block  $V_{j,N}$  contains  $\bar{l}_j$  terms such that any of its two consecutive terms are separated by distance  $k$ .

For  $j = 1, 2, \dots, k$ , define  $p_j = \bar{l}_j/N$ . It follows from (42) that

$$\sum_{j=1}^k p_j = (1/N) \sum_{j=1}^k \bar{l}_j = 1.$$

Also, for notational convenience we define  $U_{j+(i-1)k} = U_j^{(i)}$ . We now proceed with a series of lemmas.

*Lemma 4.1 Hoeffding [16]*: For all  $\gamma \in \mathbb{R}$

$$E \left[ \exp \left( \gamma \frac{V_N}{N} \right) \right] \leq \sum_{j=1}^k p_j E \left[ \exp \left( \gamma \frac{V_{j,N}}{\bar{l}_j} \right) \right].$$

The following lemma holds for stationary strongly mixing processes.

*Lemma 4.2*: Suppose that (38) holds and that  $|U_1| \leq d_1$  a.s., then for all real numbers  $\gamma > 0$ , for all integers  $q \geq 2$ , for all integers  $q' \geq 1$ , and for all  $j = 1, 2, \dots, k$

$$A_q^{q'}(\gamma) \equiv \left| E \left[ \prod_{i=1}^q \exp \left( \frac{\gamma U_j^{(i)}}{q'} \right) \right] - \prod_{i=1}^q E \left[ \exp \left( \frac{\gamma U_j^{(i)}}{q'} \right) \right] \right| \leq (4e^{-2\bar{\alpha}}) \exp \left( q + \frac{\gamma q d_1}{q'} - ck^\beta \right).$$

*Proof*: In this proof, we use an argument similar to that used by Bosq [7] in his proof of the Bernstein inequality for uniformly mixing processes. However, we use a completely different blocking scheme.

$$\begin{aligned} A_q^{q'}(\gamma) &\stackrel{(a)}{\leq} \left| E \left[ \prod_{i=1}^{q-1} \exp \left( \frac{\gamma U_j^{(i)}}{q'} \right) \exp \left( \frac{\gamma U_j^{(q)}}{q'} \right) \right] \right. \\ &\quad \left. - E \left[ \prod_{i=1}^{q-1} \exp \left( \frac{\gamma U_j^{(i)}}{q'} \right) \right] E \left[ \exp \left( \frac{\gamma U_j^{(q)}}{q'} \right) \right] \right| \\ &\quad + \left| E \left[ \prod_{i=1}^{q-1} \exp \left( \frac{\gamma U_j^{(i)}}{q'} \right) \right] - \prod_{i=1}^{q-1} E \left[ \exp \left( \frac{\gamma U_j^{(i)}}{q'} \right) \right] \right| \\ &\quad \cdot \left| E \left[ \exp \left( \frac{\gamma U_j^{(q)}}{q'} \right) \right] \right| \\ &\equiv B_q^{q'}(\gamma) + A_{q-1}^{q'}(\gamma) \left| E \left[ \exp \left( \frac{\gamma U_j^{(q)}}{q'} \right) \right] \right| \\ &\stackrel{(b)}{\leq} 4 \left\| \prod_{i=1}^{q-1} \exp \left( \frac{\gamma U_j^{(i)}}{q'} \right) \right\|_{\infty, P} \left\| \exp \left( \frac{\gamma U_j^{(q)}}{q'} \right) \right\|_{\infty, P} \\ &\quad \cdot \alpha(k) + A_{q-1}^{q'}(\gamma) \left\| \exp \left( \frac{\gamma U_j^{(q)}}{q'} \right) \right\|_{\infty, P} \\ &\leq 4e^{\gamma(q-1)d_1/q'} e^{\gamma d_1/q'} \alpha(k) + A_{q-1}^{q'}(\gamma) e^{\gamma d_1/q'} \\ &= 4e^{\gamma q d_1/q'} \alpha(k) + A_{q-1}^{q'}(\gamma) e^{\gamma d_1/q'} \\ &\stackrel{(c)}{\leq} 4(q-1)e^{\gamma q d_1/q'} \alpha(k) \\ &\stackrel{(d)}{\leq} 4e^{q-2} e^{\gamma q d_1/q'} \alpha(k) \\ &\leq (4e^{-2\bar{\alpha}}) e^q e^{\gamma q d_1/q'} e^{-ck^\beta} \end{aligned}$$

