

Memory-Universal Prediction of Stationary Random Processes

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

Abstract— We consider the problem of one-step-ahead prediction of a real-valued, stationary, strongly mixing random process $\{X_i\}_{i=-\infty}^{\infty}$. The best mean-square predictor of X_0 is its conditional mean given the entire infinite past $\{X_i\}_{i=-\infty}^{-1}$. Given a sequence of observations X_1, X_2, \dots, X_N , we propose estimators for the conditional mean based on sequences of parametric models of increasing memory and of increasing dimension, for example, neural networks and Legendre polynomials. The proposed estimators select both the model memory and the model dimension, in a data-driven fashion, by minimizing certain complexity regularized least squares criteria. When the underlying predictor function has a finite memory, we establish that the proposed estimators are *memory-universal*: the proposed estimators, which do not know the true memory, deliver the same statistical performance (rates of integrated mean-squared error) as that delivered by estimators that know the true memory. Furthermore, when the underlying predictor function does not have a finite memory, we establish that the estimator based on Legendre polynomials is *consistent*.

Index Terms—Bernstein inequality, complexity regularization, least-squares loss, Legendre polynomials, Markov processes, memory-universal prediction, mixing processes, model selection, neural networks, time series prediction.

I. INTRODUCTION

STATISTICAL prediction of random processes has numerous practical applications such as financial asset pricing [26], physical time series modeling [40], [54], stock price prediction [54], signal processing [58], and predictive speech coding [60]. Here, we consider the problem of one-step-ahead prediction of a real-valued, bounded, stationary random process $\{X_i\}_{i=-\infty}^{\infty}$. Probabilistically, the conditional mean of X_0 given the entire infinite past

$$X_{(-\infty, -1)} \equiv (\dots, X_{-2}, X_{-1})$$

namely, $E[X_0|X_{(-\infty, -1)}]$, is the best mean-square predictor of X_0 (Masani and Wiener [31]). Geometrically, the conditional mean $E[X_0|X_{(-\infty, -1)}]$ is the L^2 (nonlinear) projection of X_0 onto the subspace generated by the infinite past $X_{(-\infty, -1)}$. For $1 \leq p \leq \infty$, write a *predictor function* as

$$m_p(x) \equiv E[X_0|X_{(-p, -1)} = x], \quad x \in \mathbb{R}^p \quad (1)$$

Manuscript received July 16, 1996; revised June 1, 1997. This work was supported in part by the National Science Foundation under Grant DMS-97-03876.

D. S. Modha is with net.Mining, IBM Almaden Research Center, San Jose, CA 95120-6099 USA.

E. Masry is 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(98)00001-7.

where

$$X_{(-p, -1)} \equiv (X_{-p}, X_{-p+1}, \dots, X_{-1}).$$

In this paper, given a sequence of observations X_1, X_2, \dots, X_N drawn from the process $\{X_i\}_{i=-\infty}^{\infty}$, we are interested in estimating the *infinite memory* predictor function m_∞ .

We say that the predictor function m_∞ has a *finite memory*, if for some integer q , $1 \leq q < \infty$,

$$m_\infty(X_{(-\infty, -1)}) = m_q(X_{(-q, -1)}) \text{ almost surely.} \quad (2)$$

The condition (2) is satisfied, for example, by Markov processes of order q , but is mathematically weaker than the Markov property since only the first-order conditional moments are involved in (2). Under (2), the problem of estimating m_∞ reduces to that of estimating the predictor function m_q .

We would like to estimate the predictor function m_∞ using an estimator, say \hat{m}_N , that is simultaneously “memory-universal” and “consistent” as described below.

- 1) Suppose that the predictor function m_∞ has a finite memory q , and that the estimator \hat{m}_N does not know q . We say that \hat{m}_N is *memory-universal*, if a) it is a consistent estimator of $m_\infty (= m_q)$; and b) it delivers the same rate of convergence—in the integrated mean-squared-error sense—as that delivered by an estimator, say $\hat{m}_{(q, N)}$, that knows q .
- 2) Suppose that the predictor function m_∞ does not have a finite memory. We say that the (same) estimator \hat{m}_N is *consistent* if it converges to m_∞ in the sense of integrated mean-squared error.

Our notion of memory-universality is inspired by a similar notion in the theory of universal coding, see, for example, Ryabko [43] and [44]. Roughly speaking, memory-universal estimators implicitly “discover” the true unknown memory q . As an important aside, we point out that our notion of memory-universality is distinct from the notion of “universal consistency” traditionally considered in the nonparametric estimation literature where it means convergence under the weakest possible regularity constraints on the underlying process, see, for example, Algoet [2], [3], Devroye, Györfi, and Lugosi [20], Morvai, Yakowitz, and Györfi [36], and Stone [48]. In this paper, we assume that the underlying random process is bounded and exponentially strongly mixing, hence our estimators are not universally consistent in the traditional sense.

By the L^2 martingale convergence theorem [22, p. 217], the predictor function m_∞ is a mean-square limit of the

sequence of predictor functions $\{m_p\}_{p \geq 1}$. Hence, we propose the following two-step scheme for estimating m_∞ with the hope of attaining both memory-universality and consistency.

- 1) For each fixed memory $p \geq 1$, formulate an estimator $\hat{m}_{(p,N)}$ of m_p by minimizing a certain complexity regularized least squares loss.
- 2) Given the sequence $\{\hat{m}_{(p,N)}\}_{p \geq 1}$, select a memory $\tilde{p} \equiv \tilde{p}_N$ by minimizing a certain complexity regularized least squares loss, and use $\hat{m}_N \equiv \hat{m}_{(\tilde{p},N)}$ as the estimator of m_∞ .

Let us consider the first step for a fixed memory p . In general, the predictor function m_p is not a member of any finite-dimensional parametric family of functions, hence we estimate m_p using a sequence of parametric families of functions such as neural networks and Legendre polynomials. Statistical risk (measured by a certain integrated mean-squared error) in estimating m_p using a parametric model has two additive components: approximation error and estimation error. Generally speaking, a model with a larger dimension has a smaller approximation error but a larger estimation error, while a model with a smaller dimension has a smaller estimation error but a larger approximation error. Consequently, to minimize the statistical risk in estimating m_p from a list of parametric models, a tradeoff between the approximation error and the estimation error must be found. The tradeoff can be achieved by judiciously selecting the dimension of the model used to estimate m_p . Assuming that the underlying process is exponentially strongly mixing, a data-driven scheme—which minimizes a certain complexity regularized least squares loss—for selecting the model dimension was developed, in a slightly different context, in our previous work [34], which built on the results of Barron [8], [10], McCaffrey and Gallant [32], and Vapnik [51] for independent and identically distributed (i.i.d.) observations and the results of White [55] and White and Wooldridge [57] for strongly mixing observations. For other related work, in an i.i.d. setting, see Barron, Birgé, and Massart [12], Barron and Cover [13], Farago and Lugosi [23], Lugosi and Nobel [28], Lugosi and Zeger [29], [30], and Yang and Barron [61]. For a general review of the methodology employed to estimate a function from a sequence of parametric families of functions, see Vapnik [52].

Using the results of the first step as a building block, let us now consider the second step which is the central concern of this paper. The statistical risk in estimating the predictor function m_∞ using the estimator $\hat{m}_{(p,N)}$ has two additive components: the approximation error between m_∞ and m_p and the statistical risk in estimating m_p using $\hat{m}_{(p,N)}$. It follows from L^2 martingale convergence theorem that the approximation error between m_∞ and m_p is a decreasing function in the memory p . On the other hand, since m_p is a multivariate function from \mathbb{R}^p to \mathbb{R} , the statistical risk in estimating m_p is, generally speaking, an increasing function in the memory p . A tradeoff between the approximation error between m_∞ and m_p and the statistical risk in estimating m_p can be achieved by judiciously selecting the memory p . Two conceptually distinct approaches for memory selection

appear plausible: i) we may select the memory, say p_N , to be a deterministic, increasing function of the number of observations N , and use $\hat{m}_{(p_N,N)}$ as our estimator of m_∞ ; alternatively, ii) we may select the memory, say \tilde{p}_N , in a data-driven fashion, and use $\hat{m}_N = \hat{m}_{(\tilde{p}_N,N)}$ as our estimator of m_∞ . In this paper, we pursue a data-driven approach to memory selection, which, although computationally more expensive, is statistically more desirable than deterministic approaches as explained below. Suppose that the predictor function m_∞ has a finite—but unknown—memory q , then any deterministic, increasing memory p_N will asymptotically “overestimate” the true memory q , and hence, in general, the corresponding estimator $\hat{m}_{(p_N,N)}$ of m_q will not deliver a rate of convergence for the statistical risk comparable to that delivered by $\hat{m}_{(q,N)}$. In other words, although $\hat{m}_{(p_N,N)}$ may be consistent, it will not be memory-universal.

In this paper, we select the memory \tilde{p}_N , in a data-driven fashion, by minimizing a certain complexity regularized least squares loss. As the main contribution of this paper, assuming that the underlying random process is bounded and exponentially strongly mixing, we establish that the estimator $\hat{m}_N = \hat{m}_{(\tilde{p}_N,N)}$ is memory-universal if the predictor function m_∞ has a finite memory (Theorems 3.2 and 4.2, and Corollary 5.1), and is consistent even if the predictor function m_∞ does not have a finite memory (Theorems 4.3 and 5.2, and Remark 6.3). These results are distinct from the case when the underlying memory is known, and require novel formulation and analysis which have no counterpart in [34].

Previously, complexity regularization has been used, in an i.i.d. setting, to construct smoothness-universal or norm-universal estimators of a regression or density function (Barron [10], [11], Yang and Barron [61], and Barron, Birgé, and Massart [12]). In this paper, we use complexity regularization to construct memory-universal and consistent estimators of the (possibly) infinite memory predictor function.

For a further discussion of the relevant literature, see Remark 6.1.

This paper is organized as follows. In Section II, we present some notation and our basic assumptions. In Section III, we construct an estimator \hat{m}_N , for m_∞ , based on neural networks. Assuming that the predictor function m_∞ has a finite memory, we establish memory-universality of \hat{m}_N (compare Theorems 3.1 and 3.2). In Section IV, we construct an estimator \hat{m}_N , for m_∞ , based on Legendre polynomials. Assuming that the predictor function m_∞ has a finite memory, we establish memory-universality of \hat{m}_N (compare Theorems 4.1 and 4.2). Furthermore, even if the predictor function m_∞ does not have a finite memory, we establish consistency of \hat{m}_N (Theorem 4.3). In Section V, which is the conceptual and technical backbone of this paper, we present a scheme for constructing the estimator \hat{m}_N using a sequence of abstract parametric families of functions. The estimators considered in Sections III and IV are obtained by simply adapting the estimation scheme presented in Section V to neural networks and Legendre polynomials, respectively. Furthermore, in Section V, we establish abstract upper bounds, in terms of a certain deterministic index of resolvability, on the statistical risk in estimating m_∞ using \hat{m}_N (Theorem 5.2). Theorem 5.2 plays a key role in

establishing the memory-universality and consistency results stated in Sections III and IV. A discussion of our results is presented in Section VI, and the proofs of the main results are collected in Section VII.

II. PRELIMINARIES

Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary random process on a probability space (Ω, \mathcal{F}, P) . For $-\infty < i < \infty$, let $\mathcal{F}_{(i, \infty)}$ and $\mathcal{F}_{(-\infty, i]}$ denote the σ -algebras of events generated by $\{X_j, j \geq i\}$ and $\{X_j, j \leq i\}$, respectively. The process $\{X_i\}_{i=-\infty}^{\infty}$ is called *strongly mixing* [42], if

$$\sup_{A \in \mathcal{F}_{(-\infty, 0)}, B \in \mathcal{F}_{(j, \infty)}} |P[AB] - P[A]P[B]| = \alpha(j) \rightarrow 0 \quad \text{as } j \rightarrow \infty. \quad (3)$$

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

Assumption 2.1. Exponentially Strongly Mixing Property: 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 β and c are assumed to be known.

Assumption 2.1 is satisfied—with $\beta = 1$ —by important classes of processes such as certain linear (Withers [59]) and certain aperiodic, Harris-recurrent Markov processes (Athreya and Pantula [4, Theorem A] and Davydov [19, Theorem 1]). The former class includes certain Gaussian and non-Gaussian ARMA processes, while the latter class includes certain bilinear, nonlinear ARX, and ARCH processes (Doukhan [21] and Auestad and Tjøstheim [5]).

For $-\infty < i \leq j < \infty$, let $X_{(i, j)} = (X_i, X_{i+1}, \dots, X_j)$ and $X_{(-\infty, j]} = (\dots, X_{j-1}, X_j)$. Define the *effective number of observations* contained in the sequence of observations $\{X_{(i-p, i)}\}_{i=p+1}^N$, where $0 \leq p < N$, drawn from a process satisfying Assumption 2.1, by

$$N_p = \lfloor (N-p) \left[\{8(N-p)/c\}^{1/(\beta+1)} + p \right]^{-1} \rfloor \quad (4)$$

where $\lfloor u \rfloor$ ($\lceil u \rceil$) denotes the greatest (least) integer less (greater) than or equal to u . The concept of effective number of observations stems from the Craig–Bernstein inequality for the observations $\{X_{(i-p, i)}\}_{i=p+1}^N$ (see Lemma 7.1); also, see [34].

In the sequel, we will also need the following compactness assumption.

Assumption 2.2. Compactness: Assume that X_0 takes values in $[-1, 1]$.

We point out that Gaussian ARMA processes clearly do not satisfy the compactness condition in Assumption 2.2. However, certain non-Gaussian ARMA, bilinear, nonlinear ARX, and ARCH processes could have compact support, and hence could satisfy Assumption 2.2.

Let $P_{(i, j)}$ and $P_{(-\infty, j]}$ denote the marginal distributions¹ of $X_{(i, j)}$ and $X_{(-\infty, j]}$, respectively. For $1 \leq p \leq \infty$, let $L^2(P_{(1, p)})$ denote the space of all Borel measurable functions

¹ Strictly speaking, we assume that the sample space Ω is the canonical sample space $\prod_{i=-\infty}^{\infty} [-1, 1]$. Then, $P_{(-\infty, j]}$ is the restriction of the underlying probability measure P to the σ -algebra of events $\mathcal{F}_{-\infty}^j$.

$g: [-1, 1]^p \rightarrow \mathbb{R}$ that are square-integrable with respect to $P_{(1, p)}$. For $1 \leq p_1 \leq p_2 \leq \infty$, let $g_1 \in L^2(P_{(1, p_1)})$ and let $g_2 \in L^2(P_{(1, p_2)})$; then, define an *integrated squared distance* between the functions g_1 and g_2 as

$$\begin{aligned} r(g_2, g_1) &= r(g_1, g_2) \\ &= \int_{[-1, 1]^{p_2}} [g_1(x) - g_2(x, y)]^2 dP_{(1, p_2)}(x, y) \end{aligned} \quad (5)$$

where the dummy variables x and y take values in $[-1, 1]^{p_1}$ and $[-1, 1]^{(p_2-p_1)}$, respectively.

