Machine Learning, 39, 5–34, 2000.

Nonparametric Time Series Prediction Through Adaptive Model Selection*

RON MEIR† rmeir@ee.technion.ac.il
Department of Electrical Engineering, Technion, Haifa 32000, Israel

Editor: Lisa Hellerstein

Abstract. We consider the problem of one-step ahead prediction for time series generated by an underlying stationary stochastic process obeying the condition of absolute regularity, describing the mixing nature of process. We make use of recent results from the theory of empirical processes, and adapt the uniform convergence framework of Vapnik and Chervonenkis to the problem of time series prediction, obtaining finite sample bounds. Furthermore, by allowing both the model complexity and memory size to be adaptively determined by the data, we derive nonparametric rates of convergence through an extension of the method of structural risk minimization suggested by Vapnik. All our results are derived for general L_p error measures, and apply to both exponentially and algebraically mixing processes.

Keywords: time series prediction, adaptive model selection, structural risk minimization, mixing processes

1. Introduction

The problem of time series modeling and prediction has a long history, dating back to the pioneering work of Yule in 1927 (Yule, 1927). Most of the work since then until the 1970s has been concerned with parametric approaches to the problem whereby a simple, usually linear, model is fitted to the data (for a review of this approach, see for example the text-book by Brockwell and Davis (1991)). While many appealing mathematical properties of the parametric approach have been established, it has become clear over the years that the limitations of the approach are rather severe, in their imposition of a rigid structure on the process. One of the more productive solutions to this problem has been the extension of the classic nonparametric methods to the case of time series (see, for example, Györfi et al. (1989) and Bosq (1996) for a review). In this work we use the term *parametric model* to refer to any model which imposes a specific *form* on the estimated function, which is exactly known up to a finite number of parameters. Nonparametric models, on the the other hand, do not impose any structural assumptions, and can model any (smooth) underlying process.

In this work we consider a third approach to the problem of time series prediction, which although nonparametric in spirit, possesses many affinities with the parametric approach.

*This work was supported in part by a grant from the Israel Science Foundation. Support from the Ollendorff center of the department of Electrical Engineering at the Technion is also acknowledged.
†Part of this work was done while the author was visiting the Isaac Newton Institute, Cambridge, England.

The method is strongly related to the method of sieves, introduced by Grenander (1981) and studied further by Geman and Hwang (1982). This type of approach had in fact been introduced in the late 1970's by Vapnik and titled by him *Structural Risk Minimization* (SRM) (Vapnik, 1982). The basic idea behind this approach, applied so far in the context of independent data, is the construction of a sequence of models of increasing complexity, where each model within the hierarchy is usually taken to be parametric. As the complexity of the model increases, manifested by a growing complexity index, the model approximates a very rich, nonparametric, class of functions. One of the seminal contributions of the work by Vapnik and Chervonenkis (1971) was the establishment of upper bounds on the true performance of estimators within each class, based on a natural complexity index, which has come to be known as the VC dimension. These bounds contain two terms, the first of which is the empirical error on the training data, the second being a complexity term penalizing overly complex models. As can be expected, the empirical error decreases as the complexity of the class increases, while the second, complexity term, naturally increases. Thus, it was suggested in (Vapnik, 1982) that for each sample size one may obtain the best trade-off between the two terms, thus achieving the optimal guaranteed performance bounds for any sample size. Moreover, by tracking the optimal complexity for each sample size it was shown that very large families of functions may be modeled in this fashion. The major advantage of this type of approach is that on the one hand it is nonparametric in nature, in that very large classes of functions may be modeled, while at the same time being adaptive. By adaptive we refer to the following situation: assume that the function to be modeled in fact belongs to the sequence of models under consideration. In that case, one would like the estimation scheme to converge to the true model at a rate which is similar to the one that would be attained had we known the true model in advance. In fact, exactly this type of adaptivity has been demonstrated recently for the case of classification (Lugosi & Zeger, 1996), regression (Lugosi & Nobel, 1996) and data compression (Feder & Merhav, 1996). We should also note that a similar approach based on the so-called index of resolvability has been pursued by Barron and co-workers in a series of papers (Barron & Cover, 1991; Barron, 1994), with similar results. The major advantage of these approaches is that while being adaptive in the above sense, they can often be shown to achieve the minimax rates of convergence in nonparametric settings (Stone, 1982) under i.i.d. conditions, showing that they are effective estimation schemes in this regime as well.

In this work we extend the SRM idea to the case of time series. This extension is not entirely straightforward for several reasons. First, even within a single parametric model, the problem of deriving robust finite sample bounds is exacerbated by the dependence inherent in the process. Generalizing the basic tools of uniform convergence of empirical measures utilized in the i.i.d. setting requires the introduction of new methods. In particular, it should be clear that assumptions concerning the dependence structure of the process must be taken into account, and quantified in a precise manner. Second, in opposition to the case of regression, the dimension of the input vector is not fixed, as the process may possess a very long memory. Thus, any universal prediction scheme must allow the prediction to be based on potentially unlimited memory. By memory size we roughly refer to the number of past values of the process, needed to achieve the optimal prediction error; this term will be defined precisely in Section 2. Finally, the optimal balance between the complexity of

the model and the memory size used for prediction must be determined. We observe that our results bear strong affinities to the approach taken by Modha and Masry (1998), while deviating from them in scope and methodology; see Remark 7 in Section 6 for a detailed comparison. An additional related work is the one by Campi and Kumar (1998), which deals with the problem of learning dynamical systems in a stationary environment. In this case an input/output mapping of a fixed input dimension is learned, and estimation error bounds are given for the L_2 loss.

Finally, before outlining the remainder of the paper, we comment on the relevance of this work to Machine Learning. Clearly, many of the problems to which Machine Learning techniques are applied are inherently temporal in nature. Some obvious examples are stock market prediction, analysis of financial markets, monitoring and diagnosing complex control systems and speech recognition, to name but a few. Until recently most of the theoretical results within the PAC (Probably Approximately Correct) approach to learning have dealt with situations in which time played no role. In fact, the problem of extending the PAC framework to time series is the first 'open problem' mentioned in the recent monograph of Vidyasagar (1996). One approach to incorporating temporal structure in order to form better predictors, by more appropriate complexity regularization, is described in this work. In particular, the optimal memory size that should be used in order to form a predictor is in principle derivable from the procedure (see also (Modha & Masry, 1998)), given information about the mixing nature of the time series (see Section 4 for a definition of mixing). It is thus hoped that many of the successful Machine Learning approaches to modeling static data will be extended to time series, with the benefit of a solid mathematical framework. If precise knowledge of the mixing parameters is lacking, the procedure requires estimation of these parameters. Unfortunately, as far as we are aware, there is no efficient practical approach known at this stage for estimation of mixing parameters.

Another related and very fruitful line of recent research has been devoted to the so called on-line approach to learning, where very few assumptions are made about the data (see (Blum, 1996) for a recent survey). In the most extreme case, no assumptions whatsoever are made, and an attempt is made to compare the performance of various on-line algorithms to that of the best algorithm within some class. Prediction from expert advice and competitiveness with some comparison class are two well-studied examples within this broad field. While the assumptions in these latter approaches are very weak, it should be noted that they address a different question from the one studied in this work. Here we are concerned with establishing consistency and rates of convergence for general (off-line) algorithms, under specific statistical assumptions about the data, while the on-line work is usually concerned with comparing on-line performance to some other approach, for which performance bounds are usually not given. In fact, one can use the on-line approach to study how well these algorithms approximate the off-line algorithms studied here.

The remainder of the paper is organized as follows. In Section 2 we introduce the problem of time series prediction in a general context, discussing the basic trade-off between approximation and estimation. In Section 3 we present a brief review of some uniform convergence results for the case of i.i.d. data, which will serve as a basis for the derivation of results in the context of dependent data. Section 4 introduces the notion of mixing processes, and presents several results, mainly due to Yu (1994), establishing uniform laws of large numbers for

these processes. Section 5 then proceeds to consider the problem of prediction for scalar mixing stochastic processes, restricting the results to a single model class. Utilizing the results of Section 4, a particular estimator is shown to be consistent, and finite sample performance bounds are derived. We then proceed in Section 6 to consider a hierarchy of model classes as in the method of structural risk minimization, and present an algorithm which adaptively determines, for each sample size, an optimal value for both the complexity of the model and the memory size used. A short discussion and list of open questions concludes the paper in Section 7. Some of the proofs have been relegated to the appendix. We comment that in the sequel we will make use of the terms 'loss' and 'error' interchangeably.

2. The problem of time series prediction

Consider a stationary stochastic process $\bar{X} = \{\ldots, X_{-1}, X_0, X_1, \ldots\}$, where X_i is a real-valued random variable such that $|X_i| \leq B$ with probability 1, for some positive constant $B < \infty$. The problem of one-step prediction, in the expected L_p norm sense, can then be phrased as that of computing a *predictor* function $f(\cdot)$ of the infinite past such that $\mathbb{E}\{|X_0 - f(X_{-\infty}^{-1})|^p\}$ is minimal, where we use the notation $X_i^j = (X_i, X_{i+1}, \ldots, X_j)$, $j \geq i$. It is well known that for the special case $p = 2$, the optimal predictor is given by the conditional mean, $\mathbb{E}[X_0 \mid X_{-\infty}^{-1}]$. While this solution, in principle, settles the issue of optimal prediction, it does not settle the issue of actually computing the optimal predictor. First of all, note that to compute the conditional mean, the probabilistic law generating the stochastic process $\bar{X}$ must be known. Furthermore, this computation is usually intractable for non-trivial conditional densities. Finally, the requirement of knowing the full past, $X_{-\infty}^{-1}$ is of course rather stringent. In the case $p > 2$, the problem is further exacerbated in that there does not even exist a formal analytic solution as in the case $p = 2$. In this work, we consider the more practical situation, where a *finite* sub-sequence $X_1^N = (X_1, X_2, \ldots, X_N)$ is observed, and an optimal prediction is needed, conditioned on this data. Moreover, we allow for a sequence of model classes, in each of which the prediction is based on a finite number of past values. We denote this number by d, and refer to it as the *memory size*. Since the process may in principle possess infinite memory, in order to achieve full generality we may let $d \to \infty$ in order to obtain the optimal predictor. Of course this can only be done when the sample size $N \to \infty$, as the constraint $d \leq N$ must obviously be obeyed.

For each fixed value of the memory size d, we consider the problem of selecting an empirical estimator from a class of functions $\mathcal{F}_{d,n} : \mathbb{R}^d \to \mathbb{R}$, $|f| \leq B$ for $f \in \mathcal{F}_{d,n}$, where n is a complexity index of the class. For example, n may stand for the number of computational nodes in the single hidden layer of a feedforward neural network with d inputs, namely

$$\mathcal{F}_{d,n} = \left\{ f : f(x) = \sum_{i=1}^{n} c_i \sigma\left(a_i^T x + b_i\right) c_i \in \mathbb{R}, a_i \in \mathbb{R}^d, b_i \in \mathbb{R} \right\},$$

where σ is some activation function. Other classes could include radial basis functions and multi-variate splines with a variable number of knots, to name but a few.

