Nonparametric Output Prediction for Nonlinear Fading Memory Systems

S. R. Kulkarni, *Senior Member, IEEE*, and S. E. Posner

***Abstract*—The authors construct a class of elementary nonparametric output predictors of an unknown discrete-time nonlinear fading memory system. Their algorithms predict asymptotically well for every bounded input sequence, every disturbance sequence in certain classes, and every linear or nonlinear system that is continuous and asymptotically time-invariant, causal, and with fading memory. The predictor is based on k_n-nearest neighbor estimators from nonparametric statistics. It uses only previous input and noisy output data of the system without any knowledge of the structure of the unknown system, the bounds on the input, or the properties of noise. Under additional smoothness conditions the authors provide rates of convergence for the time-average errors of their scheme. Finally, they apply their results to the special case of stable linear time-invariant (LTI) systems.**

***Index Terms*—Estimation, fading, filtering memory, identification, nearest-neighbor, nonlinear, nonparametric, prediction.**

I. Introduction

We INTRODUCE an elementary algorithm which predicts the output of an unknown nonlinear discrete-time system that satisfies certain generic regularity conditions, such as continuity and approximate time-invariance, causality, and fading memory. The algorithm only uses the past observed input and noisy output data and works for every bounded input sequence, every system in the class (without parametric and/or structural assumptions), and a wide range of disturbances. In this sense, the algorithm is "universal" in the terminology of information theory and statistics. The algorithm we use to achieve an asymptotically good predictor is an adaptation of the well-known k_n-nearest neighbor algorithm which has been analyzed extensively in the nonparametric statistics, pattern classification, and information theory literature [1], [10], [15].

Most previous work on output prediction has been parametric in nature. This encompasses many important areas in linear systems theory. For example, the Kalman filter uses the parameters of a linear system to construct a predictor in the presence of unknown stochastic disturbances. Similarly, the Luenberger observer uses the state-space matrices to construct an observer of the unknown state. In adaptive control and other schemes, system parameters are estimated and a controller is tuned online accordingly. In this paper, we are concerned with *nonparametric* prediction. We construct a predictor of the output of an unknown system assuming only generic conditions, but without any knowledge (or even an estimate) of system parameters.

Our nonparametric approach is in line with the work of several authors. In particular, Greblicki *et al.* (e.g., [8], [9], [12], and [13]) consider Hammerstein and Wiener systems which are nonlinear systems composed of linear systems coupled with memoryless nonlinearities. They consider these systems driven by stationary or i.i.d. noise and show that various nonparametric schemes can be used to estimate the nonlinearity. In contrast, we impose only mild regularity assumptions on the system without assuming any particular system structure, and our algorithm works for any bounded input sequence. Surprisingly, we provide a predictor for which we prove that the pointwise prediction errors tend to zero, even with the generality of our setup. The price we pay for this generality is, of course, in the rate of convergence, which is to be expected. In order to make statements about rates of convergence, stronger assumptions must be placed on both the plant and the input; or by making statements about the time-average of the prediction errors, we need only impose stronger conditions on the plant.

The role of prediction is also typically linked to that of system identification. System identification is concerned with using an algorithm to select a model from a model class (generally by selecting the model that best explains the measured data) so that the distance between the model and the true plant is small in some metric. Traditionally, system identification is ultimately used for control. The chosen model is used to design or tune a controller for the underlying system. Some recent work in system identification has focused on the theoretical limits of identification algorithms in a worst case setting, i.e., in which the output disturbances are only assumed to be bounded (e.g., see [22], [11], [19], [3], [17], and references cited therein). We consider both worst case and stochastic noise models, but in the context of prediction. Our results hold for a broad class of nonlinear systems quite similar to that studied in [2] in the context of worst case identification. In contrast with identification results which require a "sufficiently rich" input sequence, we show that prediction can be performed for arbitrary input sequences. Also, since we do not provide explicit estimates of the unknown system itself, we do not require any topological structure on the class of systems, which is required from the

Manuscript received September 6, 1996; revised January 5, 1998. Recommended by Associate Editor, M. A. Dahleh. This work was supported in part by the National Science Foundation under NYI Grant IRI-9457645. This work was completed while S. E. Posner was with the Department of Electrical Engineering, Princeton University and the Department of Statistics, University of Toronto.

S. R. Kulkarni is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA.

S. E. Posner is with Marsh & McLennan Securities Corporation, New York, NY 10021 USA.

Publisher Item Identifier S 0018-9286(99)00554-1.

outset for identification results. However, at present, we make no claims as to how our approach is to be used for the purpose of controller design. Rather, our focus here is simply on output prediction.

In a very different context, a similar scheme is used in [18]. They consider the estimation of conditional probabilities for stationary ergodic time series by looking at similar strings from the past and averaging the next value after each string. In contrast, in this paper, we focus on prediction of the output of a nonlinear dynamical system driven by an arbitrary bounded input. A similar algorithm is also used by Farmer and Sidorowich ([6] and references therein) in the context of predicting chaotic time series, although to our knowledge performance statements such as those presented here have not been shown. We suspect that the results in this paper can be used to make rigorous performance statements of their algorithm as well. Our work is also in the spirit of the work of Feder *et al.* [7] in which they construct a finite memory predictor of the next outcome of a binary sequence. However, the specific formulations are quite different in that we are in a systems framework, we have access to the input sequence which provides information about the unknown output, and we focus on different algorithms.

In Section II we formulate the problem and precisely define the class of systems, inputs, and noise under consideration. In Section III, we introduce a class of data-dependent but elementary nonparametric estimators and show (Theorem 1) that with bounded input and noise sequences we can predict *pointwise* asymptotically well to within the level of the noise, and that for stochastic noise, we can get asymptotically zero mean square prediction error. In Section IV, we consider rates of convergence. With only additional regularity conditions on the system (Lipschitz continuity and rates on fading memory), rates of convergence for the time-average of the prediction errors can be obtained for every bounded input sequence (Theorem 2). In order to get rates on the pointwise prediction errors, additional conditions are needed on both the input and the system. We show (Theorem 3) that if we impose independence or stationarity assumptions on the inputs and the additional conditions on the system, then a uniform rate of convergence can be obtained for the pointwise prediction errors. In Section VI we consider in more detail the special case of stable linear time-invariant (LTI) systems.

II. Formulation and Preliminaries

We consider the online prediction of the output of an unknown discrete-time system based on past inputs and noisy output observations. Suppose an unknown discrete-time system H is driven by an input sequence $u_0, u_1, \cdots$. We consider the following sequential prediction problem. By time n we have observed the past inputs $u_0, \cdots, u_{n-1}$ and corresponding noisy outputs

$$z_i = y_i + e_i, \qquad i = 1, \cdots, n-1$$

where y_i is the output of H at time i and e_i represents measurement noise. We then observe the input u_n at time n and produce an estimate $\hat{y}_n$ of the uncorrupted output y_n. Our goal is to have small estimation errors as n increases. Precise conditions on the system H, the input $u_0, u_1, \cdots$, and the noise $e_0, e_1, \cdots$ are given below.