III. PREDICTOR ESTIMATION USING NEURAL NETWORKS

A. Neural Networks

We now present a sequence of parametric families of functions based on neural networks using some results of Barron [10]. 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 3.1. [10]: 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 $n \geq 1$ and $p \geq 1$. We now proceed to define a neural network with dimension (or “hidden units”) n and memory (or “time delays” or “lags”) p . Let

$$\gamma(p, n) = n(p+2) + 1 \quad (6)$$

represent the number of real-valued parameters parameterizing such a neural network. For $0 \leq i \leq n$, let $c_i \in \mathbb{R}$; for $1 \leq i \leq n$, let $a_i \in \mathbb{R}^p$; and let $b_i \in \mathbb{R}$. Let

$$\nu = (a_1, \dots, a_n; b_1, \dots, b_n; c_0, \dots, c_n)$$

represent a $\gamma(p, n)$ -dimensional parameter vector. Define a neural network with dimension n and memory p parameterized by ν as

$$f_{(p, (n, \nu))}(x) = \text{clip} \left(c_0 + \sum_{i=1}^n c_i \phi(a_i \cdot x + b_i) \right), \quad x \in \mathbb{R}^p \quad (7)$$

where

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

The function “clip” is used in (7) with the hindsight that the abstract estimation framework developed in Section V requires that the range of $f_{(p, (n, \nu))}$ be $[-1, 1]$ (see Assumption 5.1). Define

$$\tau_n = 2^{(2D_3+1)/D_3} D_2^{1/D_3} n^{(D_3+1)/(2D_3)} \quad (8)$$

where D_1 , D_2 , and D_3 are as in Assumption 3.1, and define a compact subset of $\mathbb{R}^{\gamma(p, n)}$, namely,

$$S_{(p, n)} = \left\{ \nu: c_0 \in [-1, 1], \sum_{i=1}^n |c_i| \leq C^{(p)}, \right. \\ \left. \max_{1 \leq i \leq n} \|a_i\|_1 \leq \tau_n, \max_{1 \leq i \leq n} |b_i| \leq \tau_n \right\} \quad (9)$$

Inputs: Natural numbers p , where $p < N$, k_1 , where $p \leq k_1 < N$, and $k_2(p)$;
 real numbers λ and $\{L_{(p,n)}\}_{n=1}^{k_2(p)}$;
 sets $\{S_{(p,n)}\}_{n=1}^{k_2(p)}$;
 parametric functions $\{\{f_{(p,(n,\nu))}\}_{\nu \in S_{(p,n)}}\}_{n=1}^{k_2(p)}$.

Estimation Scheme:

1. (least-squares estimation) For each dimension $1 \leq n \leq k_2(p)$, compute

$$\hat{\nu}(p, n) = \arg \min_{\nu \in S_{(p,n)}} \left\{ \frac{1}{N - k_1} \sum_{i=k_1+1}^N [X_i - f_{(p,(n,\nu))}(X_{(i-p,i-1)})]^2 \right\}.$$

2. (dimension selection criterion) Compute

$$\hat{n}_p = \arg \min_{1 \leq n \leq k_2(p)} \left\{ \frac{1}{N - k_1} \sum_{i=k_1+1}^N [X_i - f_{(p,(n,\hat{\nu}(p,n)))}(X_{(i-p,i-1)})]^2 + \lambda \frac{L_{(p,n)} + 2 \ln(n+1)}{N_{k_1}} \right\}.$$

Outputs: Write $\hat{\theta}(p) = (\hat{n}_p, \hat{\nu}(p, \hat{n}_p))$, and define the estimator parameterized by $\hat{\theta}(p)$ as $\hat{m}_{(p,N)} = f_{(p,\hat{\theta}(p))}$.

Fig. 1. Scheme for computing the estimator $\hat{m}_{(p,N)}$.

where the constant $C^{(p)}$ is made concrete in the next subsection and $\|\cdot\|_1$ denotes the l^1 norm.

B. Estimation Schemes and Memory-Universality

Assumption 3.2. Finite Memory: For some integer q , $1 \leq q < \infty$, assume that

$$m_\infty(X_{(-\infty, -1)}) = m_q(X_{(-q, -1)}) \text{ almost surely.}$$

Under Assumption 3.2, the problem of estimating the predictor function m_∞ reduces to that of estimating the predictor function m_q . We assume that m_q satisfies the following Fourier transform-type representation due to Barron [9].

For $w = (w_1, \dots, w_q)$ and $x = (x_1, \dots, x_q)$ in \mathbb{R}^q , let

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

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

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

denote the l^1 norm on \mathbb{R}^q .

Assumption 3.3. Bounded Spectral Norm: Assume that there exists a complex-valued function \tilde{m}_q on \mathbb{R}^q such that for $x \in [-1, 1]^q$, we have

$$m_q(x) - m_q(0) = \int_{\mathbb{R}^q} (e^{iw \cdot x} - 1) \tilde{m}_q(w) dw$$

and that

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

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

For a detailed discussion of Assumption 3.3, we refer the interested reader to Barron [9]. Also, see Hornik *et al.* [25] and Yukich, Stinchcombe, and White [62].

For the sake of brevity and simplicity, we assume that the constant C_q is known. If, in fact, C_q is unknown, it may be possible to modify our estimators using the ideas in Barron [10, eqs. (31) and (32)]. Specifically, we can replace the index n in Section V by a multi-index (n, C) , which is like inserting an additional minimization step (between steps 1 and 2) in Fig. 1.

Suppose that the memory q in Assumption 3.2 is *known*. In this case, by using the knowledge of the memory q , we construct an estimator $\hat{m}_{(q,N)}$ by invoking the estimation scheme presented in Fig. 1 with the following specific input values:

- let $k_1 = q$, $p = q$, and $k_2(q) \equiv k_2(q, N) \geq \lceil \sqrt{Nq} \rceil$;
- let $\lambda > 20/3$; for $1 \leq n \leq k_2(q)$, let

$$L_{(q,n)} = [n(q+2) + 1] \ln(32\tau_n c D_1 C_q (Nq)^{D_4})$$

where $D_4 \geq 1/2$, τ_n is as in (8), D_1 is as in Assumption 3.1, and C_q is as in Assumption 3.3;

- for $1 \leq n \leq k_2(q)$, let $S_{(q,n)}$ be obtained from (9) by substituting $p = q$ and $C^{(q)} = C_q$;
- for $1 \leq n \leq k_2(q)$ and for $\nu \in S_{(q,n)}$, let $f_{(q,(n,\nu))}$ be obtained from (7).

The input values presented above are selected, with hindsight, to establish Theorem 3.1.

Throughout this section, we assume that the least squares estimation step in Fig. 1 delivers the global minimum. From a strict mathematical perspective, finding the global minimum of a nonlinear least squares regression problem is computationally hard, see, for example, Farago and Lugosi [23] and Jones [27]. In practice, however, the backpropagation algorithms described in Back and Tsoi [7] and in Wan [53] started from a number of initial weights usually yield reasonably acceptable

Inputs: Natural numbers k_1 , where $k_1 < N$, and $\{k_2(p)\}_{p=1}^{k_1}$;
 real numbers λ and $\{\{L_{(p,n)}\}_{n=1}^{k_2(p)}\}_{p=1}^{k_1}$;
 sets $\{\{S_{(p,n)}\}_{n=1}^{k_2(p)}\}_{p=1}^{k_1}$;
 parametric functions $\{\{f_{(p,(n,\nu))}\}_{\nu \in S_{(p,n)}}\}_{n=1}^{k_2(p)}\}_{p=1}^{k_1}$.

Estimation Scheme:

1. For each memory $1 \leq p \leq k_1$, compute the parameter $\hat{\theta}(p) = (\hat{n}_p, \hat{\nu}(p, \hat{n}_p))$ and the estimator $\hat{m}_{(p,N)}$ by using the estimation scheme in Figure 1.
2. (memory selection criterion) Compute $\tilde{p} =$

$$\arg \min_{1 \leq p \leq k_1} \left\{ \frac{1}{N - k_1} \sum_{i=k_1+1}^N [X_i - f_{(p,\hat{\theta}(p))}(X_{(i-p,i-1)})]^2 + \lambda \frac{L_{(p,\hat{n}_p)} + 2 \ln(\hat{n}_p + 1) + 2 \ln(p + 1)}{N k_1} \right\}.$$

Output: Write the estimator parameterized by the memory \tilde{p} as $\hat{m}_N = \hat{m}_{(\tilde{p},N)}$.

Fig. 2. Scheme for computing the estimator \hat{m}_N .

results. Furthermore, various specialized hardwares are now available to considerably speed up training of neural networks, see, for example, Means *et al.* [33] and Sackinger and Graf [45].

Theorem 3.1. Memory q is Known: Suppose that Assumptions 2.1, 2.2, 3.1, and 3.3 hold. Then, for all $N_q \geq 2$

$$E[r(\hat{m}_{(q,N)}, m_q)] = O\left(\frac{\ln N}{N}\right)^{(1/2)(\beta/(\beta+1))}$$

where N_q is obtained from (4), β is as in Assumption 2.1, and r is as in (5).

The proof uses abstract upper bounds presented in Section V (namely, Theorem 5.1), and is briefly outlined in Section VII-C.

Now, suppose that the memory q in Assumption 3.2 is *unknown*. In this case, without the knowledge of the memory q , we construct an estimator \hat{m}_N by invoking the estimation scheme presented in Fig. 2 with the following specific input values:

- let $k_1 \equiv k_1(N) = o(N)$ be a function increasing to ∞ as $N \rightarrow \infty$, for example, $k_1 = \log N$; for $1 \leq p \leq k_1$, let $k_2(p) \equiv k_2(p, N) \geq \lceil \sqrt{N k_1} \rceil$;
- let $\lambda > 20/3$; for $1 \leq p \leq k_1$ and for $1 \leq n \leq k_2(p)$, let

$$L_{(p,n)} = [n(p+2) + 1] \ln(32\tau_n e D_1 C_q (N k_1)^{D_4})$$

where $D_4 \geq 1/2$, τ_n is as in (8), D_1 is as in Assumption 3.1, and C_q is as in Assumption 3.3;

- for $1 \leq p \leq k_1$ and for $1 \leq n \leq k_2(p)$, let $S_{(p,n)}$ be obtained from (9) by substituting

$$C^{(p)} = C_q; \quad (10)$$

- for $1 \leq p \leq k_1$, for $1 \leq n \leq k_2(p)$, and for $\nu \in S_{(p,n)}$, let $f_{(p,(n,\nu))}$ be as in (7).

The input values presented above are selected, with hindsight, to establish the following result.

Theorem 3.2. Memory q is Unknown: Suppose that Assumptions 2.1, 2.2, and 3.1–3.3 hold. Then, for $k_1 \geq q$ and for $N_{k_1} \geq 2$

$$E[r(\hat{m}_N, m_q)] = O\left(\frac{\ln N}{N}\right)^{(1/2)(\beta/(\beta+1))}$$

where N_{k_1} is obtained from (4), β is as in Assumption 2.1, and r is as in (5).

The proof uses abstract upper bounds presented in Section V (namely, Corollary 5.1), and can be found in Section VII-C.

Remark 3.1. Memory-Universality: Comparing Theorems 3.1 and 3.2, we find that the integrated mean-squared error in estimating m_q —when the memory q is unknown—has the same rate of convergence, in terms of upper bounds, as the corresponding error in estimating m_q —when q is known.

The dependence of our estimators on the parameter β is discussed in Remark 6.7.

By combining results of Barron [10, p. 129] and Barron, Birgé, and Massart [12, Proposition 6] with the generalized approximation results of Hornik *et al.* [25] and Yukich, Stinchcombe, and White [62], it is possible to relax Assumption 3.1 and the compactness restriction on the set of parameters $S_{(p,n)}$. We do not pursue these extensions here, since our principal focus is on memory-universal prediction of stationary random processes and not on the richness of the class of parametric functions employed to achieve this goal.

As an important aside, observe that in Theorems 3.1 and 3.2 the exponents of N in the respective rates of convergence do not depend on the memory q , that is, neural networks mitigate the curse of dimensionality in estimating the predictor function m_q which satisfies Assumption 3.3. This fact was first observed by Barron [10] in the context of regression estimation for i.i.d. observations.

IV. PREDICTOR ESTIMATION USING LEGENDRE POLYNOMIALS

To prevent a notational overload, in this section, we recycle the notations used in Section III.

A. Legendre Polynomials

Let $\{\varphi^{(i)}\}_{i \geq 1}$ denote the normalized Legendre polynomials [49] which are orthogonal with respect to the Lebesgue measure on $[-1, 1]$, where $\varphi^{(i)}$ is a polynomial of degree $(i - 1)$. Let $\mathbb{N} = \{1, 2, 3, \dots\}$. For $p \geq 1$, we now define a tensor product Legendre polynomial on $[-1, 1]^p$, indexed by a multi-integer $\mathbf{j} = (j_1, j_2, \dots, j_p) \in \mathbb{N}^p$, as

$$\varphi_{(p, \mathbf{j})}(x) = \varphi^{(j_1)}(x_1) \varphi^{(j_2)}(x_2) \cdots \varphi^{(j_p)}(x_p), \quad (11)$$

where $x = (x_1, x_2, \dots, x_p) \in \mathbb{R}^p$.

Fix $p \geq 1$ and $n \geq 1$. Let

$$\gamma(p, n) = n^p. \quad (12)$$

Let $\mathbf{1} = (1, 1, \dots, 1) \in \mathbb{N}^p$ and $\mathbf{n} = (n, n, \dots, n) \in \mathbb{N}^p$. We adopt the convention that the inequalities between multi-integers are to be interpreted component-wise. For $\mathbf{1} \leq \mathbf{j} \leq \mathbf{n}$, let $a_{\mathbf{j}} \in \mathbb{R}$. Let $\nu = (a_{\mathbf{1}}, \dots, a_{\mathbf{n}})$ represent a $\gamma(p, n)$ -dimensional parameter vector. Define a tensor product Legendre polynomial with dimension (or the largest coordinate-wise degree) $(n - 1)$ and memory (or time delays) p parameterized by ν as

$$f_{(p, (n, \nu))}(x) = \text{clip} \left(\sum_{\mathbf{j}=\mathbf{1}}^{\mathbf{n}} a_{\mathbf{j}} \varphi_{(p, \mathbf{j})}(x) \right), \quad x \in \mathbb{R}^p \quad (13)$$

where

$$\text{clip}(t) = -1_{\{t < -1\}} + t 1_{\{-1 \leq t \leq 1\}} + 1_{\{t > 1\}}$$

and $\varphi_{(p, \mathbf{j})}(x)$ is as in (11). We restrict attention to a compact subset of $\mathbb{R}^{\gamma(p, n)}$, namely,

$$S_{(p, n)} = \left\{ \nu: \sum_{\mathbf{j}=\mathbf{1}}^{\mathbf{n}} a_{\mathbf{j}}^2 \leq 2^p \right\}. \quad (14)$$

B. Estimation Schemes and Memory-Universality

In this subsection, we suppose that Assumption 3.2 holds, that is, the predictor function m_{∞} has a finite memory q . We assume that m_q satisfies the following differentiability condition.