Consider then an empirical predictor $\hat{f}_{d,n,N}(X_{i-d}^{i-1})$, $i > N$, for X_i based on the finite data vector X_1^N and depending on the d-dimensional vector X_{i-d}^{i-1}, where $\hat{f}_{d,n,N} \in \mathcal{F}_{d,n}$. It is

possible to split the error incurred by this predictor into three terms, each possessing a rather intuitive meaning. It is the competition between these terms which determines the optimal solution, for a *fixed* amount of data. First, define the loss of a predictor $f_d : \mathbb{R}^d \to \mathbb{R}$ as

$$L(f_d) = \mathbb{E}\left|X_i - f_d\left(X_{i-d}^{i-1}\right)\right|^p. \tag{1}$$

Observe that due to stationarity $L(f_d)$ is independent of i. Let f_d^* be the optimal predictor of memory size d minimizing the loss (1), namely

$$\mathbb{E}\left|X_i - f_d^*\left(X_{i-d}^{i-1}\right)\right|^p = \inf_{f:\mathbb{R}^d\to\mathbb{R}} \mathbb{E}\left|X_i - f\left(X_{i-d}^{i-1}\right)\right|^p,$$

and denote the error incurred by this function by L_d^*. We say that predictor has *finite memory* if

$$L\left(f_d^*\right) = L\left(f_\infty^*\right),$$

for some $d < \infty$, namely, the minimal prediction error may be achieved by a finite memory size d. Similarly, for the class $\mathcal{F}_{d,n}$ we denote by $f_{d,n}^*$ the optimal predictor within the class, namely

$$\mathbb{E}\left|X_i - f_{d,n}^*\left(X_{i-d}^{i-1}\right)\right|^p = \inf_{f\in\mathcal{F}_{d,n}} \mathbb{E}\left|X_i - f\left(X_{i-d}^{i-1}\right)\right)\right|^p,$$

denoting the resulting loss by $L_{d,n}^*$. We assume throughout the paper that $f_{d,n}^*$ exists, amounting to a compactness assumption about $\mathcal{F}_{d,n}$. If this assumption does not hold, we may simply add an arbitrarily small term to the r.h.s. of the equation. Observe that f_d^* need not in general belong to the class $\mathcal{F}_{d,n}$, due to its limited expressive power. Furthermore, denote by $\hat{f}_{d,n,N}$ an empirical estimator based on the finite data set X_1^N. We find it useful to express the error as a sum of three terms, each of which possesses a clear intuitive meaning,

$$L(\hat{f}_{d,n,N}) = \left(L(\hat{f}_{d,n,N}) - L_{d,n}^*\right) + \left(L_{d,n}^* - L_d^*\right) + L_d^*. \tag{2}$$

The third term, L_d^*, often referred to as the dynamic *miss-specification error*, is related to the error incurred in using a finite memory model (of memory size d), to predict a process with potentially infinite memory. We do not at present have any useful upper bounds for this term, which is related to the rate of convergence in the martingale convergence theorem, which to the best of our knowledge in unknown for the type of mixing processes we study in this work. The second term in (2), is related to the so-called *approximation error*, given by $\mathbb{E}\{|f_d^*(X_{i-d}^{i-1}) - f_{d,n}^*(X_{i-d}^{i-1})|^p\}$ to which it can be immediately related through the inequality $\|a|^p - |b|^p| \le p|a-b|\max(a,b)|^{p-1}$. This term measures the excess error incurred by selecting a function f from a class of limited complexity $\mathcal{F}_{d,n}$, while the optimal predictor of memory size d, namely f_d^*, may be arbitrarily complex. Of course, in order to bound this term we will have to make some regularity assumptions about the latter function. Finally, the first term in (2) represents the so called *estimation error*, and is the only term which depends on

the data X_1^N. Similarly to the problem of regression for i.i.d. data, we expect that the approximation and estimation terms lead to conflicting demands on the choice of the the complexity, n, of the functional class $\mathcal{F}_{d,n}$. Clearly, in order to minimize the approximation error the complexity should be made as large as possible. However, doing this will cause the estimation error to increase, because of the larger freedom in choosing a specific function in $\mathcal{F}_{d,n}$ to fit the data. However, in the case of time series there is an additional complication resulting from the fact that the misspecification error L_d^* is minimized by choosing d to be as large as possible, while this has the effect of increasing both the approximation as well as the estimation errors. We thus expect that some optimal values of d and n exist for each sample size N.

Up to this point, we have not specified how to select the empirical estimator $\hat{f}_{d,n,N}$. In this work we follow the ideas of Vapnik & Chervonenkis (1971), which have been studied extensively in the context of i.i.d observations, and restrict our selection to that function which minimizes the empirical error, given by

$$\hat{L}_N(f) = \frac{1}{N-d} \sum_{i=d+1}^{N} \left| X_i - f\left(X_{i-d}^{i-1}\right) \right|^p. \tag{3}$$

Thus, $\hat{f}_{d,n,N} = \text{argmin}_{f \in \mathcal{F}_{d,n}} \hat{L}_N(f)$. Again, we assume that $\hat{f}_{d,n,N}$ exists. It is a simple matter to modify this assumption by demanding that $\hat{f}_{d,n,N}$ only minimize the empirical error within some margin which is allowed to shrink as $N \to \infty$. For the sake of clarity we do not proceed in this direction. For this function, it is easy to establish the following result (see for example Lemma 8.2 in (Devroye, Györfi, & Lugosi, 1996), the proof of which does not depend on the independence property).

Lemma 2.1. *Let $\hat{f}_{d,n,N}$ be a function in $\mathcal{F}_{d,n}$ which minimizes the empirical error. Then*

$$L(\hat{f}_{d,n,N}) - \inf_{f \in \mathcal{F}_{d,n}} L(f) \leq 2 \sup_{f \in \mathcal{F}_{d,n}} |L(f) - \hat{L}_N(f)|.$$

It is obvious from Lemma 2.1 that the estimation error will vanish in the limit $N \to \infty$ if some form of uniform law of large numbers can be established. The latter will depend on the properties of the stochastic process $\bar{X}$, as well as on the attributes of the functional space $\mathcal{F}_{d,n}$. These issues will be addressed in Section 4. The main distinction here from the i.i.d. case, of course, is that random variables appearing in the empirical error, $\hat{L}_N(f)$, are no longer independent. It is therefore clear that some assumptions are needed regarding the stochastic process $\bar{X}$, in order that a uniform law of large numbers may be established. In any event, it is obvious that the standard approach of using randomization and symmetrization as in the i.i.d case (Pollard, 1984) will not work here. To circumvent this problem, two approaches have been proposed. The first makes use of extensions of the Bernstein inequality to dependent data (White, 1991; Modha & Masry, 1998). The second approach, to be pursued here, is based on mapping the problem onto one characterized by an i.i.d. process (Yu, 1994), and the utilization of the standard results for the latter case.

A comment is in order here concerning notation. Hatted variables will denote empirical estimates, while starred variables denote optimality with respect to the true (unknown)

distribution. Moreover, let $f(x)$ be a function defined over a domain χ. Then $L_p(Q)$ represents the Q-weighted L_p norm $\|f\|_{L_p(Q)} = (\int |f(x)|^p Q(x)\,dx)^{1/p}$, and l_p represents the empirical p'th order semi-norm $\|f\|_{l_p} = (N^{-1}\sum_{i=1}^{N} |f(X_i)|^p)^{1/p}$, where $\{X_1, \ldots, X_N\}$ is a given set of points defined over the domain χ.

3. Uniform convergence results for independent processes

A powerful tool used in recent years to establish both consistency and rates of convergence of empirical estimators, is provided by the theory of empirical processes (Pollard, 1984; Vaart & Wellner, 1996). Unfortunately, most of the results in this field are geared towards the case of memoryless processes, and are thus not directly suited to the study of time series. In this section, we summarize some of the basic results concerning uniform convergence for independent processes, and then present in Section 4 a recent result by Yu (1994) for dependent processes, which we will make extensive use of in the sequel.

We begin with a result concerning the uniform convergence of empirical measures to their expected value, for the case where the data is independent and identically distributed. Let $X \in \chi \subseteq \mathbb{R}^d$ be a vector-valued random variable, drawn according to some probability distribution P. Consider N independently drawn random variables $X^N = \{X_1, \ldots, X_N\}$, each drawn according to P. Let $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^+$ be a class of functions, and denote expectations with respect to P by $\mathbb{E}$. Furthermore, let P_N represent the empirical distribution, i.e., for any measurable set $A \subseteq \mathcal{B}(\chi)$,

$$P_N(X \in A) = \frac{1}{N}\sum_{i=1}^{N} I_A(X_i),$$

where $I_A(\cdot)$ is the indicator function for the set A. Denote expectations with respect to P_N by $\mathbb{E}_N$. Thus for any function $f(\cdot)$, $\mathbb{E}_N f = (1/N)\sum_{i=1}^{N} f(X_i)$.

A major tool for discussing uniform convergence within functional classes is the so called *covering number* of the class, which roughly measures how well the set can be covered by a finite subset of functions, using some specified distance measure. Formally we have (Pollard (1984)),

Definition 1. Let $\mathcal{F}$ be a class of real valued functions from χ to $\mathbb{R}$, and denote by ρ a semi-norm on $\mathcal{F}$. For each $\epsilon > 0$ define the covering number $\mathcal{N}(\epsilon, \mathcal{F}, \rho)$ as the smallest value of m for which there exist functions $g_1, g_2, \ldots, g_m$ (not necessarily in $\mathcal{F}$) such that $\min_j \rho(f, g_j) < \epsilon$ for every $f \in \mathcal{F}$. If no such finite m exists then $\mathcal{N}(\epsilon, \mathcal{F}, \rho) = \infty$.

In the sequel we will make extensive use of empirical covering numbers. Let $X^N = \{X_1, X_2, \ldots, X_N\}$ be points in χ and denote the empirical $l_{1,N}$ distance by

$$l_{1,N}(f, g) = \frac{1}{N}\sum_{i=1}^{N} |f(X_i) - g(X_i)|.$$

Moreover, let $\mathcal{F}(X^N) = \{(f(X_1),\ f(X_2), \ldots,\ f(X_N)) : f \in \mathcal{F}\}$. We denote the covering number of $\mathcal{F}$ with respect to the semi-norm $l_{1,N}$ by $\mathcal{N}(\epsilon, \mathcal{F}(X^N), l_{1,N})$, which clearly depends on the specific set of points $\{X_1, X_2, \ldots, X_N\}$ considered.

Since the functional classes considered here are in general uncountable, some conditions are required in order to avoid measurability problems. Following common practice, we assume throughout that all function classes are *permissible* in the sense specified in Pollard (1984). We then have,

Lemma 3.1 (*Pollard, 1984*)**.** *Let $\mathcal{F}$ be a permissible class of real-valued non-negative functions such that $f(x) \leq B$ for all $f \in \mathcal{F}$. Then*

$$\mathbb{P}\left\{X^N \in \chi^N : \sup_{f\in\mathcal{F}} |\mathbb{E}_N f - \mathbb{E} f| > \epsilon\right\} \leq 4\mathbb{E}\left\{\mathcal{N}\left(\frac{\epsilon}{16}, \mathcal{F}(X^N), l_{1,N}\right)\right\} \exp\left\{-\frac{N\epsilon^2}{128B^2}\right\},$$

where the expectation is taken with respect to a sample of size N drawn independently at random from the distribution P.

In most cases of interest for regression or time series analysis, one is actually interested in working in the space of loss functions, as in Haussler (1992). Consider then N randomly drawn pairs $\{(X_i, Y_i)\}$, $(X_i, Y_i) \in (\chi, \mathcal{Y})$, $i = 1, \ldots, N$, where each pair is drawn according to the distribution $P(X, Y)$ (we avoid cluttering the notation using $P_{X,Y}(\cdot, \cdot)$, as the particular distribution will be clear from its argument). For each $f \in \mathcal{F}$, let $\ell_f(x, y) : \mathbb{R}^{d+1} \to \mathbb{R}^+$ be a non-negative function termed the *loss function*, and define the *loss space* $\mathcal{L}_\mathcal{F}$

$$\mathcal{L}_\mathcal{F} = \left\{\ell_f(x, y) : x \in \mathbb{R}^d,\ y \in \mathbb{R},\ f \in \mathcal{F}\right\}.$$

The covering numbers for the spaces $\mathcal{F}$ and $\mathcal{L}_\mathcal{F}$ can be easily related if a certain Lipschitz condition is obeyed by the loss functions $\ell_f(x, y)$. In particular, assume that for all y, x_1, x_2 and f

$$|\ell_f(x_1, y) - \ell_f(x_2, y)| \leq \eta |f(x_1) - f(x_2)|.$$

Then it can easily be shown (Vidyasagar, 1996, Sec. 7.1.3) that

$$\mathcal{N}\big(\epsilon, \mathcal{L}_\mathcal{F}(Z^N), l_{1,N}\big) \leq \mathcal{N}\left(\frac{\epsilon}{\eta}, \mathcal{F}(X^N), l_{1,N}\right), \tag{4}$$

where $Z^N = \{Z_1, \ldots, Z_N\} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$. Note that the empirical semi-norm $l_{1,N}$ on the l.h.s. of (4) is taken with respect to both X and Y. In the case where $\ell_f(x, y) = |y - f(x)|^p$, it is easy to see that $\eta = p(2B)^{p-1}$, where we have used the inequality $||a|^p - |b|^p| \leq p|a - b||\max(a, b)|^{p-1}$. Now, let us define

$$L(f) = \mathbb{E}\{\ell_f(X, Y)\}, \tag{5}$$

$$\hat{L}_N(f) = \mathbb{E}_N\{\ell_f(X, Y)\}. \tag{6}$$

Using these definitions we may restate Lemma 3.1 in terms of loss functions as follows:

Lemma 3.2. *Let $\mathcal{F}$ be a permissible class of bounded functions, $|f| \le B$ for $f \in \mathcal{F}$ and some $0 < B < \infty$. For the class $\mathcal{L}_\mathcal{F}$ consisting of loss functions $\ell_f(x, y) = |y - f(x)|^p$, $f \in \mathcal{F}$, $|y| \le B$, there holds*

$$\mathbb{P}\left\{ Z^N \in \mathcal{Z}^N : \sup_{f\in\mathcal{F}} |\hat{L}_N(f) - L(f)| > \epsilon \right\} \le 4\mathbb{E}\left\{ \mathcal{N}\left(\frac{\epsilon}{16pB^{2p-2}}, \mathcal{F}(X^N), l_{1,2N} \right) \right\} \exp\left\{ -\frac{N\epsilon^2}{128(2B)^p} \right\},$$

where $\mathcal{Z}^N = (\chi \times \mathcal{Y})^N$, and the probability is taken with respect to the product measure on $\mathcal{Z}^N$.