We consider systems H: $\mathbb{R}^\infty \to \mathbb{R}^\infty$ where $\mathbb{R}^\infty = \{(u_0, u_1, \cdots)\colon u_i \in \mathbb{R}, i \geq 0\}$. For a subset $S \subset \{0, 1, 2, \cdots\}$, we define the projection operator P_S: $\mathbb{R}^\infty \to \mathbb{R}^{|S|}$, in the natural way. We use the notation

$$\begin{aligned}&\text{for } S = \{m, m+1, \cdots, n\}, \quad P_{[m,n]}u = (u_m, u_{m+1}, \cdots, u_n)\\ &\text{for } S = \{n, n+1, \cdots\}, \quad P_{[n,\infty)}u = (u_n, u_{n+1}, \cdots)\\ &\text{for } S = \{n\}, \quad P_n u = u_n\end{aligned}$$

for every $u = (u_0, u_1, \cdots) \in \mathbb{R}^\infty$. Note that in this paper we abuse consistency but conform to standard notation and use some lowercase letters to mean vectors in $\mathbb{R}^\infty$ (e.g., e, y, z, u, v) and other lowercase letters may be constants or whatever the context dictates. Similarly, uppercase letters may be operators, vectors, or constants depending on the context.

Define the closed ball of radius r of a Banach space $(\mathcal{X}, \|\cdot\|)$ as

$$\mathcal{B}_r\mathcal{X} = \{x \in \mathcal{X}\colon \|x\| \leq r\}.$$

We will mostly deal with the following balls:

$$\mathcal{B}_r\ell_\infty = \{u \in \ell_\infty\colon \|u\|_\infty \leq r\} \quad \text{where} \quad \|u\|_\infty = \sup_{i\geq 0} |u_i|$$

and

$$\mathcal{B}_r\mathbb{R}^L = \{u \in \mathbb{R}^L\colon \|u\|_\infty \leq r\}, \qquad L \geq 1.$$

We consider the online prediction of the output of an unknown system that satisfies certain general regularity conditions. The input may be any sequence in $\mathcal{B}_r\ell_\infty$. The measured output is corrupted with an additive disturbance sequence, $e \in \mathbb{R}^\infty$. The system model is

$$\begin{aligned} y &= H(u)\\ z &= y + e.\end{aligned}$$

We consider both deterministic and stochastic disturbance classes:

- $e \in \mathcal{B}_\delta\ell_\infty$, with $\delta \geq 0$;
- $e_0, e_1, \cdots$ i.i.d. zero mean and finite variance.

We consider systems that satisfy the following properties, where r is the same parameter as the bound on the allowable input sequences.

A1) H is continuous (but not necessarily linear) on $\mathcal{B}_r\ell_\infty$, i.e., $\|H(u) - H(v^{(n)})\|_\infty \xrightarrow{n\to\infty} 0$ if $\|u - v^{(n)}\|_\infty \xrightarrow{n\to\infty} 0$, for $u, v^{(1)}, v^{(2)}, \cdots \in \mathcal{B}_r\ell_\infty$.

A2) For each $\epsilon > 0$, there exists $T = T(\epsilon) \geq 1$ and $L = L(\epsilon) \geq 1$ such that the output at all times $n, m \geq T$ depends only on the previous L input components to within ϵ, i.e.,

$$|P_n H(u) - P_m H(v)| \leq \epsilon$$

for all $u, v \in \mathcal{B}_r\ell_\infty$ such that $P_{[m-L+1,m]}v = P_{[n-L+1,n]}u$.

Condition A1) is straightforward. Condition A2) is an asymptotically time-invariant, causality, and fading memory condition. These conditions are fairly general and contain many classes that are of common interest. For example, any stable LTI causal discrete-time system satisfies A1) and A2), as will be shown in Section VI. Also, Hammerstein and Wiener systems [8], [12], [13] satisfy A2) and, if the nonlinear element is continuous, A1).

III. Predictor and Main Result

Our first result is to construct a prediction algorithm that achieves pointwise convergence. We exhibit an algorithm which predicts the output of any unknown system in our class subjected to any bounded input. At time n, we have observed all the past inputs $(u_0, u_1, \cdots, u_{n-1})$ together with u_n, and we have observed noisy outputs $(z_0, z_1, \cdots, z_{n-1})$. We would like to estimate the uncorrupted system output, y_n, using an algorithm that produces an estimate $\hat{y}_n$ so that the prediction errors tend to zero asymptotically.

The algorithm we propose is an adaptation of the k_n-nearest neighbor estimators from nonparametric statistics. The basic idea of the algorithm is as follows. Take the most recent L_n inputs (where L_n is a data-dependent parameter specified in detail below) as a nominal vector in $\mathbb{R}^{L_n}$ and find the previous input substring of length L_n that is nearest to it in the $\mathrm{supnorm}$ sense. A natural estimate for y_n would be the output associated with the input vector that is nearest the nominal vector. That is, find a time in the past when the most recent L_n inputs were most similar to the current L_n inputs and use the observed output at that time as our prediction. The idea is that by the assumed continuity of the system, nearby inputs should produce nearby outputs. If there was no noise in our observations and if the parameter L_n was chosen wisely, we might expect that this algorithm would perform well. However, this basic idea needs to be refined in several ways.

First, since the "order" of the system is assumed unknown it is clear that we will need $L_n \to \infty$ as $n \to \infty$ if we wish to drive prediction errors to zero. The reason we have any hope of driving prediction errors to zero is due to the assumption that the system has fading memory. Actually, the fading memory assumption must be used in another way as well. To avoid the effects of initial conditions, we should not use input strings too close to time zero. Hence, the second refinement is to introduce another parameter T_n which tends to infinity as $n \to \infty$ and only search for nearby input strings that occur after time T_n. The third refinement we need results from the fact that our output observations are noisy. With random noise, the output at the time associated with the nearest input string may not necessarily give a good prediction. To average out the noise in the output observations, we could search for a number of past input strings that are close to the recent string and average the corresponding outputs to form our prediction. Thus, we introduce a third parameter, k_n, which is the number of "nearest neighbors" in the input string that we search for, and we will need $k_n \to \infty$ as $n \to \infty$. (Actually, in the case of worst case noise, averaging is not needed, and we can just take $k_n = 1$.)

