Abstract—A nonparametric time series prediction strategy with finite bounds is presented. The results are obtained by extending previous results used for estimation with the framework of the structural risk minimization. It provides error bounds for estimators formed by minimizing the empirical error over a fixed class of functions.

Index Terms—Class of functions, error bound, prediction strategy, time series.

I. INTRODUCTION

Most of the researches [1], [2] in the area of time series modeling and prediction are closely related to parametric approaches. A simple linear model is usually fitted to the data. Although many important results have been obtained, this kind of parametric approach has severe limitations in that it applies only to processes with certain rigid structural characteristics. Nonparametric approaches, on the other hand, do not impose any structural assumptions, and can model any smooth processes. They are, however, often mathematically rather complicated.

A new structural risk minimization (SRM) [3], [4] method is presented in this paper for time series prediction problem. The major advantage of this approach is that it is nonparametric in spirit. Almost any functions can be modeled with this method. At the same time it is adaptive in that if the function to be modeled belongs to the sequence of models under consideration, the estimation scheme can converge to the true model at a rate similar to the one that would be attained had we known the true model in advance. In addition, this approach can achieve the maximum rate of convergence in nonparametric settings [5] under i.i.d. condition.

The remainder of this paper is organized as follows. In Section II, the time series prediction problem is introduced in general context and the definition of mixing processes is given. Section III explains the problem of prediction for mixing stochastic processes and on the basis of some well-known results, an estimator with finite sample performance bounds is derived. Application examples are presented in Section IV. Finally, in Section V, we summarize the paper and discuss a direction for future research.

II. TIME SERIES PREDICTION AND MIXING PROCESSES

One-step prediction of a stationary stochastic process \( \mathbf{X} = \{\ldots, X_{-1}, X_0, X_1, \ldots\} \), where \( X_i \) is a random real variable such that with probability 1 \( |X_i| \leq B \) (\( B < \infty \)), can be described in the expected \( L_p \) norm sense as that of computing a predictor function \( f() \) of the infinite past such that \( E[|X_0 - f(X_{-1}^i)|^p] \) is minimal. Here \( X_i^j = (X_i, X_{i+1}, \ldots, X_j), \ j \geq i \). Although for the special case \( p = 2 \), the optimal predictor is given by the conditional mean \( E[X_0 | X_{-1}^i] \), it does not settle the issue of actually computing the optimal predictor. In the case \( p > 2 \), the problem is further complicated because there is no analytic solution as in the case \( p = 2 \). In this paper optimal prediction of finite sub-sequence \( X_{i}^{N} = (X_i, X_{i+1}, \ldots, X_N) \) is considered based on a finite number of past values. We refer to this number \( d \) (\( d < N \rightarrow \infty \)) as the memory size. The optimal prediction is obtained by selecting an empirical estimator from a class of functions \( F_{d,n} : R^d \rightarrow R \), \( |f| \leq B \) for \( f \in F_{d,n} \), where \( n \) is a complexity index of the class. The empirical predictor \( \hat{f}_{d,n,N} (X_{i-1}^j) \) for \( X_i \) is based on the finite data vector \( X_{1}^{N} \) and the \( d \)-dimensional vector \( X_{i-1}^{d} \), where \( \hat{f}_{d,n,N} \in F_{d,n} \). The predictor error is written as

\[
L(\hat{f}_{d,n,N}) = L(\hat{f}_{d,n,N} - f_d^*) + (L_{d,n} - L_d^*) + L_d
\]

where \( L_d^* \) represents the error incurred by the function \( E |X_i - f_d^*(X_{-1}^i)|^p = \inf_{f : R^d \rightarrow R} E |X_i - f(X_{-1}^i)|^p \), in which \( f_d^* \) is the optimal predictor of memory size \( d \) minimizing the following loss of predictor \( f_d : R^d \rightarrow R \)

\[
L(f_d) = E |X_i - f_d(X_{-1}^i)|^p
\]

(2)

and \( L_{d,n} \) is the loss incurred in the class \( F_{d,n} \) by the optimal predictor \( f_{d,n}^* \), namely

\[
E |X_i - f_{d,n}^*(X_{-1}^i)|^p = \inf_{f \in F_{d,n}} E |X_i - f(X_{-1}^i)|^p
\]

Finally \( \hat{f}_{d,n,N} \) is an empirical estimator on the basis of the finite data set \( X_{1}^{N} \).