where

- (a) follows from adding and subtracting terms,

(b)  $\prod_{i=1}^{q-1} \exp(\gamma U_j^{(i)}/q')$

is measurable with respect to  $\sigma\{U_{j+(i-1)k}, i = 1, 2, \dots, q-1\}$  and  $\exp(\gamma U_j^{(q)}/q')$  is measurable with

respect to  $\sigma\{U_{j+(q-1)k}\}$ . For each  $i'$ , we have  $U_{i'} = \psi(Z_{i'})$ , hence

$$\begin{aligned} & \sigma\{U_{j+(i-1)k}, i = 1, 2, \dots, q-1\} \\ & \subset \sigma\{Z_{j+(i-1)k}, i = 1, 2, \dots, q-1\} \\ & \sigma\{U_{j+(q-1)k}\} \\ & \subset \sigma\{Z_{j+(q-1)k}\}. \end{aligned}$$

Observe that the distance between the two  $\sigma$ -algebras above is  $(j+(q-1)k) - (j+(q-2)k) = k$ , hence (b) now follows by bounding the covariance  $B_q^{q'}(\gamma)$  using a mixing inequality in Hall and Heyde [15, Theorem A.5] for the process  $\{Z_i\}_{i=-\infty}^{\infty}$ . Also, note that  $\|\cdot\|_{\infty, P}$  denotes the usual essential supremum on  $(\Omega, \mathcal{F}, P)$ .

(c) Let  $u = e^{\gamma d_1/q'}$  and  $v_q = 4u^q \alpha(k)$ , then we have

$$A_q^{q'}(\gamma) \leq 4u^q \alpha(k) + u A_{q-1}^{q'}(\gamma) = v_q + u A_{q-1}^{q'}(\gamma).$$

For  $q = 2$

$$\begin{aligned} A_2^{q'}(\gamma) &= B_2^{q'}(\gamma) \\ &= |E[e^{\gamma U_j^{(1)}/q'} e^{\gamma U_j^{(2)}/q'}] \\ &\quad - E[e^{\gamma U_j^{(1)}/q'}] E[e^{\gamma U_j^{(2)}/q'}]|. \end{aligned}$$

Now, by proceeding as in step (b) above, we have

$$\begin{aligned} A_2^{q'}(\gamma) &\leq 4 \|e^{\gamma U_j^{(1)}/q'}\|_{\infty, P} \|e^{\gamma U_j^{(2)}/q'}\|_{\infty, P} \alpha(k) \\ &= 4e^{\gamma d_1 2/q'} \alpha(k) = v_2. \end{aligned}$$

It now follows from direct substitution that for  $q \geq 2$

$$\begin{aligned} A_q^{q'}(\gamma) &\leq \sum_{j=0}^{q-2} u^j v_{q-j} = \sum_{j=0}^{q-2} 4u^q \alpha(k) \\ &= 4(q-1)u^q \alpha(k). \end{aligned}$$

(d) follows since  $q \geq 2$ , and since for all  $x \geq 0$  we have that  $\ln x \leq x - 1$ .  $\square$

For all  $j = 1, 2, \dots, k$ , we now establish a bound, uniform over all indices  $j$ , on the moment generating function of  $V_{j,N}/\bar{l}_j$ .

**Lemma 4.3:** Suppose that (38) holds and that  $|U_1| \leq d_1$  a.s. and  $E[U_1] = 0$ . Let  $\bar{l}_N$  be as in (39). Then, for all  $\bar{l}_N \geq 2$ , for all  $0 < \gamma < (3l)/d_1$ , and for all  $j = 1, 2, \dots, k$

$$E \left[ \exp \left( \frac{\gamma V_{j,N}}{\bar{l}_j} \right) \right] < (1 + 4e^{-2\bar{\alpha}}) \exp \left[ \frac{\gamma^2 E|U_1|^2}{2l \left( 1 - \frac{\gamma d_1}{3l} \right)} \right].$$