At time n, given $u_0, \cdots, u_n$ and the parameters k_n, L_n, and T_n, let $\{U_j(L_n)\}_{j=T_n+L_n}^{n}$ be the set of all strings from the past input sequence after time T_n that are of length L_n. That is

$$U_j(L_n) = P_{[j-L_n+1,j]}u.$$

Note that each $U_j(L_n)$ is a vector in $\mathbb{R}^{L_n}$. Let $m_n^{[i]}$ be the index of the ith nearest neighbor (NN) of $U_n(L_n)$ (which is the most recent string of inputs of length L_n) from the set $\{U_j(L_n)\}_{j=T_n}^{n-1}$. The first NN distance $d_n(1, L_n) = d_n(1, L_n, T_n; u_0, \cdots, u_n)$ satisfies

$$\begin{aligned} d_n(1, L_n) &= \min_{T_n \le j < n} \|U_n(L_n) - U_j(L_n)\|_\infty \\ &= \|U_n(L_n) - U_{m_n^{[1]}}(L_n)\|_\infty \\ &= \max_{0 \le j < L_n} |u_{n-j} - u_{m_n^{[1]}-j}|. \end{aligned}$$

Similarly, $d_n(i, L_n) = d_n(i, L_n, T_n; u_0, \cdots, u_n)$ is the ith smallest distance between $U_n(L_n)$ and $\{U_j(L_n)\}_{j=T_n}^{n-1}$ and equals $\|U_n(L_n) - U_{m_n^{[i]}}(L_n)\|_\infty$. Consider the simple predictor

$$\hat{y}_n = \frac{1}{k_n} \sum_{i=1}^{k_n} z_{m_n^{[i]}} \tag{1}$$

where $z_{m_n^{[i]}} = P_{m_n^{[i]}} H(u) + e_{m_n^{[i]}}$ is the output observation at time $m_n^{[i]}$.

To complete the specification of the predictor, we need only specify the choice of the parameters k_n, L_n, and T_n as a function of n. Of course, to get asymptotically good predictions, the parameters need to be chosen carefully. In particular, to exploit the continuity of the system, we need the k_n nearest input strings to get closer and closer to the most recent input string as $n \to \infty$. An important quantity is the k_nth nearest neighbor distance $d_n = d_n(k_n, L_n, T_n; u_0, \cdots, u_n)$ which is the distance between the most recent string of length L_n and the k_nth nearest neighbor from past strings of length L_n occurring after time T_n. That is

$$d_n(k_n, L_n, T_n; u_0, \cdots, u_n) = \|U_n(L_n) - U_{m_n^{[k_n]}}(L_n)\|_\infty.$$

We need to make sure that $d_n \to 0$ as $n \to \infty$. Boundedness of the input is crucial in this regard. With a boundedness assumption, input strings of any fixed length L belong to a compact subset of $\mathbb{R}^L$, and it is this compactness that allows a suitable choice of parameters to make $d_n \to 0$.

However, it can be shown that with any fixed choice of the parameters k_n, L_n, and T_n, there is always a bounded input $u_0, u_1, \cdots$ for which $d_n(k_n, L_n, T_n; u_0, \cdots, u_n) \not\to 0$. This is typical of data-independent algorithms [15], [20] and is the reason one cannot get pointwise convergence with such algorithms for arbitrary inputs. In such cases, one must resort to making statements about the time-average of the prediction errors. To overcome this problem, we use suitable data-dependent choices of the algorithm parameters as in [14]. By choosing k_n, L_n, and T_n to depend on the observed input $u_0, \cdots, u_n$, we can construct an algorithm for which $k_n, L_n, T_n \to \infty$ and $d_n \to 0$. In this case, we can show that the pointwise prediction errors (and hence also the time-average errors) converge to zero.

Lemma 1: For every bounded sequence of inputs $u_0, u_1, \cdots$, if $k_n = k_n(u_0, \cdots, u_n)$, $L_n = L_n(u_0, \cdots, u_n)$, and $T_n = T_n(u_0, \cdots, u_n)$, are defined by

$$(k_n, L_n, T_n) = \underset{(k,L,T)}{\operatorname{argmin}} \frac{1}{k} + \frac{1}{L} + \frac{1}{T} + d_n(k, L, T; u_0, \cdots, u_n)$$

then we have the following.

1) $k_n(u_0, \cdots, u_n) \rightarrow_{n\rightarrow\infty} \infty$.
2) $L_n(u_0, \cdots, u_n) \rightarrow_{n\rightarrow\infty} \infty$.
3) $T_n(u_0, \cdots, u_n) \rightarrow_{n\rightarrow\infty} \infty$.
4) $d_n(k_n, L_n, T_n; u_0, \cdots, u_n) \rightarrow_{n\rightarrow\infty} 0$.

Proof: Let

$$J_n = \min_{(k,L,T)} \frac{1}{k} + \frac{1}{L} + \frac{1}{T} + d_n(k, L, T; u_0, \cdots, u_n).$$

We need only show that $\lim_{n\rightarrow\infty} J_n = 0$. We will do this by showing that for any $\epsilon > 0$, we have $J_n < \epsilon$, for all $n \geq n_0$ for some sufficiently large n_0.

Fix $\epsilon > 0$ and take $k, L, T > 4/\epsilon$. Consider the set of all consecutive inputs strings of length L after time T, denoted $u_1^{L,T}, u_2^{L,T}, \cdots$. Since $|u_i| \leq r$ for each i, the length L input strings are all elements of the hypercube $[-r, r]^L \subset \mathbb{R}^L$ which is a totally bounded set. Let $B_1, \cdots, B_{N(\epsilon/8)}$ denote balls of radius $\epsilon/8$ forming a finite cover of $[-r, r]^L$. The number of times a string $u_i^{L,T}$ falls in some ball with fewer than k previous elements from $u_1^{L,T}, \cdots, u_{i-1}^{L,T}$ is bounded by $kN(\epsilon/8)$. Hence, there is a finite n_0 such that for all $n \geq n_0$ at least k previous length L input strings fall in the same $\epsilon/8$ ball as the most recent string. In this case, at least k strings from $u_1^{L,T}, \cdots, u_{n-1}^{L,T}$ are within $\epsilon/4$ of $u_n^{L,T}$, so that $d_n(k, L, T; u_0, \cdots, u_n) < \epsilon/4$. Thus, for $n \geq n_0$ we have

$$J_n \leq \frac{1}{k} + \frac{1}{L} + \frac{1}{T} + d_n(k, L, T; u_0, \cdots, u_n) < \epsilon.$$

■

The following theorem, our main result, describes the asymptotic behavior of our data-dependent nonparametric predictor. The algorithm does not need to know any of the parameters used in the assumptions on the input, system, or noise. The proof of this result is given in Section V.