Assumption 4.1 Differentiability: For some unknown smoothness order $s \geq 1$, assume that all partial derivatives of total order s of the function m_q exist, are measurable, and are square-integrable.

In this section, we approximate the predictor function m_q using Legendre polynomials. We note that various other families of approximants such as trigonometric series, splines, neural networks, or wavelets would suffice as well.

In the sequel, we need the following technical condition.

Assumption 4.2: Assume that the marginal distribution of $X_{(1, q)}$, namely, $P_{(1, q)}$, has a uniformly bounded probability density.

Suppose that the memory q in Assumption 3.2 is *known*. In this case, by using the knowledge of the memory q , we construct an estimator $\hat{m}_{(q, N)}$ by invoking the estimation

scheme presented in Fig. 1 (see Section III) with the following specific input values:

- let $k_1 = q$, $p = q$, and $k_2(q) \equiv k_2(q, N) \geq \lceil (N_q)^{1/q} \rceil$;
- let $\lambda > 20/3$; for $1 \leq n \leq k_2(q)$, let

$$L_{(q, n)} = n^q \ln \left(2^{(q+4)/2} \sqrt{n^q (2n - 1)^q} (N_q)^{D_4} \right)$$

where $D_4 \geq 1$;

- for $1 \leq n \leq k_2(q)$, let $S_{(q, n)}$ be obtained from (14);
- for $1 \leq n \leq k_2(q)$ and for $\nu \in S_{(q, n)}$, let $f_{(q, (n, \nu))}$ be obtained from (13).

Note that the estimator $\hat{m}_{(q, N)}$ makes no use of the smoothness order s . The input values presented above are selected, with hindsight, to establish the following result.

Theorem 4.1. Memory q is Known: Suppose that Assumptions 2.1, 2.2, 4.1, and 4.2 hold. Then, for all $N_q \geq 2$

$$E[r(\hat{m}_{(q, N)}, m_q)] = O\left(\frac{\ln N}{N}\right)^{(2s/(2s+q))(\beta/(\beta+1))}$$

where N_q is obtained from (4), β is as in Assumption 2.1, and r is as in (5).

The proof uses abstract upper bounds presented in Section V (namely, Theorem 5.1), and is briefly outlined in Section VII-D.

Now, suppose that the memory q in Assumption 3.2 is *unknown*. In this case, without the knowledge of the memory q , we construct an estimator \hat{m}_N by invoking the estimation scheme presented in Fig. 2 (see Section III) with the following specific input values:

- let $k_1 \equiv k_1(N) = o(N)$ be a function increasing to ∞ as $N \rightarrow \infty$, for example, $k_1 = \log N$; for $1 \leq p \leq k_1$, let $k_2(p) \equiv k_2(p, N) \geq \lceil (N_{k_1})^{1/p} \rceil$;
- let $\lambda > 20/3$; for $1 \leq p \leq k_1$ and for $1 \leq n \leq k_2(p)$, let

$$L_{(p, n)} = n^p \ln \left(2^{(p+4)/2} \sqrt{n^p (2n - 1)^p} (N_{k_1})^{D_4} \right)$$

where $D_4 \geq 1$;

- for $1 \leq p \leq k_1$ and for $1 \leq n \leq k_2(p)$, $S_{(p, n)}$ be as in (14);
- for $1 \leq p \leq k_1$, for $1 \leq n \leq k_2(p)$, and for $\nu \in S_{(p, n)}$, let $f_{(p, (n, \nu))}$ be as in (13).

Note that the estimator \hat{m}_N makes no use of the smoothness order s . The input values presented above are selected, with hindsight, to establish the following result.

Theorem 4.2. Memory q Is Unknown: Suppose that Assumptions 2.1, 2.2, 3.2, 4.1, and 4.2 hold. Then, for $k_1 \geq q$ and for $N_{k_1} \geq 2$

$$E[r(\hat{m}_N, m_q)] = O\left(\frac{\ln N}{N}\right)^{(2s/(2s+q))(\beta/(\beta+1))}$$

where N_{k_1} is obtained from (4), β is as in Assumption 2.1, and r is as in (5).

The proof uses abstract upper bounds presented in Section V (namely Corollary 5.1), and can be found in Section VII-D.

Observe that Remark 3.1, when properly translated, continues to hold in the current context as well. The dependence of our estimators on the parameter β is discussed in Remark 6.7.

By modifying our estimators using the ideas in Barron [11], it is possible to eliminate the logarithmic factor in Theorems 4.1 and 4.2. However, for the sake of simplicity, and also since the resulting estimators are computationally more expensive, we do not pursue that direction here.

C. Consistent Estimation of m_∞

In this subsection, unlike the previous one, we do not assume that the predictor function m_∞ has a finite memory. Nonetheless, we continue to estimate the predictor function m_∞ using the estimator \hat{m}_N constructed in the previous subsection. To establish consistency of \hat{m}_N , we require the following technical condition.

Assumption 4.3: For each memory $1 \leq p < \infty$, assume that the marginal distribution of $X_{(1,p)}$, namely, $P_{(1,p)}$, has a uniformly bounded probability density.

Theorem 4.3. Consistency: Suppose that Assumptions 2.1, 2.2, and 4.3 hold. Then,

$$\lim_{N \rightarrow \infty} E[r(\hat{m}_N, m_\infty)] = 0 \quad (15)$$

where r is as in (5).

The proof uses abstract upper bounds presented in Section V (namely, Theorem 5.2), and can be found in Section VII-D.

To obtain a rate of convergence for $E[r(\hat{m}_N, m_\infty)]$ in Theorem 4.3, we first need to obtain a rate of convergence for the ‘‘approximation error’’ $r(m_p, m_\infty)$ under Assumption 2.1. To the best of our knowledge, no such results are currently known.

Since the same estimator \hat{m}_N is considered in Theorems 4.2 and 4.3, we have that if the predictor function m_∞ has a finite memory, then \hat{m}_N delivers memory-universality, and even if the predictor function does not have a finite memory, \hat{m}_N delivers consistency. Also, observe that in Theorem 4.3 no smoothness assumptions are imposed on the predictor function m_∞ .

V. ABSTRACT ESTIMATION FRAMEWORK

In this section, given a sequence of abstract parametric families of functions, we propose an estimator, say \hat{m}_N , for the predictor function m_∞ , and upper-bound the integrated mean-squared error of the estimator in terms of certain indices of resolvability. The benefit of abstraction is that we are able to capture the statistics behind the proposed estimation scheme in the most general case, in a clean, economical fashion, without worrying about the cumbersome details of the specific cases of interest.

Throughout this section, fix the number of observations $N \geq 1$.

A. Parameter Spaces and Complexities

The development in this subsection closely follows that in [34, Subsec. 3.A].

Throughout this subsection, fix a memory $1 \leq p < \infty$. For each integer $n \geq 1$, let $\gamma(p, n)$ denote a model dimension (for example, see (6) and (12)), and let $S_{(p,n)}$ denote a compact subset of $\mathbb{R}^{\gamma(p,n)}$. The set $S_{(p,n)}$ will serve as a collection of

parameters associated with the model dimension $\gamma(p, n)$ (for example, see (9) and (14)). By introducing a prior density on the set $S_{(p,n)}$ as in Barron [10, p. 129], it is possible to relax the compactness assumption.

For every $\nu \in S_{(p,n)}$, let $f_{(p,(n,\nu))}$ denote a real-valued function on $[-1, 1]^p$ parameterized by (n, ν) (for example, see (7) and (13)). The following condition is required to be able to invoke the Craig–Bernstein inequalities in Lemma 7.1.

Assumption 5.1: For each integer $n \geq 1$ and for every $\nu \in S_{(p,n)}$, assume that $f_{(p,(n,\nu))}$ takes values in $[-1, 1]$.

Owing to the ‘‘clip’’ function in (7) and (13), Assumption 5.1 is satisfied for both neural networks and Legendre polynomials.

Let $\rho_{(p,n)}$ denote a metric on $\mathbb{R}^{\gamma(p,n)}$. For $\varepsilon \in (0, 1]$, let $T_{(p,n)}(\varepsilon)$ denote an $(\varepsilon, \rho_{(p,n)})$ -net of the set $S_{(p,n)}$; in other words, for every $\nu_1 \in S_{(p,n)}$ there exists a $\nu_2 \in T_{(p,n)}(\varepsilon)$ such that $\rho_{(p,n)}(\nu_1, \nu_2) \leq \varepsilon$. Assume that $T_{(p,n)}(\varepsilon) \subset S_{(p,n)}$. Let $L_{(p,n)}(\varepsilon)$ be such that

$$\ln \#(T_{(p,n)}(\varepsilon)) \leq L_{(p,n)}(\varepsilon) \quad (16)$$

where $\ln = \log_e$ and $\#$ denotes the cardinality operator.

Example 5.1. Neural Networks: Let notations be as in Section III. Let $\varrho_{(p,n)}$ denote a metric on $\mathbb{R}^{\gamma(p,n)}$ defined as in Barron [10, eq. (19)], but by replacing d there by p . It follows from [10, Lemma 2] by using (8) that for every $0 < \varepsilon \leq 1$ and for every $C^{(p)} \geq 1$, there exists a $(\varepsilon, \varrho_{(p,n)})$ -net of $S_{(p,n)}$, namely, $T_{(p,n)}(\varepsilon)$, such that

$$\ln \#(T_{(p,n)}(\varepsilon)) \leq [n(p+2) + 1] \ln \frac{4\tau_n e}{\varepsilon/2} \equiv L_{(p,n)}(\varepsilon). \quad (17)$$

Example 5.2. Legendre Polynomials: Let notations be as in Section IV. Let $\varrho_{(p,n)}$ be simply the l^1 metric on $\mathbb{R}^{\gamma(p,n)}$. The hypersphere $S_{(p,n)}$ is contained in the hypercube $[-2^{p/2} - \varepsilon, 2^{p/2} + \varepsilon]^{n^p}$, which has volume $[2(2^{p/2} + \varepsilon)]^{n^p}$. Furthermore, the set $[-2^{p/2} - \varepsilon, 2^{p/2} + \varepsilon]^{n^p}$ can be covered by (small) hypercubes (with respect to the metric $\varrho_{(p,n)}$) with side length $(2\varepsilon/\sqrt{n^p})$. Since by assumption $0 < \varepsilon \leq 1$, there exists a $(\varepsilon, \varrho_{(p,n)})$ -net of $S_{(p,n)}$, namely, $T_{(p,n)}(\varepsilon)$, such that

$$\ln \#(T_{(p,n)}(\varepsilon)) \leq n^p \ln \frac{2^{(p+2)/2} \sqrt{n^p}}{\varepsilon/2} \equiv L_{(p,n)}(\varepsilon). \quad (18)$$

Assumption 5.2: For every $n \geq 1$, there exists a strictly increasing function (in ε) $\varpi_{(p,n)}(\cdot): (0, 1] \rightarrow (0, \infty)$ such that for all $\varepsilon \in (0, 1]$ and for all $\nu_1 \in S_{(p,n)}$ and $\nu_2 \in T_{(p,n)}(\varepsilon)$ with $\rho_{(p,n)}(\nu_1, \nu_2) \leq \varepsilon$, we have

$$\sup_{x \in [-1, 1]^p} |f_{(p,(n,\nu_1))}(x) - f_{(p,(n,\nu_2))}(x)| \leq \varpi_{(p,n)}(\varepsilon).$$

Assumption 5.2 implies that the function $\varpi_{(p,n)}$ is invertible; let $\varpi_{(p,n)}^{-1}$ denote the inverse. Observe that the inverse $\varpi_{(p,n)}^{-1}(\delta)$ is defined for all $0 < \delta \leq \varpi_{(p,n)}(1) < \infty$ and takes values in the range $(0, 1]$.

Assumption 5.2 says that the class of parametric functions

$$\Gamma(p, n) \equiv \{f_{(p,(n,\nu))}: \nu \in S_{(p,n)}\}$$

can be covered in the supremum norm over $[-1, 1]^p$ by a finite class of functions. In other words, we limit attention to

classes of parametric functions where upper bounds on the sup-norm covering numbers of $\Gamma(p, n)$ are available; also, see [8], [32], and [34]. This class is sufficient to demonstrate our main contribution on memory-universal prediction of stationary random processes. We note in passing that more general classes of parametric functions have been considered, for instance, by Barron, Birgé, and Massart [12], Lugosi and Nobel [28], and Lugosi and Zeger [29], [30] in the context of function estimation in an i.i.d. setting.