Note that by using (4) we have written the covering number in Lemma 3.2 in terms of $\mathcal{F}$ rather than $\mathcal{L}_\mathcal{F}$.

Finally, we recall a result from (Haussler, 1992), which allows for extra flexibility and improved rates of convergence under certain conditions. We make use of this result in Sections 5 and 6.

Lemma 3.3 (*Haussler, 1992, Theorem 2*)**.** *Let $\mathcal{F}$ be a permissible class of real-valued non-negative functions such that $f(x) \le B$ for all $f \in \mathcal{F}$, and assume $\nu > 0$ and $0 < \alpha < 1$. Then*

$$\mathbb{P}\left\{ X^N \in \chi^N : \sup_{f\in\mathcal{F}} \frac{|\mathbb{E}_N f - \mathbb{E} f|}{\mathbb{E}_N f + \mathbb{E} f + \nu} > \alpha \right\} \le 4\mathbb{E}\left\{ \mathcal{N}\left(\frac{\alpha\nu}{8}, \mathcal{F}(X^N), l_{1,N} \right) \right\} \exp\left\{ -\frac{N\alpha^2\nu}{16B} \right\}.$$

4. Uniform convergence results for mixing processes

Having presented in Section 3 the essential results for the case of independent samples, we now proceed to discuss the main tools needed in the case of dependent sequences. Many of the ideas as well as the notation of this section rely on the work of Yu (1994). First, it should be clear that it will not be possible to obtain rates of convergence for uniform laws of large numbers, unless some assumptions about the process $\bar{X}$ are stipulated. In this work, we follow the widely used practice in the field of time-series analysis, and restrict ourselves to the class of *mixing* processes. These are processes for which the 'future' depends only weakly on the 'past', in a sense that will now be made precise.

Definition 2. Let $\sigma_l = \sigma(X_1^l)$ and $\sigma'_{l+m} = \sigma(X_{l+m}^\infty)$, be the sigma-algebras of events generated by the random variables $X_1^l = (X_1, X_2, \ldots, X_l)$ and $X_{l+m}^\infty = (X_{l+m}, X_{l+m+1}, \ldots)$, respectively. The coefficient of absolute regularity, β_m, is given by

$$\beta_m = \sup_{l \geq 1} \mathbb{E} \sup_{B \in \sigma'_{l+m}} |P(B \mid \sigma_l) - P(B)|, \tag{7}$$

where the expectation is taken with respect to σ_l. A stochastic process is said to be absolutely regular, or β-mixing, if $\beta_m \to 0$ as $m \to \infty$.

We note that there exist many other definitions of mixing (see Doukhan (1994) for an extensive listing). The particular proof method we use, relying on the work of Yu (1994), is based on the β-mixing coefficient. Similar results apply, of course, to processes with stronger mixing conditions (see Doukhan (1994)). We note that Modha and Masry (1998) have recently derived similar results in the context of the more general α-mixing processes; however, their results only apply to exponential mixing (see definition below). In this work, we consider two types of processes for which the mixing coefficient decays to zero, namely *algebraically* mixing processes for which $\beta_m = O(m^{-r})$, $r > 0$, and *exponentially* mixing processes for which $\beta_m = O(\exp\{-bm^\kappa\})$, $b, \kappa > 0$. Since we are concerned with finite sample results, we will assume that the conditions can be phrased, for any $m > 0$, as

$$\begin{aligned} &\beta_m \leq \bar{b} m^{-r} && \text{(algebraic mixing)}, \\ &\beta_m \leq \tilde{b} \exp\{-bm^\kappa\} && \text{(exponential mixing)}, \end{aligned} \tag{8}$$

for some finite non-negative constants b, $\bar{b}$ and $\tilde{b}$. We refer to the exponents r and κ as the mixing exponents. Note also that the usual i.i.d. process may be obtained from either the exponentially or the algebraically mixing process, by taking the limit $\kappa \to \infty$ or $r \to \infty$, respectively. We summarize the above notions in the following assumption, which will be used throughout.

Assumption 4.1. The stationary β-mixing stochastic process $\bar{X} = \{X_i\}_{-\infty}^\infty$ is compactly supported, $|X_i| \leq B$ for some $0 < B < \infty$. Moreover, the mixing exponent is known.

Observe that, to the best of our knowledge, there is no practical method to determine whether a process is mixing, unless it is Gaussian, Markov etc. Thus, Assumption 4.1, stringent as it is, cannot be avoided at this point. This type of assumption is used both in the work on nonparametric prediction (Györfi et al., 1989) and in the results using complexity regularization, as in Modha & Masry (1998).

In order to motivate the mixing assumption, we recall two examples where exponential mixing has been established. First, consider the standard ARMA process described by the equation

$$\sum_{j=0}^{p} b_j X_{i-j} = \sum_{k=0}^{q} a_k \epsilon_{i-k},$$

where ϵ_i are i.i.d. mean-zero random variables. Under the conditions that the probability distribution of ϵ_i is absolutely continuous and that the zeros of the polynomial $P(Z) = \sum_{i=0}^{p} b_i z^i$ lie outside the unit circle, Mokkadem (1988) has shown that the process is exponentially β-mixing. A further example is provided by Markov processes obeying the so called Doeblin condition, which basically restricts the process from being trapped in sets of small measure (see Rosenblatt (1971)); this assumption is, however, rather stringent.

As mentioned above, in this section we follow Yu (1994) in deriving uniform laws of large numbers for mixing processes. While Yu's work was mainly geared towards the asymptotic regime, we will be concerned here with finite sample theory, and will need to modify her results accordingly. Moreover, our results differ from hers when discussing specific assumptions about functional classes and their metric entropies. Finally, Yu's paper was concerned with algebraically mixing processes for which $r \leq 1$, as a central limit theorem holds in the case $r > 1$. Since we wish to derive finite sample bounds, we cannot make use of central limit results. In this work we study both the exponential and the algebraic mixing cases.

In the remainder of this section, we outline the basic ideas in the construction of Yu (1994), which serves as the main tool in the following sections. This construction is essential to the proofs of Sections 5 and 6, and is thus expanded on. The basic idea in (Yu, 1994), as in many related approaches, involves the construction of an *independent-block* sequence, which is shown to be 'close' to the original process in a well-defined probabilistic sense. We first motivate the construction. Divide the sequence X_1^N into $2\mu_N$ blocks, each of size a_N. We assume that $2\mu_N a_N = N$, so as not to be concerned with the remainder terms, which become insignificant as the sample size increases. The blocks are then numbered according to their order in the block-sequence. For $1 \leq j \leq \mu_N$ define

$$\begin{aligned} H_j &= \{i : 2(j-1)a_N + 1 \leq i \leq (2j-1)a_N\}, \\ T_j &= \{i : (2j-1)a_N + 1 \leq i \leq (2j)a_N\}. \end{aligned} \tag{9}$$

Denote the random variables corresponding to the blocks H_j and T_j by

$$X^{(j)} = \{X_i : i \in H_j\} \quad \text{and} \quad X'^{(j)} = \{X_i : i \in T_j\}.$$

The sequence of H-blocks is then denoted by $X_{a_N} = \{X^{(j)}\}_{j=1}^{\mu_N}$. Now, construct a sequence of independently distributed blocks $\{\Xi^{(j)})\}_{j=1}^{\mu_N}$, where $\Xi^{(j)} = \{\xi_i : i \in H_j\}$, such that the sequence is independent of X_1^N and each block $\Xi^{(j)}$ has the same distribution as the block $X^{(j)}$ from the original sequence; denote this sequence by $\Xi_{a_N} = \{\Xi^{(j)}\}_{j=1}^{\mu_N}$. Because the process $\bar{X}$ is stationary, the blocks $\Xi^{(j)}$ are not only independent but also identically distributed. The basic idea in the construction of the independent block sequence is that one can show that it is 'close', in a well-defined sense to the original blocked sequence X_{a_N}. Moreover, by appropriately selecting the number of blocks, μ_N, depending on the mixing nature of the sequence, one may relate properties of the original sequence X_1^N, to those of the independent block sequence Ξ_{a_N}.

In accordance with Lemma 2.1, we observe that in order to bound the estimation error, use must be made of uniform laws of large numbers. Now, from the construction above we clearly have for a functional class $\mathcal{F} : \mathbb{R} \mapsto \mathbb{R}$,

$$\mathbb{P}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}\, f(X) \right| > \epsilon \right\}$$
$$= \mathbb{P}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{N}\sum_{j=1}^{\mu_N}\sum_{i\in H_j} f(X_i) - \frac{1}{2a_N}\mathbb{E}\sum_{i\in H_j} f(X_i) \right.\right.$$
$$\left.\left. + \frac{1}{N}\sum_{j=1}^{\mu_N}\sum_{i\in T_j} f(X_i) - \frac{1}{2a_N}\mathbb{E}\sum_{i\in T_j} f(X_i) \right| > \epsilon \right\},$$

where the summations have been split into a sum over the blocks, $1 \le j \le \mu_N$, followed by a summation over the elements within each even or odd block. Here $\mathbb{P}\{\cdot\}$ is taken with respect to the original sequence $\{X_1, \ldots, X_N\}$. In order to simplify the notation, in the sequel we omit the specific dependence of probabilities on their arguments, as this will be clear from the context. Thus, instead of $\mathbb{P}\{X^N \in \chi^N : \cdots\}$ we write $\mathbb{P}\{\cdots\}$. Let

$$f_{H_j}\big(X^{(j)}\big) = \sum_{i\in H_j} f(X_i); \qquad f_{T_j}\big(X^{(j)}\big) = \sum_{i\in T_j} f(X_i).$$

Then using $N = 2\mu_N a_N$, we easily conclude

$$\mathbb{P}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}\, f(X) \right| > \epsilon \right\}$$
$$\le \mathbb{P}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{\mu_N}\sum_{j=1}^{\mu_N} f_{H_j}\big(X^{(j)}\big) - \mathbb{E}\, f_{H_1} \right| > a_N\epsilon \right\}$$
$$+ \mathbb{P}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{\mu_N}\sum_{j=1}^{\mu_N} f_{T_j}\big(X^{(j)}\big) - \mathbb{E}\, f_{T_1} \right| > a_N\epsilon \right\}$$
$$\le 2\mathbb{P}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{\mu_N}\sum_{j=1}^{\mu_N} f_H\big(X^{(j)}\big) - \mathbb{E}\, f_H \right| > a_N\epsilon \right\}. \tag{10}$$

At this point the problem has been fully expressed in terms of the blocked process X_{a_N}, defined over the blocks $\{H_j\}$. From the above construction, recall that the process Ξ_{a_N} was defined to be block-wise independent, while possessing the same marginal distribution as $\{X_{a_N}\}$ on each block H_j. Since uniform laws of large numbers for independent sequences are available from the results quoted in Section 3, we need at this point to relate the results for X_{a_N} to those for the sequence Ξ_{a_N}. To do so, use is made of the following lemma from Yu (1994), the proof of which relies on standard mixing inequalities, which may be found in Doukhan (1994).