Theorem 1: Consider the predictor $\{\hat{y}_n\}$ given by (1) where k_n, L_n, T_n are chosen in a data-dependent manner according to Lemma 1. Then for any $u \in \mathcal{B}_r\ell_\infty$ for some $r < \infty$ and any H that satisfies A1) and A2), we have that:

1) for any $e \in \mathcal{B}_\delta\ell_\infty$
$$\limsup_{n\rightarrow\infty} |y_n - \hat{y}_n| \leq \delta$$
2) for any i.i.d. $e_0, e_1, \cdots$ such that $Ee_i = 0$ and $E|e_i|^2 < \infty$
$$\lim_{n\rightarrow\infty} E(y_n - \hat{y}_n)^2 = 0.$$

Notes:

- There is no uniform rate of convergence over the entire input class.
- The parameters used in the algorithm for the proof depends on the actual input sequence, in contrast with Theorem 2 (in the following section) in which the parameters are fixed and independent of the input sequence. Of course, the choice of parameters used in Lemma 1 is not the only one which will work. Many data-dependent schemes can achieve the conclusion of Lemma 1 and hence the result of Theorem 1.
- The upper bound for part 1 is clearly tight since errors of at least δ can be forced by the noise sequence each time.
- With arbitrary bounded inputs and without Assumption A1), no asymptotic prediction is possible.

IV. Rates of Convergence of Predicition Errors

The result of Theorem 1 shows that an appropriate data-dependent predictor provides estimates of the uncorrupted output such that the estimation errors converge to zero (for i.i.d. noise). However, it is easy to verify that no uniform rate of convergence is possible. The inability to obtain a uniform convergence rate arises from two distinct and fundamental reasons. One reason is that with arbitrary bounded inputs, one can construct input sequences such that the k_nth nearest neighbor distance converges to zero arbitrarily slowly. In fact, one can make the 1-NN distance converge to zero arbitrarily slowly. Thus, although the predicted output will be an average of outputs due to nearby inputs, we have no way of bounding how close the inputs will be at any particular time. However, even if we had such a bound, we still could not get a uniform rate of convergence due to a second reason, which involves the regularity of the unknown system. Namely, although continuity implies that nearby inputs result in nearby outputs, we need a stronger assumption, such as a Lipschitz condition, to have a hope of getting rates. Also, although the system is assumed to have fading memory, we need bounds on the rate at which the memory fades in order to get bounds on the prediction errors.

Thus, to obtain rates of convergence, we need assumptions on the inputs that allow bounding the nearest neighbor distances and conditions on the system that give stronger versions of Assumptions A1) and A2). A result of this type is provided in Theorem 3 below. However, first we give a rate result of a different sort. Namely, by considering the time-average of the prediction errors, we can obtain uniform rates on the time-average errors with assumptions only on the system (i.e., that hold for all bounded inputs). The basic ideas of this result will also be used for the pointwise rate result of Theorem 3. Interestingly, the prediction algorithms used in this section are of the same basic form as in Section III, but with the added simplification that the parameters k_n, L_n, and T_n need only satisfy certain rate conditions but can be chosen in a data-independent fashion.

The following stronger versions of Assumptions A1) and A2) will be used to get the rate results in this section.

A1′) There exists $K, \alpha > 0$ such that for all $u, v \in \mathcal{B}_r\ell_\infty$
$$\|H(u) - H(v)\|_\infty \leq K\|u - v\|_\infty^\alpha.$$

A2′) There exists $C \geq 0$, $0 < \rho < 1$, and T such that for $n, m > T$
$$|P_nH(u) - P_mH(v)| \leq C\rho^L$$

for every $u, v \in \mathcal{B}_r\ell_\infty$ such that $P_{[m-L+1,m]}v = P_{[n-L+1,n]}u$.

For example, as we will see in the final section, stable linear systems with a decay rate on the impulse response satisfy A1′) and A2′).

Fix nondecreasing sequences $k_n = \log^2 n$, $T_n = \log n$, and $L_n = \sqrt{\log((n-T_n)/k_n)}$. With these data-independent specifications on the parameters k_n, L_n, and T_n we use the simple predictor (1).

Theorem 2: Consider the predictor given by (1) where k_n, T_n, and L_n are chosen data-independently as above. Then for any $u \in \mathcal{B}_r\ell_\infty$ for some $r < \infty$ and any H that satisfies A1′) and A2′), we have that:

1) for any $e \in \mathcal{B}_\delta\ell_\infty$

$$\frac{1}{N}\sum_{n=1}^{N} |y_n - \hat{y}_n| \le \delta + \frac{\eta_1}{L_N^{2\gamma}} \le \delta + \frac{2\eta_1}{\log^\gamma N}$$

2) for any i.i.d. $e_0, e_1, \cdots$ such that $Ee_i = 0$ and $E|e_i|^2 < \infty$

$$\frac{1}{N}\sum_{n=1}^{N} E|y_n - \hat{y}_n|^2 \le \frac{\eta_2}{L_N^{4\gamma}} \le \frac{2\eta_2}{\log^{2\gamma} N}$$

where $\gamma = \min\{\alpha, -\log\rho\}$ and η_1, η_2 are well-defined constants.

Notes:

- It can be shown that with the algorithm used in the proof of Theorem 2, pointwise errors do *not* tend to zero for all bounded input sequences. The problem is that the parameters k_n, L_n, and T_n were chosen at the outset, independent of the inputs observed. In this case, one can always find a system and construct an input sequence for which the pointwise prediction errors do not converge to zero. The construction simply makes sure that the input is chosen so that the distance between the most recent input string of length L_n and its k_nth nearest neighbor does not converge to zero. This is the same the reason that a time-average criterion was required in [15].
- The algorithm is completely data-independent as well as independent of knowledge of the parameters in A1′) and A2′).
- Interestingly, the same algorithm works regardless of the noise class, although of course the mode of convergence depends on the type of noise.
- The time-average nature of the statements in the theorem arises not because of the noise but as a result of the arbitrary bounded inputs that are allowed.
- This algorithm can be readily modified to allow cases in which output data is missing. The only restriction is that the number of omissions must be $o(k_n)$.

Our next result is to obtain a pointwise rate of convergence by imposing stationarity on the input sequence in addition to the mentioned necessary assumptions on the system. We will use a simple modification of the algorithm used in Theorem 2 in order to exploit the stationarity of the input. Specifically, instead of searching for nearest neighbors over all strings $\{U_j(L_n)\}$ of length L_n, we now take only the set of *nonoverlapping* strings of length L_n from the past, i.e., we search for nearest neighbors from the set $\{U_j(L_n)\colon j = iL_n$ and $T_n + L_n \le j \le n\}$. The rest of the algorithm is the same. With this modification, we obtain the following result, which is proved in the following section.