*Proof:* For  $j = 1, 2, \dots, k$ , we have

$$\begin{aligned} & E \left[ \exp \left( \frac{\gamma V_{j,N}}{\bar{l}_j} \right) \right] \\ &= E \left[ \exp \left( \sum_{i=1}^{\bar{l}_j} \frac{\gamma U_{j+(i-1)k}}{\bar{l}_j} \right) \right] \\ &= E \left[ \prod_{i=1}^{\bar{l}_j} \exp \left( \frac{\gamma U_j^{(i)}}{\bar{l}_j} \right) \right] \end{aligned}$$

$$\begin{aligned} & \leq \prod_{i=1}^{\bar{l}_j} E \left[ \exp \left( \frac{\gamma U_j^{(i)}}{\bar{l}_j} \right) \right] + \left| E \left[ \prod_{i=1}^{\bar{l}_j} \exp \left( \frac{\gamma U_j^{(i)}}{\bar{l}_j} \right) \right] \right. \\ & \quad \left. - \prod_{i=1}^{\bar{l}_j} E \left[ \exp \left( \frac{\gamma U_j^{(i)}}{\bar{l}_j} \right) \right] \right| \end{aligned}$$

$$\stackrel{(a)}{=} \left\{ E \left[ \exp \left( \frac{\gamma U_j^{(1)}}{\bar{l}_j} \right) \right] \right\}^{\bar{l}_j} + A_{\bar{l}_j}^{\bar{l}_j}(\gamma)$$

$$\stackrel{(b)}{\leq} \exp \left[ \frac{\gamma^2 \bar{l}_j E \left| \frac{U_j^{(1)}}{\bar{l}_j} \right|^2}{2 \left( 1 - \frac{\gamma d_1}{3\bar{l}_j} \right)} \right] + (4e^{-2\bar{\alpha}}) \exp(\bar{l}_j + \gamma d_1 - ck^\beta)$$

$$\stackrel{(c)}{<} \exp \left[ \frac{\gamma^2 E|U_1|^2}{2\bar{l}_j \left( 1 - \frac{\gamma d_1}{3\bar{l}_j} \right)} \right] + (4e^{-2\bar{\alpha}}) \exp(4\bar{l}_j - ck^\beta)$$

$$\stackrel{(d)}{\leq} \exp \left[ \frac{\gamma^2 E|U_1|^2}{2l \left( 1 - \frac{\gamma d_1}{3l} \right)} \right] + (4e^{-2\bar{\alpha}}) \exp(4\bar{l}_1 - ck^\beta)$$

$$\stackrel{(e)}{\leq} \exp \left[ \frac{\gamma^2 E|U_1|^2}{2l \left( 1 - \frac{\gamma d_1}{3l} \right)} \right] + (4e^{-2\bar{\alpha}})$$

$$\stackrel{(f)}{<} (1 + 4e^{-2\bar{\alpha}}) \exp \left[ \frac{\gamma^2 E|U_1|^2}{2l \left( 1 - \frac{\gamma d_1}{3l} \right)} \right] \quad (43)$$

where

(a) follows from stationarity.

(b)  $E[U_j^{(1)}] = 0$ , and  $U_j^{(1)}/\bar{l}_j$  satisfies the Bernstein moment condition with  $K_1 = d_1/(3\bar{l}_j)$ . Hence, the result follows from Lemma A1 for all  $0 < \gamma < (3\bar{l}_j)/d_1$ . The bound on the second term follows from Lemma 4.2, and holds for all  $\bar{l}_j \geq 2$ . But  $\bar{l}_j \geq l$ . Thus it suffices that  $l \geq 2$ .

(c) Since

$$E|U_j^{(1)}/\bar{l}_j|^2 = (E|U_j|^2)/\bar{l}_j^2$$

and by stationarity  $E|U_j|^2 = E|U_1|^2$ . Also, since  $\gamma d_1 < 3\bar{l}_j$ .