The first term in the above error equation is the so-called estimation error, and is the only term that depends on the

Ge Guo and Zhan T. Yuan

Manuscript received November 28, 2001; revised July 22, 2002. This work was supported in part by the Ninth 5-year Plan Project of China under Grant No. 85-524-02-05, the President Fund of Gansu Province under Grant No. G5015-A52-012 and the Doctors Fund of GSUT under Grant No. 139903.

The authors are with the College of Electrics and Information Engineering, Gansu University of Technology, Lanzhou, 730050 China (e-mail: guog@gsut.edu.cn; yuanzt@gsut.edu.cn).
data \( X_1^N \). The second term in (1) is related to the so-called approximation error, given by 
\[ E \left| f_d^*(X_{i-d}^{i-1}) - f_d^*(X_{i-d}^{i-1}) \right|^p \]
to which it can be immediately related through the inequality 
\[ \| a \|^p - \| b \|^p \leq p \| a - b \| \| \max(a,b) \|^p - 1. \]
This term measures the excess error incurred by selecting a function \( f \) from a class of limited complexity \( F_{d,a} \), while the optimal predictor \( f_d \) of memory size \( d \) may be arbitrarily complex. The third term is often referred to as the dynamic selection of the empirical estimator in using a finite memory model of memory size \( d \) to predict a stochastic process is said to be absolutely regular, or the excess error incurred by selecting a function \( f \) from a class of limited complexity \( F_{d,a} \), while the optimal predictor \( f_d \) of memory size \( d \) may be arbitrarily complex. The third term is often referred to as the dynamic selection of the empirical estimator in using a finite memory model of memory size \( d \) to predict a process with potentially infinite memory.

In this work we follow the ideas of [4] and restrict the selection of the empirical estimator \( \hat{f}_{d,a,N} \) to the function that minimizes the empirical error
\[ \hat{L}_N(f) = \frac{1}{N-d} \sum_{j=d+1}^{N} |X_j - f(X_{j-d}^{j-1})|^p. \]
Thus \( \hat{f}_{d,a,N} \) is the arg min with respect to \( f \). We assume throughout the paper that \( f_d^* \) and \( \hat{f}_{d,a,N} \) exist.

In this paper, we follow the widely used practice in the field of time series analysis, and restrict our work to the class of mixing processes. In these processes the future streams of the time series are conditionally independent of the past given the present. This property allows us to derive error bounds for prediction of mixing processes, we first need to define a vector-valued process \( \bar{X} = \{\ldots, \bar{X}_0, \bar{X}_1, \ldots\} \), where \( \bar{X}_i = (X_{i}, X_{i-1}, \ldots, X_{i-d}) \in \mathbb{R}^{d+1} \). For this sequence the \( \beta \)-mixing coefficients obey the inequality \( \beta_m(\bar{X}) \leq \beta_{m-d}(\bar{X}) \).

In this paper we use the notation \( F_{d,a} \) for the functional classes used for estimation, as the results depend explicitly on \( d \) and \( n \). By the definition of \( \bar{X}_i \), we can utilize \( E(f(\bar{X}_i)) \) for the following loss function \( \ell_{f} : R^{d+1} \rightarrow R \) of any function \( f : R^{d} \rightarrow R \). \( \ell_{f} \left( X_{i-d}^{i-1}, X_i \right) = |X_i - f(X_{i-d}^{i-1})|^p \).

And we restrict our work within the loss space \( \hat{L}_{F_{a}} = \{ \ell_{f} : f \in F_{d,a} \} \).