Example 5.1 (continued): It follows from [10, Lemma 1], by invoking Assumption 2.2 and Part b) of Assumption 3.1, that Assumption 5.2 holds with $\varpi_{(p,n)}(\varepsilon) = 4D_1C^{(p)}\varepsilon$. For all $0 < \delta \leq \varpi_{(p,n)}(1) = 4D_1C^{(p)}$, the inverse of $\varpi_{(p,n)}$ can be written as

$$\varpi_{(p,n)}^{-1}(\delta) = \delta / (4D_1C^{(p)}), \quad (19)$$

Example 5.2 (continued): Let $\nu = (a_1, \dots, a_n) \in \mathbb{R}^{\gamma(p,n)}$ and let

$$\nu' = (a'_j, \dots, a'_n) \in \mathbb{R}^{\gamma(p,n)}$$

be such that $\varrho_{(p,n)}(\nu, \nu') \leq \varepsilon$. The following calculation shows that Assumption 5.2 holds with $\varpi_{(p,n)}(\varepsilon) = (\sqrt{2n-1})^p \varepsilon$.

$$\begin{aligned} & \sup_{x \in [-1, 1]^p} |f_{(p, (n, \nu_1))}(x) - f_{(p, (n, \nu_2))}(x)| \\ & \stackrel{\text{a)}}{\leq} \sup_{x \in [-1, 1]^p} \left| \sum_{j=1}^n (a_j - a'_j) \varphi_j(x) \right| \\ & \leq \left(\sup_{x \in [-1, 1]^p} |\varphi_n(x)| \right) \left| \sum_{j=1}^n (a_j - a'_j) \right| \\ & \leq (\sqrt{2n-1})^p \varrho_{(p,n)}(\nu, \nu') \\ & \leq (\sqrt{2n-1})^p \varepsilon \end{aligned}$$

where a) follows from (13). For all

$$0 < \delta \leq \varpi_{(p,n)}(1) = (\sqrt{2n-1})^p$$

the inverse of $\varpi_{(p,n)}$ can be written as

$$\varpi_{(p,n)}^{-1}(\delta) = \delta / (\sqrt{2n-1})^p. \quad (20)$$

Let $k_2(p)$ denote a natural number (for example, $k_2(p) \geq \lceil \sqrt{N_p} \rceil$ for neural networks and $k_2(p) \geq \lceil (N_p)^{1/p} \rceil$ for Legendre polynomials). Let $\Theta_{(p, k_2(p))}$ denote a collection of parameters of different dimensions, with the maximum dimension less than or equal to $k_2(p)$, such that each of the parameters comes packaged with the index of its dimension; formally, we write

$$\Theta_{(p, k_2(p))} = \bigcup_{n=1}^{k_2(p)} \{(n, \nu) : \nu \in S_{(p,n)}\}. \quad (21)$$

It follows from (21) that every $\theta \in \Theta_{(p, k_2(p))}$ must be of the form $\theta = (n, \nu)$ for some $1 \leq n \leq k_2(p)$ and for some $\nu \in S_{(p,n)}$; then, define

$$f_{(p, \theta)} = f_{(p, (n, \nu))} \quad (22)$$

and for every $0 < \delta \leq \varpi_{(p,n)}(1)$ define the “description complexity” of the parameter θ as

$$L^{(p)}(\theta, \delta) = 2 \ln(n+1) + L_{(p,n)}(\varpi_{(p,n)}^{-1}(\delta)) \quad (23)$$

where $\varpi_{(p,n)}$ is as in Assumption 5.2 and $L_{(p,n)}(\varpi_{(p,n)}^{-1}(\delta))$ is obtained from (16) by substituting $\varepsilon = \varpi_{(p,n)}^{-1}(\delta)$.

B. An Abstract Scheme for Computing $\hat{m}_{(p,N)}$

In this subsection, as a building block for the estimation scheme presented in the next subsection, we outline a scheme to construct the estimator $\hat{m}_{(p,N)}$. The estimation scheme presented in this subsection is conceptually the same as that presented in [34, eqs. (25), (26)], but is different in details.

For any natural number p , where $p < N$, for any natural number k_1 , where $p \leq k_1 < N$, for any natural number $k_2(p)$, for any real number δ , where²

$$0 < \delta \leq \min_{1 \leq n \leq k_2(p)} \varpi_{(p,n)}(1)$$

and for any real number λ , write

$$\hat{\theta}(p) = \arg \min_{\theta \in \Theta_{(p, k_2(p))}} \left\{ \frac{1}{N-k_1} \sum_{i=k_1+1}^N [X_i - f_{(p, \theta)}(X_{(i-p, i-1)})]^2 + \lambda \frac{L^{(p)}(\theta, \delta)}{N_{k_1}} \right\} \quad (24)$$

where $\Theta_{(p, k_2(p))}$ is as in (21), $f_{(p, \theta)}$ is as in (22), $L^{(p)}(\theta, \delta)$ is as in (23), and N_{k_1} is obtained from (4). Now, define the estimator parameterized by $\hat{\theta}(p)$ as

$$\hat{m}_{(p,N)} = f_{(p, \hat{\theta}(p))}. \quad (25)$$

We may now interpret the estimation scheme presented in Fig. 1 (see Section III) as a computationally convenient version of (24) and (25), which are analytically more convenient. For the sake of simplicity, in Fig. 1, we write $L_{(p,n)}$ instead of the complete expression $L_{(p,n)}(\varpi_{(p,n)}^{-1}(\delta))$ and we implicitly set $\delta = \delta(N) = (N_q)^{-D_1}$.

Define the *p-index of resolvability* corresponding to the estimator $\hat{m}_{(p,N)}$ as

$$\begin{aligned} R_{(p,N)}(m_p, k_1) &= \min_{\theta \in \Theta_{(p, k_2(p))}} \left\{ r(f_{(p, \theta)}, m_p) + \lambda \frac{L^{(p)}(\theta, \delta)}{N_{k_1}} \right\} \quad (26) \end{aligned}$$

where $\Theta_{(p, k_2(p))}$ is as in (21), $L^{(p)}(\theta, \delta)$ is as in (23), N_{k_1} is obtained from (4), and $r(f_{(p, \theta)}, m_p)$ is obtained from (5) by substituting $g_1 = f_{(p, \theta)}$ and $g_2 = m_p$.

Remark 5.1: The index of resolvability was first introduced by Barron and Cover [13] in the context of density estimation for i.i.d. observations, and by Barron [8] in the context of regression estimation for i.i.d. observations.

²If we let $0 < \delta \leq \varpi_{(p,n)}(1)$, then $\varpi_{(p,n)}^{-1}(\delta)$ is well defined. Thus if we let

$$0 < \delta \leq \min_{1 \leq n \leq k_2(p)} \varpi_{(p,n)}(1)$$

then $\varpi_{(p,n)}^{-1}(\delta)$ is well defined for each $1 \leq n \leq k_2(p)$.

Theorem 5.1: Let p be a natural number such that $p < N$. Set $k_1 = p$. Suppose that Assumptions 2.1 and 2.2 hold, and that Assumptions 5.1 and 5.2 hold. Then, for all natural numbers $k_2(p)$, for all real numbers

$$0 < \delta \leq \min_{1 \leq n \leq k_2(p)} \varpi_{(p,n)}(1)$$

for all $\lambda > 20/3$, and for all $N_p \geq 2$

$$E[r(\hat{m}_{(p,N)}, m_p)] < \bar{\eta} R_{(p,N)}(m_p, p) + \frac{12\delta}{\eta'} + \frac{4\tilde{\alpha}\lambda}{\eta' N_p} \quad (27)$$

where $\eta = 4/(\lambda - 8/3)$, $\eta' = (1 - \eta)$, $\bar{\eta} = (1 + \eta)/(1 - \eta)$, and $\tilde{\alpha} = (1 + 4e^{-2\bar{\alpha}})$.

The proof is briefly outlined in Section VII-B.

C. An Abstract Scheme for Computing \hat{m}_N

For any natural number k_1 , where $k_1 < N$, for any natural numbers $\{k_2(p)\}_{p=1}^{k_1}$, for any real number δ , where

$$0 < \delta \leq \min_{1 \leq p \leq k_1} \left\{ \min_{1 \leq n \leq k_2(p)} \varpi_{(p,n)}(1) \right\}$$

and for any real number λ , write

$$\tilde{p} = \arg \min_{1 \leq p \leq k_1} \left\{ \frac{1}{N - k_1} \sum_{i=k_1+1}^N [X_i - f_{(p, \hat{\theta}(p))}(X_{(i-p, i-1)})]^2 + \lambda \frac{L^{(p)}(\hat{\theta}(p), \delta) + 2 \ln(p+1)}{N_{k_1}} \right\} \quad (28)$$

where $\hat{\theta}(p)$ is as in (24) and $L^{(p)}(\hat{\theta}(p), \delta)$ is obtained from (23) by substituting $\theta = \hat{\theta}(p)$. Roughly speaking, the adaptive memory \tilde{p} is an estimator of the memory of the underlying predictor function m_∞ . We now write the estimator \hat{m}_N as

$$\hat{m}_N = \hat{m}_{(\tilde{p}, N)} = f_{(\tilde{p}, \hat{\theta}(\tilde{p}))}. \quad (29)$$

We may now interpret the estimation scheme presented in Fig. 2 (see Section III) as a computationally convenient version of (28) and (29), which are analytically more convenient. For the sake of simplicity, in Fig. 2, we write $L_{(p, \hat{n}_p)}$ instead of the complete expression $L_{(p, \hat{n}_p)}(\varpi_{(p, \hat{n}_p)}^{-1}(\delta))$ and we implicitly set $\delta = \delta(N) = (N_{k_1})^{-D_4}$.

Theorem 5.2: Let k_1 be a natural number such that $k_1 < N$. Suppose that Assumptions 2.1 and 2.2 hold, and that Assumptions 5.1 and 5.2 hold for each $1 \leq p \leq k_1$. Then, for all natural numbers $\{k_2(p)\}_{p=1}^{k_1}$, for all real numbers

$$0 < \delta \leq \min_{1 \leq p \leq k_1} \left\{ \min_{1 \leq n \leq k_2(p)} \varpi_{(p,n)}(1) \right\}$$

for all $\lambda > 20/3$, for all $N_{k_1} \geq 2$, and for all $1 \leq p \leq k_1$

$$E[r(\hat{m}_N, m_\infty)] < \bar{\eta} R_{(p,N)}(m_p, k_1) + \bar{\eta} r(m_p, m_\infty) + \bar{\eta} \lambda \frac{2 \ln(p+1)}{N_{k_1}} + \frac{12\delta}{\eta'} + \frac{4\tilde{\alpha}\lambda}{\eta' N_{k_1}}$$

where $\eta = 4/(\lambda - 8/3)$, $\eta' = (1 - \eta)$, $\bar{\eta} = (1 + \eta)/(1 - \eta)$, $\tilde{\alpha} = (1 + 4e^{-2\bar{\alpha}})$, and $R_{(p,N)}(m_p, k_1)$ is as in (26).

The proof can be found in Section VII-B.

Remark 5.2: Observe from the proofs of Theorems 5.1 and 5.2 that the bounds stated in the theorems continue to hold even when the parameters k_1 , $k_2(p)$, and δ are functions of N .

Remark 5.3: Observe that the index of resolvability (which consists of an approximation error term and an estimation error term) in Theorems 5.1 and 5.2 is multiplied by a constant $\bar{\eta} > 1$. This implies that, for each fixed N , the upper bounds established in Theorems 5.1 and 5.2 may not be the best possible—in the sense of constant multipliers. In a concept learning (or pattern recognition) setting, Lugosi and Zeger [30] avoid the problem of constant multipliers larger than 1 by using a Vapnik–Chervonenkis framework. However, in a mean-squared regression setting, a direct application of their method of analysis leads to a slower overall rate of convergence which is less desirable.

Corollary 5.1: Suppose all hypotheses of Theorem 5.2 hold. In addition, suppose that Assumption 3.2 holds and that $q \leq k_1$. Then

$$E[r(\hat{m}_N, m_q)] < \bar{\eta} R_{(q,N)}(m_q, k_1) + \bar{\eta} \lambda \frac{2 \ln(q+1)}{N_{k_1}} + \frac{12\delta}{\eta'} + \frac{4\tilde{\alpha}\lambda}{\eta' N_{k_1}} \quad (30)$$

where $R_{(q,N)}(m_q, k_1)$ is obtained from (26) by substituting $p = q$.

Proof: The corollary follows by applying Theorem 5.2 with $p = q$ (since $q \leq k_1$) and $m_\infty = m_q$ (since Assumption 3.2 holds). \square

VI. DISCUSSION

Remark 6.1. Related Works: We now discuss the relevant literature to establish a broader context for our results.

- Suppose that the process $\{X_i\}_{i=-\infty}^{\infty}$ is binary-valued. In this case, estimating the predictor function m_∞ is essentially the same as estimating the corresponding conditional distribution of X_0 given the entire infinite history $X_{(-\infty, -1)}$. The latter problem, owing to its applications in data compression, has received wide attention, for example, see Algoet [2], Cover [18], Rissanen [37], [38], and Ryabko [43], [44]. Our work fundamentally differs from the existing body of work for binary-valued processes in that, for binary-valued processes each element of the sequence $\{m_p\}_{p \geq 1}$ is finitely parameterized, while for real-valued processes considered here the elements of the sequence are not finitely parameterized.
- Suppose that the process $\{X_i\}_{i=-\infty}^{\infty}$ is real-valued, stationary, Gaussian ARMA. In this case, estimation of the predictor function m_∞ has been widely studied, for example, see Akaike [1], Bhansali [14], and Rissanen [39]. Our work fundamentally differs from the existing body of work for Gaussian ARMA processes, in that, for Gaussian ARMA processes each element of the sequence $\{m_p\}_{p \geq 1}$ is linear (in the observations) and finitely parameterized, while for stationary random processes considered here the elements of the sequence are neither linear nor finitely parameterized.

- Recently, supposing that the process $\{X_i\}_{i=-\infty}^{\infty}$ is real-valued, stationary, and ergodic, Algoet [2], [3], Morvai, Yakowitz, and Györfi [36], and Morvai, Yakowitz, and Algoet [35] proposed several nonparametric estimators of the predictor function m_{∞} , and established universal consistency of their estimators. This is distinct from memory-universality—which is the main focus of this paper.
- Recently, supposing that the process $\{X_i\}_{i=-\infty}^{\infty}$ is real-valued, stationary mixingale, Sin and White [47] proposed model selection criteria with the goal of selecting the best (in the sense of the smallest approximation error) of two abstract parametric models of the predictor function³ m_{∞} , and exemplified their model selection criteria for ARMAX-GARCH and STAR models. Although we consider a smaller class of processes, our estimators are applicable to sequences of parametric families of functions, minimize the overall statistical risk (that is, approximation error + estimation error), and are memory-universal and consistent.
- Supposing that the process $\{X_i\}_{i=-\infty}^{\infty}$ is real-valued, exponentially strongly mixing, and that the predictor function m_{∞} has a finite memory q (see (2)), Auestad and Tjøstheim [5], [6] (also see Tjøstheim [50]) and Cheng and Tong [17] proposed two-step schemes (based on the nonparametric kernel approach) to estimate the predictor function m_q , without the knowledge of the memory q . However, no analytical results are yet available for the estimators considered by Auestad and Tjøstheim, and although Cheng and Tong established the order consistency of their scheme, they did not establish, like we do, rates of convergence for the statistical risk.

Remark 6.2. General Regression Estimation Problem: Although so far we confined our attention to the simple and intuitively appealing problem of one-step-ahead prediction of stationary random processes, our results easily extend to a larger class of estimation problems as shown below. Let $\{X_i, Y_i\}_{i=-\infty}^{\infty}$ be a stationary random process such that X_0 takes values in $[-1, 1]$ and Y_0 takes values in \mathbb{R} . Let ψ be a measurable function such that $E|\psi(Y_0)|^2 < \infty$. For $1 \leq p \leq \infty$ and for $d \geq 0$, define the regression function as

$$m_p(x) \equiv E[\psi(Y_d)|X_{(-p, -1)} = x], \quad x \in \mathbb{R}^p.$$

Given a sequence of observations $\{X_i, Y_i\}_{i=1}^{N+d}$, we are interested in estimating the regression function m_{∞} . If we suppose that the process $\{X_i, Y_i\}_{i=-\infty}^{\infty}$ satisfies Assumption 2.1, suppose that $\psi(Y_0)$ takes values in $[-1, 1]$, and replace X_i in Figs. 1 and 2 (equivalently, in (24) and (28)) by $\psi(Y_{d+i})$, then we can use the resulting \hat{m}_N as our estimator of m_{∞} . Furthermore, all our results in Sections III–V continue to hold. Also, observe that by selecting various values for the function ψ , we can obtain a number of interesting special cases as follows.