Theorem 3: Consider the predictor given by (1) where k_n, T_n, and L_n are chosen data-independently and the nearest neighbors are selected from the set of nonoverlapping strings as described above. Then for any H that satisfies A1′) and A2′), we have that for any stationary $u_0, u_1, \cdots \in \mathbb{R}$, and any i.i.d. $e_0, e_1, \cdots$ such that $Ee_i = 0$ and $E|e_i|^2 < \infty$

$$E|y_n - \hat{y}_n|^2 = O\left[\left(\frac{1}{\log n}\right)^{2\gamma}\right]$$

where $\gamma = \min\{\alpha, -\log\rho\}$.

Notes:

- A similar statement can be made for *independent* inputs $u_0, u_1, \cdots$ with a weaker form of A1′) such as in [10].

V. Proofs of Theorems

A. Proof of Theorem 1

Proof: From (1), the prediction error at any time n satisfies

$$\begin{aligned} |y_n - \hat{y}_n| &= \left| P_nH(u) - \frac{1}{k_n}\sum_{i=1}^{k_n} P_{m_n^{[i]}}H(u) - \frac{1}{k_n}\sum_{i=1}^{k_n} e_{m_n^{[i]}} \right| \\ &\le \left| P_nH(u) - \frac{1}{k_n}\sum_{i=1}^{k_n} P_{m_n^{[i]}}H(u) \right| + \left| \frac{1}{k_n}\sum_{i=1}^{k_n} e_{m_n^{[i]}} \right| \\ &\equiv J_1 + J_2 \end{aligned} \tag{2}$$

where $J_1 = J_{n1}$ and $J_2 = J_{n2}$ are the two terms on the right-hand side of the inequality.

We first show that $J_1 \to 0$. Fix $\epsilon > 0$. Let $T = T(\epsilon)$ and $L = L(\epsilon)$ be as in Assumption A2). From Lemma 1, there exists N_1 such that $T_n(u_0, \cdots, u_n) \ge T(\epsilon)$ and $L_n(u_0, \cdots, u_n) \ge L(\epsilon)$ for all $n \ge N_1$.

Consider the mapping $H^{u,T,L}\colon \mathbb{R}^L \to \mathbb{R}$ defined by

$$H^{u,T,L}(v) = P_{T+L}H(u_0, \cdots, u_T, v_1, \cdots, v_L, u_{T+L+1}, \cdots).$$

That is, the mapping $H^{u,T,L}$ on input $v \in \mathbb{R}^L$ replaces the values of u between times $T+1$ and $T+L$ (inclusive) by those of v and returns the output of system H on this input at time $T+L$. By Assumption A1), $H^{u,T,L}(\cdot)$ is continuous on $[-r, r]^L$ and so compactness of the domain implies that in fact $H^{u,T,L}(\cdot)$ is uniformly continuous on $[-r, r]^L$ (e.g., see [5, Corollary 2.4.6]). Hence, there exists $\delta > 0$ such that for any $v_1, v_2 \in [-r, r]^L$ we have $|H^{u,T,L}(v_1) - H^{u,T,L}(v_2)| < \epsilon$ whenever $\|v_1 - v_2\|_\infty < \delta$. Again using Lemma 1, there exists N_2 such that $d_n(k_n, L_n, T_n; u_0, \cdots, u_n) < \delta$ for all $n \ge N_2$.

Define the vectors $v^{[i]} \in \mathcal{B}_r\ell_\infty$ as

$$\begin{aligned} P_{[0,T]}v^{[i]} &= P_{[0,T]}u \\ P_{[T+L+1,\infty)}v^{[i]} &= P_{[T+L+1,\infty)}u \\ P_{[T+1,T+L]}v^{[i]} &= P_{[m_n^{[i]}-L+1,m_n^{[i]}]}u. \end{aligned} \tag{3}$$

Also, define the vector $v^{[0]} \in \mathcal{B}_r\ell_\infty$ as

$$\begin{aligned} P_{[0,T]}v^{[0]} &= P_{[0,T]}u \\ P_{[T+L+1,\infty)}v^{[0]} &= P_{[T+L+1,\infty)}u \\ P_{[T+1,T+L]}v^{[0]} &= P_{[n-L+1,n]}u. \end{aligned}$$

Now, for any $n \geq \max\{N_1, N_2\}$, we can bound J_1 as follows:

$$\begin{aligned} J_1 &= \left| P_nH(u) - \frac{1}{k_n}\sum_{i=1}^{k_n} P_{m_n^{[i]}}H(u) \right| \\ &\leq \left| P_nH(u) - P_{T+L}H(v^{[0]}) \right| \\ &\quad + \left| P_{T+L}H(v^{[0]}) - \frac{1}{k_n}\sum_{i=1}^{k_n} P_{T+L}H(v^{[i]})) \right| \\ &\quad + \left| \frac{1}{k_n}\sum_{i=1}^{k_n} P_{T+L}H(v^{[i]})) - \frac{1}{k_n}\sum_{i=1}^{k_n} P_{m_n^{[i]}}H(u) \right| \\ &\equiv J_{1,1} + J_{1,2} + J_{1,3} \end{aligned}$$

where $J_{1,1}$, $J_{1,2}$, $J_{1,3}$ are the three terms on the right-hand side of the inequality. Since $n \geq N_1$, we have $T_n \geq T$ and $L_n \geq L$. Therefore, from A2) we get $J_{1,1} \leq \epsilon$ and $J_{1,3} \leq \epsilon$. To bound $J_{1,2}$, let $w^{[i]} = P_{[T+1,T+L]}v^{[i]}$ for $i = 0, 1, \cdots, k_n$. Then $w^{[i]} \in \mathbb{R}^L$ and are simply substrings of length L of the nearest neighbor strings. Since $n \geq N_2$ we have $d_n(k_n, L_n, T_n; u_0, \cdots, u_n) < \delta$ and since $L_n \geq L$ this implies $\|w^{[0]} - w^{[i]}\| < \delta$ for $i = 1, \cdots, k_n$. Hence

$$\begin{aligned} J_{1,2} &= \left| P_{T+L}H(v^{[0]}) - \frac{1}{k_n}\sum_{i=1}^{k_n} P_{T+L}H(v^{[i]})) \right| \\ &= \left| H^{u,T,L}(w^{[0]}) - \frac{1}{k_n}\sum_{i=1}^{k_n} H^{u,T,L}(w^{[i]}) \right| \\ &\leq \max_{1\leq i\leq k_n} |H^{u,T,L}(w^{[0]}) - H^{u,T,L}(w^{[i]})| \\ &\leq \epsilon \end{aligned}$$

by the uniform continuity of $H^{u,T,L}(\cdot)$ and the choice of δ. Thus, $J_1 \leq 3\epsilon$ for any $n \geq \max\{N_1, N_2\}$, and since $\epsilon > 0$ was arbitrary, we have that $\lim_{n\to\infty} J_1 = 0$.