Noting that \( |\ell_{f}| \leq \alpha_N(2B)^{p} \), then using the results of [6] with transformation \( f \mapsto \ell_{f} \) and \( N \mapsto \mu_N \) respectively, the problem is now phrased as deriving upper bounds on the uniform deviations
\[ P \left\{ \sup_{f \in F_{d,a}} \left| \ell_{f}(f) - \ell_{f}(\hat{f}) \right| > \epsilon \right\} \leq 2 \mu_N \beta \mu_N \beta \]
\[ \leq 8E \left( \sum_{N} a_{N} \epsilon^{2} \sum_{|j| \leq 2(2j-1) a_{N}} \right) \sum_{|j| \leq 2(2j-1) a_{N}} \frac{\mu_N \beta^{2}}{128(2B)^{2p}} \]
(6)

where
\[ \tilde{\ell}_{f} = \sum_{i \in H_{j}} \ell_{f} = \sum_{i \in H_{j}} \left| X_{j} - f(X_{j-d}^{j-1}) \right|^{p}, \]
and the semi-norm
\[ \tilde{\mu}_{N}(\tilde{\ell}_{f}, \tilde{\ell}_{g}) = \frac{1}{\mu_N} \sum_{j=1}^{N} \left| \sum_{i \in H_{j}} \ell_{f}(X_{j-i}) - \ell_{g}(X_{j-i}) \right| \]

The covering number in (6) is taken with respect to the loss class \( \tilde{L}_{F_{a}} \). Let’s then relate the covering numbers of \( L_{F_{a}} \) and \( \tilde{L}_{F_{a}} \).

Lemma 3.1: For any \( \epsilon > 0 \)
\[ N(\epsilon, L_{F_{a}}(\Xi_{a}), \tilde{\mu}_{N}) \leq N(\epsilon/2a_{N}, L_{F_{a}}(\mathbb{R}^{N}), \tilde{\mu}_{N}). \]

Proof: The following sequence of inequalities holds:
\[ \tilde{\mu}_{N}(\tilde{\ell}_{f}, \tilde{\ell}_{g}) = \frac{1}{\mu_N} \sum_{j=1}^{N} \left| \sum_{i \in H_{j}} \ell_{f}(X_{j-i}) - \ell_{g}(X_{j-i}) \right| \]

conditions on the function in and let expected error (2). Then, the result follows.

Since the connection between the covering number \( N(\varepsilon, L_e(Z^N); l_{1,N}) \) and \( \eta \varepsilon \), we have the following theorem.

**Theorem 3.1:** Let \( \bar{X} = \cdots, X_{-1}, X_0, X_1, \cdots \) be a stationary \( \beta \)-mixing stochastic process, with \( |X_i| \leq B \), and let \( F_{d,n} \) be a class of bounded functions \( f : \mathbb{R}^d \rightarrow [-B, B] \). For each sample size \( N \), let \( \hat{f}_{d,n} \) be the function in \( F_{d,n} \) which minimizes the empirical error (3), and \( f_{d,n} \) is the function in \( F_{d,n} \) minimizing the expected error (2). Then,

\[
P\{L(\hat{f}_{d,n}) - L(f_{d,n}) > \varepsilon\} \leq 8EN\left[\frac{\varepsilon}{64\mu_B(2B)^2} \exp[-\mu_B \varepsilon^2 / (128(2B)^2)] + 2\mu_B \beta_{a_{d,n}}^{-d}\right]
\]

**Proof:** By using the results in [2] and inequality (7) with \( \tilde{l}_f \leq a_{d,n}(2B)^d \) and \( \eta = p(2B)^{d-1} \), we have the following sequence of inequalities

\[
P\{L(\hat{f}_{d,n}) - L(f_{d,n}) > \varepsilon\} \leq P\left\{\sup_{f \in F_{d,n}} |\tilde{L}_n(f) - L(f)| > \varepsilon / 2 \right\}
\]

\[
\leq 2P\left\{\sup_{f \in F_{d,n}} |\tilde{E}_{\mu_B} f - E f| > \varepsilon / 2 \right\} + 2\mu_B \beta_{a_{d,n}}^{-d}
\]

\[
\leq 8E\left[\frac{\varepsilon}{64\mu_B(2B)^2} \exp[-\mu_B \varepsilon^2 / (128(2B)^2)] + 2\mu_B \beta_{a_{d,n}}^{-d}\right]
\]

\[
\leq 8E\left[\frac{\varepsilon}{64\mu_B(2B)^2} \exp[-\mu_B \varepsilon^2 / (128(2B)^2)] + 2\mu_B \beta_{a_{d,n}}^{-d}\right]
\]