(d) For all  $j = 1, 2, \dots, k$ ,  $\bar{l}_j \leq l$ . Thus we have

$$(1 - \gamma d_1/3\bar{l}_j) \geq (1 - \gamma d_1/3l).$$

(e) We require  $\exp(4\bar{l}_1 - ck^\beta) \leq 1$ , which holds if  $4\bar{l}_1 \leq ck^\beta$ . But  $\bar{l}_1 \leq ((N/k) + 1)$ . Thus the bound holds if  $4((N/k) + 1) \leq ck^\beta$ , or if  $4(N+k) \leq ck^{\beta+1}$ . Since  $(N+k) \leq 2N$ , the bound holds if  $8N \leq ck^{\beta+1}$ , or if

$\{8N/c\}^{1/(\beta+1)} \leq k$ . Select

$$k = \left\lceil \left\{ \frac{8N}{c} \right\}^{1/(\beta+1)} \right\rceil. \quad (44)$$

Since  $l = l_N = \lfloor N/k \rfloor$ , we have (39)—where we have written  $N^{(\alpha)}$  for  $l = l_N$ .

- (f) Equation (43) holds for all  $\gamma$  such that  $0 < \gamma < 3\bar{l}_j/d_1$ . To make the constraint uniform over all  $j$ , we require that  $\gamma$  satisfy  $0 < \gamma < 3l/d_1 \leq 3\bar{l}_j/d_1$ . Since  $\gamma^2 E|U_1|^2 / (2l(1-\gamma d_1/(3l))) > 0$ , the lemma follows.  $\square$

Let  $l = N^{(\alpha)}$  be as in (39). Then, combining Lemmas 4.1 and 4.3, we have for all  $0 < \gamma < (3l)/d_1$  and for all  $l_N \geq 2$

$$E \left[ \exp \left( \frac{\gamma V_N}{N} \right) \right] < (1 + 4e^{-2\bar{\alpha}}) \exp \left[ \frac{\gamma^2 E|U_1|^2}{2l \left( 1 - \frac{\gamma d_1}{3l} \right)} \right]. \quad (45)$$

We are now ready to establish the Craig–Bernstein and the Bernstein inequalities (40) and (41).

- a) *Craig–Bernstein Inequality*: For all  $\gamma$  and  $\tau$ , we apply Lemma A3 with  $W = \exp(\gamma V_N/N)$  and with  $\tau' = (\tau - \ln(1 + 4e^{-2\bar{\alpha}})) \in \mathbb{R}$ , to conclude

$$P \left\{ \exp \left( \frac{\gamma V_N}{N} \right) \geq e^{\tau - \ln(1 + 4e^{-2\bar{\alpha}})} E \left[ \exp \left( \frac{\gamma V_N}{N} \right) \right] \right\} \leq e^{-\tau + \ln(1 + 4e^{-2\bar{\alpha}})}.$$

Now, for all  $0 < \gamma < (3l)/d_1$  and for all  $l_N \geq 2$ , it follows from (45) and from Lemma A2 that

$$P \left\{ \exp \left( \frac{\gamma V_N}{N} \right) \geq e^{\tau - \ln(1 + 4e^{-2\bar{\alpha}})} (1 + 4e^{-2\bar{\alpha}}) \exp \left[ \frac{\gamma^2 E|U_1|^2}{2l \left( 1 - \frac{\gamma d_1}{3l} \right)} \right] \right\} \leq (1 + 4e^{-2\bar{\alpha}}) e^{-\tau}.$$

Since logarithm is a strictly increasing function and  $\gamma > 0$ , we have

$$P \left\{ \frac{1}{N} V_N \geq \frac{\tau}{\gamma} + \frac{\gamma E|U_1|^2}{2l \left( 1 - \frac{\gamma d_1}{3l} \right)} \right\} \leq (1 + 4e^{-2\bar{\alpha}}) e^{-\tau}.$$

Now, we set  $\gamma = 3\zeta_1 l$ . Thus for  $0 < \zeta_1 < 1/d_1$ , (40) follows—where we have written  $N^{(\alpha)}$  for  $l = l_N$ .