Lemma 4.1 (*Yu, 1994, Lemma 4.1*)**.** *Let the distributions of X_{a_N} and Ξ_{a_N} be Q and $\tilde{Q}$, respectively. Then for any measurable function h on $\mathbb{R}^{a_N\mu_N}$ with bound M,*

$$\big|\mathbb{E}_Q h\big(X_{a_N}\big) - \mathbb{E}_{\tilde{Q}} h\big(\Xi_{a_N}\big)\big| \le M\mu_N\beta_{a_N}.$$

Consider the block-independent sequence Ξ_{a_N} and define

$$\tilde{\mathbb{E}}_{\mu_N} \tilde{f} = \frac{1}{\mu_N} \sum_{j=1}^{\mu_N} \tilde{f}_{H_j}\big(\Xi^{(j)}\big),$$
$$\tilde{f}_{H_j}\big(\Xi^{(j)}\big) = \sum_{i \in H_j} f(\xi_i), \qquad j = 1, 2, \ldots, \mu_N.$$

Similarly we let $\tilde{\mathbb{E}}\,\tilde{f} = \mathbb{E}\,\tilde{f}(\Xi^{(j)})$. We use the tilde symbol to denote expectations with respect to the independent block process Ξ_{a_N}. Recall that by construction, $\{\Xi^{(j)}\}$ are independent and that $|\tilde{f}_{H_j}| \le a_N B$ if $|f| \le B$. In the remainder of the paper we use variables with a tilde above them to denote quantities related to the block sequence Ξ_{a_N}. With this notation and making use of (10) and Lemma 4.1 one obtains the following key result:

Lemma 4.2 (*Yu, 1994, Lemma 4.2*)**.** *Suppose $\mathcal{F} : \mathbb{R} \mapsto \mathbb{R}$ is a permissible class of bounded functions, $|f| \le B$ for $f \in \mathcal{F}$. Then*

$$\mathbb{P}\left\{ \sup_{f \in \mathcal{F}} |\mathbb{E}_N f - \mathbb{E} f| > \epsilon \right\} \le 2\tilde{\mathbb{P}}\left\{ \sup_{f \in \mathcal{F}} |\tilde{\mathbb{E}}_{\mu_N} \tilde{f} - \tilde{\mathbb{E}} \tilde{f}| > a_N \epsilon \right\} + 2\mu_N \beta_{a_N}. \tag{11}$$

Note that the result is slightly modified from Yu (1994), due to the different notation, and the fact that we have assumed that $N = 2\mu_N a_N$ exactly, i.e., there is no remainder term. Lemma 4.1 is the main result which will allow us in Section 6 to derive performance bounds for time series prediction.

5. Error bounds for time series prediction

In order to make use of the results of Section 4, we first need to transform the problem somewhat. We define a new vector-valued process $\bar{\bar{X}} = \{\ldots, \vec{X}_{-1}, \vec{X}_0, \vec{X}_1, \ldots\}$, where

$$\vec{X}_i = (X_i, X_{i-1}, \ldots, X_{i-d}) \in \mathbb{R}^{d+1}.$$

For this sequence the β-mixing coefficients obey the inequality

$$\beta_m(\bar{\bar{X}}) \le \beta_{m-d}(\bar{X}). \tag{12}$$

For any function $f : \mathbb{R}^d \to \mathbb{R}$, consider the loss function $\ell_f : \mathbb{R}^{d+1} \to \mathbb{R}^+$,

$$\ell_f\big(X_{i-d}^{i-1}, X_i\big) = \big|X_i - f\big(X_{i-d}^{i-1}\big)\big|^p.$$

In this section, as well as Section 6, we revert to the notation $\mathcal{F}_{d,n}$ for the functional classes used for estimation, as our results depend explicitly on d and n. In the present section, however, d and n are *fixed*, while they will be allowed to vary in Section 6. Keeping in mind the definition of $\vec{X}_i$, we may, with a slight abuse of notation, use the notation $\ell_f(\vec{X}_i)$ for

this function. As in Section 3, we will work within the loss space $\mathcal{L}_{\mathcal{F}_{d,n}} = \{\ell_f : f \in \mathcal{F}_{d,n}\}$. With view to the results in Section 4, we need to introduce a related class of functions.

Definition 3. Let $\mathcal{L}_{\mathcal{F}_{d,n}}$ be a class of real-valued functions from $\mathbb{R}^D \to \mathbb{R}$, $D = d+1$. For each $\ell_f \in \mathcal{L}_{\mathcal{F}_{d,n}}$ and $\vec{\mathbf{x}} = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_{a_N})$, $\vec{x}_i \in \mathbb{R}^D$, let $\tilde{\ell}_f(\vec{\mathbf{x}}) = \sum_{i=1}^{a_N} \ell_f(\vec{x}_i)$. Then define $\tilde{\mathcal{L}}_{\mathcal{F}_{d,n}} = \{\tilde{\ell}_f : \ell_f \in \mathcal{L}_{\mathcal{F}_{d,n}}\}$, where $\tilde{\ell}_f : \mathbb{R}^{a_N D} \to \mathbb{R}^+$.

We now proceed to derive error bounds for prediction of mixing stochastic processes. Recall that we seek upper bounds on the uniform deviations

$$\begin{aligned}&\mathbb{P}\left\{ \sup_{f \in \mathcal{F}_{d,n}} |L_N(f) - L(f)| > \epsilon \right\} \\ &\quad = \mathbb{P}\left\{ \sup_{f \in \mathcal{F}_{d,n}} \left| \frac{1}{N-d} \sum_{i=d+1}^{N} \ell_f(\vec{X}_i) - \mathbb{E}\ell_f(\vec{X}_{d+1}) \right| > \epsilon \right\}.\end{aligned}$$

We comment that in principle, we should use a sample size of $N - d$ instead of N, since the first d values do not appear in the empirical loss. Since we assume that $N \gg d$, we will not be bothered with this mathematical detail, assuming throughout a sample size of N. Furthermore, in this section we take both the memory size d and the complexity index n to be fixed. The question of determining d adaptively will be addressed in Section 6.

Using Lemma 4.2, with the transformation $f \mapsto \ell_f$, we immediately conclude that

$$\begin{aligned}&\mathbb{P}\left\{ \sup_{f \in \mathcal{F}_{d,n}} |L_N(f) - L(f)| > \epsilon \right\} \\ &\quad \leq 2\tilde{\mathbb{P}}\left\{ \sup_{f \in \mathcal{F}_{d,n}} \left| \frac{1}{\mu_N} \sum_{j=1}^{\mu_N} \tilde{\ell}_f\big(\vec{X}^{(j)}\big) - \mathbb{E}\tilde{\ell}_f \right| > a_N\epsilon \right\} + 2\mu_N \beta_{a_N - d},\end{aligned}$$

where

$$\tilde{\ell}_f\big(\vec{X}^{(j)}\big) = \sum_{i \in H_j} \ell_f(\vec{X}_i) = \sum_{i \in H_j} \big| X_i - f\left(X_{i-d}^{i-1}\right) \big|^p, \tag{13}$$

and H_j is defined in (9). Note that we have replaced β_{a_N} in Lemma 4.2 by $\beta_{a_N - d}$ because each $\vec{X}_i$ contains $d+1$ lagged values of the process $\bar{X}$.

Having transformed the problem to the block-independent process, we can use Lemma 3.1 with the transformation $N \mapsto \mu_N$, noting that $|\tilde{\ell}_f| \leq a_N(2B)^p$, to obtain

$$\begin{aligned}&\tilde{\mathbb{P}}\left\{ \sup_{f \in \mathcal{F}_{d,n}} \left| \frac{1}{\mu_N} \sum_{j=1}^{\mu_N} \tilde{\ell}_f\big(X^{(j)}\big) - \mathbb{E}\tilde{\ell}_f \right| > a_N\epsilon \right\} \\ &\quad \leq 4\mathbb{E}\big\{\mathcal{N}(a_N\epsilon/16, \tilde{\mathcal{L}}_{\mathcal{F}_{d,n}}(\Xi_{a_N}), \tilde{l}_{1,\mu_N})\big\} \exp\left\{ -\frac{\mu_N \epsilon^2}{128(2B)^{2p}} \right\},\end{aligned} \tag{14}$$

where $\tilde{\mathcal{L}}_{\mathcal{F}_{d,n}}(\Xi_{a_N}) = \{\tilde{\ell}_f(\vec{X}^{(1)}), \ldots, \tilde{\ell}_f(\vec{X}^{(\mu_N)}) : \tilde{\ell}_f \in \tilde{\mathcal{L}}_{\mathcal{F}_{d,n}}\}$, and we have used the semi-norm

$$\tilde{l}_{1,\mu_N}(\tilde{\ell}_f, \tilde{\ell}_g) = \frac{1}{\mu_N} \sum_{j=1}^{\mu_N} \left|\tilde{\ell}_f\left(\vec{X}^{(j)}\right) - \tilde{\ell}_g\left(\vec{X}^{(j)}\right)\right|.$$

The covering number in (14) is taken with respect to the loss class $\tilde{\mathcal{L}}_{\mathcal{F}_{d,n}}$. We now relate the covering numbers of $\mathcal{L}_{\mathcal{F}_{d,n}}$ and $\tilde{\mathcal{L}}_{\mathcal{F}_{d,n}}$.

Lemma 5.1. *For any* $\epsilon > 0$

$$\mathcal{N}\left(\epsilon, \tilde{\mathcal{L}}_{\mathcal{F}_{d,n}}\left(\Xi_{a_N}\right), \tilde{l}_{1,\mu_N}\right) \leq \mathcal{N}\left(\epsilon/2a_N, \mathcal{L}_{\mathcal{F}_{d,n}}(Z^N), l_{1,N}\right).$$

Proof: The following sequence of inequalities holds:

$$\begin{aligned}
\tilde{l}_{1,\mu_N}(\tilde{\ell}_f, \tilde{\ell}_g) &= \frac{1}{\mu_N} \sum_{j=1}^{\mu_N} \left|\tilde{\ell}_f\left(\vec{X}^{(j)}\right) - \tilde{\ell}_g\left(\vec{X}^{(j)}\right)\right| \\
&= \frac{1}{\mu_N} \sum_{j=1}^{\mu_N} \left| \sum_{i \in H_j} \ell_f(\vec{X}_i) - \sum_{i \in H_j} \ell_g(\vec{X}_I) \right| \\
&\leq \frac{1}{\mu_N} \sum_{j=1}^{\mu_N} \sum_{i \in H_j} |\ell_f(\vec{X}_i) - \ell_g(\vec{X}_i)| \\
&= \frac{1}{\mu_N} \sum_{k=1}^{N} |\ell_f(\vec{X}_k) - \ell_g(\vec{X}_k)| \\
&= 2a_N l_{1,N}(\ell_f, \ell_g),
\end{aligned}$$

where we have used $N = 2\mu_N a_N$. Setting $l_{1,N}(\ell_f, \ell_g) \leq \epsilon/2a_N$ the result follows. □

Since the connection between the covering numbers of $\mathcal{L}_{\mathcal{F}_{d,n}}$ and $\mathcal{F}_{d,n}$ is known through (4), we can summarize the results of this section in the following theorem.

Theorem 5.1. *Let* $\bar{X} = \{\ldots, X_1, X_0, X_1, \ldots\}$ *be a stationary* β*-mixing stochastic process, with* $|X_i| \leq B$*, and let* $\mathcal{F}_{d,n}$ *be a class of bounded functions,* $f : \mathbb{R}^d \to [-B, B]$*. For each sample size* N*, let* $\hat{f}_{d,n,N}$ *be the function in* $\mathcal{F}_{d,n}$ *which minimizes the empirical error* (3)*, and* $f^*_{d,n}$ *is the function in* $\mathcal{F}_{d,n}$ *minimizing the expected error* (1)*. Then,*

$$\begin{aligned}
&\mathbb{P}\{L(\hat{f}_{d,n,N}) - L(f^*_{d,n}) > \epsilon\} \\
&\quad \leq 8\mathbb{E}\mathcal{N}\left(\epsilon/64p(2B)^{p-1}, \mathcal{F}_{d,n}(X^N), l_{1,N}\right) \exp\left\{-\frac{\mu_N \epsilon^2}{128(2B)^{2p}}\right\} + 2\mu_N \beta_{a_N - d}. \quad (15)
\end{aligned}$$

Proof: The claim is established through a sequence of inequalities.

$$\begin{aligned}
\mathbb{P}\big\{L(\hat{f}_{d,n,N}) - L(f^*) > \epsilon\big\} &\overset{(a)}{\leq} \mathbb{P}\left\{ \sup_{f\in\mathcal{F}_{d,n}} |\hat{L}_N(f) - L(f)| > \epsilon/2 \right\} \\
&\overset{(b)}{\leq} 2\tilde{\mathbb{P}}\left\{ \sup_{f\in\mathcal{F}_{d,n}} \left|\tilde{\mathbb{E}}_{\mu_N}\ell_{\tilde{f}} - \tilde{\mathbb{E}}\ell_{\tilde{f}}\right| > \frac{a_N\epsilon}{2} \right\} + 2\mu_N\beta_{a_N-d} \\
&\overset{(c)}{\leq} 8\mathbb{E}\left\{\mathcal{N}\left(\frac{a_N}{32\epsilon}, \tilde{\mathcal{L}}_{\mathcal{F}_{d,n}}(\Xi_{a_N}), \tilde{l}_{1,\mu_N}\right)\right\} \exp\left\{-\frac{\mu_N a_N^2\epsilon^2}{128\big(a_N(2B)^p\big)^2}\right\} + 2\mu_N\beta_{a_N-d} \\
&\overset{(d)}{\leq} 8\mathbb{E}\mathcal{N}\left(\frac{\epsilon}{64}, \mathcal{L}_{\mathcal{F}_{d,n}}(Z^N), l_{1,N}\right) \exp\left\{-\frac{\mu_N\epsilon^2}{128(2B)^{2p}}\right\} + 2\mu_N\beta_{a_N-d} \\
&\overset{(e)}{\leq} 8\mathbb{E}\mathcal{N}\left(\frac{\epsilon}{64p(2B)^{p-1}}, \mathcal{F}_{d,n}(X^N), l_{1,N}\right) \exp\left\{-\frac{\mu_N\epsilon^2}{128(2B)^{2p}}\right\} + 2\mu_N\beta_{a_N-d}
\end{aligned}$$

Step (a) makes use of Lemma 2.1, in (b) we have used Lemma 4.2 and (12), and (c) relies on Lemma 3.1 and the observation that $\tilde{l}_f \leq a_N(2B)^p$ (see (31)). Steps (d) and (e) use, respectively, Lemma 5.1 and the inequality (4) with $\eta = p(2B)^{p-1}$. □

In order to guarantee that an estimator $\hat{f}_{d,n,N}$ converge asymptotically to the optimal estimator in the class, namely $f^*_{d,n}$, we introduce the following notion of (weak) consistency.