The above analysis is common to both statements of the theorem. To conclude the two statements in the theorem we need only study the behavior of J_2. For the first statement, $e \in \mathcal{B}_\delta\ell_\infty$, and hence $J_2 \leq \delta$, completing the claim. (Observe that we do not need the fact that $k_n \to \infty$ for this part.) For the second statement, observe that since the e_i are i.i.d. with zero mean and finite variance, we have

$$\begin{aligned} E[J_2^2] &= E\left| \frac{1}{k_n}\sum_{i=1}^{k_n} e_{m_n^{[i]}} \right|^2 \\ &= \frac{1}{k_n^2}\sum_{i=1}^{k_n} E|e_{m_n^{[i]}}|^2 \leq \frac{\operatorname{var}(e_0)}{k_n}. \end{aligned} \tag{4}$$

Then $E[J_2^2] \to 0$ since by Lemma 1 $k_n \to \infty$. This completes the proof.

B. Proof of Theorem 2

Proof: The pointwise prediction error is bounded by $J_{n1} + J_{n2}$ as in (2). Now using (3)

$$\begin{aligned} J_{n1} &= \left| \frac{1}{k_n}\sum_{i=1}^{k_n} (P_nH(u) - P_nH(v^{[i]}) \right. \\ &\qquad \left. + P_nH(v^{[i]}) - P_{m_n^{[i]}}H(u)) \right| \\ &\leq \left| \frac{1}{k_n}\sum_{i=1}^{k_n} (P_nH(u) - P_nH(v^{[i]})) \right| + \epsilon_n \end{aligned} \tag{5}$$

where $\epsilon_n = C\rho^{L_n}$. By A1$'$) we have that

$$\begin{aligned} J_{n1} &\leq \frac{K}{k_n}\sum_{i=1}^{k_n} d_n^\alpha(i, L_n, T_n) + \epsilon_n \\ &\leq K d_n^\alpha(k_n, L_n, T_n) + \epsilon_n. \end{aligned} \tag{6}$$

The prediction error in summary is

$$|y_n - \hat{y}_n| \leq J_{n1} + J_{n2} \leq K d_n^\alpha(k_n, L_n, T_n) + C\rho^{L_n} + J_{n2}.$$

For the first statement in the theorem, $J_{n2} \leq \delta$, and so we only need show that $(1/N)\,\Sigma_{n=1}^N J_{n1} \to 0$ at the rate specified. To this end, the main step is as follows. It was shown in [15] that

$$\begin{aligned} &\frac{1}{N}\sum_{n=1}^{N} d_n^\alpha(k_n, L_n, T_n) \\ &\quad \leq \frac{1}{N}\sum_{n=1}^{N} d_n^\alpha(k_N, L_N, T_N) \leq \beta_1\left(\frac{k_N}{N - T_N}\right)^{\alpha/L_N} \end{aligned}$$

where β_1 is a constant that depends only on r (the bound on the inputs). By construction $L_n^2 = \log((n - T_n)/k_n)$, and so we get

$$\frac{1}{N}\sum_{n=1}^{N} J_{n1} \leq \beta_1 e^{-\alpha L_n} + C\frac{1}{N}\sum_{n=1}^{N} \rho^{L_n} \leq \beta\rho_1^{L_n}$$

where $\rho_1 = \max\{\rho, e^{-\alpha}\} < 1$ and $\beta = 2\max\{\beta_1, 2C\}$. Next note that

$$\begin{aligned} \rho_1^{L_n} &= \rho_1^{\sqrt{\log((n-T_n)/k_n)}} \leq \rho_1^{\log\log((n-T_n)/k_n)} \\ &= \left(\log\frac{n - T_n}{k_n}\right)^{\log\rho_1} \\ &= \left(\log\frac{n - T_n}{k_n}\right)^{-\gamma} \leq \eta_3(\log n)^{-\gamma} \end{aligned}$$

where $\gamma = \min\{\log(1/\rho), \alpha\}$ and η_3 is a well-defined constant. The second statement in the theorem is similar. First

$$|y_n - \hat{y}_n|^2 \leq 2J_{n1}^2 + 2J_{n2}^2. \tag{7}$$

From (4), we have $E[J_{n2}^2] \leq \operatorname{var}(e_0)/k_n$. Next, as in the derivation above, we have

$$\frac{1}{N}\sum_{n=1}^{N} J_{n1}^2 \leq \beta_2\rho_1^{2L_N}$$

where β_2 is a constant. Using the definitions of L_n and k_n we get

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N} E|y_n-\hat{y}_n|^2 &\le \beta_2\rho_1^{2\sqrt{\log((N-T_N)/k_N)}}+\frac{4\,\mathrm{var}(e_0)}{k_N}\\ &\le \beta_2\left(\log\frac{N-T_N}{k_N}\right)^{-2\gamma}+\frac{4\,\mathrm{var}(e_0)}{\log^2 N}\\ &\le \frac{\beta_3}{\log^{2\gamma}\dfrac{N-T_N}{k_N}}\\ &\le \frac{\eta_2}{\log^{2\gamma}N}\end{aligned}$$

where β_3, η_2 are constants. ■

C. Proof of Theorem 3

Proof: The proof is much the same as for the second statement in Theorem 2, except that we are not taking the time-average and that the input sequence is stationary. From (7), we also have that $E[J_{n2}^2]=O(1/k_n)$ and need to bound $E[J_{n1}^2]$. From (6)

$$E[J_{n1}^2]=O(E[d_n^{2\alpha}(k_n,L_n)]+\rho^{2L_n}).$$

We now find the rate for which $E[d_n^{2\alpha}(k_n,L_n)]\to 0$. To do this we note that we have taken the set of *nonoverlapping* strings of length L_n from the past, $\{U_j(L_n)\}$. There will be n/L_n, rather than n, such strings. The key to this refinement is that this set of nonoverlapping strings are in fact stationary vectors in $\mathbb{R}^{L_n}$ and hence $d_n(k_n,L_n)$ is the k_n nearest neighbor distance between n/L_n stationary random vectors in $\mathbb{R}^{L_n}$. This is a known quantity and can be shown [15] to satisfy

$$E[d_n^{2\alpha}(k_n,L_n)]=O((k_nL_n/n)^{2\alpha/L_n})$$

which goes to zero as long as $L_n/\log(n/k_nL_n)\to 0$. This is satisfied since $L_n^2=\log n/k_n$ and $k_n=\log^2 n$. As in the proof of Theorem 2, it is now straightforward to verify the rate in the theorem.