- b) *Bernstein Inequality*: For all  $\zeta_2 > 0$ , and for all  $\gamma > 0$ , we have from Lemma A4 that

$$P \left\{ \frac{1}{N} V_N \geq \zeta_2 \right\} \leq e^{-\zeta_2 \gamma} E \left[ \exp \left( \frac{\gamma V_N}{N} \right) \right].$$

Now, for all  $0 < \gamma < (3l)/d_1$  and for all  $l_N \geq 2$ , we have from (45) that

$$P \left\{ \frac{1}{N} V_N \geq \zeta_2 \right\} \leq (1 + 4e^{-2\bar{\alpha}}) e^{-\zeta_2 \gamma} \exp \left[ \frac{\gamma^2 E|U_1|^2}{2l \left( 1 - \frac{\gamma d_1}{3l} \right)} \right].$$

Now, by substituting

$$\gamma = \frac{\zeta_2 l}{\left( E|U_1|^2 + \frac{\zeta_2 d_1}{3} \right)}$$

simplifying, and noting that the selected value for  $\gamma$  satisfies  $\gamma < (3l)/d_1$ , (41) follows—where we have written  $N^{(\alpha)}$  for  $l = l_N$ .

The proof of Theorem 4.3 is now complete.  $\square$

#### APPENDIX

Here, we state some useful, but simple, results without proofs.

*Lemma A1 Craig [9]*: Let  $W$  be a random variable such that  $E[W] = 0$ , and  $W$  satisfies the *Bernstein moment condition*, that is, for some  $K_1 > 0$

$$E|W|^k \leq \frac{\text{var}(W)}{2} k! K_1^{k-2}$$

for all  $k \geq 2$ . Then, for all  $0 < \zeta < 1/K_1$

$$E[\exp(\zeta W)] \leq \exp \left[ \frac{\zeta^2 E|W|^2}{2(1 - \zeta K_1)} \right].$$

*Remark A1*: If  $|W| \leq 3K_1$  a.s., then the Bernstein moment condition holds [28, p. 855].

*Lemma A2*: Let  $W$  be a random variable and let  $u_1, u_2, K_1 \in \mathbb{R}$  be such that  $u_1 \leq u_2$ , then

$$P\{W \geq u_1\} \leq K_1 \Rightarrow P\{W \geq u_2\} \leq K_1.$$

*Lemma A3*: Let  $W$  be a nonnegative random variable and let  $\tau' \in \mathbb{R}$ , then

$$P\{W \geq e^{\tau'} E[W]\} \leq e^{-\tau'}.$$

*Lemma A4*: Let  $W$  be a random variable and let  $u, t > 0$ , then

$$P\{W \geq u\} \leq e^{-ut} E[e^{tW}].$$

*Lemma A5*: Let  $\{W_i\}_{i=1}^q$  be random variables and let  $\{u_i\}_{i=1}^q, \{K_i\}_{i=1}^q$  be constants. If for each  $i = 1, 2, \dots, q$ ,  $P\{W_i \geq u_i\} \leq K_i$ , then

$$P \left\{ \sum_{i=1}^q W_i \geq \sum_{i=1}^q u_i \right\} \leq \sum_{i=1}^q K_i.$$

*Lemma A6 Shorack and Wellner [28, p. 862]*: Let  $W$  be a random variable such that  $E|W| < \infty$ , then

$$E[W] \leq \int_0^\infty P\{W \geq u\} du.$$

## ACKNOWLEDGMENT

The authors wish to thank three anonymous referees for their careful reading of the manuscript and for their valuable suggestions. The authors are especially grateful to the Associate Editor, Andrew Barron, for his numerous constructive suggestions which expanded the scope and significantly improved the presentation of this paper.