³Note that although, for the sake of coherence, we have paraphrased the contribution of the literature dealing with Gaussian ARMA processes [1], [14], [39] and that of Sin and White [47] as estimating the predictor function m_{∞} , they did not phrase their work as such.

- (($d + 1$)-Step-Ahead Prediction) If we set $\psi(y) = y$, and assume that Y_0 takes values in $[-1, 1]$, then we can estimate $E[Y_d|X_{(-\infty, -1)}]$.
- (Conditional Moments Estimation) If we set $\psi(y) = y^t$, where $t \geq 1$, and assume that Y_0 takes values in $[-1, 1]$, then we can estimate $E[(Y_d)^t|X_{(-\infty, -1)}]$.
- (Conditional Distribution Estimation) If we set $\psi(y) = 1_{\{y \leq z\}}$, where z is a fixed real number, then we can estimate $P\{Y_d \leq z|X_{(-\infty, -1)}\}$.

Remark 6.3. Consistent Estimation of m_{∞} Using Neural Networks: Observe that we did not establish a result analogous to Theorem 4.3 for the estimator \hat{m}_N based on neural networks. Now consider the following relatively strong condition.

Assumption 6.1: For each $p \geq 1$, assume that there exists a complex-valued function \tilde{m}_p on \mathbb{R}^p such that for $x \in [-1, 1]^p$, we have

$$m_p(x) - m_p(0) = \int_{\mathbb{R}^p} (e^{iw \cdot x} - 1) \tilde{m}_p(w) dw$$

and that

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

for some known $C'_p > 0$. Set $C_p = \max\{1, C'_p\}$. Suppose that Assumptions 2.1, 2.2, and 6.1 hold, and let \hat{m}_N be as in Section III, where we replace (10) by $C^{(p)} = C_p$. Then, it is possible to show, by proceeding as in the proof of Theorem 4.3, that

$$\hat{m}_N \rightarrow m_{\infty} \text{ as } N \rightarrow \infty$$

where convergence is in the sense of integrated mean-squared error. However, owing to the stringent nature of Assumption 6.1, such a result appears unappealing.

Remark 6.4. Compact Parameter Spaces: Throughout this paper, we restricted attention to sequences of parametric families with compact parameter spaces. This assumption is sufficient to treat the examples presented here. However, if the predictor function m_q is such that one must use a sequence of parametric families with noncompact parameter spaces to obtain the best bounds on the approximation error, then the current framework may prove wanting. It may be possible to extend our framework to more general sequences of parametric families along the directions considered, for instance, by Barron, Birgé, and Massart [12], Lugosi and Nobel [28], and Lugosi and Zeger [29], [30].

Remark 6.5. Order Consistency and Price of Memory-Universality: In Theorems 3.1 and 3.2 (see also Theorems 4.1 and 4.2), we established the memory-universality of the estimator \hat{m}_N . However, it should be noted that \hat{m}_N is, roughly speaking, $k_1 (= o(N))$ times more expensive to compute than the corresponding estimator $\hat{m}_{(q, N)}$ that does know the memory q . It is currently unknown whether the adaptive memory \tilde{p} converges to q in some sense; nevertheless, \hat{m}_N does converge to the predictor function m_q .

Remark 6.6. Conditional Density and Conditional Quantiles: Assuming that the predictor function m_∞ has a finite memory, we formulated memory-universal and consistent estimators for m_∞ . It may also be possible to apply our estimation methodology and proof techniques along with the results of Barron and Cover [13], Barron [8], Barron, Birgé, and Massart [12], and White [56] to establish memory-universality and consistency of suitable estimators of the conditional density and of various conditional quantiles of X_0 given $X_{(-\infty, -1)}$.

Remark 6.7. Dependence of our Estimators on β : The complexity term in (24) and (28) is motivated solely by the statistical risk bounds we are able to obtain and not by other information-theoretic or Bayesian considerations. As a consequence, the complexity term depends explicitly on the parameter β in Assumption 2.1. In practice, one may set $\beta = 1$, since important classes of processes satisfy Assumption 2.1 with that value [59]. Note, however, that if the true underlying β is larger than the value of β used in our estimators, then the resulting estimators will deliver a slower rate than that obtainable with the knowledge of the true β . On the other hand, if the value of β used in our estimators is larger than the true underlying β , then we are unable to quantify the statistical performance of the resulting estimators. Unfortunately, unlike the parameters “smoothness,” “norm,” model dimension, or model memory, it does not appear possible to select β in a data-driven fashion using complexity regularization. Furthermore, we are currently unaware of any algorithm for testing the exponentially strongly mixing condition.

Remark 6.8. Comparison with Nonparametric Prediction: We have from Theorem 3.1 that

$$\begin{aligned} E \int_{[-1,1]^q} [\hat{m}_{(q,N)}(x) - m_q(x)]^2 dP_{(1,q)}(x) \\ = O\left(\frac{\ln N}{N}\right)^{(2s/(2s+q))(\beta/(\beta+1))}. \end{aligned} \quad (31)$$

Now, suppose that the strong mixing coefficient decays algebraically, and that the predictor function m_q has continuous and bounded partial derivatives of total order s . Let $\tilde{m}_{(q,N)}$ denote a nonparametric kernel estimator [15], [40], [41] which uses a kernel of order s , then it is known that with an optimal deterministic choice of the corresponding bandwidth parameter

$$E[\tilde{m}_{(q,N)}(x) - m_q(x)]^2 \sim N^{-(2s/(2s+q))}, \quad x \in \mathbb{R}^q. \quad (32)$$

Directly comparing (31) and (32), we find that rate of convergence for our estimator $\hat{m}_{(q,N)}$ decreases by the factor $\beta/(\beta+1)$. However, the above comparison may be inherently unfair, since our estimator $\hat{m}_{(q,N)}$ selects the model dimension in a data-driven fashion whereas the kernel estimator $\tilde{m}_{(q,N)}$ does not select its bandwidth parameter in a data-driven fashion. A fair comparison would involve a kernel estimator which selects its bandwidth parameter in a data-driven fashion (using, say, cross-validation). Unfortunately, to our knowledge, obtaining rates of convergence results for kernel estimators with data-driven bandwidth selection, in the context of dependent observations, is currently an open problem.

VII. DERIVATIONS

A. A Sequence of Craig–Bernstein Inequalities

To furnish the key technical tool required in the proof of Theorem 5.2, we now extend the Craig–Bernstein inequality in [34, Theorem 4.3]. Specifically, in proof of Theorem 5.2, we need to analyze empirical means of the form

$$\frac{1}{N-p} \sum_{i=p+1}^N \psi(X_{(i-p,i)}) \quad (33)$$

where $N > p \geq 0$ and ψ is a Borel measurable function from \mathbb{R}^{p+1} to \mathbb{R} , using a Craig–Bernstein-type inequality. The following lemma supplies the needed sequence of inequalities for each $p \geq 0$ (the Craig–Bernstein inequality in [34, Theorem 4.3] corresponds to the case $p = 0$).

Lemma 7.1. (N, p)-Craig–Bernstein Inequality: Suppose that Assumption 2.1 holds. Let integers N and p be such that $N > p \geq 0$. Let $\psi: \mathbb{R}^{p+1} \rightarrow \mathbb{R}$ be a Borel measurable function. For each $-\infty < i < \infty$, let $U_i = \psi(X_{(i-p,i)})$. Assume that $|U_0| \leq d_1$ a.s. and that $E[U_0] = 0$. Let N_p be as in (4). Then, for all $N_p \geq 2$, for all $\tau \in \mathbb{R}$, and for all $0 < \zeta < 1/d_1$

$$\begin{aligned} P\left\{\frac{1}{N-p} \sum_{i=p+1}^N U_i \geq \frac{\tau}{3\zeta N_p} + \frac{3\zeta E|U_1|^2}{2(1-\zeta d_1)}\right\} \\ \leq (1 + 4e^{-2\bar{\alpha}})e^{-\tau}. \end{aligned}$$

Proof: The proof closely follows the proof of [34, Theorem 4.3]; here we merely point out the main points of departure. Since in our case $U_i = \psi(X_{(i-p,i)})$, the sigma-algebras of events

$$\sigma\{U_{j+(a-1)k}, a = 1, 2, \dots, q-1\}$$

and

$$\sigma\{U_{j+(q-1)k}\}$$

in item b) in the proof of [34, Lemma 4.2], are now measurable with respect to

$$\sigma\{X_{(j+(a-1)k-p, j+(a-1)k)}, a = 1, 2, \dots, q-1\}$$

and $\sigma\{X_{(j+(q-1)k-p, j+(q-1)k)}\}$, respectively. Thus the distance between the two sigma-algebras now becomes

$$(j + (q-1)k - p) - (j + (q-2)k) = k - p.$$

Consequently, in [34, Lemma 4.2], we now must apply the mixing inequality in Hall and Heyde [24, Theorem A.5] with $\alpha(k-p)$ instead of $\alpha(k)$. In other words, throughout [34, Lemma 4.2], we should replace $\alpha(k)$ by $\alpha(k-p)$, and should replace the constraint $k > 0$ by $k > p$. Also, throughout [34, Theorem 4.3], we replace k^β by $(k-p)^\beta$ and replace N by $N-p$. Finally, note that the “number of blocks” k in [34, eq. (44)] now becomes

$$k = \lceil \{8(N-p)/c\}^{1/(\beta+1)} + p \rceil$$

and hence the ‘‘block size’’ or the effective number of observations in [34, eq. (39)] now becomes

$$N_p = \lfloor (N-p)/k \rfloor \\ = \left\lfloor (N-p) \left[\{8(N-p)/c\}^{1/(\beta+1)} + p \right]^{-1} \right\rfloor$$

which is exactly the prescribed value in (4). \square

B. Proofs of Theorems 5.1 and 5.2

To establish a perspective for the method of analysis used in establishing Theorems 5.1 and 5.2, we recall a technique used by Barron [8]. Let $\{X_i, Y_i\}_{i=-\infty}^{\infty}$ be a sequence of i.i.d. random variables. Define the regression function by $f^*(x) = E[Y_0|X_0 = x]$. Given N observations $\{X_i, Y_i\}_{i=1}^N$, Barron proposed a certain estimator, say \hat{f}_N , of f^* based on an abstract sequence of parametric models, and established upper bounds on the integrated mean-squared error $E[r(\hat{f}_N, f^*)]$ by analyzing

$$\sum_{i=1}^N ([Y_i - f_{(n,\nu)}(X_i)]^2 - [Y_i - f^*(X_i)]^2) \quad (34)$$

for each parameter ν with dimension $1 \leq n \leq N$, using the classical Craig–Bernstein inequality. In [34], assuming that the process $\{X_i, Y_i\}_{i=-\infty}^{\infty}$ is exponentially strongly mixing, we analyzed (34) using the Craig–Bernstein inequality established there.

Proof of Theorem 5.1: Motivated by the above discussion, we can upper-bound the integrated mean-squared error $E[r(\hat{m}_{(p,N)}, m_p)]$ by analyzing

$$W((p, (n, \nu)), N) \equiv \sum_{i=p+1}^N ([X_i - f_{(p,(n,\nu))}(X_{i-p}, i-1)]^2 - [X_i - m_p(X_{i-p}, i-1)]^2) \quad (35)$$

for each parameter ν with a fixed memory $1 \leq p < N$ and dimension $1 \leq n \leq k_2(p, N)$. With this insight, the theorem follows by proceeding essentially as in [34, Proof of Theorem 3.1] but by using the (N, p) -Craig–Bernstein inequality in Lemma 7.1. \square

Proof of Theorem 5.2: Here, we seek upper bounds on the integrated mean-squared error $E[r(\hat{m}_N, m_\infty)]$. To motivate our method of proof, we first explain two approaches that do not work. As a first try, motivated by (34) and (35), one may attempt to directly analyze

$$\sum_{i=p+1}^N ([X_i - f_{(p,(n,\nu))}(X_{i-p}, i-1)]^2 - [X_i - m_\infty(X_{-\infty}, i-1)]^2) \quad (36)$$

for each parameter ν with memory $1 \leq p \leq k_1(N)$ and dimension $1 \leq n \leq k_2(p, N)$. But, since each term in the second sum in (36) depends on an infinite past, no meaningful Craig–Bernstein inequalities appear possible for the empirical mean (36).

As a second try, one may attempt to analyze (35) for each parameter ν with memory $1 \leq p \leq k_1(N)$ and dimension

$1 \leq n \leq k_2(p, N)$ by using the (N, p) -Craig–Bernstein inequality in Lemma 7.1. This would lead to the desired upper bounds on $E[r(\hat{m}_N, m_\infty)]$, if we select the memory as

$$\tilde{p} = \arg \min_{1 \leq p \leq k_1} \left\{ \frac{1}{N-p} W((p, (n, \nu)), N) + \lambda \frac{L^{(p)}(\hat{\theta}(p), \delta) + 2 \ln(p+1)}{N_{k_1}} \right\} \quad (37)$$

where $\hat{\theta}(p)$ and $L^{(p)}(\hat{\theta}(p), \delta)$ are as in (28). Since $W((p, (n, \nu)), N)$ in (37) depends on the predictor function m_p (see (35)), implementing (37) would require the knowledge of the sequence of predictors $\{m_p\}_{p=1}^{k_1}$ —which is not available.

As a key technical insight, here, we analyze the empirical process

$$\sum_{i=k_1+1}^N ([X_i - f_{(p,n)}(X_{i-p}, i-1)]^2 - [X_i - m_{k_1}(X_{i-k_1}, i-1)]^2) \quad (38)$$

for each parameter ν with memory $1 \leq p \leq k_1$ and dimension $1 \leq n \leq k_2(p, N)$, using the (N, k_1) -Craig–Bernstein inequality, and, consequently, obtain upper bounds on $E[r(\hat{m}_N, m_{k_1})]$. Note that the second sum in (38) has a finite memory k_1 and does not depend on p . Next, by simple probabilistic manipulations, we observe that (see Lemma 7.6)

$$E[r(\hat{m}_N, m_\infty)] = E[r(\hat{m}_N, m_{k_1})] + r(m_{k_1}, m_\infty). \quad (39)$$

Equation (39) combined with the upper bounds on $E[r(\hat{m}_N, m_{k_1})]$ leads to the desired upper bounds on $E[r(\hat{m}_N, m_\infty)]$. In other words, instead of estimating m_∞ , we estimate m_{k_1} for a growing memory k_1 (as $N \uparrow \infty$). Then, by virtue of the L^2 -martingale convergence theorem, we are automatically doing a good job in estimating m_∞ .