Definition 4. An estimator $\hat{f}_{d,n,N} \in \mathcal{F}_{d,n}$ is weakly consistent if for every $\epsilon > 0$, $\mathbb{P}\{|L(\hat{f}_{d,n,N}) - L(f^*_{d,n})| > \epsilon\} \to 0$ *as* $N \to \infty$.

Note that in Definition 4 we require only convergence of the loss $L(\hat{f}_{d,n,N})$, rather than of the estimator $\hat{f}_{d,n,N}$ itself, as is more customary in the field of parametric statistics.

Up to this point we have not specified μ_N and a_N, and the result is therefore quite general. In order to obtain weak consistency we require that that the r.h.s. of (15) converge to zero for each $\epsilon > 0$. This immediately yields the following conditions on μ_N (and thus also on a_N through the condition $2a_N\mu_N = N$), which will be related to the mixing conditions in (8).

Corollary 5.1. *Under the conditions of Theorem* 5.1, *and the added requirement that* $\mathbb{E}\mathcal{N}(\epsilon, \mathcal{F}_{d,n}(X^N), l_{1,N}) < \infty$, $\forall\epsilon > 0$, *the following choices of* μ_N *are sufficient to guarantee the weak consistency of the empirical minimizer* $\hat{f}_{d,n,N}$:

$$\mu_N = \Theta\big(N^{\kappa/(1+\kappa)}\big) \qquad (\textit{exponential mixing}), \tag{16}$$

$$\mu_N = \Theta\big(N^{s/(1+s)}\big), 0 < s < r \quad (\textit{algebraic mixing}), \tag{17}$$

where the notation $a_N = \Theta(b_N)$ *implies that there exist two finite positive constants* c_1 *and* c_2 *such that* $c_1 b_N \leq a_N \leq c_2 b_N$ *for all* N *larger than some* N_0.

Proof: Consider first the case of exponential mixing. In this case, the r.h.s. of (15) clearly converges to zero because of the finiteness of the covering number. The fastest rate of

convergence is achieved by balancing the two terms in the equation, leading to the choice $\mu_N = \Theta(N^{\kappa/(1+\kappa)})$. In the case of algebraic mixing, the second term on the r.h.s. of (15) is of the order $O(\mu_N a_N^{-r})$, where we have used $d = o(a_N)$, and dominates the first term for large N. Since $\mu_N a_N = \Theta(N)$, a sufficient condition to guarantee that this term converge to zero is that $\mu_N = \Theta(N^{s/(1+s)})$, $0 < s < r$, as was claimed. □

Remark 1. Strong consistency, i.e., $L(\hat{f}_{d,n,N}) \to L(f^*_{d,n})$ a.s.-P, can also be established under appropriate conditions. In the case of exponential mixing, strong consistency may immediately be established using Theorem 5.1 and the Borel-Cantelli Lemma. In the case of algebraic mixing, the requirement that $\sum_{N=1}^{\infty} \mu_N a_{a_N - d} < \infty$, together with the choice $\mu_N = N^{s/(1+s)}$, $0 < s < r$, and the condition $d = o(a_N)$ leads to strong consistency only if $0 < s < (r-1)/2$, implying $r > 1$. We note that for $r > 1$, Arcones and Yu (1994) have proven a central limit theorem, which is of course stronger than mere consistency.

From Theorem 5.1 we may immediately obtain a result for the expected loss. However, in order to do so, something must be assumed about the dependence of $\mathcal{N}(\epsilon, \mathcal{F}_{d,n}(X^N), l_{1,N})$ on ϵ. We recall the definition of pseudo-dimension (Pollard, 1984).

Definition 5. Let $(X, \mathcal{S})$ be a given set, and let $\mathcal{F} \subseteq [0, B]^X$ consist of functions from X to the closed interval $[0, B]$. A set $S = \{x_1, \ldots, x_n\}$ is P-shattered by $\mathcal{F}$ if there exists a real vector $c \in [0, B]^n$ such that, for every binary vector $e \in \{0, 1\}^n$, there exists a corresponding function $f_e \in \mathcal{F}$ such that

$$f_e(x_i) \geq c_i \quad \text{if } e_i = 1, \qquad \text{and } f_e(x_i) < c_i \quad \text{if } e_i = 0.$$

The pseudo-dimension of $\mathcal{F}$, denoted by $\mathrm{Pdim}(\mathcal{A})$, equals the largest integer n such that there exists a set of cardinality n that is P-shattered by $\mathcal{A}$. If no such value exists the pseudo-dimension is infinite.

The pseudo-dimension becomes useful due to the following result of Haussler and Long (1995), which relates it to the covering number.

Lemma 5.2 (*Haussler, 1995, Corollary 3*)**.** *For any set X, any probability measure P on X, any set $\mathcal{F}$ of P-measurable functions taking values in the interval $[0, B]$ with pseudo-dimension* $\mathrm{Pdim}(\mathcal{F})$, *and any $\epsilon > 0$,*

$$\mathcal{N}(\epsilon, \mathcal{F}, l_{1,N}(P)) \leq e(\mathrm{Pdim}(\mathcal{F}) + 1)\left(\frac{2eB}{\epsilon}\right)^{\mathrm{Pdim}(\mathcal{F})}.$$

A special case of Lemma 3.1 occurs when $l_{1,N}(P)$ is the empirical $l_{1,N}$ semi-norm, which is used to define the covering numbers. We make the following assumption concerning the functional space $\mathcal{F}_{d,n}$.

Assumption 5.1. The functional space $\mathcal{F}_{d,n}$ possesses a finite pseudo-dimension $\mathrm{Pdim}(\mathcal{F}_{d,n})$.

An immediate consequence of the assumption is that the covering number $\mathcal{N}(\epsilon, \mathcal{F}_{d,n}(X^N), l_{1,N})$ may be bounded from above by a term of the form $K_{d,n}\epsilon^{-\mathrm{Pdim}(\mathcal{F}_{d,n})}$. Many examples of classes with finite pseudo-dimension are known. Two recently studied examples are neural networks with the standard sigmoidal activation function (Karpinksi & Macintyre, 1997) or with piecewise polynomial activation functions (Goldberg & Jerrum, 1995). In the latter case, rather tight bounds on the pseudo-dimension have recently been derived in (Bartlett et al., 1998).

Remark 2. We have made Assumption 5.1 for convenience. It is known that there are situations where the pseudo-dimension is not the optimal quantity for computing upper bounds for the covering number (Lee et al., 1996; Lugosi & Zeger, 1995). However, in all these cases one obtains covering number bounds which behave like $O(\epsilon^{-D})$ for some generalized dimension D. If this is the case, replace the pseudo-dimension by the dimension D, and all the results below follow. Note, however, that in some cases D may depend on ϵ, and a more careful analysis is needed using the so-called fat-shattering dimension as is discussed in Bartlett et al. (1996). Furthermore, there are specific situations where the pseudo-dimension yields nearly optimal bounds for the estimation error, since in that case the pseudo-dimension is essentially equivalent to another combinatorial dimension, called the fat-shattering dimension, which gives nearly matching lower bounds on estimation error (see Bartlett et al. (1996)).

Corollary 5.2. *Under the conditions of Theorem* 5.1, *and the added requirement that* $\mathrm{Pdim}(\mathcal{F}_{d,n}) < \infty$, *there exists a finite value of* $N = N_0$, *such that for all* $N > N_0$

$$\mathbb{E}L(\hat{f}_{d,n,N}) - L\left(f^*_{d,n}\right) \leq 32\sqrt{2}(2B)^p \sqrt{\frac{\frac{1}{2}\mathrm{Pdim}(\mathcal{F}_{d,n})\log\mu_N + \log K_{d,n}}{\mu_N}} + 4(2B)^p \mu_N \beta_{a_N - d}. \tag{18}$$

The proof of Corollary 5.2 is a special case of Theorem 6.1 given in Section 6, to which we refer the reader for a proof. The explicit dependence on N may be obtained by plugging in the values of μ_N from Corollary 5.1. Under the conditions of Remark 1 one can easily derive the following upper bound

$$\mathbb{E}L(\hat{f}_{d,n,N}) \leq L\left(f^*_{d,n}\right) + O\left(\left(\frac{\log N}{N}\right)^{\frac{1}{2}\frac{\kappa}{1+\kappa}}\right),$$

for the exponentially mixing case. These rates are similar to those obtained in Modha & Masry (1998) for the case of exponentially α-mixing processes. In the case of algebraic mixing one obtains the same rate with s replacing κ, where $0 < s < (r-1)/2$.

Remark 3. We observe that if one is willing to incur an extra multiplicative cost in (18), one may in fact obtain faster rates of convergence based on Lemma 3.3, for general loss functions, as already pointed out by Haussler in ((1992), Section 2.4). In this case one can

show that the square root on the r.h.s. of (18) may be removed, at the price of multiplying the term $L(f^*)$ by a factor $\beta > 1$. One then obtains that

$$\mathbb{E}L(\hat{f}_{d,n,N}) \leq \beta L\big(f^*_{d,n}\big) + O\left(\left(\frac{\log N}{N}\right)^{\frac{\kappa}{1+\kappa}}\right) \qquad (\beta > 1).$$

It should be borne in mind, however, that this kind of result is not very helpful in situations where $L(f^*_{d,n}) - L(f^*_d) > 0$, which may occur if the class $\mathcal{F}_{d,n}$ does *not* contain the optimal predictor of memory size d, as it leads to faster rates of convergence, to a *non-optimal* value of the loss since $\beta > 1$. However, the main merit of this observation is in the nonparametric situation discussed in Section 6.

Remark 4. Further improvement is possible in the case of quadratic loss, namely $p = 2$ in (1). It has been observed by several authors (Barron & Cover, 1991; Lee et al., 1996; Lugosi & Nobel, 1996) that in this case better bounds are available than those given by Corollary 5.2, in the special case where certain regularity (in particular, convexity) conditions are obeyed by the class of functions $\mathcal{F}_{d,n}$. In the case of feedforward neural networks, Barron (1993) and Lee et al. (1996) have proposed a constructive learning algorithm, for which exact performance bounds have been established in Lee et al. (1996). In particular, the square root in (18) may be eliminated giving rise to an $O(\log \mu_N/\mu_N)$ convergence rate, instead of the rate $O(\sqrt{\log \mu_N/\mu_N})$ implied by (18). In this case, unlike the case of general $p > 2$, the approximation error term is unaffected, i.e., $\beta = 1$, as opposed to the results quoted in Remark 3. The main technical trick needed to establish this result has to do with the use of the Bernstein-Craig inequality (Craig, 1933), rather than the standard Hoeffding inequality (Hoeffding, 1963) used in the usual derivations of uniform laws of large numbers. Unfortunately, this approach does not seem to work for more general L_p norms, with which we are concerned in this paper.