VI. Linear Systems

In this section we consider a special class of systems that satisfy A1) and A2), namely the class of stable causal LTI discrete-time systems. The system model is

$$\begin{aligned} y&=h*u\\ z&=y+e \end{aligned}\tag{8}$$

where $*$ is the convolution operator, z is the measured output, e is a noise sequence, and $h\in\ell_1$.

In the following corollary of Theorem 2 we show that every linear system (8) satisfies A1) and A2), and we obtain rates of convergence for various special cases.

Corollary 1: Using the predictor from Theorem 1 we have that for any $u\in\mathcal{B}_r\ell_\infty$, any $h\in\ell_1$, and for any i.i.d. $e_0,e_1,\cdots$ such that $Ee_i=0$ and $E|e_i|^2<\infty$,

$$\lim_{n\to\infty}E|y_n-\hat{y}_n|^2=0.$$

Using modified algorithms we obtain

$$\frac{1}{N}\sum_{n=1}^{N}E|y_n-\hat{y}_n|^2=O[f(N)]\to 0$$

with rates of convergence as follows.

- If $h\in\{g\in\ell_1\colon g_k=0\ \forall k\ge L\}$ with L known, then $f(n)=O(n^{-2/(L+2)})$.
- If $h\in\{g\in\ell_1\colon \exists L, g_k=0\ \forall k\ge L\}$, then $f(n)=O(1/\log^2 n)$.
- If $h\in\{g\in\ell_1\colon |g_k|\le B\rho^k\}$, then the $f(n)=O(1/\log^2 n)$.

Proof: A1) is immediate by linearity

$$||H(u)-H(v)||_\infty=||H(u-v)||_\infty\le||h||_1||u-v||_\infty.$$

To show A2), fix $\epsilon>0$. By stability there exists $L\ge 1$ such that

$$\sum_{i=L}^{\infty}|h_i|\le\epsilon/2r.$$

Take $u,v\in\mathcal{B}_r\ell_\infty$ such that $P_{[m-L+1,m]}v=P_{[n-L+1,n]}u$, then

$$\begin{aligned}&|P_nH(u)-P_mH(v)|\\ &\quad=\left|\sum_{i=0}^{n}h_iu_{n-i}-\sum_{i=0}^{m}h_iv_{m-i}\right|\\ &\quad=\left|\sum_{i=0}^{L-1}h_i(u_{n-i}-v_{m-i})+\sum_{i=L}^{n}h_iu_{n-i}-\sum_{i=L}^{m}h_iv_{m-i}\right|\\ &\quad=\left|\sum_{i=L}^{n}h_iu_{n-i}-\sum_{i=L}^{m}h_iv_{m-i}\right|\\ &\quad\le 2r\sum_{i=L}^{\infty}|h_i|\le\epsilon.\end{aligned}$$

The convergence statement in the corollary then follows from Theorem 1, since this class of systems satisfies A1) and A2).

For rates of convergence, we need to check conditions A1′) and A2′). It is straightforward that $\alpha=1$ and that $K=||h||_1$. We now need to find a relationship between ϵ and L by looking at the restrictions on the tails of the impulse response. If the system has an impulse response of known length L, then $\rho=0$. Using $L_n=L$, the rate of convergence is then

$$O[(k_n/n)^{2/L}+1/k_n].$$

Using $k_n=n^{2/(L+2)}$ gives a rate of convergence $O(n^{-2/(L+2)})$. If L is unknown, then we use $L_n=O(\log n/\log\log n)$ which satisfies $L_n\to\infty$ and $\log n/L_n\to\infty$ and is eventually large enough so that $\Sigma_{i=L_n}^{\infty}|h_i|=0$. Using $k_n=n^{2/(L_n+2)}$ the rate is

$O(n^{-2/(L_n+2)}) = O(1/(\log^2 n))$. If $|h_k| \leq B\rho^k$ then note that for each L

$$\sum_{i=L}^{\infty} |h_i| \leq \frac{B}{1-\rho}\rho^L.$$

The result follows from Theorem 2 by using $C = B/(1-\rho)$. ■

VII. Summary and Remarks

The prediction algorithm introduced in this paper is in fact representative of a class of alternative but similar algorithms. For simplicity, we chose to illustrate our ideas using an adaptation of the nearest neighbor algorithm. However, many of the arguments can be extended easily using other techniques from the consistent nonparametric regression literature such as the kernel and partitioning estimators (e.g., [21], [12], and [20]). There has also been related work in the information theory literature on universal prediction for arbitrary sequences, although the results are somewhat different in nature than the results presented here. There, the performance is compared to the best predictor from a given class of predictors. In the present paper, we do not *a priori* restrict the form of the predictor, but instead we make an assumption on the outputs to be predicted and try to achieve asymptotically optimal prediction when the system generating the outputs is not known. It would be interesting to compare these approaches and/or attempt to obtain results in the spirit of the information theoretic results in the present formulation.

We now make some comments on the relationship between our prediction schemes to system identification. We first point out that the notion of identification of a system generally requires that some topological structure be imposed on the model class. For example, in the context of stable linear systems it is common to consider the systems to be elements of ℓ_1 (e.g., [22]). In addition, it is known that we require sufficiently rich input sequences (e.g., Galois sequences or persistency of excitation conditions) in order to identify the system (e.g., see [22], [17], and [16]). In contrast, our prediction scheme does not require any topological structure, works for nonlinear systems, and works for any bounded input sequence. The difference arises because in identification one needs to "explore the entire state space," whereas for prediction the behavior of the system only along the input sequence needs to be known. In a sense, this idea is captured in Lemma 1, which shows that for any bounded input sequence eventually there is sufficient data for prediction. In fact for nonrich input sequences (that do not excite the system in many ways) prediction should be even easier. On the other hand, it is clear that identification is impossible if data is gathered only in a small region or low-dimensional subpace of all possible inputs.

An obvious issue that needs to be addressed in more detail is the relationship between good identification and good prediction. Certainly good identification implies good prediction. For example, from the work in [22] a system estimate $\hat{h}^{(n)}$ of h is constructed whereby

$$\limsup_{n\to\infty} \|\hat{h}^{(n)} - h\|_1 \leq 2\delta$$

in the case in which the noise is uniformly bounded by δ. A predictor can be extracted from this in the obvious way

$$\hat{y}_n := P_n(\hat{h}^{(n-1)} * u)$$

leading to a prediction error of

$$\begin{aligned} |\hat{y}_n - y| &= |P_n(\hat{h}^{(n-1)} * u - h * u)| \\ &\leq \|\hat{h}^{(n-1)} * u - h * u\|_\infty \\ &\leq \|\hat{h}^{(n-1)} - h\|_1 \|u\|_\infty \leq r\|\hat{h}^{(n-1)} - h\|_1 \end{aligned}$$

if $u \in \mathcal{B}_r\ell_\infty$. This gives

$$\limsup_{n\to\infty} |\hat{y}_n - y| \leq 2r\delta. \tag{9}$$

In Theorem 1 we obtained a bound of δ on the $\limsup$ which improves the result in (9). Although good identification does imply good prediction, our prediction scheme is an improvement and works for every bounded input sequence (independent of the bound).