To make the lengths of various equations manageable, throughout this proof, we write $\eta = 4/(\lambda - 8/3)$, $\eta'' = (1 + \eta)$, $\eta' = (1 - \eta)$, $\bar{\eta} = \eta''/\eta'$, and $\bar{\alpha} = (1 + 4e^{-2\bar{\alpha}})$.

Let k_1 be a natural number such that $k_1 < N$. For each fixed $1 \leq p \leq k_1$, for each fixed $-\infty < i < \infty$, and for each fixed $\theta \in \Theta_{(p, k_2(p))}$, write

$$V_{(p, i, \theta)} = [X_i - f_{(p, \theta)}(X_{i-p}, i-1)]^2 - [X_i - m_{k_1}(X_{i-k_1}, i-1)]^2 \quad (40)$$

$$\hat{r}(f_{(p, \theta)}, m_{k_1}) = \frac{1}{N - k_1} \sum_{i=k_1+1}^N V_{(p, i, \theta)}. \quad (41)$$

We now proceed with a series of lemmas.

Lemma 7.2: Let p and k_1 be natural numbers such that $p \leq k_1 < N$. Suppose that Assumptions 5.1, 5.2, 2.1, and 2.2 hold. Then, for all

$$0 < \delta \leq \min_{1 \leq n \leq k_2(p)} \varpi_{(p, n)}(1)$$

for all $\theta \in \Theta_{(p, k_2(p))}$, for all $\lambda > 20/3$, for all $\tilde{\delta} > 0$, and for all $N_{k_1} \geq 2$

$$P \left\{ \eta' r(f_{(p, \theta)}, m_{k_1}) \geq \hat{r}(f_{(p, \theta)}, m_{k_1}) + \lambda \frac{L^{(p)}(\theta, \delta) + 2 \ln(p+1) + \ln 1/\tilde{\delta}}{N_{k_1}} \right\} \leq \tilde{\alpha} \tilde{\delta} e^{-L^{(p)}(\theta, \delta)} e^{-2 \ln(p+1)}.$$

Proof: For $-\infty < i < \infty$, write

$$U_{(p, i, \theta)} = -V_{(p, i, \theta)} + E[V_{(p, i, \theta)}] \quad (42)$$

where $V_{(p, i, \theta)}$ is as in (40), and observe that $\{U_{(p, i, \theta)}\}_{i=-\infty}^{\infty}$ are identically distributed. By invoking Assumptions 5.1 and 2.2, and by proceeding as in [8], we have that

$$\begin{aligned} E[V_{(p, 0, \theta)}] &= r(f_{(p, \theta)}, m_{k_1}) \\ E[U_{(p, 0, \theta)}] &= 0 \\ E|U_{(p, 0, \theta)}|^2 &\leq 8r(f_{(p, \theta)}, m_{k_1}) \end{aligned}$$

and

$$|U_{(p, 0, \theta)}| \leq 8.$$

Also, it follows from (41) and (42) that

$$\frac{1}{N - k_1} \sum_{i=k_1+1}^N U_{(p, i, \theta)} = -\hat{r}(f_{(p, \theta)}, m_{k_1}) + r(f_{(p, \theta)}, m_{k_1}). \quad (43)$$

Since Assumption 2.1 holds and since $\{U_{(p, i, \theta)}\}_{i=-\infty}^{\infty}$ are identically distributed, the lemma follows by applying the (N, k_1) -Craig–Bernstein inequality in Lemma 7.1 to (43) (with $d_1 = 8$, $3\zeta = 1/\lambda$, and $\tau = L^{(p)}(\theta, \delta) + 2 \ln(p+1) + \ln 1/\tilde{\delta}$) just as the $(N, 0)$ -Craig–Bernstein inequality was applied in [34, Lemma 3.1] to (29) there. \square

Lemma 7.3: Let k_1 be a natural number such that $k_1 < N$. Suppose that Assumptions 5.1 and 5.2 hold for each $1 \leq p \leq k_1$, and that Assumptions 2.1 and 2.2 hold. Then, for all $\lambda > 20/3$, for all $\tilde{\delta} > 0$, for all $N_{k_1} \geq 2$, and for all

$$0 < \delta \leq \min_{1 \leq p \leq k_1} \left\{ \min_{1 \leq n \leq k_2(p)} \varpi_{(p, n)}(1) \right\}$$

$$P \left\{ \eta' r(\hat{m}_N, m_{k_1}) \geq \hat{r}(\hat{m}_N, m_{k_1}) + \lambda \frac{L^{(\tilde{p})}(\hat{\theta}(\tilde{p}), \delta) + 2 \ln(\tilde{p}+1) + \ln 1/\tilde{\delta}}{N_{k_1}} + 12\delta \right\} < \tilde{\alpha} \tilde{\delta}. \quad (44)$$

Proof: Observe that $\hat{m}_N = f_{(\tilde{p}, \hat{\theta}(\tilde{p}))}$ and that $\hat{m}_{(p, N)} = f_{(p, \hat{\theta}(p))}$. Thus, to establish (44), one can first establish

$$P \left\{ \eta' r(\hat{m}_{(p, N)}, m_{k_1}) \geq \hat{r}(\hat{m}_{(p, N)}, m_{k_1}) + \lambda \frac{L^{(p)}(\hat{\theta}(p), \delta) + 2 \ln(p+1) + \ln 1/\tilde{\delta}}{N_{k_1}} + 12\delta \right\} < \tilde{\alpha} \tilde{\delta} e^{-2 \ln(p+1)} \quad (45)$$

for each fixed $1 \leq p \leq k_1$. Since the sets $\{\{\tilde{p} = p\}\}_{p=1}^{k_1}$ are disjoint, we can then pass from (45) to (44) using a union bound argument. However, we can establish (45) by invoking Lemma 7.2 and by proceeding essentially as in [34, Lemma 3.2]. We omit the details. \square

Let $\theta^*(p)$ be the element of the set $\Theta_{(p, k_2(p))}$, which attains the p -index of resolvability $R_{(p, N)}(m_p, k_1)$ in (26); formally, we write

$$\theta^*(p) = \arg \min_{\theta \in \Theta_{(p, k_2(p))}} \left\{ r(f_{(p, \theta)}, m_p) + \lambda \frac{L^{(p)}(\theta, \delta)}{N_{k_1}} \right\}. \quad (46)$$

Lemma 7.4: Suppose all hypotheses of Lemma 7.3 hold. Then, for each $1 \leq p \leq k_1$, we have

$$P \left\{ \eta' r(\hat{m}_N, m_{k_1}) \geq \hat{r}(f_{(p, \theta^*(p))}, m_{k_1}) + \lambda \frac{L^{(p)}(\theta^*(p), \delta) + 2 \ln(p+1) + \ln 1/\tilde{\delta}}{N_{k_1}} + 12\delta \right\} < \tilde{\alpha} \tilde{\delta}.$$

Proof: Recall the definition of \hat{r} in (40) and (41).

$$\begin{aligned} \hat{r}(\hat{m}_N, m_{k_1}) + \lambda \frac{L^{(\tilde{p})}(\hat{\theta}(\tilde{p}), \delta) + 2 \ln(\tilde{p}+1)}{N_{k_1}} &= \hat{r}(f_{(\tilde{p}, \hat{\theta}(\tilde{p}))}, m_{k_1}) + \lambda \frac{L^{(\tilde{p})}(\hat{\theta}(\tilde{p}), \delta) + 2 \ln(\tilde{p}+1)}{N_{k_1}} \\ &\stackrel{a)}{\leq} \hat{r}(f_{(p, \hat{\theta}(p))}, m_{k_1}) + \lambda \frac{L^{(p)}(\hat{\theta}(p), \delta) + 2 \ln(p+1)}{N_{k_1}} \\ &\stackrel{b)}{\leq} \hat{r}(f_{(p, \theta^*(p))}, m_{k_1}) + \lambda \frac{L^{(p)}(\theta^*(p), \delta) + 2 \ln(p+1)}{N_{k_1}} \end{aligned} \quad (47)$$

where a) follows from (28); and b) follows from (24). The lemma now follows from Lemma 7.3 and (47). \square

Lemma 7.5: Let p and k_1 be natural numbers such that $p \leq k_1 < N$. Suppose that Assumptions 5.1, 2.1, and 2.2 hold. Then, for all $\lambda > 20/3$, for all $\tilde{\delta} > 0$, and for all $N_{k_1} \geq 2$

$$P \left\{ \hat{r}(f_{(p, \theta^*(p))}, m_{k_1}) \geq \eta'' r(f_{(p, \theta^*(p))}, m_{k_1}) + \lambda \frac{\ln 1/\tilde{\delta}}{N_{k_1}} \right\} \leq \tilde{\alpha} \tilde{\delta}.$$

Proof: Let $V_{(p, i, \theta^*(p))}$ be obtained from (40) by substituting $\theta = \theta^*(p)$. For $i = k_1 + 1, k_1 + 2, \dots, N$, write

$$U_{(i, p, \theta^*(p))} = V_{(p, i, \theta^*(p))} - E[V_{(p, i, \theta^*(p))}].$$

The lemma follows by applying the (N, k_1) -Craig–Bernstein inequality in Lemma 7.1 to the sum

$$\frac{1}{N - k_1} \sum_{i=k_1+1}^N U_{(i, p, \theta^*(p))} = \hat{r}(f_{(p, \theta^*(p))}, m_{k_1}) - r(f_{(p, \theta^*(p))}, m_{k_1})$$

with $d_1 = 8$, $3\zeta = 1/\lambda$, and $\tau = \ln 1/\tilde{\delta}$ and by simplifying as in [34, Lemma 3.1]. \square

Lemma 7.6: Let $0 \leq p' \leq p'' \leq p''' \leq \infty$ and let $g \in L^2(P_{(1,p')})$, then

$$r(g, m_{p'}) + r(m_{p''}, m_{p'''}) = r(g, m_{p'''}).$$

Proof:

$$\begin{aligned} & r(g, m_{p'}) + r(m_{p''}, m_{p'''}) \\ &= E[g(X_{(1,p')}) - m_{p''}(X_{(1,p')})]^2 + E[m_{p''}(X_{(1,p')}) \\ &\quad - m_{p'''}(X_{(1,p''')})]^2 \\ &= E[X_0 - g(X_{(1,p')})]^2 - E[X_0 - m_{p''}(X_{(1,p')})]^2 \\ &\quad + E[X_0 - m_{p''}(X_{(1,p')})]^2 \\ &\quad - E[X_0 - m_{p'''}(X_{(1,p''')})]^2 \\ &= E[g(X_{(1,p')}) - m_{p'''}(X_{(1,p''')})]^2 \\ &= r(g, m_{p'''}). \end{aligned}$$

Lemma 7.7: Suppose all hypotheses of Theorem 5.2 hold. Then, for each $1 \leq p \leq k_1$, we have

$$\begin{aligned} E[r(\hat{m}_N, m_{k_1})] &< \bar{\eta} \left(R_{(p,N)}(m_p, k_1) + r(m_p, m_{k_1}) \right. \\ &\quad \left. + \lambda \frac{2 \ln(p+1)}{N_{k_1}} \right) + \frac{12\delta}{\eta'} + \frac{4\tilde{\alpha}\lambda}{\eta' N_{k_1}}. \end{aligned}$$

Proof: Combining Lemmas 7.4 and 7.5, we have

$$\begin{aligned} P \left\{ \eta' r(\hat{m}_N, m_{k_1}) \geq \eta'' r(f_{(p,\theta^*(p))}, m_{k_1}) \right. \\ \left. + \lambda \frac{L^{(p)}(\theta^*(p), \delta) + 2 \ln(p+1)}{N_{k_1}} + \lambda \frac{2 \ln 1/\tilde{\delta}}{N_{k_1}} + 12\delta \right\} \\ < 2\tilde{\alpha}. \quad (48) \end{aligned}$$

Applying Lemma 7.6 with $g = f_{(p,\theta^*(p))}$, $p' = p'' = p$, and $p''' = k_1$, we have

$$r(f_{(p,\theta^*(p))}, m_{k_1}) = r(f_{(p,\theta^*(p))}, m_p) + r(m_p, m_{k_1}). \quad (49)$$

Now, ignoring the term

$$-\eta\lambda(L^{(p)}(\theta^*(p), \delta) + 2 \ln(p+1))/(N_{k_1})$$

we have from (26), (46), (48), and (49) that

$$\begin{aligned} P \left\{ \eta' r(\hat{m}_N, m_{k_1}) \geq \eta'' R_{(p,N)}(m_p, k_1) + \eta'' r(m_p, m_{k_1}) \right. \\ \left. + \eta'' \lambda \frac{2 \ln(p+1)}{N_{k_1}} + 12\delta + 2\lambda \frac{\ln 1/\tilde{\delta}}{N_{k_1}} \right\} < 2\tilde{\alpha}\tilde{\delta}. \end{aligned}$$

By writing

$$\begin{aligned} W &= \eta' r(\hat{m}_N, m_{k_1}) - \eta'' R_{(p,N)}(m_p, k_1) \\ &\quad - \eta'' r(m_p, m_{k_1}) - \eta'' \lambda (2 \ln(p+1))/N_{k_1} - 12\delta \end{aligned}$$

and for $t > 0$ setting $\tilde{\delta} = \exp(-N_{k_1} t/(2\lambda))$, we have that

$$P\{W \geq t\} < 2\tilde{\alpha} \exp\left\{-\frac{N_{k_1} t}{2\lambda}\right\}. \quad (50)$$

It is easy to see that $|W| < \infty$, and hence $E|W| < \infty$. The lemma now follows from (50) and [34, Lemma A.6]. \square

The following upper bounds complete the proof of Theorem 5.2.

$$\begin{aligned} & E[r(\hat{m}_N, m_\infty)] \\ &\stackrel{a)}{=} E[r(\hat{m}_N, m_{k_1})] + r(m_{k_1}, m_\infty) \\ &\stackrel{b)}{\leq} \bar{\eta} \left(R_{(p,N)}(m_p, k_1) + r(m_p, m_{k_1}) + \lambda \frac{2 \ln(p+1)}{N_{k_1}} \right) \\ &\quad + \frac{12\delta}{\eta'} + \frac{4\tilde{\alpha}\lambda}{\eta' N_{k_1}} + r(m_{k_1}, m_\infty) \\ &\stackrel{c)}{\leq} \bar{\eta} R_{(p,N)}(m_p, k_1) + \bar{\eta} r(m_p, m_\infty) + \bar{\eta} \lambda \frac{2 \ln(p+1)}{N_{k_1}} \\ &\quad + \frac{12\delta}{\eta'} + \frac{4\tilde{\alpha}\lambda}{\eta' N_{k_1}} \end{aligned}$$

\square where a) follows by applying Lemma 7.6 (with $g = \hat{m}_N$, $p' = \tilde{p}$, $p'' = k_1$, and $p''' = \infty$) on a realization-by-realization basis; b) follows from Lemma 7.7 for each $1 \leq p \leq k_1$; and c) follows by applying Lemma 7.6 (with $g = m_p$, $p' = p$, $p'' = k_1$, and $p''' = \infty$) and since $\bar{\eta} > 1$. \square

C. Proofs of Theorems 3.1 and 3.2

First, in Lemma 7.8 below, we establish an upper bound on a certain index of resolvability. We will then establish Theorem 3.1 (respectively, Theorem 3.2) by combining Lemma 7.8 and Theorem 5.1 (respectively, Corollary 5.1).