## REFERENCES

- [1] D. W. K. Andrews and D. Pollard, "An introduction to functional central limit theorems for dependent stochastic processes," *Int. Statist. Rev.*, vol. 62, no. 1, pp. 119–132, 1994.
- [2] M. A. Arcones and B. Yu, "Central limit theorems for empirical and  $U$ -processes of stationary mixing sequences," *J. Theor. Probab.*, vol. 7, no. 1, 1994.
- [3] A. R. Barron, "Complexity regularization," in *Proc. NATO Advanced Study Institute on Nonparametric Functional Estimation*, G. Roussas, Ed. Dordrecht, The Netherlands: Kluwer, 1991, pp. 561–576.
- [4] ———, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 930–945, May 1993.
- [5] ———, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, pp. 115–133, 1994.
- [6] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, no. 4, pp. 1034–1054, July 1991.
- [7] D. Bosq, "Inégalité de Bernstein pour les processus stationnaires et mélangeants. Applications," *C. R. Seances de l'Acad. des Sci. Paris*, ser. A, vol. 281, pp. 1095–1098, 1975.
- [8] M. Carbon, "Inégalité de Bernstein pour les processus fortement mélangeants, non nécessairement stationnaires. Applications," *C. R. Seances de l'Acad. des Sci. Paris*, ser. I, vol. 297, pp. 303–306, 1983.
- [9] C. C. Craig, "On the Tchebycheff inequality of Bernstein," *Ann. Math. Statist.*, vol. 4, pp. 94–102, 1933.
- [10] Y. A. Davydov, "Mixing conditions for Markov chains," *Theory of Probability and Its Applications*, vol. XVIII, no. 2, pp. 312–328, 1973.
- [11] P. Doukhan, *Mixing: Properties and Examples*. New York: Springer-Verlag, 1994.
- [12] P. Doukhan, P. Massart, and E. Rio, "Invariance principles for absolutely regular empirical processes," *Ann. Inst. Henri Poincaré, Probab. Statist.*, vol. 31, no. 2, pp. 393–427, 1995.
- [13] A. Farago and G. Lugosi, "Strong universal consistency of neural network classifiers," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1146–1151, July 1993.
- [14] W. A. Fuller, *Introduction to Statistical Time Series*. New York: Wiley, 1976.
- [15] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Application*. New York: Academic Press, 1980.
- [16] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, pp. 13–30, 1963.
- [17] K. Hornik, M. B. Stinchcombe, H. White, and P. Auer, "Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives," *Neural Comput.*, vol. 6, pp. 1262–1275, 1994.
- [18] G. Lugosi and K. Zeger, "Nonparametric estimation via empirical risk minimization," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 677–687, May 1995.
- [19] ———, "Concept learning using complexity regularization," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 48–54, Jan. 1996.
- [20] E. Masry, "Strong consistency and rates for deconvolution of multivariate densities of stationary processes," *Stoch. Process. Appl.*, vol. 47, pp. 53–74, 1993.
- [21] D. F. McCaffrey and A. R. Gallant, "Convergence rates for single hidden layer feedforward networks," *Neural Networks*, vol. 7, no. 1, pp. 147–158, 1994.
- [22] D. S. Modha, "Universal prediction of stationary random processes," Ph.D. dissertation, Dept. Elec. Comput. Eng., Univ. California at San Diego, 1995.
- [23] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984.
- [24] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. Teaneck, NJ: World Scientific, 1989.
- [25] P. M. Robinson, "Nonparametric estimators for time series," *J. Time Series Anal.*, vol. 4, pp. 185–297, 1983.
- [26] G. G. Roussas, "Nonparametric regression estimation under mixing conditions," *Stoch. Process. Appl.*, vol. 36, pp. 107–116, 1990.
- [27] M. Rosenblatt, "A central limit theorem and strong mixing conditions," *Proc. Nat. Acad. Sci.*, vol. 4, pp. 43–47, 1956.
- [28] G. R. Shorack and J. A. Wellner, *Empirical Processes with Applications to Statistics*. New York: Wiley, 1986.
- [29] V. Vapnik, *Estimation of Dependences Based on Empirical Data*. New York: Springer-Verlag, 1982.
- [30] H. White, "Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings," *Neural Networks*, vol. 3, pp. 535–549, 1989.
- [31] H. White and J. M. Wooldridge, "Some results on sieve estimation with dependent observations," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proc. 5th Int. Symp. on Economic Theory and Econometrics*, W. A. Barnett, J. Powell, and G. Tauchen, Eds. New York: Cambridge Univ. Press, 1991.
- [32] C. S. Withers, "Conditions for linear processes to be strong-mixing," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 57, pp. 477–480, 1981.
- [33] B. Yu, "Rates of convergence for empirical processes of stationary mixing sequences," *Ann. Probab.*, vol. 22, no. 1, pp. 94–116, 1994.
- [34] J. E. Yukich, M. B. Stinchcombe, and H. White, "Sup-norm approximation bounds for networks through probabilistic methods," *IEEE Trans. Inform. Theory*, vol. 41, no. 4, pp. 1021–1027, 1995.