Up to this point the result is quite general because we have not specified \( a_{d,n} \) and \( \mu_B \). In order to obtain weak consistency we require that the r.h.s. of (8) converge to zero for each \( \varepsilon > 0 \). This immediately yields the following conditions on \( a_{d,n} \) and thus \( \mu_B \) through condition

\[2a_{d,n}\mu_B = N\]

where \( N = 2\mu_B a_{d,n} \). Setting \( l_{1,N}(f, \ell_g) \leq \varepsilon / 2a_{d,n} \) the following choices of \( a_{d,n} \) are sufficient to guarantee the weak consistency of the empirical minimizer \( \hat{f}_{d,n} \):

\[a_{d,n} = N^{1/2}2\Theta(N^{1/(2+\varepsilon)})\]

**Corollary 1:** Under the conditions of Theorem 3.1 and the additional requirement that \( EN(\varepsilon, F_{d,n}(X^N), l_{1,N}) < \infty \), the following choices of \( a_{d,n} \) are sufficient to guarantee the weak consistency of the empirical minimizer \( \hat{f}_{d,n} \):

\[a_{d,n} = N^{1/2}2\Theta(N^{1/(2+\varepsilon)})\]

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

**Proof:** Consider first the case of exponential mixing. In this case the r.h.s. of (8) clearly converges to zero because of the finiteness of the covering numbers. The fast rate of convergence is achieved by balancing the two terms in the equation, leading to the choice

\[a_{d,n} = N^{1/2}2\Theta(N^{1/(2+\varepsilon)})\]

under the condition \( a_{d,n} \) is known from

**Corollary 1.2:** Under the conditions of Theorem 3.1 and the added requirement that \( EN(\varepsilon, F_{d,n}(X^N), l_{1,N}) < \infty \), the following choices of \( a_{d,n} \) are sufficient to guarantee the weak consistency of the empirical minimizer \( \hat{f}_{d,n} \):

\[a_{d,n} = N^{1/2}2\Theta(N^{1/(2+\varepsilon)})\]

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 that is much faster than the rate obtained above in the case of mixing processes. In the latter case the effective sample size is reduced due to the temporal correlations, leading to the slower rates.

**Corollary 1.2:** Under the conditions of Theorem 3.1 and the added requirement that \( P \dim(F_{d,n}) < \infty \), there exists a finite value of \( N_0 \) such that for all \( N > N_0 \)

\[EL(\hat{f}_{d,n}) - L(f_{d,n}) \leq 8\sqrt{2\varepsilon}(2B)^d \leq \frac{8\sqrt{2\varepsilon}(2B)^d}{\sqrt{128\mu_B(2B)^2} + 2\mu_B \beta_{a_{d,n}}^{-d}}
\]

The proof of this corollary follows some of the ideas in [2], with appropriate modifications, to which we refer the reader for a proof. The explicit dependence on \( N \) may be obtained by substituting the values of \( \mu_B \) from Corollaries 3.1. So far, we can easily derive the following upper bound

\[EL(\hat{f}_{d,n}) - L(f_{d,n}) \leq 8\sqrt{2\varepsilon}(2B)^d \leq \frac{8\sqrt{2\varepsilon}(2B)^d}{\sqrt{128\mu_B(2B)^2} + 2\mu_B \beta_{a_{d,n}}^{-d}}
\]

for the exponentially mixing case. In the case of algebraic mixing one obtains the same rate with \( s \) replacing \( k \), where \( 0 < s < (r-1)/2 \).

**IV. APPLICATION EXAMPLES**

In this section we demonstrate the efficiency of the presented algorithm in the test case of obstacle detection.
and avoidance for a virtual mobile robot. In the beginning, the robot is set at an arbitrary location in the environment. In the absence of any initial positioning information, the localization module assumes a uniform position probability density function (PDF) along the heading direction of the robot and sets the robot in motion until the first landmark is in sight.