Lemma 7.8 A Bound on Index of Resolvability: Suppose that Assumptions 2.2 and 3.3 hold. Let k_3 be a natural number such that $k_3 \geq q$. Then, for all $k_2(q) \geq \lceil \sqrt{N_{k_3}} \rceil$, for all $\delta = (N_{k_3})^{-D_4}$, where $D_4 \geq 0$, and for all $N_{k_3} \geq 2$, we have

$$R_{(q,N)}(m_q, k_3) = O\left(\frac{\ln N_{k_3}}{N_{k_3}}\right)^{1/2}$$

where $R_{(q,N)}(m_q, k_3)$ is obtained from (26) and N_{k_3} is obtained from (4).

Proof: The proof follows by proceeding as in [34, Lemma 2.2]. We omit the details. \square

Proof of Theorem 3.1: Theorem 3.1 follows by combining Theorem 5.1 (for $p = q$) and Lemma 7.8 (for $k_3 = q$) in the manner of Theorem 3.2; we omit the details. \square

Proof of Theorem 3.2: It follows from our hypotheses that Assumptions 2.1, 2.2, 3.2 hold, and from Example 5.1 that Assumptions 5.1 and 5.2 hold for all $1 \leq p \leq k_1$. Consequently, all hypotheses of Corollary 5.1 hold, and we have for all $0 < \delta(N) \leq (4D_1 C_q)$, for all $\lambda > 20/3$, for $q \leq k_1$, and for all $N_{k_1} \geq 2$

$$\begin{aligned} E[r(\hat{m}_N, m_q)] &= O(R_{(q,N)}(m_q, k_1)) + O(\delta(N)) + O(N_{k_1}^{-1}) \\ &\stackrel{a)}{=} O\left(\frac{\ln N_{k_1}}{N_{k_1}}\right)^{1/2} + O\left(\frac{1}{(N_{k_1})^{D_4}}\right) \\ &\stackrel{b)}{=} O\left(\frac{\ln N}{N}\right)^{(1/2)(\beta/(\beta+1))} \end{aligned}$$

where a) follows by applying Lemma 7.8 with $k_3 = k_1$, $k_2(q) \geq \lceil \sqrt{N_{k_1}} \rceil$, and $\delta(N) = (N_{k_1})^{-D_4}$, where $D_4 \geq 0$; and b) follows if we let $D_4 \geq 1/2$, and from (4) by simple algebraic manipulations since $k_1 = o(N)$. \square

D. Proofs of Theorems 4.1–4.3

First, in Lemma 7.9 below, we establish an upper bound on a certain index of resolvability. We will then establish Theorem 4.1 (respectively, Theorem 4.2) by combining Lemma 7.9 and Theorem 5.1 (respectively, Corollary 5.1).

Lemma 7.9 A Bound on the Index of Resolvability: Suppose that Assumptions 2.2, 4.1, and 4.2 hold. Let k_3 be a natural number such that $k_3 \geq q$. Then, for all $k_2(q) \geq \lceil (N_{k_3})^{1/q} \rceil$, $\delta = (N_{k_3})^{-D_4}$, where $D_4 \geq 0$, and for all $N_{k_3} \geq 2$, we have

$$R_{(q,N)}(m_q, k_3) = O\left(\frac{\ln N_{k_3}}{N_{k_3}}\right)^{2s/(2s+q)}$$

where $R_{(q,N)}(m_q, k_3)$ is obtained from (26) and N_{k_3} is obtained from (4).

Proof:

$$\begin{aligned} & R_{(q,N)}(m_q, k_3) \\ & \stackrel{\text{a)}}{=} \min_{1 \leq n \leq k_2(q)} \left\{ \min_{\nu \in S_{(q,n)}} [r(f_{(q,(n,\nu))}, m_q)] \right. \\ & \quad \left. + \lambda \frac{L_{(q,n)}(\varpi_{(q,n)}^{-1}(\delta)) + 2 \ln(n+1)}{N_{k_3}} \right\} \\ & \stackrel{\text{b)}}{\leq} \min_{1 \leq n \leq k_2(q)} \left\{ \frac{K_1}{n^{2s}} \right. \\ & \quad \left. + \lambda \frac{L_{(q,n)}(\varpi_{(q,n)}^{-1}(\delta)) + 2 \ln(n+1)}{N_{k_3}} \right\} \\ & \stackrel{\text{c)}}{\leq} \min_{1 \leq n \leq \lceil (N_{k_3})^{1/q} \rceil} \left\{ \frac{K_1}{n^{2s}} + \lambda \frac{n^q}{N_{k_3}} \ln \frac{K_2 n^{q/2}}{(\varpi_{(q,n)}^{-1}(\delta))} \right. \\ & \quad \left. + \lambda \frac{2 \ln(n+1)}{N_{k_3}} \right\} \\ & \stackrel{\text{d)}}{\leq} \min_{1 \leq n \leq \lceil (N_{k_3})^{1/q} \rceil} \left\{ \frac{K_1}{n^{2s}} \right. \\ & \quad \left. + \lambda \frac{n^q}{N_{k_3}} \ln K_2 n^{q/2} (2n-1)^{q/2} (N_{k_3})^{D_4} \right. \\ & \quad \left. + \lambda \frac{2 \ln(n+1)}{N_{k_3}} \right\} \\ & \stackrel{\text{e)}}{\leq} \min_{1 \leq n \leq \lceil (N_{k_3})^{1/q} \rceil} \left\{ \frac{K_1}{n^{2s}} + \lambda \frac{n^q}{N_{k_3}} \ln K_3 (N_{k_3})^{K_4} \right. \\ & \quad \left. + \lambda \frac{2 \ln(2N_{k_3})}{N_{k_3}} \right\} \\ & \stackrel{\text{f)}}{\leq} \min_{1 \leq n \leq \lceil (N_{k_3})^{1/q} \rceil} \left\{ \frac{K_1}{n^{2s}} + K_5 \frac{n^q}{N_{k_3}} \ln K_6 N_{k_3} \right\} \\ & \stackrel{\text{g)}}{\leq} \min_{1 \leq n \leq \lceil (N_{k_3})^{1/q} \rceil} \left\{ \frac{K_1}{n^{2s}} + K_7 \frac{n^q}{N_{k_3}} \ln N_{k_3} \right\} \\ & \stackrel{\text{h)}}{\leq} (K_1 + 2K_7) \left(\frac{\ln N_{k_3}}{N_{k_3}} \right)^{2s/(2s+q)} \end{aligned}$$

where a) follows from (21), (23), and (26), where $S_{(q,n)}$ is obtained from (14), $f_{(q,(n,\nu))}$ is obtained from (13), $L_{(q,n)}$ is obtained from (18), and $\varpi_{(q,n)}^{-1}$ is obtained from (20); b) it follows from Assumption 4.2 that there exists a finite uniform

bound $M_q > 0$ on the probability density of the marginal distribution $P_{(1,q)}$. Hence

$$\begin{aligned} & \min_{\nu \in S_{(q,n)}} [r(f_{(q,(n,\nu))}, m_q)] \\ & = \min_{\nu \in S_{(q,n)}} \left[\int_{[-1,1]^q} [f_{(q,(n,\nu))}(x) - m_q(x)]^2 dP_{(1,q)}(x) \right] \\ & \leq \min_{\nu \in S_{(q,n)}} \left[M_q \int_{[-1,1]^q} [f_{(q,(n,\nu))}(x) - m_q(x)]^2 dx \right] \\ & \leq M_q \sum_{j \in \{1 \leq i \leq n\}^c \subset \mathbb{N}^q} (b_{(q,j)})^2 \end{aligned} \quad (51)$$

where

$$b_{(q,j)} = \int_{[-1,1]^q} m_q(x) \varphi_{(q,j)}(x) dx$$

and the polynomial $\varphi_{(q,j)}$ is obtained from (11). Now, obtaining upper bounds on the tail term in (51) is a standard exercise in multivariate approximation theory. Specifically, under Assumption 4.1, it can be shown that

$$\sum_{j \in \{1 \leq i \leq n\}^c \subset \mathbb{N}^q} (b_{(q,j)})^2 \leq \frac{K'_1}{n^{2s}}$$

see, for example, Canuto and Quarteroni [16] or Sheu [46, Theorem 4.2]. Finally, set $K_1 = M_q K'_1$; c) follows from (18) by setting $K_2 = 2^{(q+2)/2}$, and also since $k_2(q) \geq \lceil (N_{k_3})^{1/q} \rceil$; d) follows by setting $\delta = (N_{k_3})^{-D_4}$ for some $D_4 \geq 0$; e) since $n \leq \lceil (N_{k_3})^{1/q} \rceil$, follows by setting $K_3 = K_2 2^{3q/2}$ and $K_4 = (D_4 + 1)$; f) follows by setting $K_5 = \max\{K_4 \lambda, 2\lambda\}$ and by setting $K_6 = \max\{K_3^{1/K_4}, 2\}$; g) follows by setting $K_7 = 2K_5(\max\{\ln K_6, 1\})$; h) follows by setting $n = \lceil (N_{k_3}/(\ln N_{k_3}))^{1/(2s+q)} \rceil$, which takes values in the set $\{1, 2, \dots, \lceil (N_{k_3})^{1/q} \rceil\}$ for $N_{k_3} \geq 2$. \square

Proof of Theorem 4.1: Theorem 4.1 follows by combining Theorem 5.1 (for $p = q$) and Lemma 7.9 (for $k_3 = q$) in the manner of Theorem 4.2; we omit the details. \square

Proof of Theorem 4.2: It follows from our hypotheses that Assumptions 2.1, 2.2, and 3.2 hold, and from Example 5.1 that Assumptions 5.1 and 5.2 hold for all $1 \leq p \leq k_1$. Consequently, all hypotheses of Corollary 5.1 hold, and we have for all $0 < \delta(N) \leq \sqrt{3}$, for all $\lambda > 20/3$, for $q \leq k_1$, and for all $N_{k_1} \geq 2$

$$\begin{aligned} E[r(\hat{\mu}_N, m_q)] & = O(R_{(q,N)}(m_q, k_1)) + O(\delta(N)) + O(N_{k_1}^{-1}) \\ & \stackrel{\text{a)}}{=} O\left(\frac{\ln N_{k_1}}{N_{k_1}}\right)^{2s/(2s+q)} + O\left(\frac{1}{(N_{k_1})^{D_4}}\right) \\ & \stackrel{\text{b)}}{=} O\left(\frac{\ln N}{N}\right)^{(2s/(2s+q))(\beta/(\beta+1))} \end{aligned}$$

where a) follows by applying Lemma 7.9 with $k_3 = k_1$, $k_2(q) \geq \lceil (N_{k_1})^{1/q} \rceil$, and $\delta(N) = (N_{k_1})^{-D_4}$, where $D_4 \geq 0$; and b) follows if we let $D_4 \geq 1$, and from (4) by simple algebraic manipulations since $k_1 = o(N)$. \square

Proof of Theorem 4.3: Choose a small $\xi > 0$. We know by the L^2 martingale convergence theorem that $r(m_p, m_\infty)$ monotonically decreases to 0 as $p \rightarrow \infty$. Hence, there exists an integer \bar{p} such that

$$r(m_{\bar{p}}, m_\infty) \leq \xi/(2\bar{\eta}) \quad (52)$$

where constant $\bar{\eta}$ is as in the hypothesis of Theorem 5.2.

For $\mathbf{j} \in \mathbb{N}^{\bar{p}}$, define

$$b_{(\bar{p}, \mathbf{j})} = \int_{[-1, 1]^{\bar{p}}} m_{\bar{p}}(x) \varphi_{(\bar{p}, \mathbf{j})}(x) dx \quad (53)$$

where the polynomial $\varphi_{(\bar{p}, \mathbf{j})}(x)$ is obtained from (11). Write

$$\nu_n = (b_{(\bar{p}, \mathbf{1})}, \dots, b_{(\bar{p}, \mathbf{n})}) \quad (54)$$

where $\mathbf{1} = (1, 1, \dots, 1) \in \mathbb{N}^{\bar{p}}$ and $\mathbf{n} = (n, n, \dots, n) \in \mathbb{N}^{\bar{p}}$. It follows from Parseval's identity that

$$\sum_{\mathbf{j} \in \mathbb{N}^{\bar{p}}} (b_{(\bar{p}, \mathbf{j})})^2 = \int_{[-1, 1]^{\bar{p}}} (m_{\bar{p}}(x))^2 dx \leq 2^{\bar{p}} \quad (55)$$

where the last inequality follows since the range of $m_{\bar{p}}$ is $[-1, 1]$. Since the polynomial system $\{\varphi_{(\bar{p}, \mathbf{j})}\}_{\mathbf{j} \in \mathbb{N}^{\bar{p}}}$ is complete and orthonormal for the space of measurable, square-integrable (with respect to the Lebesgue measure) functions on $[-1, 1]^{\bar{p}}$, there exists a dimension \bar{n} such that

$$\int_{[-1, 1]^{\bar{p}}} \left[\sum_{\mathbf{j}=1}^{\bar{n}} b_{(\bar{p}, \mathbf{j})} \varphi_{(\bar{p}, \mathbf{j})}(x) - m_{\bar{p}}(x) \right]^2 dx \leq \xi/(2\bar{\eta}M_{\bar{p}}) \quad (56)$$

where $\bar{\mathbf{n}} = (\bar{n}, \bar{n}, \dots, \bar{n}) \in \mathbb{N}^{\bar{p}}$ and $M_{\bar{p}}$ denotes the uniform bound, which is finite by Assumption 4.3, on the probability density of the marginal distribution $P_{(1, \bar{p})}$. Since clip is continuous, we have from (13) that

$$\begin{aligned} & \int_{[-1, 1]^{\bar{p}}} [f_{(\bar{p}, (\bar{n}, \nu_{\bar{n}}))}(x) - m_{\bar{p}}(x)]^2 dx \\ & \leq \int_{[-1, 1]^{\bar{p}}} \left[\sum_{\mathbf{j}=1}^{\bar{n}} b_{(\bar{p}, \mathbf{j})} \varphi_{(\bar{p}, \mathbf{j})}(x) - m_{\bar{p}}(x) \right]^2 dx. \end{aligned} \quad (57)$$