For the converse, we argue that our predictor when subjected to a "rich" input leads to a notion of identification. We first need to impose a topology on the system class. Take a σ-compact subset $(\mathcal{X}, \rho)$ of nonlinear systems that satisfy A1) and A2). Define the class of input sequences $\mathcal{U}$ that have the following "richness" property:

- $\mathcal{U}$ is the set of all bounded inputs $u \in \mathcal{B}_r\ell_\infty$ such that for any $G, H \in (\mathcal{X}, \rho)$, we have that
$$\limsup_{n\to\infty} |P_nG(u) - P_nH(u)| \leq \delta$$
implies that
$$\rho(G, H) \leq \delta.$$

Note that in the case of stable linear systems, $\mathcal{U}$ contains the set of Galois sequences. Now if we apply an input belonging to $\mathcal{U}$, then our prediction algorithm is in fact excluding systems that are 2δ away from the true system. We suspect that an Occam's Razor-type algorithm similar to that used by Tse *et al.* [22] can be used to construct an identifier for these systems. It seems that the class of "rich" inputs leading to good identification are precisely those inputs for which the prediction algorithm has slow convergence properties. The question of finding interesting nonlinear system classes with interesting "rich" input classes remains to be seen.

A somewhat different question involves the connection between this prediction algorithm and adaptive control. At present no claims are made as to how some version of the present algorithm might be used for control, but one might expect that if one can predict well one should also be able to control well. In fact, one might argue that in control (as with prediction) one is interested in the behavior of the system on the current input rather than on all possible inputs as in system identification. However, the central problem in this case is selecting an appropriate input (to satisfy the control objective) from the family of all admissible inputs.

It is precisely this selection problem that seems to require some sort of search over the input space which is more like the identification problem. Nevertheless, it may be possible to devise a strategy that alternates or trades-off learning phases with control phases, particularly if suitable assumptions restricting the behavior of the system are imposed. It may be interesting to pursue such directions.

REFERENCES

[1] T. M. Cover, "Estimation by the nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 50–55, Jan. 1968.

[2] M. A. Dahleh, E. D. Sontag, D. N. C. Tse, and J. N. Tsitsiklis, "Worst-case identification of nonlinear fading memory systems," *Automatica*, vol. 31, pp. 503–508, 1995.

[3] M. A. Dahleh, T. Theodosopoulos, and J. N. Tsitsiklis, "The sample complexity of worst-case identification for f.i.r. linear systems," *Syst. Contr. Lett.*, vol. 20, pp. 157–166, Mar. 1993.

[4] L. Devroye, "Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 61, pp. 467–481, 1982.

[5] R. M. Dudley, *Real Analysis and Probability*. New York: Chapman & Hall, 1989.

[6] J. D. Farmer and J. J. Sidorowich, "Exploiting chaos to predict the future and reduce noise," *Evolution, Learning, and Cognition*. Singapore: World Scientific, 1988, pp. 265–289.

[7] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1258–1270, July 1992.

[8] W. Greblicki, "Nonparametric identification of Wiener systems by orthogonal series," *IEEE Trans. Automat. Contr.*, vol. 30, pp. 2077–2086, 1994.

[9] W. Greblicki and M. Pawlak, "Dynamic system identification with order statistics," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1474–1489, 1994.

[10] L. Gyørfi, "The rate of convergence of k_n-NN regression estimates and classification rules," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 362–364, May 1981.

[11] A. J. Helmicki, C. A. Jacobson, and C. N. Nett, "Control oriented system identification: A worst-case/deterministic approach in $\mathcal{H}_\infty$," *IEEE Trans. Automat. Contr.*, vol. 36, Oct. 1991.

[12] A. Krzyżak, "On estimation of a class of nonlinear systems by the kernel regression estimate," *IEEE Trans. Inform. Theory*, vol. 36, pp. 141–152, 1990.

[13] ——, "Identification of nonlinear systems by recursive kernel regression estimates," *Int. J. Syst. Sci.*, vol. 24, pp. 577–598, 1993.

[14] S. R. Kulkarni, S. E. Posner, and S. Sandilya "Data-dependent $k_n - NN$ estimators consistent for arbitrary processes," in *Proc. IEEE Int. Symp. Information Theory*, 1998, p. 388.

[15] S. R. Kulkarni and S. E. Posner, "Rates of convergence of nearest neighbor estimation under arbitrary sampling," *IEEE Trans. Inform. Theory*, July 1995, pp. 1028–1039.

[16] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[17] P. M. Mäkilä, "Robust identification and Galois sequences," *Int. J. Contr.*, vol. 54, pp. 1189–1200, 1991.

[18] G. Morvai, S. Yakowitz, and L. Györfi, "Nonparametric inferences for ergodic stationary time series," *Ann. Stats.* vol. 24, pp. 370–379, 1996.

[19] K. Poolla and A. Tikku, "On the time complexity of worst-case system identification," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 944–950, May 1994.

[20] S. E. Posner, "Nonparametric estimation, regression, and prediction under minimal regularity conditions," Ph.D. dissertation, Dept. Electrical Engineering, Princeton Univ., 1995.

[21] C. J. Stone, "Consistent nonparametric regression," *Ann. Stat.*, vol. 8, pp. 1348–1360, 1977.

[22] D. N. C. Tse, M. A. Dahleh, and J. N. Tsitsiklis, "Optimal asymptotic identification under bounded disturbances," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 1176–1190, Aug. 1993.

S. R. Kulkarni (M'91–SM'96), for photograph and biography, see p. 607 of the May 1998 issue of this TRANSACTIONS.

S. E. Posner received the BA.Sc. degree in engineering science and the MA.Sc degree in electrical engineering from the University of Toronto in 1990 and 1992, respectively, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, in 1995.

From July 1995 to December 1996, he was an Assistant Professor of Statistics at the University of Toronto, where he conducted research in derivatives pricing and investment theory. From December 1996 to June 1998 he was a Quant on the fixed-income derivatives desk at ING Barings Securities, New York. He is currently Vice President of Marsh & McLennan Securities Corp.

Dr. Posner was the recipient of the Best Paper Award in Investments from the Southern Finance Association.