In order to support the new prediction method, the simulated robot carries many virtual sensors. All the signals are fed to a Kalman filter implemented for this set of sensors. The estimated robot position displacement is available at any time. Mathematically, the position displacement can be described by the following equation

\[
\sum_{j=0}^{p} a_j X_{t-j} = \sum_{k=0}^{q} b_k \epsilon_{t-k}
\]

where \( \epsilon_i \) are i.i.d. mean-zero random variables, \( X_t \) is robot position displacement, \( a_j \) and \( b_k \) are coefficients to be determined, \( p \) and \( q \) are data level. The probability distribution of \( \epsilon_i \) is assumed to be absolutely continuous. This model is exponentially \( \beta \)-mixing under the condition that the zeros of the polynomial \( A(z) = \sum_{j=0}^{p} a_j z^j \) lie outside the unit circle in \( z \) plane.

In simulations both Yule’s method and the present optimal prediction algorithm were studied. Fig. 1 shows the comparison of exploration results obtained using different position prediction methods. The robot position in Fig. 1(a) is estimated using Yule’s method while the present prediction method is used in Fig. 1(b). It is clear from Fig. 1(a) that the robot successfully avoids all the obstacles in the environment and reaches the target in the end. But the performance, namely the position prediction error and the exploration path, is clearly unsatisfactory. In Fig. 1(b), however, when the present prediction algorithm is used, the robot performs the same task much better.

Other successful applications include mould monitoring and diagnosis in a continuous casting line and a digital controller developed for optical disk drive, to name but a few. For more information about the details of the applications mentioned above, the readers are referred to [7] and [8].

V. CONCLUSION

In this paper, a nonparametric time series prediction strategy with finite bounds is presented. The results are obtained by extending previous results used for estimation with the framework of the structural risk minimization. It provides error bounds for estimators formed by minimizing the empirical error over a fixed class of functions. The major advantage of this approach is that it can adaptively achieve universal consistency and good rates of convergence in nonparametric settings, while retaining parametric rates of convergence in special situations.

It is clear, however, that the complexity of the class of functions plays a crucial role in the procedure. If the covering numbers are too large, 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 possible approach to this shortcoming is to consider 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 an approach is one of the issues need to be addressed in our future work.
ACKNOWLEDGMENT

The Authors would like to thank Ron Meir for his help in mathematical analysis of the result. Also, we thank the editors of IJECE and the three anonymous reviewers for their constructive suggestions and help in the preparation of the final paper.

REFERENCES


**Ge Guo** was born in 1972. He received the B.E. degree in automatic instrument and apparatus in 1994, and the Ph.D. degree in control theory and control engineering in 1998, both degrees are from Northeastern University, Shenyang, Liaoning, P. R. China.

In 1999, he joined the Department of Electrics and Information Engineering at Lanzhou University, Lanzhou, Gansu, as a research fellow for short periods. He was soon promoted to Associate Professor in Gansu University of Technology, Lanzhou, in 1999. In 2000, he was honored Trans-Century Pioneers in Science by the Educational Board of Gansu Province. His research interests include control theory and application, intelligent computing and robotics.

Dr. Guo is a member of the Association of Automation, China.

**Z. T. Yuan** was born in 1961. He received the B.E. and M.E. degrees in industrial automation from Xi’an Jiaotong University, Shaanxi, in 1985 and 1992, respectively.

Since 1992 he has been with the Department of Electrics and Information Engineering at Gansu University of Technology, Lanzhou, Gansu. With his outstanding work, he was appointed Dean of the Department of Electrics and Information Engineering. Soon after, he was promoted to Professor in 1996 and vice-president of the university in 1999. In 2000, he was honored Trans-Century Pioneers in Science by the Educational Board of Gansu Province. His research interests include information engineering, software and protocol engineering.

Prof. Yuan was awarded several times the Prize of Advancement in Science and Technology sponsored by Gansu province for his constructive research work. He possesses membership and leadership of several scientific associations of China.