We have from Assumption 4.3, (56), and (57) that

$$\begin{aligned} & r(f_{(\bar{p}, (\bar{n}, \nu_{\bar{n}}))}, m_{\bar{p}}) \\ & \leq M_{\bar{p}} \int_{[-1, 1]^{\bar{p}}} [f_{(\bar{p}, (\bar{n}, \nu_{\bar{n}}))}(x) - m_{\bar{p}}(x)]^2 dx \\ & \leq \xi/(2\bar{\eta}). \end{aligned} \quad (58)$$

The following sequence of upper bounds essentially completes the proof.

$$\begin{aligned} & E[r(\hat{m}_N, m_\infty)] \\ & \stackrel{\text{a)}}{\leq} \bar{\eta} R_{(\bar{p}, N)}(m_{\bar{p}}, k_1) + \bar{\eta} r(m_{\bar{p}}, m_\infty) + O(\delta(N)) \\ & \quad + O(1/N_{k_1}) \\ & \stackrel{\text{b)}}{\leq} \bar{\eta} R_{(\bar{p}, N)}(m_{\bar{p}}, k_1) + \xi/2 + O(1/N_{k_1}) \end{aligned}$$

$$\begin{aligned} & \stackrel{\text{c)}}{=} \bar{\eta} \min_{1 \leq n \leq k_2(\bar{p})} \left\{ \min_{\nu \in S_{(\bar{p}, n)}} [r(f_{(\bar{p}, (n, \nu))}, m_{\bar{p}})] \right. \\ & \quad \left. + \lambda \frac{L_{(\bar{p}, n)}(\varpi_{(\bar{p}, n)}^{-1}(\delta(N))) + 2 \ln(n+1)}{N_{k_1}} \right\} \\ & \quad + \xi/2 + O(1/N_{k_1}) \\ & \stackrel{\text{d)}}{\leq} \bar{\eta} \min_{\nu \in S_{(\bar{p}, \bar{n})}} [r(f_{(\bar{p}, (\bar{n}, \nu))}, m_{\bar{p}})] \\ & \quad + \bar{\eta} \lambda \frac{L_{(\bar{p}, \bar{n})}(\varpi_{(\bar{p}, \bar{n})}^{-1}(\delta(N))) + 2 \ln(\bar{n}+1)}{N_{k_1}} \\ & \quad + \xi/2 + O(1/N_{k_1}) \\ & \stackrel{\text{e)}}{\leq} \bar{\eta} r(f_{(\bar{p}, (\bar{n}, \nu_{\bar{n}}))}, m_{\bar{p}}) \\ & \quad + \bar{\eta} \lambda \frac{L_{(\bar{p}, \bar{n})}(\varpi_{(\bar{p}, \bar{n})}^{-1}(\delta(N))) + 2 \ln(\bar{n}+1)}{N_{k_1}} \\ & \quad + \xi/2 + O(1/N_{k_1}) \\ & \stackrel{\text{f)}}{\leq} \xi + \bar{\eta} \lambda \\ & \quad \cdot \frac{\bar{\eta} \ln[2^{(\bar{p}+2)/2} \bar{\eta}^{\bar{p}/2} (2\bar{n}+1)^{\bar{p}/2} (N_{k_1})^{D_4}] + 2 \ln(\bar{n}+1)}{N_{k_1}} \\ & \quad + O(1/N_{k_1}) \\ & \stackrel{\text{g)}}{\leq} \xi + O\left(\frac{\ln N_{k_1}}{N_{k_1}}\right) \end{aligned} \quad (59)$$

where a) follows by invoking Theorem 5.2 for $p = \bar{p}$, for all $0 < \delta(N) \leq \sqrt{3}$, for all $\lambda > 20/3$, for all $k_2(\bar{p}, N)$, for all $N_{k_1} \geq 2$, and for all large N such that $k_1(N) \geq \bar{p}$; b) follows from (52) and by setting $\delta(N) = (N_{k_1})^{-D_4}$, where $D_4 \geq 1$; c) follows from (21), (23), and (26), where $S_{(\bar{p}, n)}$ is obtained from (14), $f_{(\bar{p}, (\bar{n}, \nu_{\bar{n}}))}$ is obtained from (13), $L_{(\bar{p}, \bar{n})}$ is obtained from (18), and $\varpi_{(\bar{p}, \bar{n})}^{-1}$ is obtained from (20); d) holds for all large N such that $k_1(N) \geq \bar{p}$ and $k_2(\bar{p}, N) \geq \bar{n}$; e) follows since we have from (14), (54), and (55), that $\nu_{\bar{n}} \in S_{(\bar{p}, \bar{n})}$; f) follows from (58) and from (18) and (20); and g) follows by simple algebraic manipulations.

Since we may choose ξ as small as desired, and since $N_{k_1} \rightarrow \infty$ (since $k_1 = o(N)$), the theorem follows from (59). \square

ACKNOWLEDGMENT

The authors wish to thank S. Altekar, R. Hecht-Nielsen, H. Mhaskar, J. Rissanen, and H. White for valuable discussions.

REFERENCES

- [1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.
- [2] P. H. Algoet, "Universal schemes for prediction, gambling, and portfolio selection," *Ann. Probab.*, vol. 20, no. 2, pp. 901–941, 1992. Correction: *ibid.*, vol. 23, pp. 474–478, 1995.
- [3] ———, "The strong law of large numbers for sequential decisions under uncertainty," *IEEE Trans. Inform. Theory*, vol. 40, pp. 609–633, May 1994.
- [4] K. B. Athreya and S. G. Pantula, "Mixing properties of Harris chains and autoregressive processes," *J. Appl. Probab.*, vol. 23, pp. 880–892, 1986.
- [5] B. Auestad and D. Tjøstheim, "Identification of nonlinear time series: First order characterization and order determination," *Biometrika*, vol. 77, no. 4, pp. 669–687, 1990.

- [6] ———, “Functional identification in nonlinear time series,” in *Proceedings NATO Advanced Study Institute on Nonparametric Functional Estimation*, G. Roussas, Ed. Dordrecht, The Netherlands: Kluwer, 1991.
- [7] A. D. Back and A. C. Tsoi, “FIR and IIR synapses, a new neural network architecture for time series modeling,” *Neural Comput.*, vol. 3, pp. 375–385, 1991.
- [8] A. R. Barron, “Complexity regularization,” in *Proceedings NATO Advanced Study Institute on Nonparametric Functional Estimation*, G. Roussas, Ed. Dordrecht, The Netherlands: Kluwer, 1991, pp. 561–576.
- [9] ———, “Universal approximation bounds for superpositions of a sigmoidal function,” *IEEE Trans. Inform. Theory*, vol. 39, pp. 930–945, May 1993.
- [10] ———, “Approximation and estimation bounds for artificial neural networks,” *Mach. Learn.*, vol. 14, pp. 115–133, 1994.
- [11] ———, “Asymptotically optimal model selection and neural nets,” in *Proc. 1994 IEEE-IMS Workshop on Information Theory and Statistics*. New York: IEEE Press, Oct. 1994, p. 35.
- [12] A. R. Barron, L. Birgé, and P. Massart, “Risk bounds for model selection via penalization,” 1995, preprint.
- [13] A. R. Barron and T. M. Cover, “Minimum complexity density estimation,” *IEEE Trans. Inform. Theory*, vol. 37, pp. 1034–1054, July 1991.
- [14] R. J. Bhansali, “Order selection for linear time series models: A review,” in *Developments in Time Series Analysis: In Honour of M. B. Priestley*, T. Subba Rao, Ed. New York: Chapman and Hall, 1993, pp. 50–66.
- [15] D. Bosq, *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. New York: Springer-Verlag, 1996.
- [16] C. Canuto and A. Quarteroni, “Approximation results for orthogonal polynomials in Sobolev spaces,” *Math. of Computation*, vol. 38, pp. 67–86, 1982.
- [17] B. Cheng and H. Tong, “On consistent nonparametric order determination and chaos,” *J. Roy. Statist. Soc., B*, vol. 54, no. 2, pp. 427–449, 1992.
- [18] T. M. Cover, “Open problems in information theory,” in *Proc. 1975 IEEE-USSR Joint Workshop on Information Theory*, (Moscow, USSR, Dec. 15–19, 1975). New York: IEEE Press, 1975.
- [19] Y. A. Davydov, “Mixing conditions for Markov chains,” *Theory Probab. Applications*, vol. XVIII, no. 2, pp. 312–328, 1973.
- [20] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.
- [21] P. Doukhan, *Mixing: Properties and Examples*. New York: Springer-Verlag, 1994.
- [22] R. Durrett, *Probability: Theory and Examples*. Pacific Grove, CA: Wadsworth & Brooks, 1991.
- [23] A. Farago and G. Lugosi, “Strong universal consistency of neural network classifiers,” *IEEE Trans. Inform. Theory*, vol. 39, pp. 1146–1151, July 1993.
- [24] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Application*. New York: Academic, 1980.
- [25] 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.
- [26] J. M. Hutchinson, A. W. Lo, and T. Poggio, “A nonparametric approach to pricing and hedging derivative securities via learning networks,” *J. Finance*, vol. XLIX, no. 3, July 1994.
- [27] L. K. Jones, “The computational intractability of training sigmoidal neural networks,” *IEEE Trans. Inform. Theory*, vol. 43, pp. 167–173, Jan. 1997.
- [28] G. Lugosi and A. Nobel, “Adaptive model selection using empirical complexities,” 1995, preprint.
- [29] G. Lugosi and K. Zeger, “Nonparametric estimation via empirical risk minimization,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 677–687, May 1995.
- [30] ———, “Concept learning using complexity regularization,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 48–54, Jan. 1996.
- [31] P. Masani and N. Wiener, “Non-linear prediction,” in *Probability and Statistics: The Harald Cramér Volume*, U. Grenander, Ed. Stockholm, Sweden: Almqvist and Wiksell, 1959, pp. 190–212.
- [32] D. F. McCaffrey and A. R. Gallant, “Convergence rates for single hidden layer feedforward networks,” *Neural Net.*, vol. 7, no. 1, pp. 147–158, 1994.
- [33] R. W. Means, B. Wallach, D. Busby, and R. Lengel, Jr., “Bispectrum signal processing on HNC’s SIMD numerical array processor (SNAP),” in *1993 Proc. Supercomputing* (Portland, OR, Nov. 15–19, 1993), pp. 535–537; Los Alamitos, CA: IEEE Comput. Soc. Press, 1993.
- [34] D. S. Modha and E. Masry, “Minimum complexity regression estimation with weakly dependent observations,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 2133–2145, Nov. 1996.
- [35] G. Morvai, S. Yakowitz, and P. Algoet, “Weakly convergent nonparametric forecasting of nonparametric stationary time series,” 1996, submitted for publication.
- [36] G. Morvai, S. Yakowitz, and L. Györfi, “Nonparametric inferences for ergodic, stationary time series,” *Ann. Statist.*, vol. 24, no. 1, pp. 370–379, 1996.
- [37] J. Rissanen, “A universal data compression system,” *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 656–664, Sept. 1983.
- [38] ———, “Complexity of strings in the class of Markov sources,” *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 526–532, July 1986.
- [39] ———, *Stochastic Complexity in Statistical Inquiry*. Teaneck, NJ: World Scientific, 1989.
- [40] P. M. Robinson, “Nonparametric estimators for time series,” *J. Time Series Anal.*, vol. 4, pp. 185–297, 1983.
- [41] G. G. Roussas, “Nonparametric regression estimation under mixing conditions,” *Stochastic Process. Appl.*, vol. 36, pp. 107–116, 1990.
- [42] M. Rosenblatt, “A central limit theorem and strong mixing conditions,” *Proc. Nat. Acad. Sci.*, vol. 4, pp. 43–47, 1956.
- [43] B. Y. Ryabko, “Twice-universal coding,” *Probl. Inform. Transm.*, vol. 20, pp. 173–177, July–Sept. 1984.
- [44] ———, “Prediction of random sequences and universal coding,” *Probl. Inform. Transm.*, vol. 24, pp. 87–96, Apr.–June 1988.
- [45] E. Sackinger and H. P. Graf, “A board system for high-speed image analysis and neural networks,” *IEEE Trans. Neural Net.*, vol. 7, pp. 214–221, Jan. 1996.
- [46] C.-H. Sheu, “Density estimation with Kullback–Leibler loss,” Ph.D. dissertation, Dept. Statistics, Univ. Illinois at Urbana-Champaign, 1990.
- [47] C.-Y. Sin and H. White, “Information criteria for selecting possibly misspecified parametric models,” 1995, preprint.
- [48] C. J. Stone, “Consistent nonparametric regression (with discussion),” *Ann. Statist.*, vol. 5, pp. 549–645, 1977.
- [49] G. Szegő, *Orthogonal Polynomials*. New York: American Math. Soc., 1939.
- [50] D. Tjøstheim, “Non-linear time series: A selective review,” *Scandinavian J. Statist.*, vol. 21, pp. 97–130, 1994.
- [51] V. Vapnik, *Estimation of Dependences Based on Empirical Data*. New York: Springer-Verlag, 1982.
- [52] ———, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
- [53] E. A. Wan, “Discrete time neural networks,” *J. Appl. Intell.*, vol. 3, pp. 91–105, 1993.
- [54] A. S. Weigend and N. A. Gershenfeld, “The future of time series: Learning and understanding,” in *Time Series Prediction: Forecasting the Future and Understanding the Past: Proc. NATO Advanced Research Workshop on Comparative Time Series Analysis*, A. S. Weigend and N. A. Gershenfeld, Eds. Reading, MA: Addison-Wesley, 1994.
- [55] H. White, “Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings,” *Neural Net.*, vol. 3, pp. 535–549, 1989.
- [56] ———, “Nonparametric estimation of conditional quantiles using neural networks,” in *Computing Science and Statistics, Statistics of Many Parameters: Curves, Images, Spatial Models, Proc. 22nd Symp. on the Interface*, C. Page and R. LePage, Eds. New York: Springer-Verlag, 1992, pp. 190–199.
- [57] 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. in Economic Theory and Econometrics*, W. A. Barnett, J. Powell, and G. Tauchen, Eds. New York: Cambridge Univ. Press, 1991.
- [58] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.
- [59] C. S. Withers, “Conditions for linear processes to be strong-mixing,” *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 57, pp. 477–480, 1981.
- [60] L. Wu, M. Niranjan, and F. Fallside, “Fully vector-quantized neural network-based code-excited nonlinear predictive speech coding,” *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 482–489, Oct. 1994.
- [61] Y. Yang and A. R. Barron, “An asymptotic property of model selection criteria,” 1995, preprint.
- [62] J. E. Yukich, M. B. Stinchcombe, and H. White, “Sup-norm approximation bounds for networks through probabilistic methods,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 1021–1027, July 1995.