Finally, it is worth commenting on the 'predictability' of the sequence. When the sequence is i.i.d., previous values do not tell us anything about the next value. However, in this case the estimation error, which essentially measures the finite sample effects, decays at a rate $O(\sqrt{\log N/N})$, which is much faster than the rate obtained above in the case of mixing processes. In the latter case the 'effective' sample size, $N^{\kappa/(1+\kappa)}$ (for exponential mixing; similarly for algebraic mixing), is reduced due to the temporal correlations, leading to the slower rates.

6. Structural risk minimization for time series

The results in Section 5 provide error bounds for estimators formed by minimizing the empirical error over a fixed class of functions. It is clear that the complexity of the class of functions plays a crucial role in the procedure. If the class is too rich, manifested by very large covering numbers, clearly the estimation error term will be very large. On the other hand, biasing the class of functions by restricting its complexity, leads to poor approximation rates. A well-known strategy for overcoming this dilemma is obtained by considering

a hierarchy of functional classes with increasing complexity. For any given sample size, the optimal trade-off between estimation and approximation can then be determined by balancing the two terms. Such a procedure was developed in the late seventies by Vapnik (1982), and termed by him *structural risk minimization* (SRM). Other more recent approaches, collectively termed complexity regularization, have been extensively studied in recent years e.g. (Lugosi & Zeger, 1996; Barron et al., 1996). It should be borne in mind, however, that in the context of time series there is an added complexity, that does not exist in the case of regression for i.i.d. data. Recall that the results derived in Section 5 assumed some fixed lag vector d. In general the optimal value of d is unknown, and could in fact be infinite. In order to achieve optimal performance in a nonparametric setting, it is crucial that the size of the memory be chosen adaptively as well. This added complexity needs to be incorporated into the SRM framework, if optimal performance in the face of unknown memory size is to be achieved.

Let $\mathcal{F}_{d,n}$, $d, n \in \mathbb{N}$ be a sequence of functions, and define

$$\mathcal{F} = \bigcup_{d=1}^{\infty} \bigcup_{n=1}^{\infty} \mathcal{F}_{d,n}.$$

Keeping in mind the definition of the covering numbers $\mathcal{N}(\epsilon, \mathcal{F}(X^N), l_{1,N})$, utilized throughout the previous sections, we find it useful to define an upper bound on these numbers, which does not depend on the specific data observed. The existence of this bound is guaranteed by Lemma 5.2 and Assumption 5.1. For any $d, n \in \mathbb{N}$ and $\epsilon > 0$ let

$$\mathcal{N}_1(\epsilon, \mathcal{F}_{d,n}) = \sup_{X^N \in \mathcal{X}^N} \mathcal{N}\big(\epsilon, \mathcal{F}_{d,n}(X^N), l_{1,N}\big). \tag{19}$$

We observe in passing that Lugosi and Nobel (1996) have recently considered situations where the pseudo-dimension $\mathrm{Pdim}(\mathcal{F}_{d,n})$ is unknown, and the covering number is estimated empirically from the data. Although this line of research is potentially very useful, we do not pursue it here, but rather assume that upper bounds on the pseudo-dimensions of $\mathcal{F}_{d,n}$ are known, as is the case for various classes of functions such as neural networks, radial basis functions etc. (see examples in Vidyasagar (1996)).

In line with the, by now classic, approach outlined in Vapnik (1982) we introduce a new empirical function, which takes into account both the empirical error and the complexity costs penalizing overly complex models (large n and overly large memory size d). Let

$$\hat{\hat{L}}_{d,n,N}(f) = \hat{L}_N(f) + \Delta_{d,n,N}(\epsilon) + \Delta_{d,N}, \tag{20}$$

where $\hat{L}_N(f)$ is the empirical error (3) of the predictor f. We have introduced the 'double-hat' notation, as in $\hat{\hat{L}}$, to emphasize the two-level estimation process involved. The complexity penalties Δ are given by,

$$\Delta_{d,n,N}(\epsilon) = \sqrt{\frac{\log \mathcal{N}_1(\epsilon, \mathcal{F}_{d,n}) + c_n}{\mu_N / 32(2B)^{2p}}} \tag{21}$$

$$\Delta_{d,N} = \sqrt{\frac{c'_d}{\mu_N/32(2B)^{2p}}}. \tag{22}$$

The specific form and constants in these definitions are chosen with hindsight, so as to achieve the optimal rates of convergence in Theorem 6.1 below. The constants c_n and c'_d are positive constants obeying $\sum_{n=1}^{\infty} e^{-c_n} \leq 1$ and similarly for c'_d. A possible choice is $c_n = 2\log n + 1$ and $c'_d = 2\log d + 1$. The value of μ_N can be chosen in accordance with Corollary 5.1.

Let $\hat{f}_{d,n,N}$ be the minimizer of the empirical loss $\hat{L}_N(f)$ within the class of functions $\mathcal{F}_{d,n}$, namely

$$\hat{L}_N(\hat{f}_{d,n,N}) = \min_{f \in \mathcal{F}_{d,n}} \hat{L}_N(f).$$

We assume that the classes $\mathcal{F}_{d,n}$ are compact, so that such a minimizer exists. Observe that less stringent conditions, of the form $\hat{L}_N(\hat{f}_{d,n,N}) \leq \min_{f \in \mathcal{F}_{d,n}} \hat{L}_N(f) + \epsilon_N$ for an appropriate ϵ_N may be used, but lead to the same results, and add little to the generality. Further, let $\hat{\hat{f}}_N$ be the function in $\mathcal{F}_N$ minimizing the complexity penalized loss (20), namely

$$\hat{\hat{L}}_{d,n,N}(\hat{\hat{f}}_N) = \min_{d \geq 1} \min_{n \geq 1} \hat{\hat{L}}_{d,n,N}(\hat{f}_{d,n,N}). \tag{23}$$

We now present the basic result establishing the consistency of the structural risk minimization approach for time series, together with upper bounds on its performance. As in Section 5, we assume that the pseudo-dimension of each class $\mathcal{F}_{d,n}$ is finite, motivating the assumption:

Assumption 6.1. The covering number of each class $\mathcal{F}_{d,n}$ can be bounded from above by

$$\mathcal{N}_1(\epsilon, \mathcal{F}_{d,n}) \leq K_{d,n}\epsilon^{-\gamma_{d,n}},$$

where $\gamma_{d,n} = \mathrm{Pdim}(\mathcal{F}_{d,n}) < \infty$ for any d and n.

We recall Remark 2, keeping in mind that in cases where the pseudo-dimension does not provide a tight upper bound on the covering number we may replace $\gamma_{d,n}$ by some other generalized dimension, such as the fat-shattering dimension at an appropriate scale. This modification does not affect the arguments below. Before presenting the main result of this section, we make an assumption concerning the size of the memory d.

Assumption 6.2. For each value of N, $0 \leq d \leq a_N^{1-\epsilon}$ for some $0<\epsilon<1$, implying $d = o(a_N)$.

For each d, let N_0 be the (finite) value of N, such that by Assumption 6.2 $d < a_N/2$ for $N > N_0$; such a value exists, since from Corollary 5.1 a_N becomes arbitrarily large with

increasing N. The following lemma, proved in the appendix, is crucial in establishing the main result of this section.

Lemma 6.1. *Let $\bar{X} = \{\ldots, X_1, X_0, X_1, \ldots\}$ be a stationary β-mixing stochastic process, with $|X_i| \leq B, 0 < B < \infty$. Furthermore, let $\mathcal{F}_{d,n}$ be a class of bounded functions, $f : \mathbb{R}^d \to [-B, B]$, and let Assumption* 6.2 *hold, implying that $d \leq a_N/2$ for N larger than some N_0. For each N let $\hat{\hat{f}}_N$ be the predictor selected according to the SRM condition* (23)*. Let $\epsilon > 0$ be given and assume further that $32p(2B)^{p-1}\epsilon \leq \Delta_{d,n,N}(\epsilon) + \Delta_{d,N} \leq t/4$. Then for $N > N_0$*

$$\mathbb{P}\left\{L(\hat{\hat{f}}_N) - \inf_{f\in\mathcal{F}_{d,n}} L(f) > t\right\} \leq 8e^{-\mu_N t^2/128(2B)^{2p}} + 8\mathcal{N}_1\left(t/256p(2B)^{p-1}, \mathcal{F}_{d,n}\right)e^{-\mu_N t^2/2048(2B)^{2p}} + 4(2B)^p\mu_N\beta_{a_N/2}.$$

We then have the main result of this section; the proof is given in the appendix.

Theorem 6.1. *Let $\bar{X} = \{\ldots, X_1, X_0, X_1, \ldots\}$ be a stationary β-mixing stochastic process, with $|X_i| \leq B$, and let $\mathcal{F}_{d,n}$ be a class of bounded functions, $f : \mathbb{R}^d \to [-B, B]$. Then, for $N > N_0$, and given Assumption* 6.2*, the expected loss of the function $\hat{\hat{f}}_N$, selected according to the SRM principle* (23)*, is upper bounded by*

$$\mathbb{E}L(\hat{\hat{f}}_N) \leq \min_{d,n}\left\{\inf_{f\in\mathcal{F}_{d,n}} L(f) + c_1\sqrt{\frac{\frac{1}{2}\gamma_{d,n}\log\mu_N + \log K_{d,n} + c_n}{\mu_N}} + \frac{c_2}{\sqrt{\mu_N}}\right\} + 4(2B)^p\mu_N\beta_{a_N/2},$$

where $c_1 = 32\sqrt{2}(2B)^p$ and $c_2 = (640 + 32\sqrt{c'_d})\sqrt{2}(2B)^p$, and the constants c_n and c'_d are defined in (21) *and* (23)*, respectively.*

Remark 5. Observe that similarly to Remark 3, better rates can be obtained at the price of multiplying the approximation error term in Theorem 6.1 by a constant larger than 1. In this case we obtain a bound of the form

$$\mathbb{E}L(\hat{\hat{f}}_N) \leq \min_{d,n}\left\{\beta\inf_{f\in\mathcal{F}_{d,n}} L(f) + c'_1\frac{\gamma_{d,n}\log\mu_N}{\mu_N}\right\} + 4(2B)^p\mu_N\beta_{a_N/2}, \tag{24}$$

where $\beta > 1$, and c'_1 and c'_2 are given constants. The proof of this result relies on Lemma 3.3, and is very similar to that given below for Theorem 6.1. We recall that the improved rate is only relevant in the situation where $(\min_{d,n}\inf_{f\in\mathcal{F}_{d,n}} L(f) - L(f^*_\infty))$ can be made to vanish, which may not be the case in practical situations where the complexity of the approximating class may be restricted to some finite values of d and n.

Remark 6. The results derived in 6.1 demonstrate that there is an almost optimal trade-off between approximation and estimation, in the following sense. Even if $\inf_{f\in\mathcal{F}_{d,n}} L(f)$ attained some minimal value for some finite *known* values of n and d (remaining constant

for larger values), there would still be an additional loss incurred in the estimation, due to the finiteness of the sample size. What Theorem 6.1 establishes is that the additional loss resulting from the extra degrees of freedom resulting from the variable memory size d and model complexity n do not affect the performance. This result can be viewed as an extension of existing results for i.i.d. data, both for the case of binary classification (Lugosi & Zeger, 1996) and regression (Lugosi & Nobel, 1996).

Remark 7. We compare our results to those derived recently by Modha and Masry (1998). These authors also considered the problem of time series prediction in the context of mixing stochastic processes, deriving similar finite sample bounds in a nonparametric setting. We observe that their results differ from ours in several important ways. First, the framework used in Modha and Masry (1998) was based on Barron and Cover's notion of the index of resolvability (Barron & Cover, 1991), which entails rather different assumptions on the covering numbers of the functional classes used for estimation. In particular, the compactness of the parameter domain used by the estimator seems crucial for that approach. In the context of the covering numbers used in this work, all that is needed is the finiteness of the pseudo-dimension of the functional class, together with a boundedness condition on the function, a much weaker condition in general than boundedness of parameters. Moreover, the latter condition renders the establishment of approximation error bounds rather difficult. The reader is referred to Vidyasagar (1996) for several examples of cases where the pseudo-dimension is finite even though the parameters are unbounded. Second, our work holds for general L_p norms rather than the the L_2 norm studied in Modha and Masry (1998). For the case $p = 2$, we obtain the same rates under the further condition of convexity alluded to in Remark 4. Third, the results established in this work hold for both exponential and algebraic mixing stochastic processes, as opposed to the case of exponential mixing studied in Modha and Masry (1998). However, our results relate to β-mixing processes, a stronger condition than the one used in Modha and Masry (1998). Two interesting open problems remain, which have not yet been answered for either approach. First, the assumption of the boundedness of the process is somewhat restrictive, especially as this assumption is not required in the context of nonparametric time series prediction in general. Second, the development of an adaptive method to determine the mixing coefficients β is of importance if a fully adaptive algorithm is required. There do not seem to exist at present any approaches which address this issue satisfactorily.

Finally, we consider the nonparametric rates of convergence achieved within the adaptive scheme proposed in the present section. We refer the reader to the books by Györfi et al. (1989) and Bosq (1996) for surveys of the field of nonparametric time series prediction. We consider a typical result established in Györfi et al. (1989) (see Section 3.4) for compactly supported ϕ-mixing processes. Since ϕ-mixing implies β-mixing, the result is applicable in the context of this paper. Let

$$R_d(\mathbf{x}) = \mathbb{E}\big(X_i \,\big|\, X_{i-d}^{i-1} = \mathbf{x}\big),$$

and denote by $R_{d,N}(\mathbf{x})$ a nonparametric estimator for $R_d(\mathbf{x})$ obtained using the Nadayara-Watson kernel estimator with width h and a sample of size N. Note that $R_d(\mathbf{x})$ is only

optimal as a predictor when using the quadratic loss. Assume further that $R_d(\mathbf{x})$ belongs to a Lipschitz space of k times continuously differentiable functions over the compact domain G, such that their k-th order derivatives are Lipschitz continuous of order 1. Then from Theorem 3.4.4 and Remark 3.3.5 in Györfi et al. (1989), we have that

$$\sup_{\mathbf{x}\in G} |R_d(\mathbf{x}) - R_{d,N}(\mathbf{x})| = O\Big(h^{k+1} + \big(m_N \log N \big/ N h^d\big)^{1/2}\Big) \qquad \text{(a.s.)},$$

where $m_N \sim (\log N)^{1/\kappa}$ for exponentially mixing processes and $m_N \sim n^{1/(1+r)}$ for algebraically mixing processes with exponent r. Setting h to its optimal value, minimizing the sum of the two terms we find that

$$\sup_{\mathbf{x}\in G} |R_d(\mathbf{x}) - R_{d,N}(\mathbf{x})| = O\left(\left(\frac{(\log N)^{\kappa/(1+\kappa)}}{N}\right)^{\frac{k+1}{2(k+1)+d}}\right) \qquad \text{(a.s.)},$$

for exponentially mixing processes and

$$\sup_{\mathbf{x}\in G} |R_d(\mathbf{x}) - R_{d,N}(\mathbf{x})| \leq O\left(\left(\frac{(\log N)^{(1+r)/r}}{N}\right)^{\frac{r}{1+r}\frac{k+1}{2(k+1)+d}}\right) \qquad \text{(a.s.)},$$

in the case of algebraic mixing.

Turning now to the results of this section, we need establish similar rates of convergence in a nonparametric setting. Assume, for example, that the optimal predictor of memory size d belongs to a Sobolev space consisting of functions with square integrable $(k+1)$-th order derivatives. The reason for considering $k+1$ derivatives rather than k is related to our wish to compare the results for Sobolev space to those for the Lipschitz space discussed above. As is demonstrated in Section 2.9 of Devore & Lorentz (1993), the Lipschitz space with k derivatives is isomorphic to the Sobolev space with $k+1$ integrable derivatives. Furthermore, assume that the functional space $\mathcal{F}_{d,n}$ is such that $\inf_{f\in\mathcal{F}_{d,n}} L(f) \leq cn^{-(k+1)/d}$ for any f in the Sobolev space. This type of result is well known for spline functions, and has recently been demonstrated for neural networks (Mhaskar, 1996) and mixture of expert architectures (Zeevi et al., 1998). Using the results of Theorem 6.1, and assuming that the optimal memory size d is known, as in the nonparametric setting above, we can compute the value for the complexity index n which yields fastest rates of convergence. Making use of the values of μ_N from Corollary 5.1 and Remark 1 we obtain, after some algebra, that

$$\mathbb{E}L(\hat{\hat{f}}_N) - L_d^* = O\left(\frac{\log N}{N}\right)^{\frac{\eta}{1+\eta}\frac{k+1}{2(k+1)+qd}}, \tag{25}$$

where L_d^* is the error incurred by the optimal predictor of memory size d, and where η equals s in the case of algebraic mixing and κ for exponential mixing (see Remark 1). Observe that the results quoted above for the kernel method, are given in terms of the distance $|R_d(\mathbf{x}) - R_{d,N}(\mathbf{x})|$, rather than the error itself, as in (25). However, in the quadratic case $p = 2$, for which the kernel method results are given, it is easy to establish (keeping in mind the

boundedness of the variables) that $\mathbb{E}|R_{d,N}(\mathbf{x}) - Y|^2 - |R_d(\mathbf{x}) - Y|^2 \leq 4B|R_{d,N}(\mathbf{x}) - R_d(\mathbf{x})|$, implying similar rates of convergence for the error difference, as in (25). In the situation where the improved rates (24) are used one may remove the factor of 2 multiplying $(k+1)$ in the denominator in (25); however, in this case the convergence is to βL_d^*, $\beta > 1$, rather than L_d^*.

In the derivation we have assumed that $\gamma_{d,n} = \text{Pdim}(\mathcal{F}_{d,n}) \propto n^q$ for some positive value of q, which is a typical situation (see examples in Vidyasagar (1996)). For example, we have recently shown (Bartlett, Maiorov, & Meir, 1998) that $q = 1+\epsilon$ ($\epsilon > 0$ arbitrarily small) for feedforward neural networks composed of piecewise polynomial activation functions, while Karpinski and Macintyre (1997) have established $q = 4$ for networks constructed using the standard sigmoidal activation function. Note that in these two cases the approximation errors have also been shown to be of the same order (Maiorov & Meir, 1999).

Furthermore, note the extra factor $\kappa/(1+\kappa)$ which multiplies the exponent, and leads to slower rates of convergence, as compared to the nonparametric results, in the case of exponential mixing. Note also that for this result to hold in the case of algebraic mixing we needed to assume that $r > 1$, in accordance with Remark 1. In summary then, we observe that our results in the nonparametric situation are in general not as tight as those attained by the classic kernel methods, a point also observed in Modha and Masry (1998). One of the major open problems then in the field of nonparametric time series prediction through adaptive model selection would be to achieve optimal (minimax) rates of convergence. Of course it should be borne in mind that the approach has the great advantage of adaptivity, in that parametric rates of convergence are achieved if the underlying structure is simple, as discussed in Remark 6. Note that in the special case $p = 2$, faster nonparametric rates are achieved under the special conditions discussed in Remark 4.

It should be kept in mind, however, that the above nonparametric rates of convergence were derived under the assumption that the memory size d is finite. In situations where the memory size is unbounded, a cost must be added for using a finite value of d. This point is made explicit in (2), where the additional term L_d^* is added, to take account of the loss incurred by using a finite value of d. Unfortunately, in order to compute an upper bound on this term, rates of convergence in the martingale convergence theorem would be needed; to the best of our knowledge such results are unknown at present for mixing processes.

7. Concluding remarks

We have presented bounds on the error incurred in nonparametric time-series prediction by sequences of parametric models, characterized by well-behaved metric entropies. This work extends previous results which make more demanding assumptions concerning boundedness of parameters and smoothness of the functions used for estimation. Our results were derived within the framework of the structural risk minimization approach pioneered by Vapnik, and were extended to the case of time-series by taking into account a complexity controlled adaptive memory size, based on the mixing relationships assumed to hold for the underlying stochastic process. The general approach has the potential advantage of achieving universal consistency and good rates of convergence in nonparametric settings, while retaining parametric rates of convergence in special situations. This adaptivity advantage has been

established for absolutely regular mixing processes, and rates of convergence have been established in cases where the memory size d is finite, but not necessarily known in advance.

There remain several issues, which need to be addressed in future work. First, the adaptive algorithm, as well as the bounds rely heavily on a knowledge of the mixing nature of the process. As mentioned in Section 4, no tools are currently available for establishing mixing properties, which are therefore more of a theoretician's dream than a practical tool. Second, rates of convergence were derived only for the case where the optimal memory size d is finite, albeit unknown. It is a challenging problem to derive convergence rates in the general case where the memory size may be infinite. Third, it is an open question to establish minimaxity of the nonparametric rates of convergence derived in Section 6, similarly to the results in the i.i.d. setting. Finally, and at a more technical level, we have assumed throughout that the stochastic process and prediction functions are bounded, clearly an impractical assumption in real life. We believe, however, that this assumption can be eliminated using the techniques of van der Geer (1987), a topic which is currently being pursued.

Appendix

Proof of Lemma 6.1: We split the problem into two components.

$$\begin{aligned}\mathbb{P}\left\{L(\hat{\hat{f}}_N) - \inf_{f\in\mathcal{F}_{d,n}} L(f) > t\right\} &< \mathbb{P}\left\{L(\hat{\hat{f}}) - \min_{d,n} \hat{\hat{L}}_{d,n,N}(\hat{f}_{d,n,N}) > \frac{t}{2}\right\} \\ &\quad + \mathbb{P}\left\{\min_{d,n} \hat{\hat{L}}_{d,n,N}(\hat{f}_{d,n,N}) - \inf_{f\in\mathcal{F}_{d,n}} L(f) > \frac{t}{2}\right\} \\ &\equiv J_1 + J_2. \end{aligned} \tag{26}$$

We deal separately with each of the terms. Let $\hat{d}$ and $\hat{n}$ be the values of d and n which minimize $\hat{\hat{L}}_{d,n,N}(\hat{f}_{d,n,N})$, i.e., $\hat{\hat{f}}_N = \hat{f}_{\hat{d},\hat{n},N}$. Then

$$\begin{aligned} J_1 &= \mathbb{P}\left\{L(\hat{f}_{\hat{d},\hat{n},N}) - \hat{\hat{L}}_{\hat{d},\hat{n},N}(\hat{f}_{\hat{d},\hat{n},N}) > \frac{t}{2}\right\} \\ &\le \mathbb{P}\left\{L(\hat{f}_{\hat{d},\hat{n},N}) - \hat{L}_N(\hat{f}_{\hat{d},\hat{n},N}) > \frac{t}{2} + \Delta_{\hat{d},\hat{n},N}(\epsilon) + \Delta_{\hat{d},N}\right\} \\ &\le \mathbb{P}\left\{\sup_{f\in\mathcal{F}_{\hat{d},\hat{n}}} |L(f) - \hat{L}_N(f)| > \frac{t}{2} + \Delta_{\hat{d},\hat{n},N}(\epsilon) + \Delta_{\hat{d},N}\right\} \end{aligned}$$

where we used the definition (20). We then have from the union bound that

$$\begin{aligned} J_1 &\le \mathbb{P}\left\{\max_{d,n}\left(\sup_{f\in\mathcal{F}_{d,n}} |L(f) - \hat{L}_N(f)| > \frac{t}{2} + \Delta_{d,n,N}(\epsilon) + \Delta_{d,N}\right)\right\} \\ &\le \sum_{d,n} \mathbb{P}\left\{\sup_{f\in\mathcal{F}_{d,n}} |L(f) - \hat{L}_N(f)| > \frac{t}{2} + \Delta_{d,n,N}(\epsilon) + \Delta_{d,N}\right\}. \end{aligned}$$

Using the notation $u_{d,n,N} = t + 2\Delta_{d,n,N}(\epsilon) + 2\Delta_{d,N}$, we then have from the proof of Theorem 5.1 and (19)

$$J_1 \leq 8 \sum_{d,n} \mathcal{N}\big(u_{d,n,N}/64p(2B)^{p-1}, \mathcal{F}_{d,n}\big)\exp\left\{-\frac{\mu_N u_{d,n,N}^2}{128(2B)^{2p}}\right\} + 2\mu_N \beta_{a_N/2}.$$

We note that we have replaced the term β_{a_N-d} appearing in 5.1 by $\beta_{a_N/2}$. This is allowed since we assumed that $d \leq a_N/2$, and β_m is monotonically non-increasing in m. Substituting the value of $u_{d,n,N}$ and making use of the assumption $\Delta_{d,n,N(\epsilon)} + \Delta_{d,N} > 32p(2B)^{p-1}\epsilon$, we then find

$$\begin{aligned}
J_1 &\leq 8 \sum_{d,n} \mathcal{N}_1\big(\epsilon, \mathcal{F}_{d,n}\big) \exp\left\{-\frac{\mu_N}{128(2B)^{2p}}[t + 2\Delta_{d,n,N}(\epsilon) + 2\Delta_{d,N}]^2\right\} + 2\mu_N \beta_{a_N/2} \\
&\leq 8 \sum_{d,n} \mathcal{N}_1\big(\epsilon, \mathcal{F}_{d,n}\big) \exp\left\{-\frac{\mu_N}{128(2B)^{2p}}\big[t^2 + 4\Delta_{d,n,N}^2(\epsilon) + 4\Delta_{d,N}^2\big]\right\} + 2\mu_N \beta_{a_N/2} \\
&\leq 8 \sum_{d,n} \exp\left\{-\frac{\mu_N t^2}{128(2B)^{2p}} - c_n - c_d'\right\} + 2\mu_N \beta_{a_N/2}, \\
&\leq 8 \exp\left\{-\frac{\mu_N t^2}{128(2B)^{2p}}\right\} + 2\mu_N \beta_{a_N/2}.
\end{aligned}$$

where use has been made of the postulated summability properties of the sequences $\{e^{-c_d'}\}$ and $\{e^{-c_n}\}$, and (21) and (22).

In order to conclude the proof we need to consider the second term in (26). Following similar reasoning to that used above, and making use of the assumption $\Delta_{d,n,N(\epsilon)} + \Delta_{d,N} \leq t/4$, we have:

$$\begin{aligned}
J_2 &= \mathbb{P}\left\{\hat{L}_N(\hat{f}_{\hat{d},\hat{n},N}) + \Delta_{\hat{d},\hat{n},N}(\epsilon) + \Delta_{d,N} - L_{d,n}^* > \frac{t}{2}\right\} \\
&\leq \mathbb{P}\left\{\hat{L}_N(\hat{f}_{\hat{d},\hat{n},N}) - L_{d,n}^* > \frac{t}{4}\right\} \\
&\leq \mathbb{P}\left\{\hat{L}_N\big(f_{d,n}^*\big) - L\big(f_{d,n}^*\big) > \frac{t}{4}\right\} \\
&\leq \mathbb{P}\left\{\sup_{f \in \mathcal{F}_{d,n}} |\hat{L}_N(f) - L(f)| > \frac{t}{4}\right\} \\
&\leq 8\mathcal{N}_1\big(t/256p(2B)^{p-1}, \mathcal{F}_{d,n}\big)\exp\left\{-\frac{\mu_N t^2}{2048(2B)^{2p}}\right\} + 2\mu_N \beta_{a_N/2},
\end{aligned}$$

where again Theorem 5.1 has been used. The result then follows on combining the upper bounds on J_1 and J_2. □

Proof of Theorem 6.1: Our proof follows some of the ideas in Lugosi and Nobel (1996), with appropriate modifications. Obviously

$$\mathbb{E}L(\hat{\hat{f}}_N) = \min_{d,n}\left\{\left(\mathbb{E}L(\hat{\hat{f}}_N) - L^*_{d,n}\right) + L^*_{d,n}\right\},$$

where $L^*_{d,n} = \inf_{f\in\mathcal{F}_{d,n}} L(f)$. Using the notation

$$\Gamma_{d,n,N} = L(\hat{\hat{f}}_N) - L^*_{d,n},$$

we have

$$\begin{aligned}\mathbb{E}\Gamma_{d,n,N} &= \mathbb{E}\left\{L(\hat{\hat{f}}_N) - L^*_{d,n}\right\}\\ &= \int_0^\infty \mathbb{P}\left\{L(\hat{\hat{f}}_N) - L^*_{d,n} > t\right\} dt\\ &\le u + \int_u^{(2B)^p} \mathbb{P}\left\{L(\hat{\hat{f}}_N) - L^*_{d,n} > t\right\} dt,\end{aligned}$$

where we have used the fact that $|X_i - f(X_{i-d}^{i-1})|^p \le (2B)^p$, implying $L(f) = \mathbb{E}|X_i - f(X_{i-d}^{i-1})|^p \le (2B)^p$. Substituting the results of Lemma 6.1 and defining $a = 1/256p(2B)^{p-1}$, we then obtain

$$\begin{aligned}\mathbb{E}\Gamma_{d,n,N} &\le u + \int_u^{(2B)^p}\left[8e^{-\frac{\mu_N t^2}{128(2B)^{2p}}} + 8\mathcal{N}_1(at,\mathcal{F}_{d,n})e^{-\frac{\mu_N t^2}{2048(2B)^{2p}}}\right]dt + 4(2B)^p\mu_N\beta_{a_N/2}\\ &\le u + \frac{16}{\sqrt{\mu_N/128(2B)^{2p}}}\exp\left\{-\frac{\mu_N u^2}{128(2B)^{2p}}\right\}\\ &\quad + \frac{16\mathcal{N}_1(u/256p(2B)^{p-1},\mathcal{F}_{d,n})}{\sqrt{\mu_N/2048(2B)^{2p}}}\exp\left\{-\frac{\mu_N u^2}{2048(2B)^{2p}}\right\} + 4(2B)^p\mu_N\beta_{a_N/2},\end{aligned}$$

where we have used $\int_u^\infty \exp(-\beta t^2)\,dt \le 2\exp(-\beta u^2)/\sqrt{\beta}$.

Choosing $u = 8\Delta_{d,n,N}(\epsilon) + 8\Delta_{d,N}$ and using the condition $u > 256p(2B)^{p-1}\epsilon$ and the choice $\epsilon = 1/\sqrt{\mu_N}$, we obtain the following result after some algebra

$$\begin{aligned}\mathbb{E}\Gamma_{d,n,N} &\le 32\sqrt{2}(2B)^p\sqrt{\frac{\frac{1}{2}\gamma_{d,n}\,\log\mu_N + \log K_{d,n} + c_n}{\mu_N}}\\ &\quad + \frac{(640 + 32\sqrt{c'_d})\sqrt{2}(2B)^p}{\sqrt{\mu_N}} + 4(2B)^p\mu_N\beta_{a_N/2},\end{aligned}$$

where use has been made of (21) and (22). The claim follows upon using the definition of $\Gamma_{d,n,N}$. □

Acknowledgment

The author is grateful to Neri Merhav and Assaf Zeevi for their careful reading of the manuscript, and for their constructive and very helpful comments. Helpful comments from two anonymous reviewers and from the editor are also gratefully acknowledged.

References

Arcones, M.A. & Yu, B. (1994). Central limit theorems for empirical and u-proceses of stationary mixing sequences. *J. Theoretical Prob.*, *7*(1), 47–71.

Barron, A.R. (1993). Universal approximation bound for superpositions of a sigmoidal function. *IEEE Trans. Inf. Th.*, *39*, 930–945.

Barron, A. (1994). Approximation and Estimation Bounds for Artificial Neural Networks. *Machine Learning*, *4*, 115–133.

Barron, A., Birgé, L., & Massart, P. (1999). Risk Bounds for Model Selection via Penalization. *Th. and Related Fields*, *113*(3), 301–413.

Barron, A. & Cover, T. (1991). Minimum Complexity Density Estimation. *IEEE Trans. Inf. Theory*, *37*(4), 1034–1054.

Bartlett, P.L., Long, P.M., & Williamson, R.C. (1996). Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, *52*(3), 434–452.

Bartlett, P.L., Maiorov, V., & Meir, R. (1998). Almost linear vc dimension bounds for piecwise polynomial networks. *Neural Computation*, *10*, 2159–2173.

Blum, A. (1996). *On-line algorithms in machine learning*. Technical Report, Carnegie Mellon University, Presented at Dagstuhl Workshop on On-Line Algorithms. http://www.cs.cmu.edu/People/avrim/Papers/survey.ps.gz.

Bosq, D. (1996). *Nonparametric statistics for stochastic processes: Estimation and Prediction.* New York: Springer Verlag.

Brockwell, P.J. & Davis, R.A. (1991). *Time series: theory and methods*. New York: Springer Verlag.

Campi, M.C. & Kumar, P.R. (1998). Learning dynamical systems in a stationary environment. *Systems and Control Letters*, *34*, 125–132.

Craig, C.C. (1993). On the Tchebycheff Inequality of Bernstein. *Ann. Math. Statis.*, *4*, 94–102.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. New York: Springer Verlag.

Devore, R.A. & Lorentz, G.G. (1993). *Constructive approximation*. New York: Springer Verlag.

Doukhan, P. (1994). *Mixing—properties and examples*. New York: Springer Verlag.

Feder, M. & Merhav, N. (1996). Hierarchical universal coding. *IEEE Trans. Inf. Th.*, *42*(5), 1354–1364.

Geman, S. & Hwang, C.R. (1982). Nonparameteric Maximum Likelihood Estimation by the Method of Sieves. *Annals of Statistics*, *10*(2), 401–414.

Goldberg, P.W. & Jerrum, M.R. (1995). Bounding the VC Dimension of Concept Classes Parameterized by Real Numbers. *Machine Learning*, *18*, 131–148.

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

Györfi, L., Härdle, W., Sarda, P., & Vieu, P. (1989). *Nonparametric curve estimation from time series*. New York: Springer Verlag.

Haussler, D. (1992). Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications. *Information and Computation*, *100*, 78–150.

Haussler, D. (1995). Sphere Packing Numbers for Subsets of the Boolean n-Cube with Bounded Vapnik-Chervonenkis Dimesnion. *J. Combinatorial Theory, Series A*, *69*, 217–232.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statis. Assoc.*, *58*, 13–30.

Karpinksi, M. & Macintyre, A. (1997). Polynomial Bounds for VC Dimension of Sigmoidal and General Pfaffian Neural Networks. *Journal of Computer and System Science*, *54*, 169–176.

Lee, W.S., Bartlett, P.S., & Williamson, R.C. (1996). Efficient Agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Inf. Theory*, *42*(6), 2118–2132.

Lugosi, G. & Nobel, A. (1996). Adaptive Model Selection Using Empirical Complexities, preprint.
Lugosi, G. & Zeger, K. (1995). Nonparametric Estimation via Empirical Risk Minimization. *IEEE Trans. Inf. Theory*, *41*(3), 677–687.
Lugosi, G. & Zeger, K. (1996). Concept learning using complexity regularization. *IEEE Trans. Inf. Theory*, *42*(1), 48–54.
Maiorov, V.E. & Meir, R. (in press). On the near optimality of the stochastic approximation of smooth functions by neural network. *Advances in Computational Mathematics*.
Mhaskar, H. (1996). Neural Networks for Optimal Approximation of Smooth and Analytic Functions. *Neural Computation*, *8*(1), 164–177.
Modha, D. & Masry, E. (1998). Memory Universal Prediction of Stationary Random Processes. *IEEE Trans. Inf. Th.*, *44*(1), 117–133.
Mokkadem, A. (1988). Mixing Properties of ARMA Processes. *Stochastics Proc. Appl.*, *29*, 309–315.
Pollard, D. (1984). *Convergence of empirical processes*. New York: Springer Verlag.
Rosenblatt, M. (1971). *Markov processes: structure and asymptotic behavior*. New York: Springer Verlag.
Stone, C.J. (1982). Optimal global rates of convergence for nonparametric estimators. *Annals of Statistics*, *10*, 1040–1053.
van de Geer, S. (1987). A new approach to least-squares estimation, with applications. *The Annals of Statistics*, *15*(2), 587–602.
van der Vaart, A.W. & Wellner, J.A. (1996). *Weak convergence and empirical processes*. New York: Springer Verlag.
Vapnik, V.N. (1982). *Estimation of dependences based on empirical data*. New York: Springer Verlag.
Vapnik, V. & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, *16*, 264–280.
Vidyasagar, M. (1996). *A theory of learning and generalization*. New York: Springer Verlag.
White, H. (1991). Some results on sieve estimation with dependent observations. In J. Powell, W.A. Barnett, & G. Tau Chen (Eds.), *Nonparametric and semi-parametric methods in econometrics and statistics*. Cambridge University Press.
Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *Annals of probability*, *22*, 94–116.
Yule, G.U. (1927). On a Method of Investigating Periodicities in Disturbed Series with Special Reference to Wölfer's Sunspot Numbers. *Philos. Trans. Roy. Soc.*, *A226*, 267–298.
Zeevi, A., Meir, R., & Maiorov, V. (1998). Error bounds for functional approximation and estimation using mixtures of experts. *IEEE Trans. Information Theory*, *44*(3), 1010–1025.

Received April 17, 1998
Accepted May 18, 1999
Final manuscript May 17, 1999